Using the Bayesian Normal Gamma prior to identify associated sequence variants.

Elizabeth M. Boggis

Submitted for the degree of Doctor of Philosophy

School of Mathematics and Statistics

November 2014

Supervisors: Dr. Kevin Walters and Dr. Marta Milo

University of Sheffield

Acknowledgements

When I started my PhD I had a very poor grasp of genetics and eQTL analysis. Many thanks go to my supervisors for their help and patience while I obtained this knowledge and for getting me to the point I'm at today. Marta, thank you for being able to point me in the right direction and for sharing your biological and genetics knowledge with me on many occasions when I have felt completely lost and confused. Kevin, thank you for helping me to solve the many problems I have had both statistical and to do with the data. As a supervisory team you have both given me help and support to find my way through the many problems I have encountered during my PhD. I could not have done this without your encouragement and expertise.

Thanks also go to the University of Sheffield for funding this project and the many conferences I have been to. Conferences have made the whole experience much more rewarding!

There are many more people at the University of Sheffield that have helped and supported me thoughout my PhD. These include many staff and other PhD students who have made me smile and kept me entertained during lunchtimes and Friday evenings in the pub! Thanks go especially to Amy for keeping me sane while I was coding the NG for the first time!

I would also like to thank Nicolo Fusi and Wei-Yu Lin who have both provided me with help on certain aspects of analysing data, mainly annotating and pre-processing data. Also to Steve Scott at Google for providing the Spike and slab code we have used throughout this thesis. Thank you all for your help.

I must also thank my family. My mum and dad, Sonia and Jim Boggis, have always supported me in whatever decision I have made. I have needed their support more than ever during my PhD. They have always been at the end of the phone to cheer me up when things are going badly, always been happy to read through my work, even when they don't understand it. Their encouragement, support, feedback and straight talking has been invaluable. I'm very lucky to have such a supportive family. Thank you all for the interest you have shown in my work. You can all go back to rant free conversations with me now I hope!

Abstract

The Normal Gamma prior, a Bayesian adaptive shrinkage method which is implemented using MCMC, is compared to other statistical methods as an eQTL approach to identifying causal or associated genetic mutations. The methods are compared on simulated data, where the results show the Normal Gamma prior to be a far superior method. On human data it is more difficult to assess the results for accuracy, but we can conclude that the Normal Gamma prior highlights SNPs in concordance with other methods. We also note that the Normal Gamma prior, although enforcing very harsh shrinkage, reports many less false positive SNPs than other methods.

We develop the Normal Gamma prior to include functional information which we use to differentially penalise synonymous and non-synonymous SNPs, as well as intronic, intergenic, splicing, UTR3 and other SNPs where necessary. In initial simulation studies, the prior distribution penalises synonymous SNPs on average more than non-synonymous SNPs. Further developments increase the penalisation on intronic, intergenic, UTR3, synonymous and other SNPs more than splicing and non-synonymous SNPs due to larger functional significance scores for the latter. The effect of this on the differential shrinkage between the two sets of SNPs can be seen in the posterior rankings and effect size estimates. We believe that this differential shrinkage form of the Normal Gamma prior is a powerful tool for detecting causal or associated SNPs, and has been shown to increase the posterior mean effect size estimates for causal SNPs with respect to the standard Normal Gamma, as well as increasing the ranking of validated causal SNPs (with respect to the standard Normal Gamma).

Publications

E. M. Boggis, M. Milo and K. Walters. *Exploiting Adaptive Bayesian Regression Shrinkage to Identify Exome Sequence Variants Associated with Gene Expression.* The Contribution of Young Researchers to Bayesian Statistics, Springer Proceedings in Mathematics & Statistics, Springer International Publishing, 63:135-138, 2014. ISBN: 978-3-319-02083-9.

E. M. Boggis, M. Milo and K. Walters. *Exploiting Bayesian shrinkage within a linear model framework to identify exome sequence variants associated with gene expression.* Poster presented at Joint 21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 12th European Conference on Computational Biology (ECCB) 2013, 19 - 23 July 2013, Berlin, Germany. Poster number N6. F1000Poster reference: F1000Posters 2013, 4: 925 (poster).

Thesis summary

This thesis can be split into five sections. The first section covers Chapters 1 and 2, and is an introduction to the field, area and data used within the thesis. Chapters 3 and 4 investigate statistical methods that could be applied to the eQTL data. Chapters 5 and 6 present the calculations and application of the standard Normal Gamma, while the remaining Chapters 7 - 10 present the inclusion of functional information in the Normal Gamma prior hierarchy. Chapter 11 concludes the thesis.

Chapter 1 introduces the basics from the field of genetics and aims to contextualise the need for the model developed in this thesis.

Chapter 2 introduces the datasets used throughout the thesis. It aims to give a very brief background to the data production techniques.

Chapter 3 compares a selection of Bayesian shrinkage methods, highlighting those that we will use throughout this thesis.

Chapter 4 compares the results of running six statistical methods on simulated eQTL datasets. The aim of the chapter is to assess which statistical method will be used throughout the remainder of the thesis.

Chapter 5 contains all the calculations required to fully implement the Normal Gamma prior hierarchy.

Chapter 6 presents the results from the application of the Normal Gamma to the Yeast dataset, as well as to two human datasets, Hulse and Fairfax.

Chapter 7 develops the Normal Gamma prior to include functional information relating to synonymous and non-synonymous SNPs and reports the simulation results based on the initial development. A different transformation of the FS score which encodes the function information for synonymous and non-synonymous SNPs for inclusion in the Normal Gamma prior hierarchy is also explored here, along with the associated results on simulated data.

Chapter 8 extends the Normal Gamma with functional information to include seven groups of functional information priors, compared to the two groups (synonymous and non-synonymous) used previously. It contains all the required calculations for implementing this development to the Normal Gamma.

Chapter 9 reports the simulation results for the Normal Gamma with the seven functional information groups described in Chapter 8.

Chapter 10 compares results from the Normal Gamma with and without functional information on subsets of the two human datasets, Hulse and Fairfax.

Chapter 11 concludes the thesis, discussing the results that have been obtained throughout the thesis and their application, as well as ideas for further development to the Normal Gamma prior.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, we combine sophisticated statistical methodology and complex eQTL datasets (expression of quantitative trait loci,data which consists of gene expression and genotype data) with the aim of reducing the search space for causal/associated genetic sequence variants.

This chapter will cover basic biology and genetics needed to understand the project, as well as a brief overview of current techniques used to analyse genetics data. The remaining chapters in this thesis investigate and compare statistical methods on eQTL data. We choose to develop the statistical method based on applying the Normal Gamma prior hierarchy to a standard linear model. We develop this structure to include functional information. We then test both the standard Normal Gamma model, and the developed Normal Gamma model on both simulated and real datasets.

1.1 Genetics Background

Genetics is the study of genes, the inheritance and the variations and mutations within living organisms. Genetic epidemiology is the study of the role of genetic factors such as mutations in determining health and disease within families and populations. Genetic epidemiology includes the interaction between both genetic and environmental factors. Genetic epidemiology has evolved with new data and further insight into the human genome. Historically genetic epidemiologists identified monogenic diseases, now they are working, as we are, on more complex, polygenic diseases. The data used by genetic epidemiologists has progressed from familial data used for linkage studies, through to GWA (genome wide association) data using tagSNPs and case-control disease status, to sequencing data where every DNA base is read. These data are combined with either case-control/disease status or a corresponding gene expression value in current GWA/fine mapping studies or eQTL (expression of quantitative trait locus) studies respectively.

1.2 Background Biology

The structure of DNA

The information in the human genome is stored in the four chemical bases that make up DNA; adenine (A), thymine (T), cytosine (C), and guanine (G). These always bind in specific pairs, A and T, and C and G. This pairing is the key to DNA replication or copying. DNA is arranged in a double helix form, where each strand consists of a sequence of *nucleotides*. A nucleotide is any one of the four bases of DNA (A, T, C or G) chemically bound to a sugar molecule and a phosphate molecule. Only one strand of the DNA needs to be known as the other is predefined by the specific pairwise binding of DNA.

DNA is not just one long string of bases A, T, C and G, it is split into 22 pairs of *chromosomes* or *autosomes* and 1 pair of sex chromosomes. Chromosomes are numbered according to their size, with the exception of chromosomes 21 and 22 where chromosome 21 has now been found to be smaller than chromosome 22. The present estimation, from Ensembl [Flicek et al., 2012], for the size of chromosome 1, the largest chromosome, is that it has approximately 249 million bases (also referred to as nucleotides) and chromosome 21, the smallest chromosome, has 48 million bases.

Chromosomes are subdivided into *genes* and intergenic regions. The genes contain many introns and exons. Each gene codes for a particular protein. Introns are the non-coding part of DNA. They separate the exons and are found inside and outside *genes*, but they are not *transcribed*. Exons are the coding parts of the DNA. They are only found within genes and are transcribed and translated to make proteins.

We have two copies of each chromosome that can be different. One is inherited from each parent. If the two alleles at each location match the reference allele then the individual is *homozygous wildtype* at that allele. If both alleles match the alternative or minor allele (the least common allele in the population) the individual is *homozygous SNP*, if the two alleles do not match each other, the individual is *heterozygous*.

From DNA to Proteins

In order to make proteins, DNA is transcribed into mRNA, which is smaller than DNA. Proteins are created from DNA via the two processes known as *transcription* and *translation*, see Figure 1.1. Transcription takes place in the nucleus of the cell which is where all our DNA is stored. During transcription the relevant gene(s) are transcribed into single-stranded *mRNA*. The mRNA is small enough to leave the nucleus of the cell and travel to the cell's cytoplasm where translation takes place. A *ribosome*, the protein synthesising part of the cell, reads the mRNA sequence from the start *codon* AUG until it reaches one of the stop codons (UAA, UGA or UAG). A codon is a set of three bases of mRNA that specify a particular amino acid. It is the sequence of amino acid that makes the protein, allowing the cell to carry out its function.

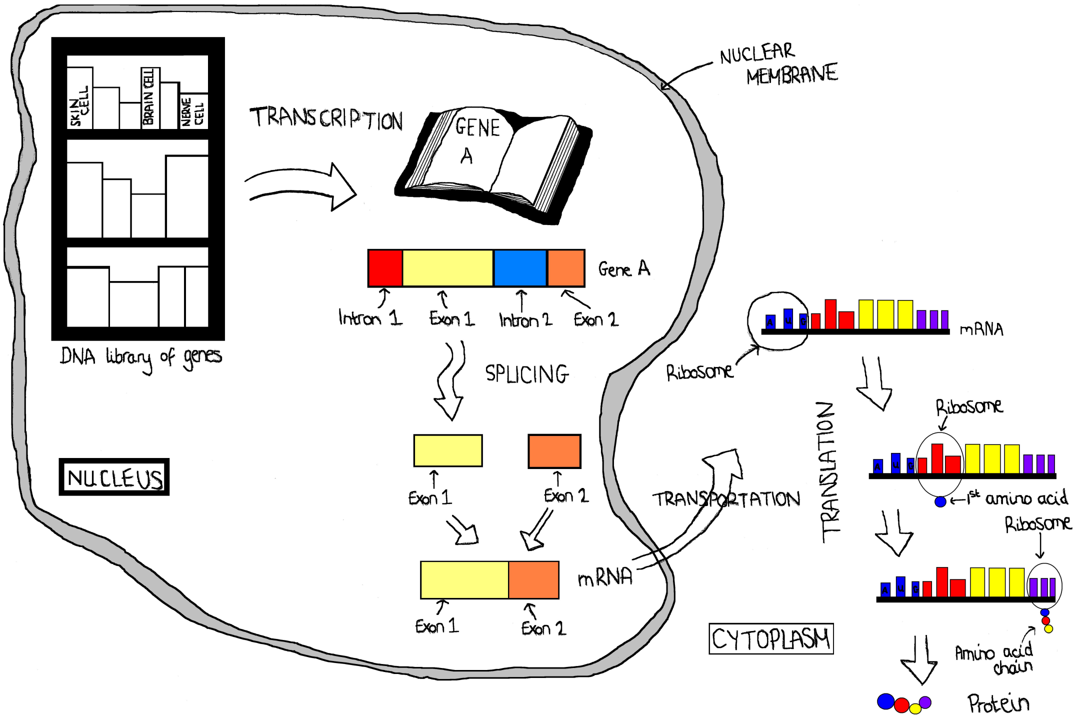


Figure 1.1: In the nucleus of every cell all our DNA is stored. There are only certain parts of the DNA code, specific genes, that are required in any cell. This diagram shows a simplification of the process of transcription which selects the relevant gene for that cell and makes mRNA, full details can be found in Alberts et al. [2008]. This mRNA then moves from the nucleus of the cell to the cytoplasm of the cell, where it is translated by a ribosome which analyses codons and creates an amino acid chain. Once a stop codon is reached by the ribosome, it stops translation and a protein is formed. It is this protein that allows the cell to carry out its function.

There are four different bases in RNA (A, U, C and G), hence there are $4 \times 4 \times 4 = 64$ possible different codons, given that each codon comprises of three bases. However, as there are only 20 amino acids, and of these we know that three of these are stop codons (UAA, UGA and UAG) and one is a start codon (AUG), this means that many codons code for the same amino acid.

DNA Replication

Not all healthy cells live forever, and so they must pass on their genetic information to the new cells. The process of duplicating a cell is called *replication*. Replication aims to create an identical copy of the cell and the DNA within it. This process is very efficient due to the pairwise binding of DNA. To replicate, the DNA double helix unwinds itself and splits into two separate strands. Free nucleotide molecules then attach themselves to their partner base on the single strand DNA to create two identical copies of the DNA, see Figure 1.2. While the DNA is unwinding, the cell nucleus divides, each nucleus containing a single strand of the DNA. Once the DNA has been replicated in the new nuclei, the cell divides completely.

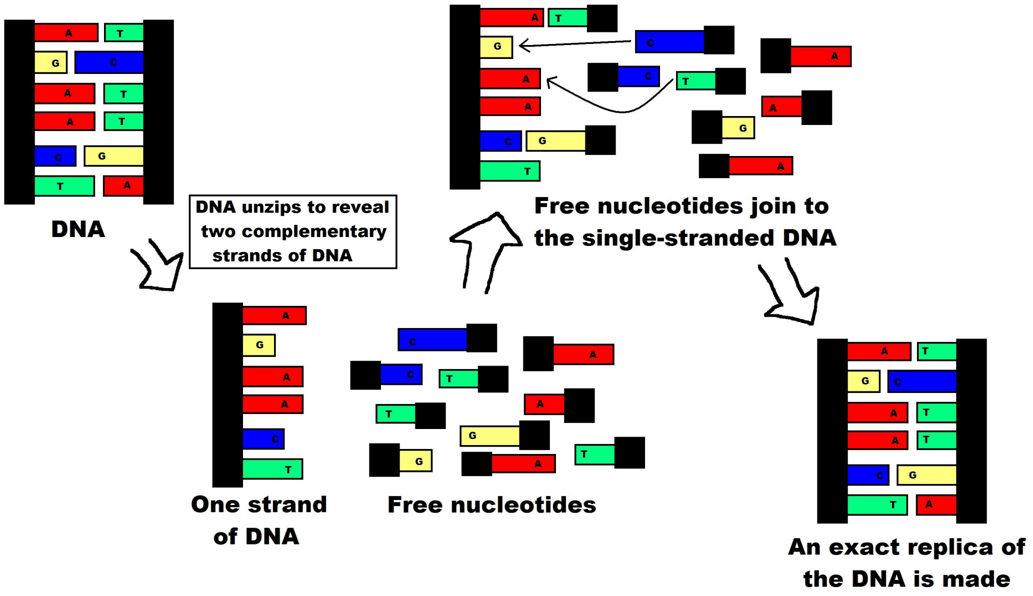

DNA
DNA unzips to reveal two complementary strands of DNA
Free nucleotides join to the single-stranded DNA
One strand of DNA
Free nucleotides
An exact replica of the DNA is made

Figure 1.2: DNA replication is the process of DNA copying itself exactly. This diagram explains, for one of the two strands, how this happens. The process happens simultaneously for the second strand, and so at the end, there are two exact copies of the DNA. This process is only possible as the DNA binds in specific pairs, A (red) and T (green), and C (blue) and G (yellow).

1.2.1 Mutations.

During replication, mistakes or *mutations* occur, some of which are not corrected. It is stated in Alberts et al. [2008] that uncorrected mutations occur at a frequency of approximately one mutation per 10^9 bases, or, on average, three times during the replication of the entire string of DNA (3.2×10^9 bases long). DNA has a very sophisticated repair method that checks all newly made DNA for errors after replication. However this is not infallible and so mutations do occur. The effects of mutations differ based on where the mutation is located, and on what the mutation is. Cells have the ability to kill themselves if they are too badly mutated to function correctly, this is known as *apoptosis.* This prevents some serious mutations from replicating. However, it again is not infallible and as such *somatic* mutations, mutations that are not hereditary, do occur.

A mutation that leads to a single base change is called a *single nucleotide polymorphism/variant, SNP/SNV* (pronounced snip/sniv), see Figure 1.3. Mutations that occur in the coding regions of the DNA, the exons, are labelled according to the change that occurred at DNA base level, and their effect on the protein or amino acid. *Synonymous* mutations do not directly change the protein as the amino acids remain unchanged; this occurs quite regularly due to the number of of codons that code for the same amino acid. However, if a mutation causes a change in an amino acid and therefore in a protein, it is known as a *non-synonymous* mutation, some of which have a serious effect on the individual.

Other main types of mutations are *indels* and *CNVs.* Indels are insertions and deletions of various numbers of bases in the DNA sequence, see Figures 1.4 and 1.5 respectively. CNVs, or copy number variants are changes in the number of a repeated nucleotide or sequence of nucleotides of any size, see Figure 1.6.

Mutations that do not occur in the coding region of the DNA are labelled according to their location, for example mutations found in the intergenic regions are known as intergenic mutations. Similarly for intronic, splicing, UTR3′, UTR5′, upstream, downstream and all other types of mutation.

1.2.2 Gene Expression.

The genes that are expressed in a certain cell are those that are transcribed and translated into proteins in that cell. It is the expression of genes that enables the human body to function. Genes that are transcribed can have two types of effects; *cis-acting* and *trans-acting.* A cis-acting gene affects the immediate vicinity of the gene only, whereas a trans-acting gene can have an effect in many other locations at varying distances from the gene. Genes send signals to other genes via the proteins that they encode. We measure these signals when measuring gene expression. It captures the amount of mRNA which is used to create the proteins for message carrying.

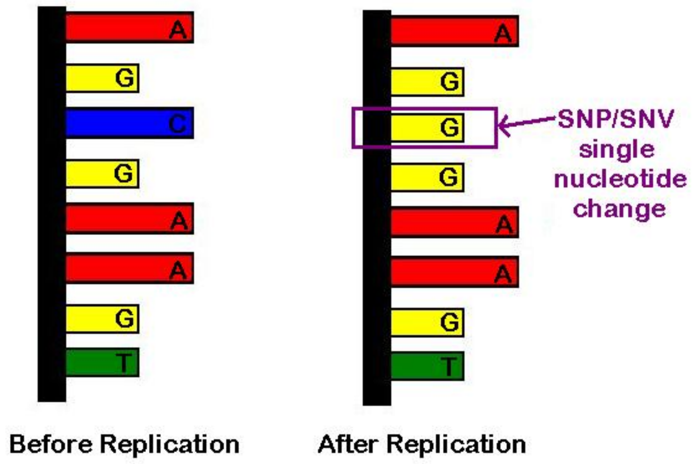


Figure 1.3: During replication, the process of copying DNA, mutations can occur. If a single nucleotide is mis-copied then the result is a SNP (Single Nucleotide Polymorphism) or SNV (Single Nucleotide Variant). This figure demonstrates pictorially a single strand of DNA pre- and post-replication, highlighting the single nucleotide change.

1.2.3 The Human Genome and Exome.

The human genome is reported as a 2m long string of DNA contained in each cell, made up of 3.2×10^9 nucleotides. The human exome, the set of all exons, is the coding part of the DNA. It comprises 1.5% of the human genome [Alberts et al., 2008].

In the last decade, understanding of the role of the exome has developed such that any non-exonic regions of the DNA are no longer thought of as just 'junk DNA'. It is now known that intergenic and intronic regions contain splice sites, transcription factors, and other regulatory regions of the genome which are important for the functionality of the genome. Palazzo and Gregory [2014] present a review summarising the extent to which 'junk DNA' has been shown to exist/not exist over the past decade or more. Projects such as ENCODE [The ENCODE Project Consortium, 2012] are trying to capture this information in publicly available databases.

Not only has our understanding of the genome improved, but also our understanding of epigenetic or environmental factors and they way they interact with the interpretation of our DNA. These epigenetic effects are not only linked to the exome, but are more frequently linked to non-exonic regions.

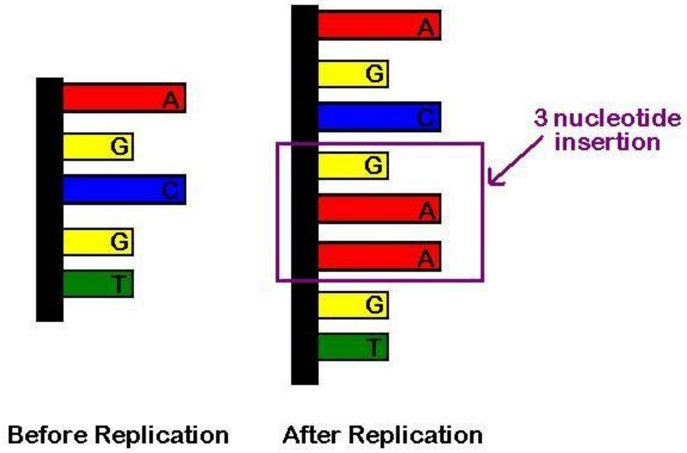


Figure 1.4: During replication, the process of copying DNA, mutations can occur. If one or more nucleotides are inserted into the DNA sequence, this is known as an insertion. Depending on the number of nucleotides inserted, the effect can be very different. This figure demonstrates pictorially a single strand of DNA pre- and post-replication, highlighting a 3 nucleotide insertion.

1.3 Current statistical analysis techniques for genetics data

Current statistical techniques focus on identifying associated sequence variants rather than causal sequence variants, although the ultimate aim is to identify truly causal variants. An associated variant is often one that appears to be having an effect that is linked to a disease, but that has not been biologically shown to lead to the disease. Any mutation (SNP) that has been biologically validated will be referred to as causal. Any SNP that has not been biologically validated but that has been identified by statistical methods as playing a role in disease will be referred to as associated. These associated SNPs may subsequently be found through biological testing to only be in high LD (linkage disequilibrium (correlation)) with the truly causal SNP, and to have no actual effect themselves.

Genome Wide Association Studies (GWAS), unlike sequencing studies, typically use tagSNPs followed by imputation to infer information about disease associated SNPs in the whole genome. TagSNPs are SNPs which are representative of a region of the genome and can be used to infer information about the entire genome, or a particular region of interest. In general, GWAS typically involve around 1 million genotyped SNPs which increases to tens of millions of SNPs post imputation [Bush and Moore, 2012], [Cantor et al., 2010]. The numbers of individuals can also reach tens of thousands of cases and controls

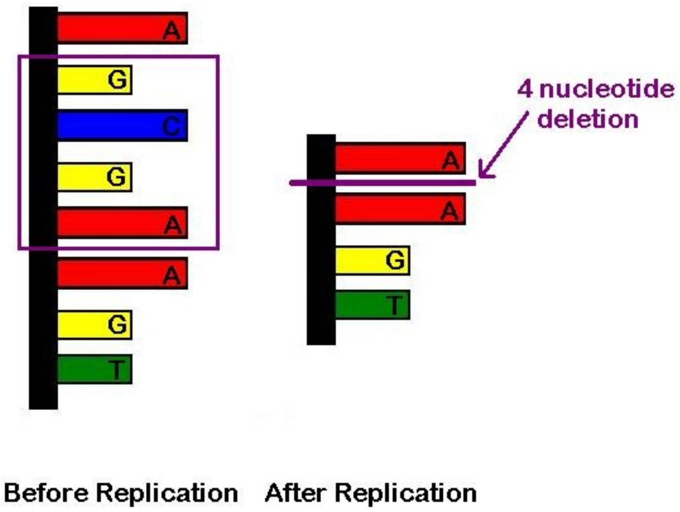


Figure 1.5: During replication, the process of copying DNA, mutations can occur. If one or more nucleotides are deleted from the DNA sequence, this is known as a deletion. Depending on the number of nucleotides deleted, the effect can be very different. This figure demonstrates pictorially a single strand of DNA pre- and post-replication, highlighting a 4 nucleotide deletion.

for large consortia, although smaller GWAS tend to include thousands of individuals [Lourdusamy et al., 2012].

The main difference between the eQTL (expression of quantitative trait locus) study that we perform throughout this thesis and a GWAS is that GWAS tend to use a combination of disease status (case/control) paired with SNP information, whereas eQTL studies use gene expression paired with SNP data. This leads to modelling differences as GWAS use a binary outcome variable and eQTL studies use a continuous outcome variable.

1.4 eQTL

An eQTL is a QTL (quantitative trail locus) that is associated with gene expression. This means that we use gene expression as our quantitative trait for the particular locus. eQTLs, as with most genomic regions, can have cis- and/or trans- acting effects. Cis-acting effects are those that have an effect in the vicinity of where they are located. Trans-acting effects are those that act at a distance, i.e. affect a different gene to they reside in.

eQTL studies are increasing in number at present as the cost of sequencing and gene expression quantification decrease. They are being used to study polygenic or multifactorial diseases which cannot be well understood using single SNP methods. eQTL analyses require paired data consisting of gene expression

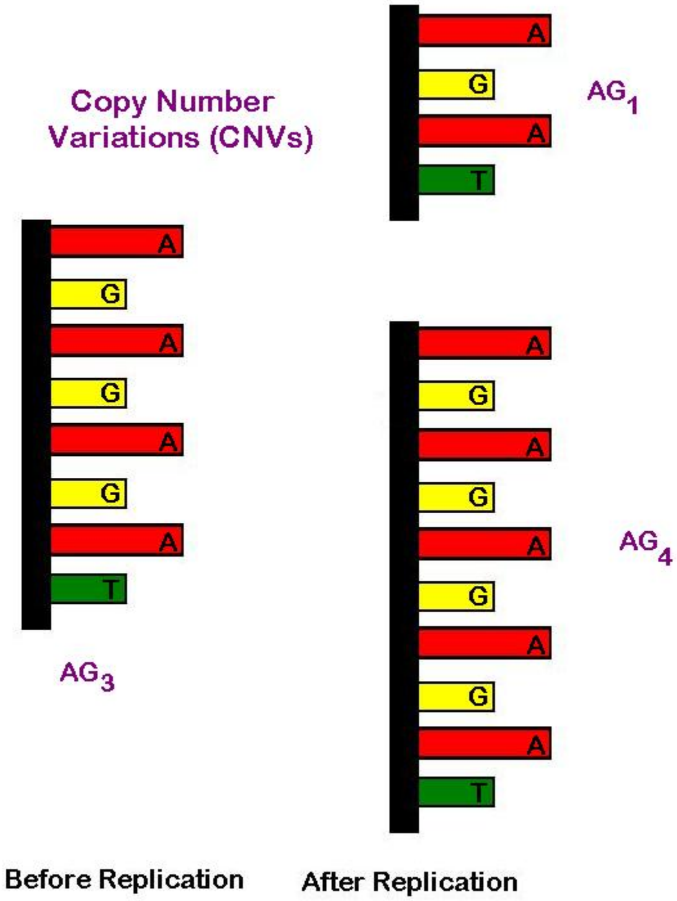


Figure 1.6: During replication, the process of copying DNA, mutations can occur. If a repetitive sequence of DNA is being copied and a mistake is made in the number of times the repetitive sequence is copied, we have a CNV (copy number variant). This figure demonstrates pictorially a single strand of DNA pre- and post-replication, highlighting two possible CNVs of the AG repeated sequence. The subscript number indicates the number of the repeats. In our example, AG_4 indicates the AG pair is repeated 4 times, while AG_1 indicates the AG pair is repeated only once.

and sequence level data [Franke and Jansen, 2009], [Suthram et al., 2008]. eQTL studies tend to have small n (number of individuals), large p (number of SNPs) due to the large number of SNPs in the human genome. This means that eQTL studies, as with GWAS, are computationally intensive and require the use of non-standard statistical techniques.

Many simple diseases are genetically quite well understood, given linkage analysis and familial studies that have previously been carried out. It is now the complex, polygenic diseases we aim to understand for diagnostic and treatment purposes.

This thesis reviews possible multivariate statistical methods that can be applied to eQTL data (see Chapter 3) and focuses on developing the Normal Gamma prior to include functional information which can be used to prioritise SNPs a priori (see Chapters 7 - 10).

1.5 Software and Computing tools

Hapgen2

Hapgen2 is the latest version of the Hapgen software [Su et al., 2011]. It simulates case-control datasets at different SNP markers. It uses either HapMap [The International HapMap Consortium, 2005] or the 1000 Genomes [Altshuler et al., 2010] database to obtain data with a realistic LD (linkage disequilibrium or correlation) structure between the SNPs at the location chosen. The software requires inputs on the specific region to simulate data for, the number of cases and controls, and any specific details on the causal SNPs required, such as the location of the causal SNP(s) and their corresponding effect size(s). The output contains details on this information.

Impute2

Impute2 [Howie et al., 2009] is a sister program to Hapgen2. It is used to impute missing genotype data. For any dataset with given subjects and SNPs a genotype can be imputed with varying degree of certainty. Impute2 uses either HapMap [The International HapMap Consortium, 2005] or the 1000 Genomes [Altshuler et al., 2010] database as well as the LD structure and genotypes in the sample that are not missing to provide a best estimate for the missing data. SNPs for the whole cohort or simply missing SNPs for particular individuals can also be imputed. The SNPs that are imputed are only those that are stated in the set of input files specified. The output contains information on the probability of the imputed SNPs being not homozygous wildtype. It also contains a measure relating to the information with which the imputation of the missing genotype has been made (the *info score*). This is used by researchers as a quality control metric.

R

R [R Development Core Team, 2008] was used for some data processing and statistical analyses.

MATLAB

Much of the data processing and analysis was performed off-line using the commercial software package MATLAB [The MathWorks Inc.].

Iceberg

Much of the work carried out was very computationally expensive and so was carried out on Iceberg, the Unix based High Performance Computing Cluster provided by The University of Sheffield. This includes the simulation and imputation carried out using HAPGEN2 and IMPUTE2, as well as some of the analyses in R and all analysis in MATLAB.

1.6 Introduction to Original Research

In this thesis we compare and contrast a spectrum of statistical methods with the aim of selecting one that is better than the others at detecting causal or associated sequence variants. The first section of this thesis, Chapter 3, investigates possible methods we could use, before assessing via simulation their effectiveness. Having decided the Normal Gamma method performs best based on results in Chapter 4, Chapter 5 goes on to replicate the calculations required to implement the Normal Gamma, before assessing it on real Yeast and Human data in Chapter 6.

The Normal Gamma prior is then extended in Chapters 7, 8 and 9 to include functional information with the aim of prioritising groups of SNPs with a priori more chance of being causal. Simulations are then used to assess this inclusion. Finally, the extended model is applied to the same Human datasets as the basic model with the results compared to one another in Chapter 10.

The application of the six statistical models to the eQTL data is a new approach, both for the models and for the analysis of the data. The inclusion of functional information to improve a model is not new, but the SNP specific nature of the inclusion we propose is very different to the current methods proposed.

The developments and extension to the Normal Gamma prior hierarchy show that the model prioritises SNPs that are known to be causal/associated with higher rank than the standard model. This is important for the field of genetics and mathematical biology as it shows that the introduction of a complex statistical model does help to reduce the search space for causal/associated sequence variants.

If the results from the model can be combined with expert knowledge, a much more effective search space can be targeted when attempting to understand complex diseases.

Chapter 2

An overview of data generation and preprocessing

In this chapter we will describe the data that is used to test and evaluate the models used in this thesis. There will be details on how the data was produced and a brief overview of the analysis carried out to obtain the input values that we use.

2.1 Introduction to the Datasets

There are three origins of the data used in this thesis - synthetic data that we simulate ourselves, yeast data produced for Zhu et al. [2008], and human data from Hulse and Cai [2013] and Fairfax et al. [2012]. The data will be used to compare the effectiveness at detecting causal SNPs, SNPs which have been biologically validated in the literature, and the flexibility of many statistical models to different eQTL datasets. Here we investigate sample sizes, data production techniques and over arching methods for preprocessing the data.

Datasets with respect to the Normal Gamma implementation

Our implementation of the NG (Normal Gamma prior hierarchy) is a gene-by-gene approach. Because of computational constraints, we cannot practically test every gene, hence we select a subset of genes to test. We cannot test every SNP in the genome, hence we also select a subset of SNPs to test for associations to disease.

We remove SNPs where all individuals have identical genotypes. Not removing these SNPs would lead to identifiability problems between the SNPs and the intercept or background gene expression level α. In some cases, the yeast data for example, there is a different baseline set up for the data and so we introduce an indicator variable to represent the change. This could be an environmental effect or some other such confounding factor that is known to effect gene expression for one set of individuals but not another.

As we cannot assess differential expression with data for only one group of individuals, we select the genes to test our data on based on what has been published in the literature. Some genes are reported in the literature as having an associated/causal SNP within them. We initially include only the exonic SNPs from each chromosome that the gene is located on. Where this leads to small numbers of SNPs, we also include all SNPs with an annotation that includes our required gene name. The latter is also used when the published SNPs are not exonic.

2.1.1 Simulated data

The two pairs of simulated datasets are generated using HapGen2 [Su et al., 2011] to generate the SNPs and R [R Development Core Team, 2008] to calculate the corresponding gene expression. We label these datasets 1A, 1B, 2A and 2B. When simulating the two HapGen simulated datasets, the correlation structure in our region of interest is taken into account. More details are given in Section 4.1.1, page 51 and Section 4.2.1, page 57 for datasets 1 and 2 respectively. HapGen2 uses sequence data from HapMap [The International HapMap Consortium, 2005] and the 1000 Genomes project [Altshuler et al., 2010] to understand the structure of the genomic mutations and their frequency in the population. This genomic structure is reflected in the data produced.

2.1.2 Yeast data

The Yeast dataset [Zhu et al., 2008] comprises of 218 yeast grown in two different conditions, ethanol (109 samples) and glucose (109 samples) which we need to take into account when doing the eQTL modelling. The yeast genome is far simpler than the human genome. It is both smaller and has more clearly defined functions. The data was generated to investigate gene-environment interactions, hence is assessing the effect of the two growth conditions for the yeast.

The yeast genotyping data is produced using Oligonucleotide Microarrays, using a method described in Winzeler et al. [1998]. In this dataset, the resulting genetic map of 3312 markers covered $> 99\%$ of the genome. The gene expression is measured using Expression Profiling by Agilent Microarrays.

We select genes in our dataset that were reported in two or three of the published articles referenced in Figure 4B of Lee et al. [2009]. We state these genes in Table 2.1. The gene/hotspot location pairings are taken from three sources; Lee et al. [2009], which uses a Bayesian method that includes functional information, Yvert et al. [2003] which uses clustering and linkage analysis and Zhu et al. [2008] which uses yeast regulatory networks. Two or all three of these methods agree on 11 genes that we choose to compare our statistical methods on, see Table 2.1. Lee et al. [2009] report an eQTL hotspot location corresponding to each gene listed, for example the hotspot for YCR040W is defined to be Chr3:230,000. For analysis purposes, we define the target SNPs to be any Yeast SNP that is within the hotspot region. We define this hotspot region to include all SNPs within $\pm 10kb$ of the given hotspot location, which does not define the location of one particular SNP. We chose a narrow region for the hotspot SNPs to restrict the number of SNPs we define as target SNPs, given that only the hotspot location is reported, and not the location of a particular causal SNP. We only know that the location has been shown to be associated to the gene via eQTL analysis in two or three of the studies.

The design of the experiment which produced the yeast data requires the inclusion of a binary blocking variable, $\alpha_{environ}$ to represent the different environments, glucose or ethanol. We therefore use the following gene-by-gene model $y_j = \alpha + \alpha_{environ} + \sum_{i=1}^{p} \beta_i X_{i,j} + \epsilon_j$, where y_i is the gene expression for individual j, α is the background gene expression level, p is total number of SNPs, β_i is the effect size of SNP i and $X_{i,j}$ is the genotype for SNP i for individual j in all the statistical methods where possible.

We summarise the Yeast data as follows in Table 2.1.

2.1.3 Hulse Data

The Hulse dataset [Hulse and Cai, 2013] comprises of gene expression data from GSE6536 [Stranger et al., 2007] and GSE11582 [Choy et al., 2008] matched by individual to sequence data from HapMap release 28 [The International HapMap Consortium, 2005]. The data in GSE6536 is Illumina Sentrix Human-6 Expression BeadChip on RNA extracted from lymphoblastoid cell lines (LCLs). The data in GSE11582 is Affymetrix Human Genome U133A Array on RNA from 269 cell lines which have been densely genotyped by the International HapMap Project [The International HapMap Consortium, 2005]. We choose to use only GSE6536 to prevent introducing confounding effects by combining data from two different platforms.

The Hulse study [Hulse and Cai, 2013] explores the genome wide association between genetic variants and gene expression variability in humans, which they denote expression variability QTL (evQTL). The study finds 218 genes which are involved in cis-acting evQTLs, 8 of which are validated using genotype data

Gene name	n	p	Target SNPs (freq of the minor allele within the sample)
YBR158W	218	1802	148 (0.404), 149 (0.385), 153 (0.413), 154 (0.422),155(0.413), 156 (0.404)
YBR162C	218	1802	148 (0.404), 149 (0.385), 153 (0.413), 154 (0.422),155(0.413), 156 (0.404)
YCL009C	218	1802	205 (0.459), 206 (0.45), 207 (0.45), 208 (0.45), 209 (0.468)
YCR040W	218	1802	217 (0.477)
YHR005C	218	1802	750 (0.385), 751 (0.367), 752 (0.376)
YHR005C-A	218	1802	750 (0.385), 751 (0.367), 752 (0.376)
YLR256W	218	1802	1232 (0.422), 1233 (0.394), 1235 (0.413), 1236 (0.44), 1237 (0.45)
YLR442C	218	1802	1316 (0.413), 1317 (0.477), 1318 (0.486), 1319 (0.477), 1320 (0.229), 1321 (0.239), 1322 (0.248)
YNL088W	218	1802	1513 (0.394)
YOL084W	218	1802	1600 (0.44), 1601 (0.45)
YOR125C	218	1802	1678 (0.459), 1679 (0.44), 1680 (0.45), 1682 (0.45)

Table 2.1: Table showing the genes, number of individuals n, number of SNPs p and the target SNPs and the frequency of the minor allele within the sample for the Yeast genes from Lee et al. [2009].

from the 1000 Genomes Project [Altshuler et al., 2010]. These 8 validated genes and SNPs denoted in Table 1 of Hulse and Cai [2013] are those that we will select from to compare statistical models, see Table 2.2 for the final list of seven genes we use. There are many other genes in which one or more SNPs associated to an evQTL have been located (166 in GSE6536 and 60 in GSE11582). We omit to use these genes as they have not be validated.

There have been SNPs associated to an evQTL detected in intronic regions in multiple genes, including IL6, ADCY1, PLOD2 and SNX7. The latter, SNX7 also has both synonymous and non-synonymous SNPs (rs2019213 and rs35296149 respectively) deemed to be associated to the evQTL.

This dataset is also used to compare associated SNPs found using evQTLs and eQTLs. evQTL models treat the variability in gene expression as the response y, whereas eQTL models use the gene expression value. The reported loci we use for the Hulse dataset focus on genes and SNPs associated to evQTLs. In Hulse and Cai [2013] a direct comparison of results from evQTL and eQTL studies has been highlighted. This shows differing levels of concordance between eQTL and evQTL results in difference scenarios.

When recoding the HapMap genotype data for this dataset from haplotype to $\{0, 1\}$ representing homozygous wildtype and either heterozygous or homozygous SNP respectively, we find there are unread/undefined bases that have no defined genotype. In these cases we use imputation to obtain an estimate of genotype. We use Impute2 [Howie et al., 2009] to estimate the expected number of minor alleles. We then use this expected number to calculate the probability of the genotype not being homozygous wildtype. We use this probability as the genotype for our SNP. We keep imputed SNPs with an info score greater than 0.3, the lowest value suggested in the Impute2 documentation. This excludes approximately, on average 9% of the SNPs. If we had chosen to include SNPs with an info score greater than 0.5, the upper value suggested in the Impute2 documentation, we would have excluded an extra approximately 2% of SNPs.

The genes to analyse were chosen based on the 8 validated genes in Table 1 of Hulse and Cai [2013]. Only 7 of the genes were identified using the data from GSE6536. The gene we omit is FERMT2, as it did not have any recorded gene expression value. Hulse and Cai [2013] state that SNX7 has causal SNPs that are both synonymous and non-synonymous (rs2019213 and rs35296149 respectively). Neither SNP was included in our analysis; the former rs2019213 being excluded as all individuals in the analysis were homozygous wildtype at this location and the latter rs35296149 does not appear in the HapMap dataset used.

To use the genes and SNPs defined in Table 2.2, we annotate the SNPs using ANNOVAR and retain any SNPs with annotations containing the gene in question, irrespective of its location in the genome. For this reason we include intronic, intergenic and other types of SNP. We chose to do this for this dataset due to the small number of exonic SNPs that were found in the targeted genes.

We summarise the Hulse data as follows in Table 2.2.

Gene name	n	p
ADCY1	39	351
CTNNA2	38	2919
DAAM2	39	149
IL6	39	189
PLOD2	39	1488
SNX7	39	389
TNFRSF11B	39	379

Table 2.2: Table showing the number of SNPs p and individuals n for each gene in the Hulse dataset [Hulse and Cai, 2013].

2.1.4 Fairfax Data

The Fairfax dataset [Fairfax et al., 2012] was produced to allow the study of paired purified primary monocytes and B-cells with the aim of identifying cis- and trans-acting eQTLs. Results aim to identify effects unique to monocytes or B-cells via cell specific eQTLs. Monocytes and B-cells were chosen to help identify relevance to immunity and inflammation.

The data were obtained using Illumina Human HT-12 v4 BeadChip for the genome-wide expression profiling and Illumina HumanOmniExpress-12v1.0 BeadChips for the genotyping.

In the published study, after quality control, the eQTL analysis was performed at 651210 markers for 283 individuals. As a gene-by-gene analysis was not carried out here, the number of genes is not stated.

We use the genes that are reported in Figure 6b of Fairfax et al. [2012] to assess the statistical methods. This gives 8 scenarios of gene expression and SNPs to assess, see Table 2.3. In each gene there is only one validated causal SNP. As with the Hulse dataset, there are missing genotype data. To obtain complete data, we use Impute2 to estimate the genotype of any missing value. We use only those imputed genotypes where the info score is greater than 0.3.

Following our processing of the data we retain a different number of SNPs and individuals for each gene, see Table 2.3 for the details on the final numbers of SNPs and individuals for each gene. The numbers of SNPs are dependent on the proportion of missing data and subsequent imputation results. In this case, we keep all exonic SNPs on the chromosome where the gene of interest is located. We also include the SNPs that are reported as causal regardless of the location (exonic, intronic, or otherwise).

The results in Figure 6b of Fairfax et al. [2012] show that the number of copies of the rare/minor/alternative allele has a clear effect on gene expression, therefore we recode the SNPs as $\{0, 1, 2\}$ to reflect the additive effect of the number of minor alleles.

We summarise the Fairfax data as follows in Table 2.3.

2.2 Genotyping and Sequencing methods

Genotyping refers to determining which genetic variants an individual has. This is often done using chips or arrays that contain known, specific genomic sequences aimed at identifying the presence or absence of particular mutations. Sequencing refers to determining every base in a given length strand of DNA. There are many different processes used for both genotyping and sequencing.

Gene name	n	p	Causal SNP (sample frequency)
ERAP2 bcell	244	792	rs10044354 (0.689)
ERAP2 mono	244	792	rs10044354 (0.689)
CARD9 mono	243	511	rs4266763 (0.650)
FADS1 bcell	243	1076	rs174548 (0.514)
RBM6 bcell	243	932	rs1061474 (0.675)
RBM6 mono	243	932	rs1061474 (0.675)
CD40 mono	243	468	rs4810485 (0.428)
FAM167A bcell	244	551	rs13277113 (0.934)

Table 2.3: Table showing the genes and gene expression type (monocytes (mono) or B-cell (bcell)), number of individuals n, number of SNPs p and the causal SNP and the sample frequency of the minor allele (MAF) for the Fairfax genes from Fairfax et al. [2012].

Sequencing is carried out on many short sequences of DNA, rather than the whole length. Hundreds or thousands of copies of the DNA are produced using PCR prior to sequencing. The shorter the reads that are sequenced, the more accurate the sequence. However, accurately mapping the short sequences is more difficult than with longer reads due to multiple matches. Sequencing is often measured using fluorescence released when a base binds to the reference sequence. Nucleotides are washed through the sample one base at a time, and fluorescence measured.

Oligonucleotide microarrays are where short single strand DNA molecules (oligonucleotides) are spotted onto a microarray containing synthetic oligonucleotide probes. Genes are usually represented by a probeset, a set of multiple probes. The probeset is designed to map to either a specific region of the transcript targeted or to any non-specific coding region of the genome. Only a single sample can be measured per chip. Gene expression is quantified as the strength of the hybridisation between the reference and sample which is given by the fluorescence emitted by the sample when illuminated with a laser post hybridisation.

Illumina Genotyping BeadChip Arrays use the same bead array concept as for gene expression arrays. BeadChip technology relies on the attachment of oligonucleotides (short, single strand DNA molecules) to silica beads. The beads are then deposited into wells. The proportion of binding of the samples is measured by the released colour. This is used to define the genotype.

2.3 Gene Expression Quantification

Gene expression technologies do not read base-by-base, they estimate the quantities of expressed genes present, via quantification of light emission, on a log_2 scale. Expression profiling is used as a broad title to cover all methods of measuring gene expression, excluding RNA-seq which is a relatively new technique applied to the sequencing of RNA.

Microarrays use an array containing many DNA samples with attached fluorescent probes. The expression levels of hundreds or thousands of genes within a cell can be measured using the amount of fluorescence released as the mRNA binds in each spot on the array. Using computer packages designed for analysing the output intensities of microarrays, a gene expression profile for each sample can be produced using the annotation file for each probe.

Illumina gene expression BeadChips use 79-base oligonucleotides targeting particular genes. Other arrays and platforms use different length sequences. The gene specific probes are attached to the beads on the arrays. As the required gene sequence binds with these specific probes, fluorescence is released. The intensity of this is measured by a scanner and recorded and decoded to provide abundance or gene expression level.

Affymetrix produce three main types of arrays. The main differences between the $3'$ IVT (in vitro transcription) arrays, the HTA (human transcriptome arrays) and the Exon arrays are their preparation and targets.

All Affymetrix Arrays measure the fluorescence intensity output from scanning the binding intensities of the probes within the array. There are 11 probes in each probeset, and these work in pairs; one Mismatch (MM) and one Perfect match (PM) probe per probeset. The MM probe has the complementary base at the 13^{th} position (the middle of the 25 base long probes) and so is expected to bind less well than the PM probe, giving a quantification of background binding and spurious hybridisation. When the probes bind to the RNA they release a fluorescence that is attached to the RNA being screened. It is the intensity of this fluorescence that is used to quantify the gene expression level.

RNA-seq is also known as whole transcriptome shotgun sequencing. It uses the capabilities of the Next Generation Sequencing (NGS) technologies to provide a snapshot of RNA presence and quantity at a given time in a given location on the genome. The quantity and presence of RNA is constantly evolving dependent on the process the cell is undergoing at that time.

RNA-seq is flexible in that it can use mRNA transcripts, total RNA and many types of small RNA including miRNA and tRNA. This allows different snapshots of an individual to be taken. RNA-seq cannot provide a single value to represent the gene expression level as microarrays do. RNA-seq provides a count of each transcript and isoform present. This represents a problem in the type of analysis we will perform as we require the data to be summarised by a single gene expression value. Tools are currently in development for this but are yet to be widely used.

2.4 Processing Raw Genotyping/Sequence data

Post sequencing, we have reads of particular lengths based on the sequencing platform used. These need to be mapped and aligned to the reference genome. The latest human reference genome at the start of this project was hg19 and was released in February 2009. The version of the reference genome used is important as mutations get renamed and added between versions.

Shorter reads are more accurate but are harder to map uniquely due to repetition in the genome. Errors during mapping and alignment can lead to false findings that cannot be replicated.

The amount of sequencing is measured by the coverage. ‘X’ (pronounced times) coverage tells you the number of reads that cover each base and ‘%’ coverage tells you the overall proportion of the genome that has been sequenced.

2.5 Analysis of gene expression data

Initial analysis of gene expression data consists of simple analysis methods designed to help understand and describe the data. After initial analysis has been carried out, techniques can be applied to investigate features of the data that have been highlighted.

Differential expression is the first stage of analysis to explore gene expression data when comparing case-control or two groups (e.g. two extreme phenotypes) of gene expression values. It is the difference between the $\log_2$ gene expression values. This quantity is also known as the fold change. Differentially expressed genes can help to understand the mechanisms of disease, assuming that SNPs affect gene expression which affects phenotype.

The significance of differential expression can be calculated using statistical tests such a t-tests or equivalent and analysing the p-value or q-value. The q-value is the false discovery rate (FDR) equivalent of the p-value. For an individual hypothesis test, the q-value is the minimum FDR at which the test is defined to be significant.

Clustering analysis is a mathematical tool for grouping objects with others that are more closely related to one another. Eisen et al. [1998] were the first to apply clustering techniques to gene expression data. The use of hierarchical clustering techniques allows grouping of genes with similar expression patterns. A dendogram is used to represent the correlation between fold changes of genes and individuals, showing at which point the genes were clustered. To visualise the results of clustering, heatmaps are used. Heatmaps are a tiled array of genes and individuals, the colour in the blocks is scaled to represent differential expression (positive and negative). Analysis of clustering and of correlation of genes and fold changes helps with understanding the mechanism of disease.

Pathway analysis is a mapping of genes onto a signalling network map that defines how gene expression signals are modulated and/or regulated. PANTHER [Thomas et al., 2003] and DAVID [Huang et al., 2008] are two popular tools for this. Understanding the signalling pathway that genes are involved in can help to identify causes of disease.

A pathway can be either genetic or biological. A genetic pathway marks the interactions between groups of genes that depend on each others functions in order to function themselves. A biological pathway, as used by PANTHER, is a set of actions between molecules that can lead a cell to reproduce or change in a certain way. Pathways can lead to the construction of new molecules, the switching on and/or off of genes or by encouraging a cell to move to another part of the body by interfering with the signals sent from cell to cell via the biological pathways. An example of published work using pathway analysis can be found in Emmert-Streib and Glazko [2011].

2.6 Computation, Data Processing, Data manipulation.

When using any genetic data it is often very computationally and time intensive to reformat the data from its original format to a usable format. In the case of the eQTL data, we have to maintain a match between the SNP and expression data as well as reformatting the data.

Before beginning the process of matching the gene expression and SNP data both need to be annotated. For gene expression, we need to annotate the probe names so that each gene is referred to by name. It is also helpful if the location of the gene can also be obtained. Annotating gene expression data from arrays is relatively simple. With each gene expression tool there is an annotation file. This is a file that contains information for each probe, i.e. each row has information on the probe name, the associated gene name, the location of the gene, any other names, current known function as well as many other pieces of information. Annotating SNPs is much less simple. We are often only provided with an rs number and the two alleles obtained after sequencing. In order to carry out any analysis we need to know the reference alleles, and it is often helpful to know the location (position and gene) for each SNP. There are online tools that carry out this analysis but these are very restricted in the number of SNPs (rs numbers) that can be included at any one time. When analysing SNPs across the whole genome this will take a considerable amount of time. I found the most efficient way to do this was to download the HapMap data and use UNIX to match the two files.

Once the data has been processed, as above, we reformat the data for analysis. To begin with we match the pairs of expression and SNPs to an individual. Where there are multiple gene expression values from varying probes or repeats, we take the maximum of the values. This is an area where there is no definitive answer [MAQC Consortium, 2006], [Miller et al., 2011]. For us, we chose to take the maximum of multiple values as all probes are designed to bind to a specific tissue, hence we select the 'best' binding which should occur in the 'correct' tissue. Hence, where choosing the mean can be misleading, choosing the maximum value represents the signal in the tissue we want to measure, according to the binding strength. Different probes can be measuring expression levels for different tissues which can lead to very different expression values due to suitability of the data to the chip. To overcome the uncertainty associated with the mean and non specific binding, we take the maximum expression over all values.

We next recode the SNPs as $\{0, 1\}$ or $\{0, 1, 2\}$. For ease of interpretation we use the $\{0, 1\}$ coding. This allows simple comparisons between the effect sizes of each SNP for both the Hulse and the Yeast dataset. In the Fairfax dataset where there is clear evidence that the allele count affects gene expression, we use the additive mode of inheritance $\{0, 1, 2\}$ rather than the dominant $\{0, 1\}$ coding.

In most datasets there will be missing SNPs for some individuals. In order to use the data most effectively the missing SNPs need to be imputed. To impute missing data using Impute2 [Howie et al., 2009], we initially remove any individuals or SNPs with $> 5\%$ missing data. A limit of 5% was chosen as it reduces the amount of missing data to impute, hence increasing the accuracy of the imputed data. The info scores of the imputed SNPs increased and became more usable once the poor quality individuals and SNPs were removed.

To use Impute2 [Howie et al., 2009] we have to reformat the data we have and also generate other reference files for the software, a legend and a strand file. Post imputation, we keep only the imputed SNPs with an info score greater than

0.3. This is the lower end of the range of info scores used in publication according to the Impute2 documentation. Any SNPs which have not been fully imputed due to poor quality scores are then discarded. We treat the genotype of our imputed SNPs as the probability of the SNP being not homozygous wildtype, i.e. coded as 1 in $\{0, 1\}$ coding. We scale and centre the newly imputed SNP matrix prior to running the Normal Gamma.

Data processing and manipulation plays a key role in any analysis. With genetics data, the quantity and complexity of the data makes this a much more onerous process than in other fields. Here we have tried to convey a small amount of the processing that has taken place to obtain the results in this thesis.

2.7 The Functional Significance Score

We use the Functional Significance (FS) score to represent the functional information that we will include in the Normal Gamma framework in Chapter 7 onwards. We use this score as a robust summary of the deleterious effects of different SNPs. In this section we give the details on how the score is calculated.

The FS score [Lee and Shatkay, 2009] is a score which combines information on the deleterious effects of SNPs from 16 publicly available web services and databases. The deleterious effect of a SNP is recorded in $\delta_{i,j}$ for each element in $\mathbb{F}$, where

$$\mathbb{F} = \{\text{protein coding, splicing regulation,transcriptional regulation, post-transcriptional regulation}\}.$$

and $i = 1, \ldots, p$ is the SNP number, $j = 1, \ldots, q$ is the tool number. As not every tool examines all four features in the set $\mathbb{F}$, we define $F_{j,k} = 1$ if tool j examines the deleterious effect of feature k ($k \in \{1, 2, 3, 4\}$).

Each tool is given a reliability score (TR), which is calculated from

$$TR_j = Pr(Y_i = 1|\delta_{i,j} = 1),$$

where $Y_i = 1$ if SNP i is deleterious and 0 otherwise, and $\delta_{i,j} = 1$ when tool j predicts SNP i to be deleterious.

The confidence score given by the tool for each $\delta_{i,j}$ is recorded as $S_{i,j}$. $S_{i,j}$ are not measured on the same scale by the different tools and can therefore be difficult to interpret. Lee and Shatkay [2009] propose normalising this confidence score to $[0,1]$ using

$$\bar{S}_{i,j} = \frac{1}{2}\left(\delta_{i,j} + (1 - C_{i,j})\frac{(S_{i,j} - \min_i S_{i,j})}{(\max_i S_{i,j} - \min_i S_{i,j})}\right),$$

where

$$C_{i,j} = \begin{cases} 1 & \text{if } X_i \text{ resides on a non-conserved regulatory site} \\ 0 & \text{otherwise.} \end{cases}$$

$C_{i,j}$ is included here to take into account whether the region is conserved or not. A highly conserved region will often contain important functional genetic information. Its function is also often better understood as it is easier to characterise. The normalised confidence score, denoted $\bar{S}_{i,j}$, ensures all confidence scores are within the $[0,1]$ range. If a SNP is predicted to be deleterious, the score is normalised to $(0.5,1]$, and if it is not predicted as deleterious the score is normalised to $[0,0.5)$. A score of 0.5 represents uncertainty about the deleterious nature of the SNP.

The FS score combines all of this information regarding the confidence of the deleterious effect, or otherwise, into one FS score for each SNP i which is calculated from

$$FS_i = \max_{k\in\mathbb{F}} \frac{\sum_{j=1}^{q}(F_{j,k})(TR_j)(\delta_{i,j})(\bar{S}_{i,j})}{\sum_{j=1}^{q}(F_{j,k})(TR_j)}. \tag{2.1}$$

Lee and Shatkay [2009] have produced an online resource containing FS scores for 112,949 SNPs, of which 1,399 are known to be disease related. It is this complete set of FS scores for all SNPs that we obtain from the authors, that we use to define our functional information.

2.8 Conclusion

In this chapter we have described the datasets that are going to be used throughout this thesis, as well as briefly commenting on the different types of data production that have been used to produce the data. Using published data, we could not influence the data production techniques used, and so we have to assume the best techniques were used when all factors, including cost, were taken into account. Having a brief understanding of the different techniques used to produce the data is essential for understanding the best ways to model and analyse the data. It also allows the user to obtain the maximum information from the datasets.

Chapter 3 searches the literature for statistical methods that, taking into account the information in this Chapter, may perform well when trying to detect causal sequence variants from eQTL datasets.

Chapter 3

Review of shrinkage inducing statistical methods for eQTL mapping

In this chapter we describe statistical methods that can be used to shrink parameter estimates towards 0 (partially or completely) in a linear model framework, and review methods that allow functional information to be incorporated. Shrinkage is essential for eQTL analyses as biologically, we know that many SNPs have little to no effect on gene expression. We consider linear models that relate gene expression of a particular gene to the genotypes of a set of SNPs, often in a given genomic region of different types, for example exonic/intronic/intergenic SNPs. Crucially, these models will readily allow for inclusion of functional information into the model.

To characterise functional information, we use the FS score [Lee and Shatkay, 2009] described in Section 2.7, page 24. This combines functional information, such as protein coding, splicing regulation, transcriptional regulation and post-transcriptional regulation, focusing on the deleterious effect of individual SNPs from multiple on-line resources. The functional effects of each SNP are then combined using weights which reflect the importance of the feature and reliability of the on-line resource. The resultant score is constrained to $[0, 1]$ where 0 represents no deleterious effect, 1 represents a highly deleterious effect and 0.5 represents no knowledge.

We use a standard linear model to represent gene expression for simplicity of modelling, even though gene expression regulation is non-linear in the tails due to saturation, particularly when measuring gene expression using arrays. Other models could be used to better represent the true distribution of gene expression values, however for modelling purposes, a linear model framework provides adequate results in relation to the simplification of the modelling process.

In this chapter we assess different techniques for estimating the regression coefficients in the following gene-by-gene linear model:

$$y_j = \alpha + \beta_1 X_{1,j} + \ldots + \beta_i X_{i,j} + \ldots + \beta_p X_{p,j} + \epsilon_j, \tag{3.1}$$

where y_j is the gene expression value for individual j, α represents the background gene expression (which is gene specific), β_i is the effect size for SNP i and $X_{i,j}$ represents SNP i for individual j and ϵ represents the noise term. This model forms the foundation for all the following techniques.

3.1 Shrinkage

Shrinkage occurs naturally within a Bayesian context because the posterior distribution combines information from the likelihood and the prior distribution. Careful choice of the prior distribution, with a lot of mass near to zero, can induce sparse statistical models. In the context of eQTL data, we know many SNPs have no, or negligible effects on gene expression, so shrinkage of some parameter estimates towards 0 is desirable.

Penalisation of the model parameters is used to reduce the model complexity, often by shrinking parameter estimates towards 0. Outside of the Bayesian framework, penalisation is used to induce shrinkage in parameter estimates. Historically the mean squared error, $(\mathbf{y} - \hat{\mathbf{y}})^2$, i.e. the squared difference between the observed and the estimated values, was used to penalise parameters. This approach does not lead to sparse models. More recently, with the advent of increasingly large datasets, L_1 and L_2 penalised regularisation methods have been used as they shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. See Wu et al. [2009], Yi and Xu [2008] and Ma et al. [2007] for examples of using L_1 penalised regularisation methods in a genetics context, and Malo et al. [2008], Piepho [2009] and Whittaker et al. [2000] for examples of using L_2 penalised regularisation methods in a genetics context.

The two key penalisation terms, L_1 and L_2, are defined as follows.

$$L_1 = ||\beta_1|| = \sum_{i=1}^{n} |\beta_i| \tag{3.2}$$

$$L_2 = ||\beta_2||^2 = \sum_{i=1}^{n} \beta_i^2. \tag{3.3}$$

Penalisation based on the L_1 norm drives many parameters to zero, however

L_2 norm penalisation is not specifically designed to achieve sparsity. The L_1 norm has a rotated hyper-cube (diamond in 2D) shaped penalty with vertices on the axes. This means that the likelihood contours often first intersect the penalty contours at the axes. This geometrical property forces many parameter estimates to be 0. The hyper-sphere (circular in 2D) penalty of the L_2 norm does not have the same geometric properties as the L_1 norm, meaning that the likelihood contours and the penalty contours are less likely to intersect on the axes. As a result, the L_2 norm does not frequently force parameter estimates to 0. Details of the regularisation can be found in Ng [2004] with a clear graphical representation in Shi et al. [2013] and Fu [1998].

3.2 Statistical approaches to modelling eQTL data without including functional information

We can divide the statistical approaches that can be used to model eQTL data without functional information into two categories; models that use a fully Bayesian approach (MCMC) and methods that use a MAP estimation approach, reporting only the posterior mode. Within the fully Bayesian approach, there are two categories of priors; variable selection priors which use priors with point masses at 0 (which can induce strict sparsity), and other continuous shrinkage priors with a lot of mass near to 0. The latter shrink many parameter estimates close to, but not equal to zero.

Bayesian continuous shrinkage prior distributions tend to have a sharp mode at 0, with the mass in the tails influencing the amount of shrinkage applied to large estimates. Variable selection methods use an indicator variable for each parameter to select variables to include in the model.

3.2.1 Fully Bayesian approaches

There are four fully Bayesian approaches that we describe here; piMASS [Guan and Stephens, 2011], Spike and slab [Ishwaran and Rao, 2005], the Normal Gamma [Griffin and Brown, 2010] and the Bayesian Lasso [Leng et al., 2014]. Having described these methods we compare and contrast them.

piMASS

piMASS [Guan and Stephens, 2011] is a form of Bayesian Variable Selection Regression that reports the posterior probability of inclusion (having a non-zero regression coefficient) as its measure of association as well as estimating the regression coefficients/effect sizes conditional on being in the model. piMASS uses a standard linear regression set-up

$$\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \tau, \boldsymbol{\beta}, \mathbf{X} \sim N_n(\boldsymbol{\mu} + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \tau^{-1} I_n), \tag{3.4}$$

with $\mathbf{y}$ an $n \times 1$ vector of gene expression values, X an $n \times p$ matrix of genotypes, $\boldsymbol{\beta}$ the vector of effect sizes (the regression coefficients), $\boldsymbol{\mu}$ the background gene expression, γ is the binary indicator variable defining inclusion in the model and τ the precision parameter. The hierarchical structure of the prior distributions, as defined in Guan and Stephens [2011] are:

$$\tau \sim Ga\left(\frac{\lambda}{2}, \frac{\kappa}{2}\right), \tag{3.5}$$

$$\mu|\tau \sim N\left(0, \frac{\sigma_\mu^2}{\tau}\right), \tag{3.6}$$

$$\gamma_j \sim \text{Bernoulli}(\pi), \tag{3.7}$$

$$\boldsymbol{\beta}_\gamma|\tau, \boldsymbol{\gamma} \sim N_{|\gamma|}\left(0, \frac{\sigma_a^2}{\tau} I_{|\gamma|}\right), \tag{3.8}$$

$$\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\gamma} \sim \delta_{\beta=0}, \tag{3.9}$$

where $|\gamma| := \sum_j \gamma_j$, $\delta_{\beta=0}$ represents the point mass at 0, and $\boldsymbol{\beta}_{-\gamma}$ denotes the β coefficients for which $\gamma_j = 0$. Hyperparameters are defined in Guan and Stephens [2011] to induce sparsity

$$\log(\pi) \sim U\left(\log\left(\frac{1}{p}\right), \log\left(\frac{M}{p}\right)\right) \tag{3.10}$$

and the typical size of non-zero coefficients (σ_a^2) is defined as

$$\sigma_a^2(h, \gamma) = \frac{h}{1-h} \frac{1}{\sum_{j:\gamma_j=1} s_j}, \tag{3.11}$$

where s_j is the variance of covariate/SNP j and h is an approximation to the proportion of variance explained.

Equation 3.10 is chosen such that the upper and lower limits of π correspond to an expectation of 1 and M SNPs being included in the model. M is defined as 400 in Guan and Stephens [2011] due to computational restrictions. Note that $M \leq p$ as p is the total number of SNPs, and M represents the number of SNPs we expect to be included in the model.

piMASS uses Metropolis-Hastings updating to explore the parameter space when estimating the posterior mean effect sizes. At each iteration of the MCMC, addition, deletion or switching (of nearby SNPs only) is applied to a single SNP in the dataset. This updates the set of SNPs included in the model at each iteration. Proposing local moves with occasional large steps improves convergence, while focusing on SNPs with strong marginal associations helps select which SNPs to include into the model. The estimates of β conditional on inclusion and the posterior inclusion probability (PIP) are calculated where possible using the Rao-Blackwellisation technique to reduce the Monte Carlo variation. The Rao-Blackwellisation technique calculates the expectation of an estimator ($P(\gamma_j = 1|\mathbf{y})$ in this case) conditional on other parameters ($\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, τ, h and π, excluding the parameter values that correspond to j in this case). The Rao-Blackwell theorem states that the conditional expectation is typically a better estimator and is never a worse estimator as it is optimal by the mean-squared-error or other similar criteria, which leads to a reduction in the sampling variance in comparison to counting the proportion of $\gamma_j = 1$ in the MCMC chain. Full details of the general application can be found in Casella and Robert [1996] and McKeague and Wefelmeyer [2000] with details specific to piMASS found in section 3.1 of Guan and Stephens [2011].

The default priors for piMASS are believed to be suitable for general use in most GWAS applications. We therefore use the default parameters of piMASS for comparison. It is noted that piMASS performs poorly when there are multiple correlated SNPs that are far apart on a chromosome. This is because piMASS favours the inclusions and exclusions of SNPs that are close together on the chromosome due to the general LD structure of chromosomes.

Given that we believe many SNPs have little to no effect, a selection method seems appropriate for this type of dataset.

Spike and slab

The Spike and slab prior was initially proposed by Mitchell and Beauchamp [1988] and involves designing a hierarchy of priors over the parameter and model space that selectively shrink only those effects that are near to zero. It has been suggested by Ishwaran and Rao [2005] that, in order to prevent the likelihood from swamping the prior information, the responses should be scaled by a factor of $\sqrt{n}$ to ensure that the effect of the prior is visible in the posterior estimates.

The standard hierarchical set-up for the Spike and slab prior can be seen in Equations 3.12 and 3.13.

$$y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2) \qquad \text{for } i = 1, \dots, n \tag{3.12}$$

$$\beta|\gamma \sim N_p(\mathbf{0}, \Gamma) \qquad \text{where } \Gamma = \text{diag}(\gamma_1, \dots, \gamma_p). \tag{3.13}$$

There are also prior distributions on γ and σ^2. The shrinkage is determined by the hypervariances γ_i; small hypervariances shrink coefficients towards zero.

There are many different adaptations of the Spike and slab prior that use different prior distributions. Ishwaran and Rao [2005] adapt the standard set-up above to use continuous bimodal priors as well as rescaling the Spike and slab prior to make it sample size invariant, see Equations 3.14-3.17. In this set up, $y_i^* = \frac{y_i}{\sqrt{n}}$.

$$\begin{aligned} y_i^*|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2 n) && \text{for } i = 1, \ldots, n && (3.14)\\ \beta_j|J_j, \tau_j^2 &\sim N(\mathbf{0}, J_j\tau_j^2) && \text{for } j = 1, \ldots, p && (3.15)\\ J_j|v_0, w &\sim (1-w)\delta_{J_j=v_0} + w\delta_{J_j=1} &&&& (3.16)\\ \tau_j^{-2}|a_1, a_2 &\sim Ga(a_1, a_2) &&&& (3.17)\\ w &\sim U(0,1) &&&& (3.18)\\ \sigma^2 &\sim Ga(b_1, b_2), &&&& (3.19) \end{aligned}$$

where $b_1 = b_2 = 0.0001$, and a_1, a_2 are chosen such that the bimodal prior has peak at v_0, a small value close to 0, with a right continuous tail. In order to find the posterior estimates of this, Ishwaran and Rao [2005] advise exploiting conjugacy of the prior distributions and the likelihood to use the Gibbs updating algorithm.

We use the traditional Spike and slab selection method that use a mixture prior distribution consisting of the Normal distribution and point mass at 0, as described in Equations 3.12 and 3.13. We use the Spike and slab method in our analysis as it is possible, with the implementation we are using, to specify the prior inclusion probability of each SNP. We also select this method due to the flexibility regarding changing which of the SNPs are included at each iteration of the algorithm. We assess the number of SNPs we expect within a given dataset to have non-zero coefficient and set the prior inclusion probability accordingly. For our simulation results, we use the prior inclusion probability of 0.05 for all SNPs unless otherwise stated.

The Normal Gamma prior

The Normal Gamma (NG) prior [Griffin and Brown, 2010] uses the principles of the Spike and slab [Mitchell and Beauchamp, 1988] to shrink parameter estimates in a similar method to the Bayesian Lasso [Park and Casella, 2008]. The double exponential prior formation of the Bayesian Lasso is a special case of the Normal Gamma prior. Griffin and Brown [2010] propose a hierarchical structure for the parameters β_i which, conditional on ψ_i are assumed to originate from a Normal distribution with variance parameter ψ_i. This variance is assumed to follow a Gamma distribution with hyperparameters λ and γ. The parameters are updated over time using a Metropolis-Hastings within Gibbs Sampler approach. All details of this method are explained in detail in Chapter 5.

The parameters of the Normal Gamma are all heavily inter-related; there is not one single parameter that directly controls the shrinkage. The marginal prior variance of β, $2\lambda\gamma^2$ has expectation, defined by Griffin and Brown [2010], to be M. M represents an empirical estimate of the variance of the least squares estimates. This is used to determine the variability of the β's, hence this value controls the majority of the shrinkage in the NG.

We state the prior hierarchy as follows, in Equations 3.20-3.23, with uninformative priors on α and σ^2.

$$\pi(\lambda) \sim Ex\left(\frac{1}{2}\right) \tag{3.20}$$

$$\pi(\gamma^{-2}|\lambda) \sim Ga\left(2, \frac{M}{2\lambda}\right) \tag{3.21}$$

$$\pi(\psi_i|\lambda, \gamma^{-2}) \sim Ga\left(\lambda, \frac{1}{2\gamma^2}\right) \tag{3.22}$$

$$\pi(\beta_i|\psi_i) \sim N(0, \psi_i) \tag{3.23}$$

where $\pi(a)$ represents the prior distribution on a and M is a fixed scalar defined as $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$ where $\hat{\boldsymbol{\beta}}$ is the least squares (LS) estimate of $\boldsymbol{\beta}$ when X is non-singular. When X is singular, or when $p > n - 1$, M is redefined as $\frac{1}{n}\sum_{i=1}^{p} \hat{\beta}_i^2$, where $\hat{\beta}$ is the minimum length least squares (MLLS) estimate, see Section 3.4.1.

With the response variable defined as $\mathbf{y} = (y_1, \ldots, y_n)$, the likelihood is defined as follows in Equation 3.24.

$$f(\mathbf{y}|\lambda, \gamma^{-2}, \boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}) \sim N_n(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta}, \sigma^2\mathbf{I_n}), \tag{3.24}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the parameter vector representing the effects of genetic variants, and N_n represents the multivariate normal (MVN) of dimension n.

The key property of the Normal Gamma is that the prior structure is adaptive in the sense that it has fatter tails than the Lasso/Bayesian Lasso priors, and so it shrinks larger parameter values less than smaller ones. This makes is highly suited to eQTL data, even though it is reportedly computationally and time expensive.

The Bayesian Lasso

The Bayesian Lasso [Park and Casella, 2008] uses a conditional Laplace prior and a non-informative, scale invariant marginal prior on the variance parameter. The full Bayesian Lasso uses MCMC and can be represented by the following hierarchical form:

$$\begin{aligned}
\mathbf{y}|\mu, \mathbf{X}, \beta, \sigma^2 &\sim N_n(\mu + \mathbf{X}\beta, \sigma^2\mathbf{I}_n),\\
\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N_p(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau),\\
\mathbf{D}_\tau &= diag(\tau_1^2, \ldots, \tau_p^2),\\
\tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} exp\left(\frac{-\lambda^2\tau_j^2}{2}\right),\\
\tau_1^2, \ldots, \tau_p^2 &> 0,
\end{aligned}$$

where X represents the genotype matrix, and $\mathbf{y}$ the gene expression vector as previously defined for piMASS. μ is given an independent, uniform prior and σ^2 a non-informative, scale invariant prior.

One challenge with the Bayesian method is the choice of λ. Park and Casella [2008] suggest using one of two methods, an empirical Bayes method, or selecting an appropriate hyper prior for λ.

Park and Casella [2008] find that for the diabetes dataset they use, there is good comparability between the results of the Lasso and the Bayesian Lasso, although the Bayesian Lasso is more computationally expensive.

Comparison of the fully Bayesian methods

piMASS and Spike and slab both use variable selection through indicator variables. This gives truly sparse models. In eQTL datasets, sparsity is key to truly reflecting the knowledge that only a few SNPs affect gene expression. piMASS applies selection methods to the data. Using proximity of the SNPs and their marginal associations with respect to gene expression, piMASS selects a different SNP set in every iteration. Spike and slab uses no such assumptions for SNP inclusion. The variables/covariates are included based on random sampling and previous inclusion results.

The Bayesian Lasso is a special case of the Normal Gamma. The Normal Gamma prior with $\lambda = 1$ is equivalent to the Bayesian Lasso. This means that the Normal Gamma is a generalisation of the Bayesian Lasso.

Comparing across the selection methods, piMASS and Spike and slab, and the shrinkage methods, the Normal Gamma and Bayesian Lasso, we notice strong similarities of the prior hierarchies, with the only noticeable difference being the inclusion or omission of an indicator variable. This highlights the fundamental Bayesian prior structure for variable selection/shrinkage in a linear model framework.

In this eQTL setting, we highlight in particular that piMASS takes into account features of genetics data such as proximity of SNPs to one another - a feature that is highly likely to affect association given the strong LD found in the human genome. This should give piMASS an advantage over other methods for detecting causal SNPs when they are in high LD blocks, but might limit its success with identifying trans-acting SNPs.

3.2.2 MAP estimation approaches

We describe two closely related MAP estimation techniques; HyperLasso [Hoggart et al., 2008] and the Lasso [Tibshirani, 1996]. We then investigate these methods for similarities and differences.

HyperLasso

HyperLasso [Hoggart et al., 2008] is a Bayesian approach that determines MAP (maximum-a-posteriori) estimates of the parameters rather than sampling from the full posterior distribution. It uses variable selection in a logistic regression model with each covariate representing a SNP. The prior distribution of the HyperLasso, the NEG (Normal Exponential Gamma), is a continuous distribution with a sharp mode at zero and can have heavy tails. It is a generalisation of the double exponential (DE) distribution prior. The marginal prior for $\boldsymbol{\beta}$ is given by

$$NEG(\boldsymbol{\beta}|\lambda,\gamma) = \int_0^\infty \int_0^\infty N(\boldsymbol{\beta}|0,\sigma^2)Ga(\sigma^2|1,\psi)Ga(\psi|\lambda\gamma^2)d\sigma^2 d\psi \tag{3.25}$$

$$= \kappa \mathrm{exp}\left(\frac{\boldsymbol{\beta}^2}{4\gamma^2}\right) D_{-2\lambda-1}\left(\frac{|\boldsymbol{\beta}|}{\gamma}\right), \tag{3.26}$$

where D is the parabolic cylindrical function and κ is the integrating constant. The MAP estimation using the posterior mode sets some of the regression coefficients to 0. The prior distribution leads to increased shrinkage on parameters close to 0 with minimal shrinkage on those variables that are selected by the model. This effect is caused by the heavy tails of the NEG prior distribution, similarly to the NG prior. The MAP estimates depend on the initial values of the two prior hyperparameters.

The posterior of the HyperLasso is not always unimodal, especially in the $n < p$ case, and the order in which the coefficients are updated also affects the

MAP estimate, particularly in the case of highly correlated SNPs. Multiple runs of the algorithm are used to explore the possibility of multiple posterior modes.

HyperLasso is widely used on and has the default parameters set to maximise performance on GWAS data. eQTL data has a continuous normal response variable, gene expression, whereas GWAS data has a binary response representing phenotype/disease status.

HyperLasso is a sparsity-inducing selection method which makes it a good method for this type of data. The computational efficiency and tailoring of this model for genetics data make it invaluable. We note that the program cannot be used with genotype values that are different to 0, 1, or 2.

The difficulty in using the HyperLasso (HL) is in the choice of hyperparameters. The results are highly sensitive to the chosen hyperparameter values. The authors provide some guidance about the choice of parameter values, but this is limited to the case-control setting. We experiment with different values of the parameters, but to maximise the AUC of the ROC curve we use the default parameters and specify the shape parameter to be 0.1. We tested a range of values for the shape parameter from 100 to 0.001 and we found the results to be similar. Using these parameter values does not make the most of the selection property of the HL.

Lasso

The Lasso [Tibshirani, 1996] estimates linear regression coefficients through an L_1 constrained least squares penalisation to achieve the following minimisation:

$$\hat{\beta}_L = \text{argmin}_\beta \left\{ (y - X\beta)^T (y - X\beta) + \lambda \sum |\beta_j| \right\}. \tag{3.27}$$

for some $\lambda \geq 0$ and where the columns of X are standardised.

The Lasso uses the double exponential prior on the p regression coefficients, defined as

$$\pi(\beta|\tau) = \left(\frac{\tau}{2}\right)^2 \exp(-\tau \sum |\beta|), \tag{3.28}$$

and the likelihood is defined as

$$f(y|\beta, \sigma^2) \sim N_n(y|X\beta, \sigma^2 I_n). \tag{3.29}$$

The posterior for β is

$$\begin{aligned} P(\beta|X,y,\tau,\sigma^2) &= \left(\frac{\tau}{2}\right)^2 \exp(-\tau \sum|\beta_i|) \\ &\quad \times \frac{1}{2\pi}(\sigma I_n)^{-1}\exp\left(-\frac{1}{2}(y-X\beta)^T(\sigma^2 I_n)^{-1}(y-X\beta)\right) \\ &= \left(\frac{\tau}{2}\right)^2 \frac{1}{2\pi\sigma}\exp\left(-\tau\sum|\beta_i| - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right) \\ &\propto \exp\left(-\tau\sum|\beta_i| - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left\{2\tau\sigma^2\sum|\beta_i| + (y-X\beta)^T(y-X\beta)\right\}\right) \end{aligned}$$

Take the logarithm

$$\begin{aligned} \log\left\{P(\beta|X,y,\tau,\sigma^2)\right\} &= -\frac{1}{2\sigma^2}\left\{2\tau\sigma^2\sum|\beta_i| + (y-X\beta)^T(y-X\beta)\right\} \\ &\propto \lambda\sum|\beta_i| + (y-X\beta)^T(y-X\beta). \end{aligned} \tag{3.30}$$

So the LASSO estimates can be interpreted as the posterior mode estimate when the regression parameters all have independent, identically distributed Laplace (double-exponential) priors.

Comparison of MAP estimation approaches

The Lasso is a special case of the HyperLasso. The HyperLasso has the more flexible NEG prior compared to the constrained Laplacian (or double exponential) prior of the Lasso. This means the HyperLasso is a generalisation of the Lasso, and so we only use the HyperLasso to compare this approach.

3.3 Statistical approaches to modelling eQTL data including functional information

In this section we focus on two statistical methods (Lirnet [Lee et al., 2009] and SBFA [Parts et al., 2011]) that include functional information to inform the parameter estimates; we describe them in detail and then compare and contrast them. We choose to assess only these two methods as they cover two very different methods of incorporating functional information but both within the regression framework.

Lirnet

When carrying out an eQTL study, the key challenge is that the number of candidate regulators, often highly correlated, is enormous relative to the amount of available data. This makes robustly identifying the correct regulator very difficult. For this reason, Lee et al. [2009] split the data into two categories, g-regulators and e-regulators. These are defined as follows.

- The value of a g-regulator or genotype regulator represents genetic mutations (SNPs) on a chromosomal region. This is used to denote homozygous wildtype, heterozygous or homozygous SNP at the genetic location. We denote this $f_{n,k}$ for SNP n and feature k.

- The value of an e-regulator or expression regulator is the expression level of genes that are known to have regulatory roles; the values represent the activity levels or the size of the effect of genes that might regulate a module. An e-regulator is denoted $g_{r,k}$ for regulator r and feature k. A module is a cluster of genes where it is assumed that the expression of the genes within the module are all governed by the same regulatory program. The e-regulator takes into account that the expression of one gene can have a regulatory effect on the expression of a subsequent gene.

Lirnet is a complex algorithm that can be summarised as follows. The following stages are repeated iteratively until convergence, as demonstrated in the pictorial representation in Figure 3.1. Throughout this section explaining Lirnet, it may be helpful to refer back to Figure 3.1 to clarify the order of the stages of the algorithm.

1. Use the current estimates of the regulatory priors for features k, α_k and β_k, to calculate C_R from $C_R = C_1 \times P(\text{regulator } r \text{ is causal}) + C_0 \times (1 - P(\text{regulator } r \text{ is causal}))$, where $P(\text{regulator } r \text{ is causal})$ is a non-linear function of α_k and β_k, and C_1 and C_0 are defined by cross-validation.

2. Use the minimisation term, Equation 3.32, to estimate the parameters $w_{m,r}$, with α_k and β_k fixed. This means that the residual sum of squares term, as well as the L_1 and L_2 terms are all penalising the choice of parameters $w_{m,r}$ (for module m and regulator r).

3. Fix the parameters $w_{m,r}$ and use the minimisation term, Equation 3.32, to create new estimates for α_k and β_k. Here C_R is written as a non-linear function of α_k and β_k and so is minimised, as well as the θ term. As the $w_{m,r}$ are fixed, the L_1 and θ terms are the only terms involved in this minimisation.

The weighted sum of the parameters, $\sum_k \alpha_k g_{r,k}$ for e-regulators and $\sum_k \beta_k f_{n,k}$ for g-regulators is used to calculate $P(\text{regulator } r \text{ is causal})$, which is in turn

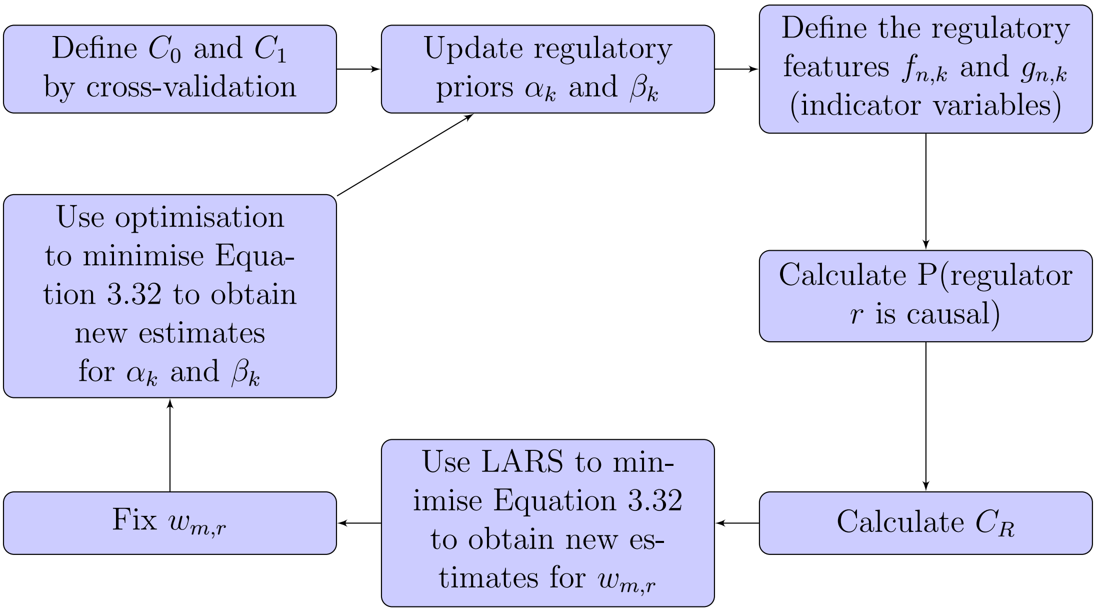

Define C_0 and C_1 by cross-validation
Update regulatory priors α_k and β_k
Define the regulatory features $f_{n,k}$ and $g_{n,k}$ (indicator variables)
Calculate P(regulator r is causal)
Calculate C_R
Use LARS to minimise Equation 3.32 to obtain new estimates for $w_{m,r}$
Fix $w_{m,r}$
Use optimisation to minimise Equation 3.32 to obtain new estimates for α_k and β_k

Figure 3.1: Flowchart showing the Lirnet algorithm.

used to calculate C_R which is a non-linear function of α_k and β_k. C_R is fixed in the minimisation of $w_{m,r}$, but is optimised when minimising for α and β. The model that is being fitted to the data by Lirnet for predicting the gene expression value y is a simple linear combination of potential regulators, x_i, defined as

$$y_{m,g} = w_{m,1}x_1 + w_{m,2}x_2 + w_{m,3}x_3 + \ldots + w_{m,n}x_n + \epsilon_{m,g} \tag{3.31}$$

for gene g in module m, where $\epsilon_g \sim N(0, \sigma^2)$. Lee et al. [2009] include penalisation terms to prevent overfitting of the model, such that the parameters are iteratively estimated according to 3.1 using Equation 3.32.

$$\begin{aligned} \mathit{minimse} \quad & \textstyle\sum_{\text{module } m} \left(\sum_{\text{gene } g} (y_{m,g} - \sum_{\text{regulator } r} w_{m,r} x_r)^2 \right) \\ & + \textstyle\sum_{\text{module } m} \left(\sum_{\text{regulator } r} C_R |w_{m,r}| \right) \\ & + \textstyle\sum_{\text{module } m} \left(D \sum_{\text{regulator } r} w_{m,r}^2 \right) \\ & + E \textstyle\sum_k \theta_k^2, \end{aligned} \tag{3.32}$$

where $\theta = \{\alpha\} \cup \{\beta\}$ and D and E are scalars predefined using 10-fold cross validation.

The minimisation expression, Equation 3.32, contains the standard residual sum of squares and an L_1 and an L_2 penalisation term, as well as a term for the combined L_2 penalisation of the α's and β's. This encourages the model not to overfit and take into account the correlated variables but also to consider carefully which features to "switch on".

We note that D does not need to depend on the regulators r in the same way as C_R, this is due to the L_2 regularisation being used for correlated data, and the functionality of the L_2 regularisation compared to the L_1 regularisation. As the SNPs are highly unlikely to be excluded from the model using the L_2 regularisation, it is not important to force the values as far from zero as possible, and so the weighting of the sum can be outside the sum over the regulators. However, it needs to be inside the sum over the modules as a module is dependant on the number of SNPs inside it. For the C_R weighting inside the sum over the regulators, this is necessary as any SNPs with small parameters are likely to be excluded, and so by multiplying by C_R, which is large when a SNP has a high probability of being causal, we force the value further from zero, therefore reducing the chances of the SNP being excluded from the model. The final coefficient E is simply weighting the amount of penalisation given to the sum of the squares of the elements in the union of the parameters α and β.

The parameters α and β, defined by Lee et al. [2009] as regulatory priors, are used to include prior knowledge to the model in order to improve the prediction ability of the eQTL model. The regulatory priors state the importance given to each of the regulatory features. For simplicity the regulatory features, $g_{r,k}$ and $f_{n,k}$, are taken to be indicator variables.

In order to reduce the computational power required *modules* are created. The *modules* used in Lirnet refer to groups of genes where the expression of the target gene in each module is regulated by the same regulatory program. This means that the genes within a module are highly correlated as they exhibit the same expression and regulatory responses to other regulators, i.e. they are co-expressed and co-regulated. Some of these modules contain multiple genes, while others contain only 1 or 2 genes.

It is possible for a SNP to have many regulatory features that are deemed to be important, this means that there may be many significant parameters β_k. Lirnet uses LASSO to introduce sparsity into the regulatory programs to satisfy the biological requirement that only a small number of regulators r should have non-zero weight. The sigmoid function is introduced to put an upper bound on the $\sum_k \alpha_k g_{r,k}$ for e-regulators and $\sum_k \beta_k f_{n,k}$ for g-regulators. This prevents the saturation effect of regulatory features. The sigmoid or logistic function is defined as $sigmoid(x) = \frac{1}{1+exp(-x)}$.

The probability of a SNP being causal is defined as follows for a g-regulator and e-regulator respectively. It includes the sigmoid function to prevent saturation.

$$P(\text{SNP } n \text{ causes variation in gene expression levels}) = \text{ sigmoid}\left(\sum_k \beta_k f_{n,k}\right). \tag{3.33}$$

$$P(\text{e-regulator } r \text{ is causal}) = \text{ sigmoid}\left(\sum_k \alpha_k g_{n,k}\right). \tag{3.34}$$

If tagSNPs, a SNP representing a region of the genome with high linkage disequilibrium, are used in the model then each SNP is representative of a region rather than a specific base, and so we calculate the probability that a region is causal as in Equation 3.35. Lee et al. [2009] states this is not the only method for aggregating the contributions of all SNPs in a region, but the double sigmoid function prevents an unbounded increase in the regulatory potential, while the multiplication and subtraction scale the output to $[0, 1]$.

$$P(\text{Region } r \text{ is causal}) = \text{ sigmoid}\left(\sum_{n\in\{\text{SNPs in region } r\}}\left(2 \times \text{sigmoid}\left(\sum_k \beta_k f_{n,k}\right) - 1\right)\right). \tag{3.35}$$

Once the probabilities have been calculated, C_R is then estimated. The estimation of C_R uses C_0 and C_1 which are predefined by cross-validation, with $C_1 > C_0$. C_R is defined as

$$C_R = C_1 \times P(\text{regulator } r \text{ is causal}) \; + \; C_0 \times (1 - P(\text{regulator } r \text{ is causal})),$$

where C_1 is the maximum regularisation parameter and C_0 is the minimum regularisation parameter. Notice here that as the probability that a regulator r is causal increases, the value for C_R increases linearly, see Figure 3.2. The reason for using the parameter C_R is that it allows a method for increasing the weight given to SNPs that have a high probability of being causal even if the parameter $w_{m,r}$ is estimated as being small. This is essential to prevent causal SNPs being excluded from the model by the L_1 regularisation term. This may arise in the situation where there are multiple SNPs that are highly correlated because of linkage disequilibrium, and the SNP we are looking at is the causal SNP. We would know this based on the regulatory features of the SNPs.

Once we have calculated C_R, we use Equation 3.32 to estimate the parameters $w_{m,r}$ based on the fixed values of α_k and β_k from the previous iteration. This will ensure that the parameter estimates are sparse due to the L_1 regularisation term. However, multiplying the $w_{m,r}$ by C_R in the L_1 term will ensure that SNPs with a high probability of being causal should not be excluded from the model. This step of the application of the minimisation includes minimisation

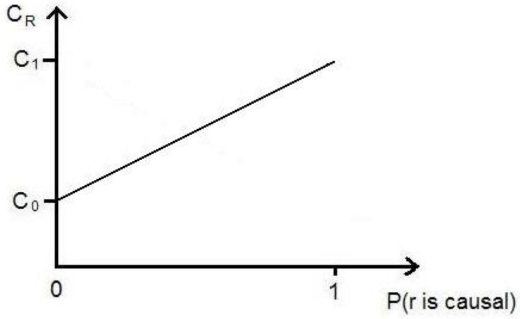


Figure 3.2: The relationship between the probability of a regulator r being causal and the regularisation parameters.

with respect to the residual sum of squares, the L_1 and the L_2 terms.

Having minimised Equation 3.32 with respect to $w_{m,r}$, we begin the second use of the penalised regression with the following minimisation in order to re-estimate the values for our parameters α and β, fixing $w_{m,r}$ from the previous iteration. In Equation 3.32, we minimise with respect to parameters α and β only. These are in the C_R and θ terms only, where C_R is a non-linear function of α and β.

Lirnet outputs the regulatory programs for modules used in the analysis, those where $w_{m,r} \neq 0$; the ranking of causal sequence variants based on the regulatory potential C_R; and regulatory priors β, the information about mutations that induce downstream effects.

Overall, Lirnet aims to learn the regulatory potential of each genetic variant using as much information from the literature and the data as possible. It incorporates functional information in the form of indicator variables called regulatory features. These are included in the modelling of the regulatory potential to reflect the biological knowledge in the effect size estimates or regulatory potential of each feature.

SBFA

Sparse Bayesian Factor Analysis (SBFA) [Parts et al., 2011] incorporates functional information into the prior distribution in the sparse factor analysis model used to infer intermediate, or unobserved phenotypes that routinely influence transcript levels of multiple genes based on the gene expression levels. This information is then used to enhance the ability of the eQTL model to detect causal SNPs by informing interaction effects between the genotypes, unobserved phenotypes and factor effects such as environmental conditions.

SBFA is based on the assumption that gene expression levels are influenced by the effects of the locus genotype, intermediate unobserved factors and the interaction between the two. The model, see Equation 3.36, assumes that the effects are additive.

$$y_{g,j} = \mu_g + \sum_{n=1}^{N} \theta_{g,n} s_{n,j} + \sum_{k=1}^{K} w_{g,k} x_{k,j} + \sum_{k=1}^{K} \sum_{n=1}^{N} \phi_{g,k,n}(s_{n,j} x_{k,j}) + \psi_{g,j}, \qquad (3.36)$$

where $y_{g,j}$ is the gene expression level for individual j for gene g, μ_g represents the mean gene expression level, $\sum_{n=1}^{N} \theta_{g,n} s_{n,j}$ represents the SNP or genotype effect for the N SNPs with weights $\theta_{g,n}$ and genotype $s_{n,j}$, $\sum_{k=1}^{K} w_{g,k} x_{k,j}$ represents the factor effect of the latent factors $x_{k,j}$ which are calculated based on the SNPs $s_{n,j}$ for the K factors with weights $w_{g,k}$, $\sum_{k=1}^{K} \sum_{n=1}^{N} \phi_{g,k,n}(s_{n,j} x_{k,j})$ represents the interaction between the SNPs and the factors with weights $\phi_{g,k,n}$, and $\psi_{g,j} \sim N(0, \sigma^2)$ represents noise.

The latent factor activations X activate the intermediate or unobserved factors; these can be associated to SNPs via the relationship

$$x_{k,j} = \mu_k + \sum_{n=1}^{N} \beta_{k,n} s_{n,j} + \epsilon_{k,j}, \qquad (3.37)$$

where μ_k is the background effect, $\beta_{k,n}$ are the association weights, $s_{n,j}$ are the SNPs and $\epsilon_{k,j}$ is the observation noise.

To estimate the parameters, a two step approach is used:

Step1: Factor inference: infer latent factors $X = (x_1, \ldots, x_k)$ and weights $\mathbf{W} = \{w_{g,k}\}$ from expression levels alone, ignoring the effects of the SNP via the association and interaction effects. This can only be approximated using SBFA.

Step2: Association and interaction testing: conditional on the state of the inferred factors, significance of the associations of factors to SNPs (see Equation 3.37) and SNP-gene-factor interactions (see Equation 3.36) are tested.

In Step 1, factors are inferred using an SBFA model, where the factor model is

$$y_{g,j} = \sum_{k=1}^{K} w_{g,k} x_{k,j} + \psi_{g,j}, \qquad (3.38)$$

which is simply Equation 3.36 with the direct genetic associations and the interactions removed. This explains the observed gene expression $y_{g,j}$ using latent factors/random variables.

A binary indicator variable $z_{g,k}$ is introduced to encode whether factor k regulates gene g, such that

$$z_{g,k} = \begin{cases} 1 & \text{if factor } k \text{ regulates gene } g \\ 0 & \text{otherwise} \end{cases}$$

The prior distributions on the weights are defined using this indicator variable as follows:

$$P(w_{g,k}|z_{g,k} = 0) = N(w_{g,k}|0, \sigma_0^2) \tag{3.39}$$

$$P(w_{g,k}|z_{g,k} = 1) = N(w_{g,k}|0, 1), \tag{3.40}$$

where σ_0^2 is chosen to be small, hence driving the weights to 0 and inducing sparsity. Parts et al. [2011] used $\sigma_0^2 = 10^{-4}$ in their simulations.

Functional information from two online databases containing information about functional characteristics of Yeast data, KEGG [Kanehisa and Goto, 2000] and Yeastract [Teixeira et al., 2006] is encoded as a Bernoulli prior on $z_{g,k}$. This is defined as:

$$\pi_{g,k} = P(z_{g,k} = 1) = \begin{cases} \nu_0 & \text{if there is a link} \\ 1 - \nu_1 & \text{if there is no link} \end{cases}$$

where ν_0 is the false negative rate and ν_1 is the false positive rate of the observed prior information. Parts et al. [2011] use $\nu_0 = 0.06$ and $\nu_1 = 0.001$ for Yeastract prior information, and $\nu_0 = 0.0001$ and $\nu_1 = 0.001$ for KEGG data. This information is used to generate the prior on the weights $w_{g,k}$ which is defined as:

$$P(w_{g,k}|\pi_{g,k}) = \pi_{g,k} N(w_{g,k}|0, 1) + (1 - \pi_{g,k}) N(w_{g,k}|0, \sigma_0^2). \tag{3.41}$$

The remaining prior distributions for the latent factors X, the per gene noise ψ_g and the precision τ_g are defined as follows:

$$x_{k,j} \sim N(0, 1) \tag{3.42}$$

$$\psi_g \sim N(0, \frac{1}{\tau_g}) \tag{3.43}$$

$$\tau_g \sim Ga(\tau_g|a_\tau, b_\tau). \tag{3.44}$$

Parts et al. [2011] use $a_\tau = b_\tau = 0.001$ to give an uninformative prior on the precision.

In part two, standard marker regression is used to calculate the test statistics for association and interaction effects involving the inferred factor activations. A LOD score distribution for significance of association and interaction weights was calculated by permuting the data. The local FDRs (false discovery rates) or Q-values for the association and interaction states were also calculated. Uncertainty is incorporated by recalculating Q-values, adjusted p-values found by optimising the false discovery rate approach, for multiple random restarts of the model and then combining this information into a statistic to assess the overall significance of one particular effect.

Comparison of statistical methods including functional information

We aim to include functional information that is SNP specific, and that can be used to prioritise SNPs which are, a priori, more likely to be causal. SBFA is currently specific to Yeast data but could be applied to human data, given the availability of online functional information databases. Many of those that are available at present, such as ENCODE, contain a lot of missing data and many variables are highly correlated. This makes adapting SBFA very challenging. SBFA infers many latent factors as the authors believe these to be important when assessing causal sequence variants. We are only interested in using the observed data we have, and prioritising this based on SNP specific functional information. This makes SBFA unsuitable for our aim. Similarly, Lirnet incorporates a lot of functional information but not on a SNP specific level. Grouping SNPs and genes is computationally more efficient but we aim to find a model that performs specific selection of causal SNPs. Lirnet does not provide a framework for this. We summarise these two methods in Table 3.1, showing the advantages, disadvantages and outputs from these two methods.

We believe that these methods give ideas for including functional information, but within the specific requirements we are aiming to meet, using one of the previously mentioned statistical methods and adapting it to include functional information will provide better results than adapting either of these methods.

3.4 Statistical Models to test on eQTL data

The statistical models assessed here all provide a suitable framework for analysing eQTL data. However, for including functional information in the form of the FS score, Lirnet and SBFA already include functional information in a different way. Neither use functional information that is entirely SNP specific. We think this is important for enhancing detection of associated SNPs. The HyperLasso

Lirnet	
Output	A set of regulatory programs for each module. These are the weights on each regulator in a module that regulate gene expression.
	A ranking of causal/associated SNPs based on the regulatory potential of SNPs and genes.
	Information on the regulatory prior which can provide insight into which SNPs lead to downstream, phenotypic effects.
Advantages	Includes multiple SNPs and other genes as regulators therefore can account for interactions.
	Includes information on the functionality of the genome.
Disadvantages	Requires a lot of information on the functionality of the genome as input.
	Complex method with no clear areas for development.

SBFA	
Output	Genotype-factor interactions based on gene expression.
	Effect of strain and environment based on changes in gene expression.
Advantages	Includes functional information.
	Accounts for interactions between genotype and environmental and other unobserved factors.
Disadvantages	Model tailored to yeast data.
	The model doesn't transfer easily to humans.
	The model relies on too many inferences from generic yeast functional data.

Table 3.1: A summary of the two eQTL methods, Lirnet and SBFA, that include functional information.

is a generalisation of the Lasso, which has the NEG prior rather than the double exponential (DE) prior. As such we will use only HyperLasso to represent Bayesian MAP estimation techniques on eQTL data. The Bayesian Lasso is a special case of the Normal Gamma prior and so we have no need to test both of these. Hence we will use only the Normal Gamma prior to assess Bayesian shrinkage methods. We compare the prior distributions of the NG and the NEG with respect to the DE prior in Figure 3.3. As in Hoggart et al. [2008], we define the γ parameter of the NEG such that the density of the DE prior and the NEG prior is the same at 0. We cannot do this for the NG, and so we define the marginal prior variance of $\beta_i = 2\lambda\gamma^2 = 2$ as Griffin and Brown [2010] have done, and vary λ such that this relationship ($\gamma = \sqrt{\frac{1}{\lambda}}$) remains constant. We include the DE prior with $\xi = 10$ on both plots for comparison.

Figure 3.3 shows how the tails of the NEG and NG are similar but that the central point of the NG appears much narrower. This means that the NG will induce more shrinkage close to 0 compared to the NEG. Both will enforce less

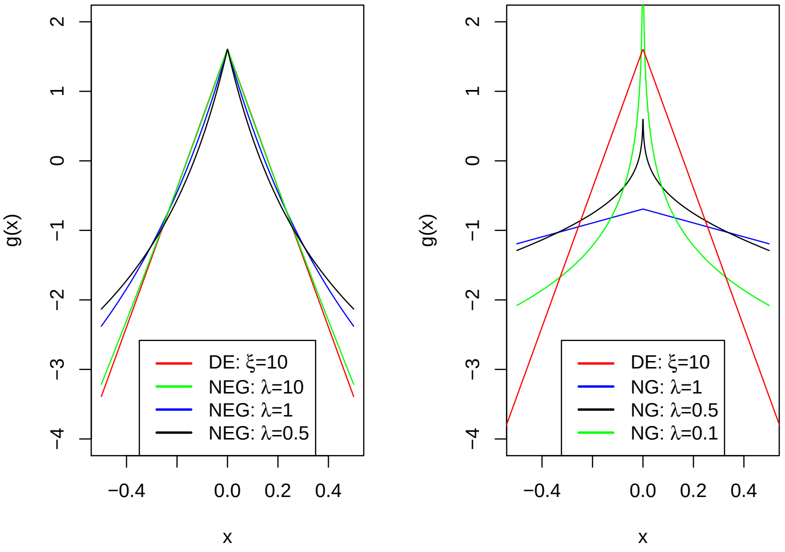


Figure 3.3: A comparison, with respect to the double exponential (DE) prior of the LASSO, for the Normal Gamma and the Normal Exponential Gamma with different parameters. **Left:** The NEG (Normal Exponential Gamma) prior distribution for HyperLasso. The γ parameter has been chosen such that the density of the NEG and DE priors are the same at 0. **Right:** The NG (Normal Gamma) prior distribution. The γ parameter has been chosen such that the marginal prior variance of β_i, which is defined as $2\lambda\gamma^2$ is fixed to be 2.

shrinkage in the tails.

piMASS and the Spike and slab both use fully Bayesian selection approaches, however they are quite different and therefore we include both when testing models on eQTL data. Dependent on the simulation results in Chapter 4, we will decide which of these methods to develop further and to adapt to include functional information.

We feel that it is important to assess commonly used statistical methods for their ability to detect associated sequence variants. As a results we will also include the univariate likelihood ratio test and the least squares (LS) or minimum length least squares (MLLS) estimates to provide a full comparison between univariate, multivariate, frequentist and Bayesian approaches.

3.4.1 Least Squares

The standard least squares (LS) estimates are widely used in frequentist statistics for estimation of parameters $\hat{\beta}$ in a standard linear model. The LS estimates for the standard linear model $y = \alpha + \beta X + \epsilon$ are

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{3.45}$$

However, the LS estimates can only be calculated when $n > p$. In the case of eQTL data, we have $p > n$. As a result we have to use the minimum length least squares (MLLS) estimates of β.

The MLLS estimates are, according to Choi [2006], the unique solutions of minimising $\beta \in \text{argmin}||X\beta - y||_2$ with respect to β. This is equivalent to minimising the residual sum of squares, $(y - X\beta)^T(y - X\beta)$. The minimum length solution of this, or of $X\beta = y$, is unique and is also referred to as the pseudo-inverse solution. The formal definition of this is equivalent to the LS estimate and is defined as:

$$\hat{\beta} = (X^T X)^{\dagger} X^T y,$$

where $A^{\dagger}$ is the pseudo-inverse of A.

The pseudo-inverse is one of many generalised inverses that are discussed in detail by Ben-Israel and Greville [2003]. Generalised inverses do not always have a unique solution, although the pseudo-inverse we are calculating here is unique [Choi, 2006].

3.4.2 Univariate Likelihood Ratio Test

A likelihood ratio (LR) test is a univariate method to compare two nested models. In our case we compare the null model $y = \alpha + \epsilon$ to the single SNP model $y = \alpha + \beta_A X_A$, where β_A is a scalar and X_A is the column vector of the distribution of SNP A across all individuals. We use this set-up unless other factors need to be added to reduce confounding as in the yeast data. We use the LR test statistic values when plotting the ROC curves.

3.5 Conclusion

In this chapter we have assessed different statistical methods that may be suitable for detecting causal or associated sequence variants in eQTL. We have concluded that there are four Bayesian multivariate statistical methods piMASS, SS, HL and the NG that we will compare the results from on simulated and real Yeast and Human eQTL datasets. For completeness of possible statistical approaches, including methods that are currently used, we will also compare the LR test and the LS/MLLS estimates on simulated and real eQTL datasets. The results from the simulation study are presented in Chapter 4.

Chapter 4

Comparing the performance of statistical methods for eQTL detection via simulation

Using two simulated datasets, we show differences between the performance of six statistical methods, piMASS, HyperLasso, Normal Gamma, Spike and slab, minimum length least squares (MLLS) and the LR test on eQTL data with the number of SNPs p greater than the number of individuals n. The results of these simulation studies will inform our decision about which statistical method to adapt to include functional information. The performance of some of these statistical methods depends on the minor allele frequencies (MAF) of the causal SNP.

4.1 HapGen Simulated dataset 1 - larger effect sizes

4.1.1 Simulating the data

We use HapGen2 [Su et al., 2011] to simulate data using the actual SNP correlation structure from human haplotypes. HapGen2 generates DNA sequences based on the minor allele frequency (MAF) and linkage disequilibrium (LD) structure of the reference dataset. Here we use only the control samples generated from the European haplotypes of the August 2010 release of the 1000 genomes data [Altshuler et al., 2010]. In this dataset we simulate 6 causal SNPs per dataset to have a range of effect sizes; the large causal SNP has effect size 1.5, the others have effect size simulated from a $N(0.5, 0.1^2)$ distribution. In the case of $\{0, 1\}$ coding, the effect size estimates the additional increase in gene expression due to having a mutation at that location. These values for effect sizes were chosen to be large in terms of realistic values such that we could initially assess the statistical models where the results were clear. An effect size of 0.5 means that someone who has the SNP has gene expression value that is approximately 0.5 higher than someone who does not have the SNP. We simulate 9 datasets with these identical causal SNPs and effect sizes giving 54 causal SNPs out of $631 \times 9 = 5679$ SNPs based on 631 total SNPs per dataset.

The 9 datasets are simulated from the region around the CASPASE8 gene on chromosome 2 - a region widely believed to be associated to breast cancer and melanoma (Barrett et al. [2011], Camp et al. [2012]). A 200kbase region (from 201566128 to 201766128 in the human genome 19 build of chromosome 2) surrounding the CASPASE8 gene was used for simulations. This region has mixed LD block sizes and strengths of LD allowing us to evaluate the effectiveness of the different methods at detecting causal SNPs in different size and strength LD blocks.

The causal SNPs were chosen such that there were 3 causal SNPs in high LD, one with a large effect size (1.5) and two with small effect sizes (simulated from $N(0.5, 0.1^2)$; 2 causal SNPs in low LD, both with small effect sizes (simulated from $N(0.5, 0.1^2)$); and 1 causal SNP in a low LD block with a small effect size (simulated from $N(0.5, 0.1^2)$). SNPs in low LD have $r^2 < 0.1$, and SNPs in high LD have $r^2 > 0.6$. There are two datasets simulated - one with all causal SNPs having population MAF approximately 0.2 (HapGen dataset 1A) and the other with population MAF approximately 0.02 (HapGen dataset 1B). Both have the same SNP set up, but with different causal SNPs due to the different MAF we require for each SNP.

We simulate the i^{th} gene expression (y_i) as $y_i = \sum_{j=1}^{p=631} X_{ij}\beta_j + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$, X_{ij} represents the j^{th} genotype for person i, and β_j represents the effect size of the j^{th} SNP. We control the effect sizes for HapGen dataset 1 such that a large effect size is $\beta_i = 1.5$ and a small effect size is $\beta_i \sim N(0.5, 0.1^2)$. We treat y_i as having been centred in our simulations, such that there is no background expression ($\alpha = 0$). We use the dominant modelling of SNPs, where 0 represents the homozygous wildtype genotype and 1 represents the other two genotypes. The variability modelled into the gene expression represents the noise in real data. The effect sizes and number of SNPs in the region are chosen to be consistent with other simulated datasets, see Kang et al. [2012], Wu et al. [2011], Petersen et al. [2013].

4.1.2 Results

The data simulated using HapGen2 [Su et al., 2011] has been analysed with all the methods we are comparing (piMASS, HyperLasso, Normal Gamma, MLLS, LR test, Spike and slab). We use ROC curves to compare the performance of the methods.

Possible estimators for piMASS

We assessed three possible estimators for piMASS - the posterior mean of β, the posterior inclusion probability and the posterior inclusion probabilities based on Rao-Blackwellisation. See the subsection describing piMASS in Section 3.2.1, page 29 for details. The posterior mean of β performs best in terms of the AUC and so we choose to use this to summarise piMASS.

Possible summary statistics for Spike and slab

We also compare two posterior estimates for the Spike and slab - the posterior mean of all iterations post burn-in and the percentage of non-zero coefficient estimates post burn-in. The posterior mean of all iterations gives the largest AUC based on the ROC curve and so we use this as the summary statistic for the Spike and slab.

Possible summary statistics for the Normal Gamma

In Appendix B, we compare summary statistics for the Normal Gamma prior on a basic simulated dataset that omits the correlation structure between SNPs. We compare many different statistics using the mean, median, 95^{th} percentile, interquartile range, credible interval containing 0 and variance and several combinations thereof. The summary statistic giving the greatest AUC of the ROC curve was the posterior mean. We therefore use this as our summary statistic for summarising the Normal Gamma as optimally as possibly with respect to the ranks in the AUC of a ROC curve.

When assessing the credible intervals that do not contain 0, we notice that the SNPs with the highest posterior mean effect sizes are those whose credible interval does not contain 0. We therefore assess the posterior mean effect sizes when defining which SNPs to suggest for biological validation.

A summary of the summary statistics used

For the likelihood ratio test we use the likelihood ratio value to plot the ROC curves but report the effect size based on the least squares estimate from the single SNP model. There is a direct relationship between the effect size estimate from the LS estimate on the single SNP model and the LR test value, hence the ROC curves are identical with either statistic. For HyperLasso we use the MAP estimate output from the model, and for the LS or MLLS we use the coefficient estimate that is calculated.

As discussed in Chapter 4 and Appendix B, there are different posterior summary statistics that we could use for piMASS, NG and Spike and slab. We choose the statistic that maximises the area under a ROC curve. We find that the statistics that maximises the AUC is the most simple statistic, such that the posterior mean of all sampled values post burn-in is the statistic we use to summarise these methods.

Formal association testing versus ranking

Throughout this thesis based on the results found in Appendix B, we choose to rank the SNPs based on posterior mean effect size, or equivalent for the other non fully Bayesian approaches, and propose the top 5 to be those SNPs of interest.

If we had decided to do formal association testing, this would have been very difficult using all 6 methods. For the NG we could use the posterior credible interval as shown in Appendix B, page 175. The LR test already performs a statistical test so we can use the results to define formal association. HyperLasso is a selection method, so to define formal association we could select only those $\beta_i \neq 0$. This would require very careful and detailed parameter choices. For piMASS and Spike and slab we could also use the posterior credible interval. For the LS we could perform a nested test comparing the full model and the reduced model with only one SNP removed, similar to the LR test. But for the MLLS we could not perform such a test as only the point estimate is produced.

In the frequentist tests the critical values would need to be adjusted to account for multiple testing and then the SNPs that are selected will depend on the adjustment used such as the family-wise error rate (FWER) or false discovery rate (FDR). The MLLS does not allow for any formal testing, and would have to be omitted.

We choose to rank the estimates from each method based on a given summary statistic. We then propose a certain number or percentage of the top SNPs for further testing. We believe that this is more flexible and robust to the changes in effect size based on different genes. Defining a threshold that can be used across all genes for all numbers of individuals and SNPs would be challenging. By ranking the SNPs, this makes the output more approachable for non-statisticians. It gives scope for disease experts to adjust the results from the model to better fit the biological knowledge which is omitted by the model.

Many other techniques could be applied to summarising the MCMC. We look at these in more detail in the Discussion (Chapter 11), but other ideas to identify associated SNPs could include calculating a t-statistic to compare the posterior distribution to the prior distribution using both the mean and the variance. Comparing the prior and posterior distributions for statistically significant differences may lead to better results, especially if we create an artificial interval to define 0. One of the difficulties with many of the summary statistics tested was that there were a lot of posterior values that were 0 (to 4 d.p.). This is mainly caused by the harsh shrinkage from the prior distribution and the lack of overwhelming evidence in the likelihood.

Assessing the results

Using ROC curves, for dataset 1A (MAF 0.2), all methods perform well with the minimum AUC of 0.6236 being for the LR test, see Figure 4.1. Similarly for dataset 1B (MAF 0.02) all methods perform well, with the minimum AUC being 0.6915 for Least Squares, see Figure 4.2. We perform DeLong's test [DeLong et al., 1988] to test for significant differences between the AUCs. The p-values from DeLong's test, and the AUC of the ROCs can be found in Tables 4.1 and 4.2 respectively.

We notice that the univariate LR test performs worst for dataset 1A (MAF 0.2) (AUC 0.6236) but performs well for dataset 1B (MAF 0.02) (AUC 0.8608). NG and Spike and slab perform better for dataset 1A than 1B (AUCs 0.9841 and 0.96, and 0.9853 and 0.9236 respectively).

For dataset 1A, there is no statistical difference between the AUC for the Normal Gamma and Spike and slab, and the LR test and MLLS estimates. All other pairwise tests reveal high levels of statistical significance. For dataset 1B, there is no statistical difference between the AUC for Spike and slab and piMASS, piMASS and the NG, and HL and LS (MLLS).

	AUC	HL	LS	LR	NG	piMASS	S&S
HL	0.7806	.	0.07199	0.00233	7.356×10^{-7}	0.004027	6.543×10^{-7}
LS	0.684		.	0.3294	1.131×10^{-11}	1.365×10^{-5}	2.682×10^{-12}
LR	0.6236			.	4.398×10^{-15}	5.13×10^{-9}	1.797×10^{-14}
NG	0.9841				.	9.704×10^{-8}	0.824
piMASS	0.9016					.	6.031×10^{-8}
S&S	0.9853						.

Table 4.1: This table reports the AUCs and the p-values from DeLong's test for ROC curves for the 6 statistical methods we are comparing on the Hapgen simulated dataset 1A. This dataset has 54 causal SNPs simulated according to the CASPASE8 region with a MAF approximately equal to 0.2.

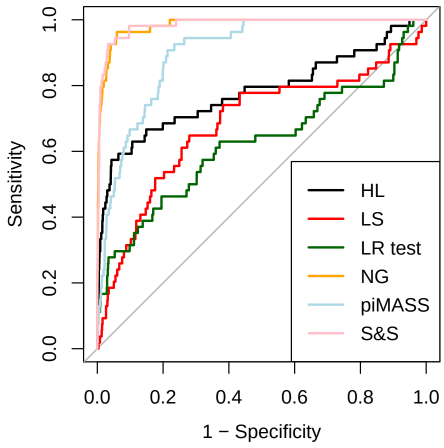


Figure 4.1: ROC curve comparing different statistical methods (HL, LS (MLLS), LR test, NG, piMASS and Spike and slab (S&S)) for detecting simulated causal SNPs. The data includes 5679 SNPs of which 54 are simulated to be causal SNPs, 6 of which have effect size 1.5 with the remaining causal SNPs having effect size simulated from a $N(0.5, 0.1^2)$ distribution. The causal SNPs, simulated using data from a subset of the CASPASE8 region, have a MAF of approximately 0.2 in the population (HapGen dataset 1A).

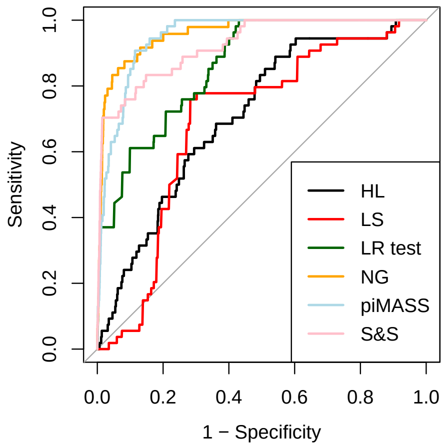


Figure 4.2: ROC curve comparing different statistical methods (HL, LS (MLLS), LR test, NG, piMASS and Spike and slab (S&S)) for detecting simulated causal SNPs. The data includes 5679 SNPs of which 54 are simulated to be causal SNPs, 6 of which have effect size 1.5 with the remaining causal SNPs having effect size simulated from a $N(0.5, 0.1^2)$ distribution. The causal SNPs, simulated using data from a subset of the CASPASE8 region, have a MAF of approximately 0.2 in the population (HapGen dataset 1B).

4.1.3 MCMC Plots

When assessing the Normal Gamma for convergence we investigate the posterior density and trace plots. Due to the irregular prior density on our SNP effect

	AUC	HL	LS	LR	NG	piMASS	S&S
HL	0.7089	.	0.6314	1.732×10^{-6}	$< 2.2 \times 10^{-16}$	3.191×10^{-15}	8.076×10^{-13}
LS	0.6915		.	4.393×10^{-8}	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	2.34×10^{-15}
LR	0.8608			.	2.717×10^{-7}	3.569×10^{-6}	5.717×10^{-5}
NG	0.96				.	0.139	0.0009261
piMASS	0.9496					.	0.06099
S&S	0.9236						.

Table 4.2: This table reports the AUCs and the p-values from DeLong's test for ROC curves for the 6 statistical methods we are comparing on the Hapgen simulated dataset 1B. This dataset has 54 causal SNPs simulated according to the CASPASE8 region with a MAF approximately equal to 0.02.

sizes β_i, we expect to see many posterior estimates very close to 0. This means our posterior histograms will have a large spike at 0, and also the trace plots may appear quite sporadic as the estimates move between approximately 0, and other values. Using only dataset 1B (MAF 0.02) we show plots of 15,000 iterations in the form of histograms and corresponding trace plots, both excluding burn-in for six SNPs. In Figure 4.3, we see the examples of 6 typical posterior histograms from the Normal Gamma prior, and in Figure 4.4, we see the corresponding trace plots. Those SNPs we have selected give trace plots and histograms that are representative of the shapes of the posterior distributions that we see for all histograms and trace plots. The top right plot in Figure 4.3 and Figure 4.4 shows the posterior distribution and corresponding trace plot for a causal SNP with moderate effect size (approximately 0.6).

4.1.4 Discussion of results

In this set of results the Normal Gamma is performing very well, as are the Spike and slab and piMASS. In the MAF 0.02 case (dataset 1B), the LR test also performs very well. These results show that the Normal Gamma in particular performs well, in terms of SNP ranking, for the simulated eQTL data with both large and small MAFs in the presence of moderate to large effect sizes.

4.2 HapGen simulated dataset 2 - smaller effect sizes

4.2.1 Simulating the data

This dataset is simulated as in Section 4.1.1. We use SNP data simulated using HapGen2 [Su et al., 2011] and change only the causal effect sizes. In this dataset we simulate all 6 causal SNPs per sub-dataset to have an effect size of 0.4. We calculate the gene expression y as in Section 4.1.1. For the results of

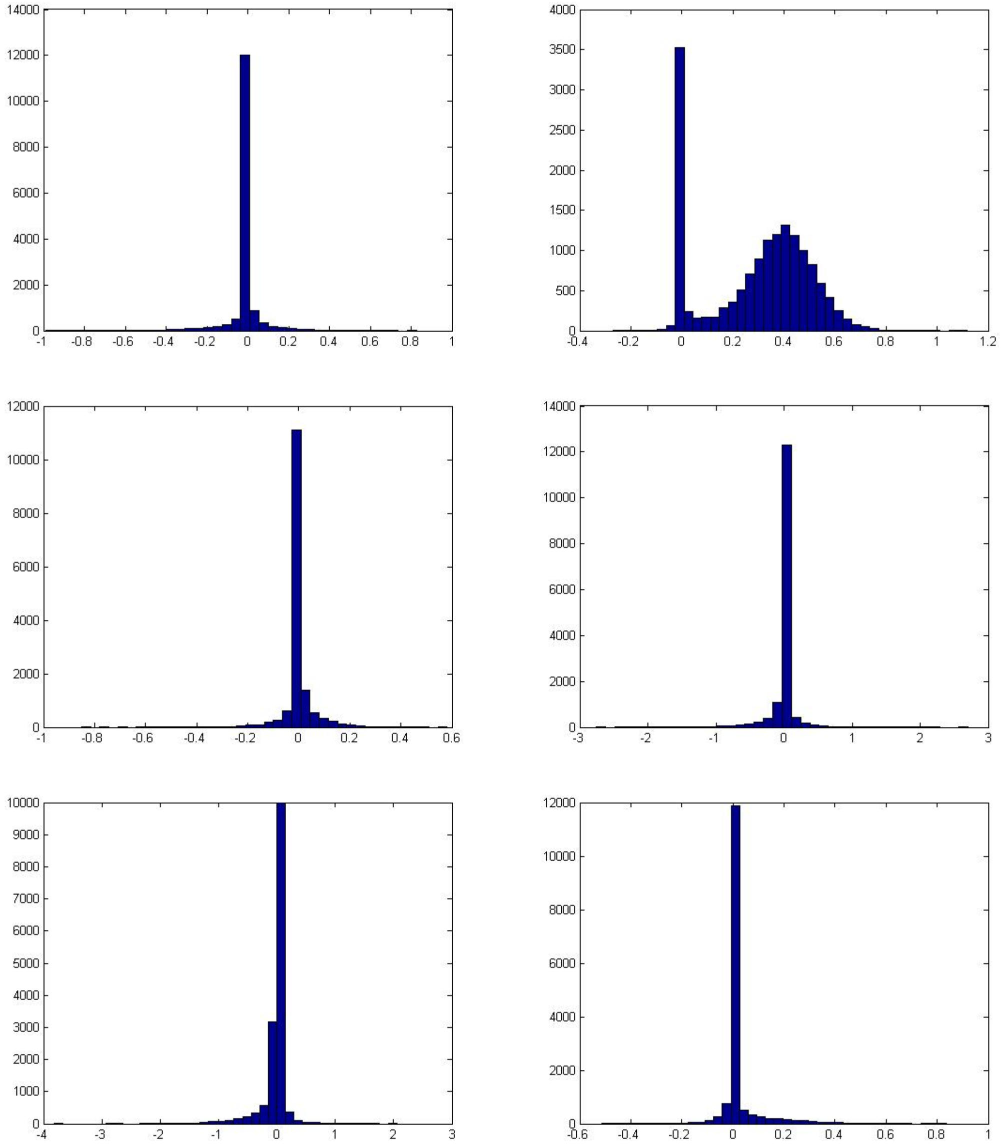


Figure 4.3: An example of the posterior distributions in the form of a histogram for 6 SNPs from the Normal Gamma prior for dataset 1B, with 15,000 iterations (after the burn-in has been removed). The top right corner plot is for a causal SNP with effect size approximately 0.6.

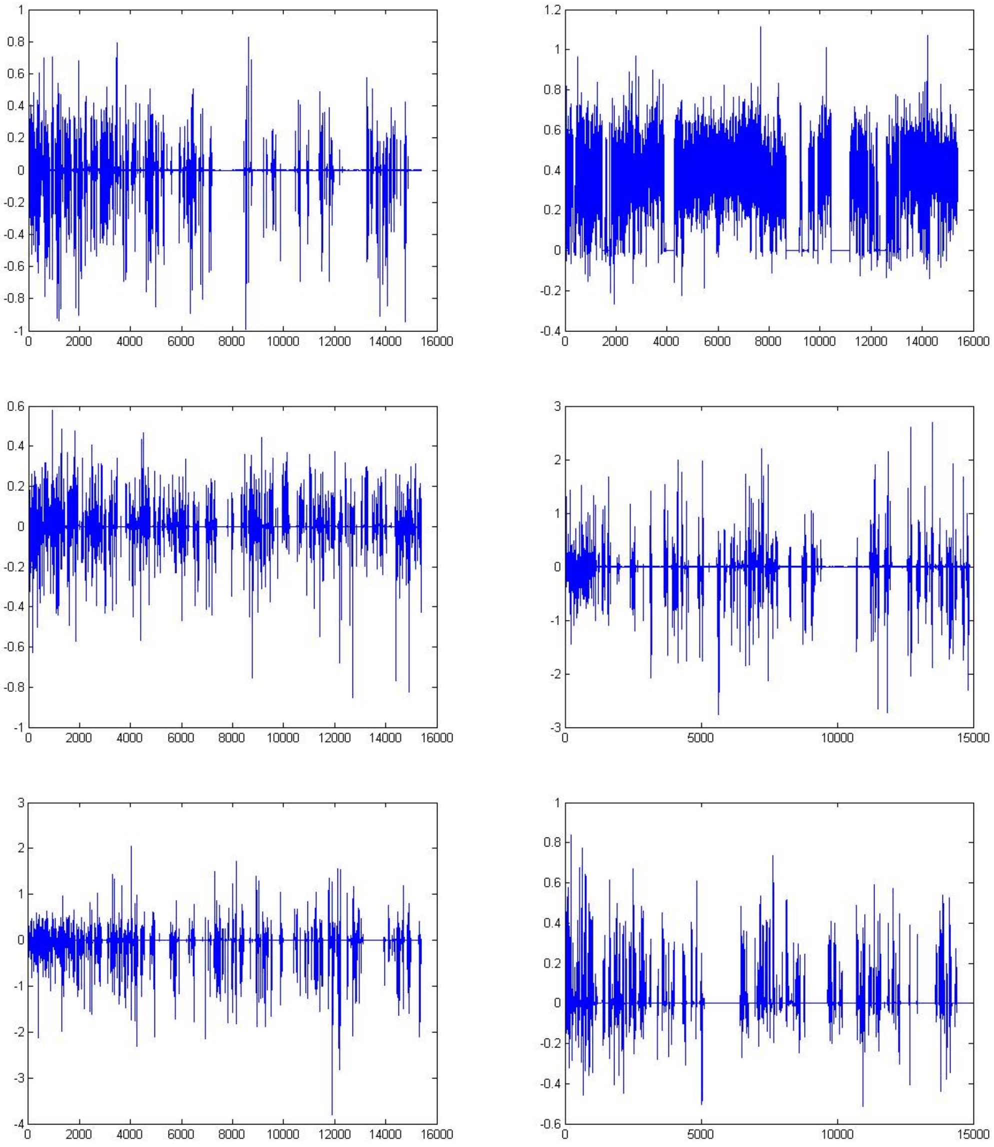


Figure 4.4: An example of the MCMC trace plots for 6 SNPs from the Normal Gamma prior for dataset 1B, with 15,000 iterations and the burn-in excluded. The top right corner trace plot is for a causal SNP with effect size approximately 0.6.

this dataset, we use 9 datasets giving 54 causal SNPs out of 631 × 9 = 5679 SNPs.

The aim of using these two datasets is to make the causal SNPs more difficult to detect. This should split the performance of the methods more clearly. We label the dataset for MAF 0.2 dataset 2A and the dataset for MAF 0.02 dataset 2B.

4.2.2 Results

Again we use the ROC cuves to assess the comparison of methods. For dataset 2A (MAF 0.2), Figure 4.5, the NG and Spike and slab perform very well. Table 4.3 reports the AUC for each statistical method, and reports the p-value from DeLong's test [DeLong et al., 1988] with the null hypothesis that the AUC of ROC1 is the same as the AUC of ROC2, where ROC1 is taken as the reference ROC curve. We notice that the AUCs for LS and LR are statistically similar, with poor AUCs; 0.6917 and 0.6040 respectively. These methods have AUCs that are statistically significantly different to all other methods. We also notice that piMASS is statistically significantly different to all methods with maximum p-value 0.01068, except the HL where there is no statistical difference with p-value (0.52).

	AUC	HL	LS	LR	NG	piMASS	S&S
HL	0.7809	.	0.05224	9.99×10^{-5}	2.769×10^{-8}	0.5152	1.865×10^{-5}
LS	0.6917		.	0.0615	3.691×10^{-11}	0.01068	2.549×10^{-9}
LR	0.604			.	$< 2.2 \times 10^{-16}$	3.510×10^{-8}	$< 2.2 \times 10^{-16}$
NG	0.9702				.	6.151×10^{-8}	0.02775
piMASS	0.8093					.	0.0001813
S&S	0.9418						.

Table 4.3: This table reports the AUCs and the p-values from DeLong's test for ROC curves for the 6 statistical methods we are comparing on the Hapgen simulated dataset 2A. This dataset has 54 causal SNPs simulated according to the CASPASE8 region with a MAF approximately equal to 0.2.

	AUC	HL	LS	LR	NG	piMASS	S&S
HL	0.6667	.	0.3932	0.08166	1.508×10^{-13}	6.676×10^{-10}	7.431×10^{-12}
LS	0.6343		.	0.004153	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
LR	0.7365			.	7.49×10^{-10}	9.98×10^{-6}	1.487×10^{-8}
NG	0.9322				.	0.01395	0.05946
piMASS	0.8999					.	0.5063
S&S	0.9119						.

Table 4.4: This table reports the AUCs and the p-values from DeLong's test for ROC curves for the 6 statistical methods we are comparing on the Hapgen simulated dataset 2B. This dataset has 54 causal SNPs simulated according to the CASPASE8 region with a MAF approximately equal to 0.02.

For dataset 2B (MAF 0.02), Figure 4.6, we notice that piMASS performs much better than in dataset 2A and has results comparable to the NG and

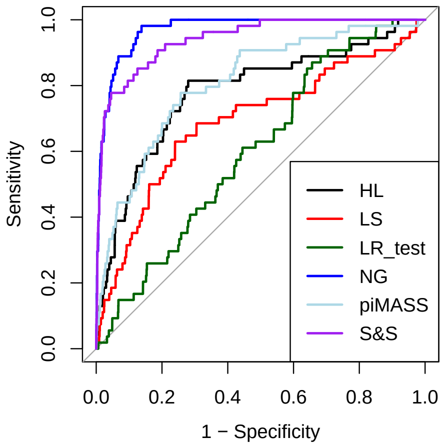


Figure 4.5: ROC curve comparing the different statistical methods for detecting causal SNPs on our simulated data with all 54 causal SNPs having effect size 0.4. The data was simulated using HapGen2 Su et al. [2011], targetting SNPs in the CASPASE8 region. Causal SNPs with a MAF of approximately 0.2 in the population (HapGen simulated data 2A).

Spike and slab. HL, LS and LR are poor in comparison to the NG. In dataset 2B where the causal SNPs have MAF 0.02 (rarer variants), LS and LR appear better than in dataset 2A where the causal SNPs have MAF 0.2 (more common variants). HL has the reverse performance, it performs better in the case of a rarer variant.

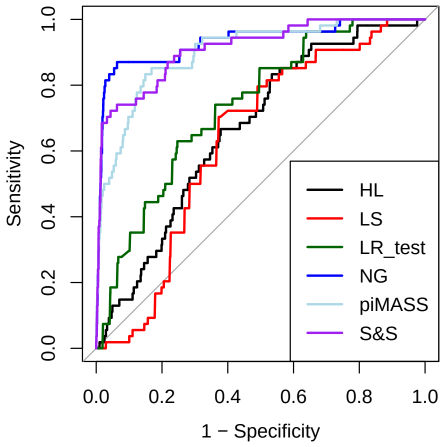


Figure 4.6: ROC curve comparing the different statistical methods for detecting causal SNPs on our simulated data with all 54 causal SNPs having effect size 0.4. The data was simulated using HapGen2 Su et al. [2011], targetting SNPs in the CASPASE8 region. Causal SNPs with a MAF of approximately 0.02 in the population (HapGen simulated data 2B).

4.2.3 Discussion of results

In both cases, for datasets 2A and 2B (MAF 0.2 and MAF 0.02 respectively), the Normal Gamma prior is performing very well, as does the Spike and slab. HyperLasso has performance that is comparable to the non-Bayesian methods. As expected, from datasets 1A and 1B, the fully Bayesian methods of the NG, Spike and slab, and piMASS are consistently outperforming the LS and LR test.

We notice, from these results, that piMASS, LR and LS seem to improve as the frequency of the causal/associated mutation increases, whereas the NG, HL and Spike and slab seem to become slightly worse, although the NG and Spike and slab still perform very well.

Given the results in this section, it is clear that the Normal Gamma and Spike and slab perform consistently well on datasets 1A, 1B, 2A and 2B. An MCMC approach using the Normal Gamma prior has the advantage that prior functional genomic information can easily be incorporated into its prior structure. For this reason we choose the Normal Gamma to adapt to include functional information. We begin by assessing the computational requirements of this method and verifying the convergence.

4.3 Convergence, Computational Time and Effect of the prior

To check convergence of the Normal Gamma we use the R-hat statistic of Brooks and Gelman [1998]. We check that the rankings of the Normal Gamma posterior estimates remain the same using the AUCs of ROC curves. We use DeLong's test to test that there is no statistically significant difference between AUCs of these ROC curves. We include details of computational time as a comparator to the other statistical methods (HyperLasso, piMASS, LS/MLLS, LR, Spike and slab), some of which perform a trade-off between statistical model accuracy and computational time. For example, HyperLasso uses the MAP estimate based on the local maximum rather than verifying it is a global maximum. It also provides no estimate of the uncertainty in the posterior distribution. We also assess the effect of the prior on the posterior results of the Normal Gamma.

4.3.1 Checking convergence using R-hat

We check convergence of the Markov chains using the R-hat statistic, see Brooks and Gelman [1998]. This diagnostic is not as stringent as other diagnostics. We choose not to thin the Markov chain as there is debate as to the effectiveness of it, see Link and Eaton [2012], in comparison to increasing the chain length. Given we choose not to thin the Markov chain, we prefer to run the chain for longer and retain all iterations.

The Gelman and Brooks R-hat convergence statistic

The R-hat convergence diagnostic investigates the scaled, weighted difference between the within and between chain variances. It is also known as the potential scale reduction factor. It is defined in Brooks and Gelman [1998] as

$$\hat{R} = \sqrt{\frac{V}{W}}, \tag{4.1}$$

where $\theta_{i,j}$ is the i^{th} element of the j^{th} chain, n is the number of iterations for each of the m chains, $\bar{\theta}_j$ is the mean of all i elements of chain j (the within chain mean).

$$\begin{aligned}
s_j^2 =& \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij} - \bar{\theta}_j\right)^2 && && (4.2)\\
W =& \frac{1}{m}\sum_{j=1}^{m} s_j^2 && \text{Mean chain variance} && (4.3)\\
\bar{\bar{\theta}} =& \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j && && (4.4)\\
B =& \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2 && \text{Between chain variance} && (4.5)\\
V =& \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B. && && (4.6)
\end{aligned}$$

The minimum number of runs of a chain for meaningful interpretation of this diagnostic is stated as 10. Brooks and Gelman [1998] suggests an R-hat value of < 1.1 is advisable for convergence.

Application

We test this once for each dataset, using 10 runs of the NG. We tabulate the maximum R-hat values for all parameters in the Normal Gamma for one simulated dataset in Table 4.5. The simulated data requires more than 20,000 iterations for convergence. At 20,000 iterations, most parameters have converged. The only parameters not converged are γ^2 and λ with R-hat values of 1.147808 and 1.235675 respectively. By 30,000 iterations, with a 5,000 iteration burn-in these have converged. The convergence diagnostic shows that the between chain variance is very small which implies that the within chain variances must be similar across chains. This suggests that the stationary distribution has been reached at each run of the chain.

R hat for simulated data (MAF 0.02, dataset 2B)			
	100,000 iterations	30,000 iterations	20,000 iterations
Maximum	1.018263	1.021728	1.235675

Table 4.5: Table showing the maximum R-hat values for all the Normal Gamma parameters for the simulated dataset run with different numbers of iterations. The burn-in was kept constant at 5,000 iteration. The R-hat statistic was calculated using 10 datasets.

Replicability of the Normal Gamma Prior.

We check the replicability of the Normal Gamma by reporting summary statistics for the AUCs.

The mean AUCs for the 10 plots are 0.7847, 0.7909 and 0.78148 respectively for 20,000, 30,000 and 100,000 iterations, with the minimum AUCs 0.7693, 0.7643 and 0.7795, the maximum AUCs 0.8091, 0.7909 and 0.7837 and the standard deviations 0.0147, 0.0127 and 0.00122 respectively. Clearly the AUCs for different numbers of iterations and within each number of iterations are all very similar. We omit the ROC curves as the 6 causal SNPs out of 631 total SNPs do not provide us with a good visual representation. Using the AUCs, we conclude that the Normal Gamma is replicable in as much as the SNP rankings are similar between chains.

We check, using DeLong's test for comparing AUCs, that there is no statistically significant difference between the ROC curves for 100,000 iterations, 30,000 iterations or 20,000 iterations. We find that there is no statistically significant difference between any pairwise combination of the 10 ROC curves for 100,000 iterations or 30,000 iterations with the minimum pairwise p-value being 0.6785 and 0.2205 respectively. At 20,000 iterations, a number of pairwise comparisons return a minimum p-value of 0.04835 which is bordering on statistical significance.

The R-hat statistic is reported for all real data analyses in later chapters. Visual checks are also carried out for these analyses in later chapters but details are omitted where no issues are identified.

4.3.2 Computational time

At present, computational requirements are an important feature of any software/algorithm. For comparison of the Normal Gamma to other statistical methods we assess computational time. HyperLasso uses the MAP estimate which doesn't sample from the joint posterior distribution, but is quicker to compute. The Normal Gamma and Spike and slab do sample from the joint posterior distribution and so are computationally intensive. Here we assess whether the NG is prohibitively time consuming.

When checking the convergence we also monitor the time taken for the Normal Gamma to run. The mean and range of times taken for $n = 300$ and $p = 631$, and different chain lengths are tabulated in Table 4.6.

Computational time for simulated data (MAF 0.02, dataset 2B)			
	100,000 iterations	30,000 iterations	20,000 iterations
Time	15 hrs (7 hrs-22 hrs)	5.5 hrs (3 hrs-7 hrs)	4.2 hrs (2 hrs - 6 hrs)

Table 4.6: The computational time taken for each of the simulated datasets with different numbers of iterations used to calculate the R-hat convergence statistic. There are 300 individuals and 631 SNPs in the dataset.

We notice an almost linear increase in time taken with the number of iterations when the number of individuals (n) and the number of SNPs (p) remain constant as is expected. For the Normal Gamma prior on simulated data for the same number of iterations, there is a large range in computational time. This may be due to different values being sampled from the full conditional distributions which can lead to different parts of the code being run, see Section 5.5 for details of implementation of the NG; or it may be due to memory demands when storing the MCMC iterations. The difference in time to process the Normal Gamma may also be due to using Iceberg, the University of Sheffield high performance computing (HPC) facility [The University of Sheffield]. This processes many jobs at a time and so demands from other processes on the nodes of the system can also lead to increased computational time.

Compared to other methods, the Normal Gamma is much slower. In certain cases the Spike and slab can be similarly slow but HL, LS, piMASS and LR test all run in less than 1 hour for all datasets. However for datasets of this size, the time taken is not prohibitively long.

4.4 Information in the likelihood as a function of sample size

We assess, using a ROC curve, see Figure 4.7, how the Normal Gamma prior changes its performance with a change in the number of individuals on HapGen simulated dataset 2B from $n = 300$ to $n = 100$ and $n = 50$ individuals while maintaining the same number of total and causal SNPs. This allows us to understand how much information is in the likelihood as the sample size changes. We expect to see a decrease in the AUC when the number of individuals decreases.

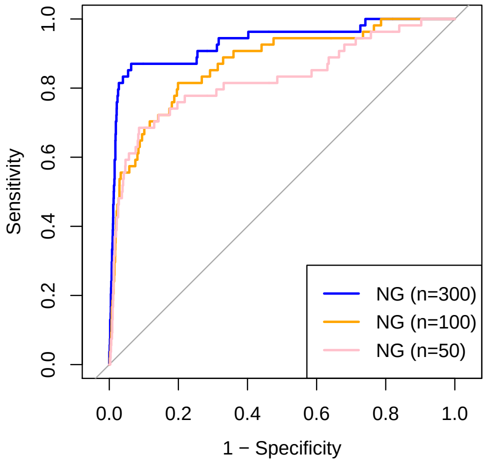


Figure 4.7: ROC curve comparing the different statistical methods for detecting causal SNPs on our simulated data with all 54 causal SNPs having effect size 0.4. The data was simulated using HapGen2 Su et al. [2011], targetting SNPs in the CASPASE8 region. Causal SNPs with a MAF of approximately 0.02 in the population (HapGen simulated data 2B).

The AUC for the Normal Gamma with $n = 300$ is 0.9322. This decreases to 0.8704 when there are 100 individuals, and decreases further to 0.8263 with only 50 individuals. All these AUCs show a respectable performance from the Normal Gamma, even compared to other methods with $n = 300$.

4.5 Conclusion

In this chapter, we have seen how the Normal Gamma performs very well on our simulated data. We notice that its performance is similar to other methods when the MAF of the causal SNPs is 0.02. In the case where the MAF is 0.2, the Normal Gamma and the Spike and slab prior are both superior to other methods. In the worst cases for the Normal Gamma, it is still comparable to the next best method.

We assessed convergence of the Normal Gamma visually in the form of a ROC curve and formally in terms of the R hat convergence diagnostic.

We do not vary the MAF (minor allele frequency) of the causal SNPs within a dataset. This could affect the ability to detect causal /associated SNPs but we cannot draw conclusions on this using our simulated datasets. Further understanding of these methods when run on eQTL data would require the investigation of these extra features.

Due to the performance and the ease of incorporating functional information, we choose the Normal Gamma as our method to develop, and continue to use the same comparison methods where possible.

Chapter 5

Implementing MCMC using the Normal Gamma Prior

We use the Normal Gamma prior as the basis for our eQTL study based on the results in Chapter 4 and the adaptive shrinkage framework of the model. In this chapter we describe, in detail, the prior structure of the Normal Gamma prior [Griffin and Brown, 2010], and how it can be implemented as computationally efficiently as possible.

We verify, and correct where necessary, the calculations stated in Griffin and Brown [2010] without derivation.

Throughout this chapter, and the thesis, we adopt the convention that variables in bold represent vectors, and variables that are capitalised represent matrices.

We can represent the relationship between the variables in the Normal Gamma using a DAG (directed acyclic graph). This can be found in Figure 5.1. The pseudo-code that we use to implement the Normal Gamma can be found in Appendix D.3.

5.1 Prior Structure.

The Normal Gamma prior, defined by Griffin and Brown [2010] and in Equations 5.1-5.6, is applied to a standard linear model of the form $\mathbf{y} = \alpha\mathbf{1}_n + X\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$, α represents background gene expression, $\mathbf{y}$ the response vector of gene expression, X the matrix of genotypes and $\boldsymbol{\beta}$ the vector of effect sizes.

Note that, in this chapter, n denotes the number of individuals, p denotes the number of SNPs, $X_{j,i}$ denotes genotype of the i^{th} SNP for individual j, β_i represents the effect size of SNP i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ represents a $p \times 1$

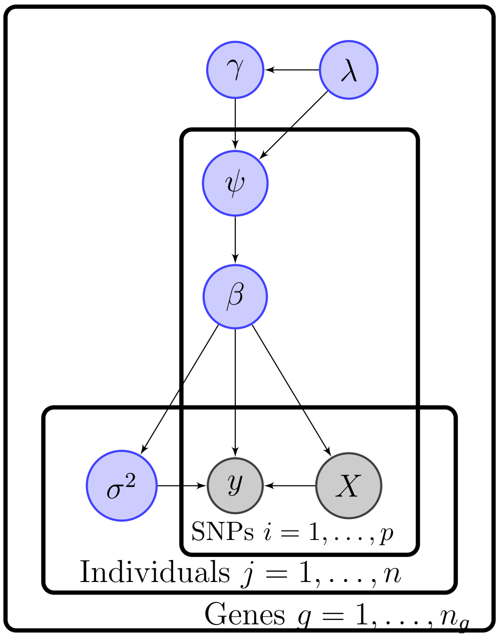


Figure 5.1: DAG representing the relationships between the variables in the NG model. The grey shaded nodes represent observed variables, the plates represent the loops and the arrows represent the relationships between the parameters.

vector of SNP effect sizes, y_j represents gene expression for individual j, with $\mathbf{y} = (y_1, \ldots, y_n)^T$ an $n \times 1$ vector of gene expressions, ψ_i represent the prior variance associated with β_i, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^T$.

$$\pi(\alpha) \propto 1 \tag{5.1}$$

$$\pi(\sigma^{-2}) \propto 1 \tag{5.2}$$

$$\pi(\lambda) \sim Ex\left(\frac{1}{2}\right) \tag{5.3}$$

$$\pi(\gamma^{-2}|\lambda) \sim Ga\left(2, \frac{M}{2\lambda}\right) \tag{5.4}$$

$$\pi(\psi_i|\lambda, \gamma^{-2}) \sim Ga\left(\lambda, \frac{1}{2\gamma^2}\right) \tag{5.5}$$

$$\pi(\beta_i|\psi_i) \sim N(0, \psi_i) \tag{5.6}$$

where $\pi(a)$ represents the prior distribution on a and M is a fixed scalar defined as $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$ where $\hat{\boldsymbol{\beta}}$ is the least squares (LS) estimate of $\boldsymbol{\beta}$ when X is non-singular. When X is singular, or when $p > n - 1$, M is redefined as $\frac{1}{n}\sum_{i=1}^{p} \hat{\beta}_i^2$, where $\hat{\beta}$ is the minimum length least squares (MLLS) estimate, see Section 3.4.1. The parametrisation of the Gamma distribution using shape and rate parameters is defined in Appendix A.

Griffin and Brown [2010] choose the constant M which represents the expectation of the marginal prior variance of β ($E[\pi(\text{var}(\beta_i|\lambda,\gamma^2))] = E[2\lambda\gamma^2] = M$). M represents an empirical estimate of the variance of the least squares estimates of β_i. It is used to control the amount of shrinkage enforced by the Normal Gamma by controlling the range of values for ψ. When including functional information within the Normal Gamma framework, we use this ability to control the amount of shrinkage enforced by the prior to prioritise SNPs with more evidence of a deleterious effect.

With the response variable defined as $\mathbf{y} = (y_1, \dots, y_n)$, the likelihood is defined as follows in Equation 5.7.

$$f(\mathbf{y}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2}) \sim N_n(\mathbf{y}-\alpha\mathbf{1}_n - X\boldsymbol{\beta}, \sigma^2\mathbf{I_n}), \tag{5.7}$$

where $\boldsymbol{\beta} = (\beta_1,\beta_2,\dots,\beta_p)^T$ is the parameter vector representing the effects of genetic variants, and N_n represents the multivariate normal (MVN) in n dimensions.

In the Normal Gamma prior hierarchy, we calculate the marginal prior variance of β_i using the law of total variance giving:

$$\begin{aligned}\text{var}(\beta_i|\lambda,\gamma) &= E_{\psi_i}[\text{var}(\beta_i|\lambda,\gamma,\psi_i)] + \text{var}_{\psi_i}(E[\beta_i|\lambda,\gamma,\psi_i]) \\ &= E_{\psi_i}[\psi_i|\lambda,\gamma] + 0 \\ &= 2\lambda\gamma^2.\end{aligned}$$

Hence we state $\text{var}(\pi(\beta_i|\lambda,\gamma)) = 2\lambda\gamma^2$. This is given an $IG(2,M)$ distribution so that $E[\pi(\text{var}(\beta_i|\lambda,\gamma))] = M$. M, defined as $\frac{1}{p}\sum_{i=1}^{p}\hat{\beta}_i^2$, provides an approximate estimate of the variance of the LS/MLLS estimates. Note that by the properties of the Gamma and Inverse-Gamma distributions $2\lambda\gamma^2 \sim IG(2,M) \implies \gamma^{-2}|\lambda \sim Ga\left(2,\frac{M}{2\lambda}\right)$.

In Griffin and Brown [2010] $\pi(\lambda)$ is stated as taking an $Ex(1)$ distribution but tuned to have an $E\left(\frac{1}{2}\right)$ distribution. For flexibility we leave it as $\pi(\lambda)$ when calculating the full conditional distributions.

5.2 Calculating the Full Conditional Distributions.

In this section we check and correct, where necessary, the calculations for the full conditional distributions stated without derivation in Griffin and Brown [2010].

In order to calculate the full conditional distribution for each parameter, we write down the joint posterior distribution and select all the terms involving our given parameter. To use Gibbs Sampling we need the full conditional distributions to have the form of a standard distribution, otherwise we calculate the acceptance probability for Metropolis-Hastings updating.

The joint posterior distribution is calculated using Bayes' Theorem as follows.

$$\begin{aligned}
&f(\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2}|\mathbf{y})\\
&\propto f(\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2},\mathbf{y})\\
&\propto \pi(\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2})f(\mathbf{y}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2})\\
&\propto \pi(\alpha)\pi(\sigma^{-2})\pi(\lambda)\pi(\gamma^{-2}|\lambda)\pi(\boldsymbol{\psi}|\gamma^{-2},\lambda)\pi(\boldsymbol{\beta}|\boldsymbol{\psi})f(\mathbf{y}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2}),
\end{aligned}$$

where $f(\mathbf{y}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\sigma^{-2})$ is the likelihood and $\pi(.)$ represents the prior distribution of the parameter.

Using the prior distributions, Equations 5.1-5.6, with the parametrisation of the Gamma distribution given in Appendix A, the joint posterior distribution is defined as:

$$\begin{aligned}
&\pi(\lambda)\times\frac{\left(\frac{M}{2\lambda}\right)^2}{\Gamma(2)}\left(\gamma^{-2}\right)^{2-1}\exp\left(\frac{-M}{2\lambda}\gamma^{-2}\right)\\
&\times\frac{1}{\Gamma(\lambda)}\left(\frac{1}{2\gamma^2}\right)^{\lambda}(\psi_i)^{\lambda-1}\exp\left(-\psi_i\frac{1}{2\gamma^2}\right)\\
&\times\prod_{i=1}^{p}\psi_i^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\left(\mathrm{diag}\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right)\\
&\times\left(\sigma^{-2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y}-\alpha\mathbf{1}_n-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1}_n-X\boldsymbol{\beta})\right).
\end{aligned}\tag{5.8}$$

5.2.1 Full Conditional Distribution for σ^{-2}.

Selecting the terms of Equation 5.8 involving σ^{-2}, we calculate the full conditional distribution up to proportionality as follows.

$$\begin{aligned}
&f(\sigma^{-2}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\mathbf{y})\\
&\qquad\propto(\sigma^2I_n)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\alpha-X\boldsymbol{\beta})^T(\sigma^2I_n)^{-1}(\mathbf{y}-\alpha-X\boldsymbol{\beta})\right)\\
&\qquad\propto\sigma^{-n}\exp\left(-\frac{1}{2}(\mathbf{y}-\alpha-X\boldsymbol{\beta})^T\frac{1}{\sigma^2}I_n(\mathbf{y}-\alpha-X\boldsymbol{\beta})\right)\\
&\qquad\propto(\sigma^{-2})^{\frac{n}{2}}\exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y}-\alpha-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha-X\boldsymbol{\beta})\right).
\end{aligned}\tag{5.9}$$

This is proportional to a Gamma distribution of the form:

$$\sigma^{-2}|\lambda, \gamma^{-2}, \boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \mathbf{y} \sim Ga\left(\frac{n}{2}+1,\ \frac{(\mathbf{y}-\alpha-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha-X\boldsymbol{\beta})}{2}\right). \quad (5.10)$$

There is a discrepancy between the shape parameters of the Gamma distributions in our calculations $\left(\frac{n}{2}+1\right)$ and those in [Griffin and Brown, 2010] $\left(\frac{n}{2}\right)$. We use our version for the full conditional distribution.

5.2.2 Full Conditional Distribution for ψ_i.

Selecting the terms of Equation 5.8 involving ψ_i, we calculate the full conditional distribution up to proportionality as follows.

$$\begin{aligned}
&f(\psi_i|\sigma^{-2}\lambda, \gamma^{-2}, \alpha, \beta_i, \mathbf{y}) \\
&\quad \propto \prod_{i=1}^{p}\left(\frac{1}{\Gamma(\lambda)}\left(\frac{1}{2\gamma^2}\right)^{\lambda}(\psi_i)^{\lambda-1}\exp\left(-\psi_i\frac{1}{2\gamma^2}\right)\frac{1}{\sqrt{\psi_i}}\exp\left(-\frac{1}{2\psi_i}\beta_i^2\right)\right) \\
&\quad \propto \prod_{i=1}^{p}\left((\psi_i)^{\lambda-1}\psi_i^{-\frac{1}{2}}\right)\exp\left(-\frac{1}{2\gamma^2}\sum_{i=1}^{p}\psi_i-\frac{1}{2}\sum_{i=1}^{p}\frac{\beta_i^2}{\psi_i}\right) \\
&\quad \propto \prod_{i=1}^{p}\left((\psi_i)^{\lambda-1-\frac{1}{2}}\right)\exp\left(-\frac{1}{2}\left(\frac{1}{\gamma^2}\sum_{i=1}^{p}\psi_i+\sum_{i=1}^{p}\frac{\beta_i^2}{\psi_i}\right)\right). \quad (5.11)
\end{aligned}$$

This is proportional to a Generalised Inverse Gaussian distribution of the form:

$$\psi_i|\sigma^{-2}\lambda, \gamma^{-2}, \alpha, \beta_i, \mathbf{y} \sim GIG\left(\lambda-\frac{1}{2},\ \frac{1}{\gamma^2},\ \beta_i^2\right). \quad (5.12)$$

To avoid confusion, we specify the parametrisation of the Generalised Inverse Gaussian density we use in Appendix A.

5.2.3 Full Conditional Distribution for γ^{-2}.

Selecting the terms of Equation 5.8 involving γ, we calculate the full conditional distribution up to proportionality as follows.

$$\begin{aligned}
&f(\gamma^{-2}|\sigma^{-2},\lambda,\psi_i,\alpha,\beta_i,\mathbf{y})\\
&\propto \left(\frac{M}{2\lambda}\right)^2 \frac{1}{\gamma^2}\exp\left(-\frac{M}{2\lambda}\frac{1}{\gamma^2}\right)\times\prod_{i=1}^{p}\left(\frac{1}{\Gamma(\lambda)}\left(\frac{1}{2\gamma^2}\right)^{\lambda}(\psi_i)^{\lambda-1}\exp\left(-\psi_i\frac{1}{2\gamma^2}\right)\right)\\
&\propto \frac{1}{\gamma^2}\exp\left(-\frac{M}{2\lambda\gamma^2}\right)\left(\frac{1}{\gamma^2}\right)^{p\lambda}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{\psi_i}{\gamma^2}\right)\\
&\propto (\gamma^{-2})^{p\lambda+1}\exp\left(-\gamma^{-2}\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right)\right). \qquad (5.13)
\end{aligned}$$

This is proportional to a Gamma distribution of the form:

$$\gamma^{-2}|\sigma^{-2},\lambda,\psi_i,\alpha,\beta_i,\mathbf{y}\sim Ga\left(p\lambda+2,\ \frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right). \qquad (5.14)$$

5.2.4 Full Conditional Distribution for $\boldsymbol{\phi}=(\alpha,\boldsymbol{\beta})^T$.

We update α and $\boldsymbol{\beta}$ simultaneously by updating $\boldsymbol{\phi}=(\alpha,\boldsymbol{\beta})^T$. We therefore select the terms of Equation 5.8 involving α and $\boldsymbol{\beta}$ and use these terms to calculate the full conditional distribution.

As in Griffin and Brown [2010], we begin by defining $X^*=[1:X]$. This is our full design matrix for $\boldsymbol{\phi}$ and takes into account both α and $\boldsymbol{\beta}$. We also define

$$\Lambda = diag\left(0,\frac{1}{\psi_1},\frac{1}{\psi_2},\ldots,\frac{1}{\psi_p}\right). \qquad (5.15)$$

$$\begin{aligned}
&f(\boldsymbol{\phi}|\gamma^{-2},\sigma^{-2},\lambda,\psi_i,\mathbf{y})\\
&\propto\prod_{i=1}^{p}\psi_i^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\left(diag\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right)\\
&\times(\sigma^{-2})^{\frac{n}{2}}\exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})\right)\\
&\propto\exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\left(diag\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right)\exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})\right)\\
&=\exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}^T diag\left(\frac{1}{\psi_i}\right)\boldsymbol{\beta}+\sigma^{-2}(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})\right]\right)\\
&=\exp\left(-\frac{1}{2}\left[\boldsymbol{\phi}^T\Lambda\boldsymbol{\phi}+\sigma^{-2}(\mathbf{y}-X^*\boldsymbol{\phi})^T(\mathbf{y}-X^*\boldsymbol{\phi})\right]\right)\\
&\propto\exp\left(-\frac{1}{2\sigma^2}\left(\sigma^2\boldsymbol{\phi}^T\Lambda\boldsymbol{\phi}+\boldsymbol{\phi}^T X^{*T}X^*\boldsymbol{\phi}-2\boldsymbol{\phi}^T X^{*T}\mathbf{y}\right)\right)
\end{aligned}$$

since $\boldsymbol{\phi}^T X^{*T}\mathbf{y} = \mathbf{y}^T X^* \boldsymbol{\phi}$ is a scalar

$$\propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\phi}^T\left(\sigma^2\Lambda + X^{*T}X^*\right)\boldsymbol{\phi} - 2\boldsymbol{\phi}^T X^{*T}\mathbf{y}\right)\right)$$

Let $\left(\sigma^2\Lambda + X^{*T}X^*\right) = A$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\left(A\boldsymbol{\phi} - 2\boldsymbol{\phi}^T X^{*T}\mathbf{y}\right)\right)$$

By completing the square, we obtain

$$\propto \exp\left(-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\phi} - A^{-1}X^{*T}\mathbf{y}\right)^T A\left(\boldsymbol{\phi} - A^{-1}X^{*T}\mathbf{y}\right)\right)\right). \tag{5.16}$$

This takes the form of a Multivariate Normal distribution with mean

$$\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1} X^{*T}\mathbf{y}$$

and variance

$$\left(\sigma^{-2}A\right)^{-1} = \left(\sigma^{-2}\left(\sigma^2\Lambda + X^{*T}X^*\right)\right)^{-1} = \sigma^2\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1}.$$

Hence the full conditional distribution for $\boldsymbol{\phi}$ is:

$$\boldsymbol{\phi}|\gamma^{-2}, \sigma^{-2}, \lambda, \boldsymbol{\psi}, \mathbf{y} \sim N_{p+1}\left(\left(X^{*T}X^* + \sigma^2\Lambda\right)^{-1} X^{*T}\mathbf{y}, \sigma^2\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1}\right). \tag{5.17}$$

5.2.5 Full Conditional Distribution for λ.

Selecting the terms of Equation 5.8 involving λ, we calculate the full conditional distribution up to proportionality as follows.

$$\begin{aligned}
&f(\lambda|\gamma^{-2}, \boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y}) \\
&\quad \propto \pi(\lambda) \times \prod_{i=1}^{p} \frac{1}{\Gamma(\lambda)}\left(\frac{1}{2\gamma^2}\right)^{\lambda} (\psi_i)^{\lambda-1} \exp\left(-\psi_i \frac{1}{2\gamma^2}\right) \times \\
&\quad \frac{\left(\frac{M}{2\lambda}\right)^2}{\Gamma(2)} \left(\gamma^{-2}\right)^{2-1} \exp\left(\frac{-M}{2\lambda}\gamma^{-2}\right) \\
&\quad \propto \pi(\lambda)\left(\frac{M}{2\lambda}\right)^2 \exp\left(\frac{-M}{2\lambda}\gamma^{-2}\right)\left(\frac{1}{\Gamma(\lambda)}\right)^p \left(\frac{1}{2\gamma^2}\right)^{p\lambda} \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \\
&\quad \propto \pi(\lambda) \frac{1}{(\Gamma(\lambda))^p \left(2\gamma^2\right)^{p\lambda}} \left(\frac{M}{2\lambda}\right)^2 \exp\left(\frac{-M}{2\lambda}\gamma^{-2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda-1}
\end{aligned} \tag{5.18}$$

This is not a standard distribution up to proportionality and so it is not possible to update λ using Gibbs Sampler, we therefore have to use the Metropolis-Hastings algorithm.

Having calculated the full conditional distribution, this allows us to calculate the acceptance probability for our proposed λ, which we call λ'. The acceptance probability is defined as:

$$\min\left(1, \frac{f(\lambda'|\gamma^{-2}, \boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})}{f(\lambda|\gamma^{-2}, \boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})} \frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)}\right) \tag{5.19}$$

In order to calculate this, we take the minimum of 1 and the ratio between the full conditional distribution for λ evaluated at $\lambda = \lambda'$ and the full conditional of λ evaluated at $\lambda = \lambda$ as well as the proposal distributions for $\lambda|\lambda'$ and vice-versa.

We know that the prior for $\gamma^{-2}|\lambda$ is derived from $E[\pi(\psi|\lambda, \gamma)] = 2\lambda\gamma^2 \sim IG(2, M)$. We therefore need to take this into account when we update λ. Hence we define $\gamma'^2 = \frac{2\lambda\gamma^2}{2\lambda'}$ based on the newly updated value for γ^2, and update as if we were updating λ and γ^{-2} jointly.

This gives an updated joint full conditional distribution for λ and γ^2 to be:

$$\begin{aligned}
&f(\lambda, \gamma^{-2}|\boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y}) \\
&\quad\propto \pi(\lambda) \times \prod_{i=1}^{p} \frac{1}{\Gamma(\lambda)} \left(\frac{1}{2\gamma^2}\right)^{\lambda} (\psi_i)^{\lambda-1} \exp\left(-\psi_i \frac{1}{2\gamma^2}\right) \times \left(\frac{M}{2\lambda}\right)^2 (\gamma^{-2}) \exp\left(\frac{-M}{2\lambda\gamma^2}\right) \\
&\quad\propto \pi(\lambda) \left(\frac{1}{\Gamma(\lambda)}\right)^p \left(\frac{1}{2\gamma^2}\right)^{p\lambda} \exp\left(-\sum_{i=1}^{p} \psi_i \frac{1}{2\gamma^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \times \left(\frac{1}{2\lambda}\right) \left(\frac{M^2}{2\lambda\gamma^2}\right) \exp\left(\frac{-M}{2\lambda\gamma^2}\right) \\
&\quad\propto \pi(\lambda) \frac{1}{(\Gamma(\lambda))^p (2\gamma^2)^{p\lambda}} \exp\left(-\sum_{i=1}^{p} \psi_i \frac{1}{2\gamma^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \times \left(\frac{1}{2\lambda}\right) \left(\frac{M^2}{2\lambda\gamma^2}\right) \exp\left(\frac{-M}{2\lambda\gamma^2}\right).
\end{aligned} \tag{5.20}$$

We can then use the ratio of this updated full conditional distributed evaluated at λ' and γ' divided by the ratio evaluated at λ and γ to derive the first part of the acceptance probability for λ in the Metropolis-Hasting updating.

The first part of the acceptance probability, the ratio of the full conditional distributions given in Equation 5.20 is:

$$\begin{aligned}
&\frac{f(\lambda', \gamma'^2|\boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})}{f(\lambda, \gamma^2|\boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})} \\
&\quad= \frac{\pi(\lambda') \frac{1}{(\Gamma(\lambda'))^p (2\gamma'^2)^{p\lambda'}} \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma'^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda'-1} \left(\frac{1}{2\lambda'}\right) \left(\frac{M^2}{2\lambda'\gamma'^2}\right) \exp\left(\frac{-M}{2\lambda'\gamma'^2}\right)}{\pi(\lambda) \frac{1}{(\Gamma(\lambda))^p (2\gamma^2)^{p\lambda}} \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \left(\frac{1}{2\lambda}\right) \left(\frac{M^2}{2\lambda\gamma^2}\right) \exp\left(\frac{-M}{2\lambda\gamma^2}\right)}
\end{aligned}$$

Substituting $2\lambda\gamma^2 = 2\lambda'\gamma'^2$

$$\begin{aligned}
&\quad= \frac{\lambda\pi(\lambda') \frac{1}{(\Gamma(\lambda'))^p (2\gamma'^2)^{p\lambda'}} \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma'^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda'-1} \left(\frac{M^2}{2\lambda'\gamma'^2}\right) \exp\left(\frac{-M}{2\lambda'\gamma'^2}\right)}{\lambda'\pi(\lambda) \frac{1}{(\Gamma(\lambda))^p (2\gamma^2)^{p\lambda}} \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma^2}\right) \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \left(\frac{M^2}{2\lambda'\gamma'^2}\right) \exp\left(\frac{-M}{2\lambda'\gamma'^2}\right)} \\
&\quad= \frac{\lambda}{\lambda'} \frac{\pi(\lambda')}{\pi(\lambda)} \left(\frac{\Gamma(\lambda)}{\Gamma(\lambda')}\right)^p \frac{(2\gamma^2)^{p\lambda}}{(2\gamma'^2)^{p\lambda'}} \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma'^2} + \frac{\sum_{i=1}^{p}\psi_i}{2\gamma^2}\right) \left(\prod_{i=1}^{p}(\psi_i)\right)^{\lambda'-\lambda}.
\end{aligned} \tag{5.21}$$

The proposal value λ' is defined in Griffin and Brown [2010] as $\lambda' = exp(\sigma^2_\lambda z)\lambda$, where $\lambda > 0$ and z is a value sampled from the standard normal distribution. This is symmetric so is not needed to be included in the acceptance probability. σ^2_λ is chosen to ensure that the Markov Chain explores the space effectively without proposing too many jumps that are rejected. Specifically, σ^2_λ is, according to Griffin and Brown [2010], chosen such that the acceptance rate is approximately 20%-30%. Simulations led us to use $\sigma^2_\lambda = 0.05$ in all our MCMC routines.

We now define the acceptance probability of λ to be:

$$\min\left\{1, \frac{f(\lambda', {\gamma'}^2|\boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})}{f(\lambda, \gamma^2|\boldsymbol{\psi}, \alpha, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{y})} \frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)}\right\}$$
$$= \min\left\{1, \frac{\lambda'}{\lambda}\frac{\pi(\lambda')}{\pi(\lambda)}\left(\frac{\Gamma(\lambda)}{\Gamma(\lambda')}\right)^p \frac{(2\gamma^2)^{p\lambda}}{(2\gamma'^2)^{p\lambda'}} \exp\left(-\frac{\sum_{i=1}^p \psi_i}{2{\gamma'}^2} + \frac{\sum_{i=1}^p \psi_i}{2\gamma^2}\right)\left(\prod_{i=1}^p (\psi_i)\right)^{\lambda'-\lambda}\right\} \tag{5.22}$$

For the purposes of coding the Normal Gamma, we use the prior distribution for λ suggested by Griffin and Brown [2010], $\pi(\lambda) \sim Ex(\frac{1}{2})$. Based on our results in Chapter 4, this prior distribution works well for our simulated datasets. The acceptance probability stated here is different to the published version in Griffin and Brown [2010].

5.3 Proving the posterior is proper

As we have two improper priors on σ^{-2} and α we need to ensure that the posterior is proper before continuing. Initially we can define requirements such that all the full conditional distributions are valid. This means that for the posterior to be proper, we require the following:

1. $\frac{n}{2} + 1 > 0$
2. $\frac{(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})}{2} > 0$
3. $p\lambda + 2 > 0$
4. $\frac{M}{2\lambda} + \frac{1}{2}\sum_{i=1}^p \psi_i > 0$
5. $\sigma^2\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1} > 0$.

These constrains ensure that none of the full conditionals are improper. This does not ensure the posterior is proper, but if these conditions are not met then the posterior is improper.

We now take the joint posterior distribution defined in Equation 5.8 and integrate this with respect to all parameters to show that the posterior is proper.

We begin by rearranging the joint posterior and stating it up to proportionality.

$$\begin{aligned}
&\pi(\lambda) \times \frac{\left(\frac{M}{2\lambda}\right)^2}{\Gamma(2)} \left(\gamma^{-2}\right)^{2-1} \exp\left(\frac{-M}{2\lambda}\gamma^{-2}\right) \\
&\times \frac{1}{\Gamma(\lambda)} \left(\frac{1}{2\gamma^2}\right)^{\lambda} (\psi_i)^{\lambda-1} \exp\left(-\psi_i \frac{1}{2\gamma^2}\right) \\
&\times \prod_{i=1}^{p} \psi_i^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \left(\mathrm{diag}\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right) \\
&\times \left(\sigma^{-2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})^T(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})\right) \\
&\propto \pi(\lambda) \left(\frac{1}{\lambda\gamma}\right)^2 \exp\left(-\frac{M}{2\lambda\gamma^2}\right) \left(\frac{1}{\Gamma(\lambda)}\right)^p \left(\frac{1}{2\gamma^2}\right)^{p\lambda} \exp\left(-\sum_{i=1}^{p}\psi_i \frac{1}{2\gamma^2}\right) \\
&\times \prod_{i=1}^{p} \psi_i^{\lambda-1-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \left(\mathrm{diag}\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right) \\
&\times \left(\sigma^{-2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})^T(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})\right) \\
&\propto \pi(\lambda) \left(\frac{1}{\lambda\gamma}\right)^2 \exp\left(-\frac{M}{2\lambda\gamma^2}\right) \left(\frac{1}{\Gamma(\lambda)}\right)^p \left(\frac{1}{2\gamma^2}\right)^{p\lambda} \exp\left(-\sum_{i=1}^{p}\psi_i \frac{1}{2\gamma^2}\right) \\
&\times \prod_{i=1}^{p} \psi_i^{\lambda-1-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \left(\mathrm{diag}\left(\frac{1}{\psi_i}\right)\right)\boldsymbol{\beta}\right) \\
&\times \left(\sigma^{-2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})^T(\mathbf{y} - \alpha\mathbf{1}_n - X\boldsymbol{\beta})\right).
\end{aligned}$$

If we use the rearranging and simplification of combining $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ from Equation 5.16, we define the joint posterior as follows, defining $\left(\sigma^2\Lambda + X^{*T}X^*\right) = A$.

$$\begin{aligned}
&\propto \pi(\lambda) \left(\frac{1}{\lambda\gamma}\right)^2 \exp\left(-\frac{M}{2\lambda\gamma^2}\right) \left(\frac{1}{\Gamma(\lambda)}\right)^p \left(\frac{1}{2\gamma^2}\right)^{p\lambda} \exp\left(-\sum_{i=1}^{p}\psi_i \frac{1}{2\gamma^2}\right) \left(\sigma^{-2}\right)^{\frac{n}{2}} \\
&\times \left(\prod_{i=1}^{p} \psi_i^{\lambda-1-\frac{1}{2}}\right) \exp\left(-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\phi} - A^{-1}X^{*T}\mathbf{y}\right)^T A \left(\boldsymbol{\phi} - A^{-1}X^{*T}\mathbf{y}\right)\right)\right).
\end{aligned}$$

We are now ready to integrate the joint posterior density.

We begin by integrating with respect to $\boldsymbol{\phi}$ using the full conditional distribution for $\boldsymbol{\phi}$ to calculate the value of the integral, giving the following. We maintain this up to proportionality, removing only the constants.

$$\pi(\lambda)\left(\frac{1}{\lambda\gamma}\right)^{2}\exp\left(-\frac{M}{2\lambda\gamma^{2}}\right)\left(\frac{1}{\Gamma(\lambda)}\right)^{p}\left(\frac{1}{2\gamma^{2}}\right)^{p\lambda}\exp\left(-\sum_{i=1}^{p}\psi_{i}\frac{1}{2\gamma^{2}}\right)\left(\sigma^{-2}\right)^{\frac{n}{2}}$$
$$\times\left(\prod_{i=1}^{p}\psi_{i}^{\lambda-1-\frac{1}{2}}\right)\det\left(\sigma^{-2}A\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^{2}}y^{T}y\right).$$

Replacing A and simplifying, we have:

$$\pi(\lambda)\left(\frac{1}{\lambda\gamma}\right)^{2}\exp\left(-\frac{M}{2\lambda\gamma^{2}}\right)\left(\frac{1}{\Gamma(\lambda)}\right)^{p}\left(\frac{1}{2\gamma^{2}}\right)^{p\lambda}\exp\left(-\sum_{i=1}^{p}\psi_{i}\frac{1}{2\gamma^{2}}\right)\left(\sigma^{-2}\right)^{\frac{n}{2}}$$
$$\times\left(\prod_{i=1}^{p}\psi_{i}^{\lambda-1-\frac{1}{2}}\right)\det\left(\sigma^{2}\left(\sigma^{2}\Lambda+X^{*T}X^{*}\right)^{-1}\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^{2}}y^{T}y\right).$$

We now choose to integrate with respect to γ^{-2}. Again we use the full conditional distribution to enable us to calculate the value of the integral which helps us to obtain the following.

$$=\pi(\lambda)\left(\frac{1}{\lambda}\right)^{2}\left(\frac{1}{\Gamma(\lambda)}\right)^{p}\left(\sigma^{-2}\right)^{\frac{n}{2}+\frac{p+1}{2}}\left(\prod_{i=1}^{p}\psi_{i}^{\lambda-1-\frac{1}{2}}\right)\det\left(\sigma^{2}\left(\sigma^{2}\Lambda+X^{*T}X^{*}\right)^{-1}\right)^{-\frac{1}{2}}$$
$$\times\,\Gamma(p\lambda+2)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_{i}\right)^{-(p\lambda+2)}\exp\left(-\frac{1}{2\sigma^{2}}y^{T}y\right).$$

At this point there are no more standard distributions we can use to help us to integrate out the remaining parameters (λ, σ^{-2} and ψ_i). We therefore bound as many parts of the function as possible to enable us to integrate the remaining parts of the posterior.

We aim to bound the following parts of the joint posterior for the reasons explained.

1. $\det\left(\sigma^{2}\left(\sigma^{2}\Lambda+X^{*T}X^{*}\right)^{-1}\right)^{-\frac{1}{2}}$ has to be bounded to remove the dependence on σ^2 and ψ_i (the diagonal elements of Λ).

2. $\dfrac{\Gamma(p\lambda+2)}{(\Gamma(\lambda))^{p}}$ needs to be bounded so that we can integrate with respect to λ.

Firstly we can bound the determinant as follows:

$$\det\left(\sigma^{2}\left(\sigma^{2}\Lambda+X^{*T}X^{*}\right)^{-1}\right)^{-\frac{1}{2}}\leq\left(\sigma^{-2}\right)^{\frac{p+1}{2}}\det\left(\left(\sigma^{2}\Lambda+X^{*T}X^{*}\right)^{-1}\right)^{-\frac{1}{2}}.$$

Next we bound $\det\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1}$. We know that the maximum entry of $X^{*T}X^*$ is $(p+1)$ when using the $\{0,1\}$ coding for SNPs or $2(p+1)$ when using the $\{0,1,2\}$ coding. We also know that $\sigma^2\Lambda \geq 0$ as Λ is a diagonal matrix with diagonal elements of $\{0, \frac{1}{\psi_1}, \frac{1}{\psi_2}, \dots, \frac{1}{\psi_p}\}$ where $\psi_i > 0$ and $\sigma^2 > 0$ by definition. We can therefore bound the matrix using this information and Hadamard's inequality such that

$$\det\left(\left(\sigma^2\Lambda + X^{*T}X^*\right)^{-1}\right) \leq \left(2(p+1)\right)^{p+1}(p+1)^{\frac{p+1}{2}}.$$

This removes the dependence on ψ_i and σ^2 from our posterior distribution.

Using the upper bounding, we can now integrate the following with respect to σ^{-2}.

$$\pi(\lambda)\left(\frac{1}{\lambda}\right)^2\left(\frac{1}{\Gamma(\lambda)}\right)^p\left(\sigma^{-2}\right)^{\frac{n}{2}+\frac{p+1}{2}}\left(\prod_{i=1}^{p}\psi_i^{\lambda-1-\frac{1}{2}}\right)\Gamma(p\lambda+2)$$
$$\times\exp\left(-\frac{1}{2\sigma^2}y^Ty\right)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right)^{-(p\lambda+2)}.$$

We obtain the following up to proportionality (having removed anything that is constant with respect to ψ_i or λ) using the similarity to the Gamma distribution to enable us to integrate out σ^{-2}. We note that $\frac{n+p+3}{2} > 0$ and $\frac{1}{2}y^Ty$ is positive definite by definition meaning that the integral is well defined.

$$\pi(\lambda)\left(\frac{1}{\lambda}\right)^2\left(\frac{1}{\Gamma(\lambda)}\right)^p\left(\prod_{i=1}^{p}\psi_i^{\lambda-1-\frac{1}{2}}\right)\Gamma(p\lambda+2)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right)^{-(p\lambda+2)}.$$

We now need to bound the following to enable us to integrate with respect to λ and ψ_i.

1. $\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right)^{-(p\lambda+2)}$ needs to be bounded so we can integrate with respect to ψ_i.

2. $\dfrac{\Gamma(p\lambda+2)}{(\Gamma(\lambda))^p}$ needs to be bounded so that we can integrate with respect to λ.

To bound the Gamma function, we use the Stirling formula, which states:

$$\Gamma(x+1) \approx \frac{x^x\exp\{-x\}}{\sqrt{2\pi x}}.$$

This gives:

$$\begin{aligned}
\frac{\Gamma(p\lambda+2)}{(\Gamma(\lambda))^p} &\approx \frac{\dfrac{(p\lambda+2-1)^{(p\lambda+2-1)}\exp\{-(p\lambda+2-1)\}}{\sqrt{2\pi(p\lambda+2-1)}}}{\left(\dfrac{(\lambda-1)^{(\lambda-1)}\exp\{-(\lambda-1)\}}{\sqrt{2\pi(\lambda-1)}}\right)^p} \\
&\propto \frac{\dfrac{(p\lambda+1)^{(p\lambda+1)}\exp\{-(p\lambda+1)\}}{\sqrt{p\lambda+1}}}{\left(\dfrac{(\lambda-1)^{(\lambda-1)}\exp\{-(\lambda-1)\}}{\sqrt{\lambda-1}}\right)^p} \\
&= \frac{(p\lambda+1)^{(p\lambda+1)}}{(\lambda-1)^{(\lambda-1)}}\frac{\sqrt{\lambda-1}}{\sqrt{p\lambda+1}}\frac{\exp\{-(p\lambda+1)\}}{(\exp\{-(\lambda-1)\})^p} \\
&= \frac{(p\lambda+1)^{(p\lambda+1)}}{(\lambda-1)^{(\lambda-1)}}\frac{\sqrt{\lambda-1}}{\sqrt{p\lambda+1}}\frac{\exp\{-(p\lambda+1)\}}{\exp\{-p(\lambda-1)\}} \\
&= \frac{(p\lambda+1)^{(p\lambda+1)}}{(\lambda-1)^{(\lambda-1)}}\frac{\sqrt{\lambda-1}}{\sqrt{p\lambda+1}}\exp\{-(p\lambda+1)-p(\lambda-1)\} \\
&= \frac{(p\lambda+1)^{(p\lambda+1)}}{(\lambda-1)^{(\lambda-1)}}\frac{\sqrt{\lambda-1}}{\sqrt{p\lambda+1}}\exp\{-2p\lambda\} \\
&= \frac{(p\lambda+1)^{(p\lambda+1)}}{(\lambda-1)^{(\lambda-1)}}\left(\frac{\lambda-1}{p\lambda+1}\right)^{\frac{1}{2}}\exp\{-2p\lambda\} \\
&= \frac{(p\lambda+1)^{(p\lambda+\frac{1}{2})}}{(\lambda-1)^{(\lambda-\frac{1}{2})}}\exp\{-2p\lambda\}.
\end{aligned}$$

We still need to bound this to enable us to integrate with respect to λ. However, at this stage we have the following to integrate with respect to λ and ψ_i.

$$\pi(\lambda)\left(\frac{1}{\lambda}\right)^2\left(\prod_{i=1}^p \psi_i^{\lambda-1-\frac{1}{2}}\right)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^p\psi_i\right)^{-(p\lambda+2)}\frac{(p\lambda+1)^{(p\lambda+\frac{1}{2})}}{(\lambda-1)^{(\lambda-\frac{1}{2})}}\exp\{-2p\lambda\}.$$

We also know that $\pi(\lambda)\sim \text{Ex}\left(\frac{1}{2}\right)$, hence we can include this to give:

$$\begin{aligned}
&\exp\left(-\frac{1}{2}\lambda\right)\left(\frac{1}{\lambda}\right)^2\left(\prod_{i=1}^p \psi_i^{\lambda-1-\frac{1}{2}}\right)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^p\psi_i\right)^{-(p\lambda+2)}\frac{(p\lambda+1)^{(p\lambda+\frac{1}{2})}}{(\lambda-1)^{(\lambda-\frac{1}{2})}}\exp\left(-2p\lambda\right) \\
&= \left(\prod_{i=1}^p \psi_i^{\lambda-1-\frac{1}{2}}\right)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^p\psi_i\right)^{-(p\lambda+2)}\frac{(p\lambda+1)^{(p\lambda+\frac{1}{2})}}{\lambda^2(\lambda-1)^{(\lambda-\frac{1}{2})}}\exp\left(-2p\lambda-\frac{1}{2}\lambda\right) \\
&= \left(\prod_{i=1}^p \psi_i^{\lambda-1-\frac{1}{2}}\right)\left(\frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^p\psi_i\right)^{-(p\lambda+2)}\frac{(p\lambda+1)^{(p\lambda+\frac{1}{2})}}{\lambda^2(\lambda-1)^{(\lambda-\frac{1}{2})}}\exp\left(-\lambda\left(2p+\frac{1}{2}\right)\right).
\end{aligned}$$

When integrating with respect to λ, the integral diverges due to the terms of the form λ^λ. The integral with respect to ψ_i is more difficult to calculate as there is a term involving $\prod_{i=1}^{p}\psi_i$ and in the exponent there is the $\sum_{i=1}^{p}\psi_i$. However, as the form of the terms involving ψ_i do not follow the form of any standard distribution, when we bound the integral and substitute in the limits of 0 and ∞, the value of the integral will be of the form $\infty - 0$. Hence this integral is infinite when we integrate with respect to both λ and ψ_i, in either order.

However, we note that, due to bounding, this is the upper limit of the integral of the joint posterior density, and hence we know that the value of the integral of the join density is strictly less than infinity. Therefore the posterior is proper.

5.4 Summary of Full Conditional Distributions for Normal Gamma.

The full conditional distributions for σ^{-2}, ψ_i, γ^{-2}, and $\boldsymbol{\phi}$ are as follows:

$$\sigma^{-2}|\lambda,\gamma^{-2},\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\mathbf{y} \sim Ga\left(\frac{n}{2}+1,\ \frac{(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})^T(\mathbf{y}-\alpha\mathbf{1_n}-X\boldsymbol{\beta})}{2}\right).$$

$$\psi_i|\sigma^{-2}\lambda,\gamma^{-2},\alpha,\boldsymbol{\beta},\mathbf{y} \sim GIG\left(\lambda-\frac{1}{2},\frac{1}{\gamma^2},\beta_i^2\right).$$

$$\gamma^{-2}|\sigma^{-2},\lambda,\boldsymbol{\psi},\alpha,\boldsymbol{\beta},\mathbf{y} \sim Ga\left(p\lambda+2,\ \frac{M}{2\lambda}+\frac{1}{2}\sum_{i=1}^{p}\psi_i\right).$$

$$\boldsymbol{\phi}|\gamma^{-2},\sigma^{-2},\lambda,\boldsymbol{\psi},\mathbf{y} \sim N_{p+1}\left(\left(X^{*T}X^{*}+\sigma^2\Lambda\right)^{-1}X^{*T}y,\sigma^2\left(\sigma^2\Lambda+X^{*T}X^{*}\right)^{-1}\right).$$

The full conditional for λ cannot be updated using Gibbs sampling as the distributions are not conjugate. We therefore update using the Metropolis-Hastings acceptance probability

$$= \min\left\{1,\frac{\lambda'}{\lambda}\frac{\pi(\lambda')}{\pi(\lambda)}\left(\frac{\Gamma(\lambda)}{\Gamma(\lambda')}\right)^{p}\frac{(2\gamma^2)^{p\lambda}}{(2\gamma'^2)^{p\lambda'}}exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2\gamma'^2}+\frac{\sum_{i=1}^{p}\psi_i}{2\gamma^2}\right)\left(\prod_{i=1}^{p}(\psi_i)\right)^{\lambda'-\lambda}\right\} \tag{5.23}$$

where $\lambda' = \exp(\sigma_\lambda^2 z)\lambda$ is the proposal value for λ and $\gamma'^2 = \frac{2\lambda\gamma^2}{2\lambda'}$ is the corresponding value of γ^2 given the proposed value λ'.

5.5 Implementation of the Normal Gamma Prior.

When implementing the Normal Gamma, we can improve the computational burden using mathematical reformulation of the full conditional distributions. We need to consider the numerical accuracy of the software and reformulate the full conditionals to take this into account. At times we have to use modified methods that mitigate these computational limitations. Many of these reformulations are stated, without derivation, in Griffin and Brown [2010].

5.5.1 Improving the accuracy when sampling from the Full Conditional for $\boldsymbol{\phi}$.

To increase sampling efficiency when sampling from the full conditional distribution for $\boldsymbol{\phi}$, Equation 5.17, page 75, we standardise the multivariate normal using a Cholesky decomposition. This allows us to sample from the standardised MVN. We standardise the MVN using the linear transformation property.

The Linear Transformation Property states that if $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{x}+\mathbf{b} \sim N_q(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^T)$, where $\mathbf{A}$ is a $q \times p$ matrix, and $\mathbf{b}$ is a $q \times 1$ vector.

To standardise $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we first centre X so that $\mathbf{X} - \boldsymbol{\mu} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. We then use the Cholesky decomposition of the positive definite matrix $\boldsymbol{\Sigma}$ into

$$\boldsymbol{\Sigma} = C^T C \implies (C^T)^{-1}\boldsymbol{\Sigma}C^{-1} = I_p. \tag{5.24}$$

Since C is invertible, we have

$$CC^{-1} = I \implies (C^{-1})^T C^T = I \implies (C^{-1})^T = (C^T)^{-1} \implies ((C^T)^{-1})^T = C^{-1} \tag{5.25}$$

Using the linear transformation property on $\mathbf{X} - \boldsymbol{\mu} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ yields

$$\begin{aligned}
(C^T)^{-1}(\mathbf{X} - \boldsymbol{\mu}) &\sim N_p(0, (C^T)^{-1}\boldsymbol{\Sigma}((C^T)^{-1})^T) \\
&\sim N_p(0, (C^T)^{-1}\boldsymbol{\Sigma}C^{-1}) && \text{using Equation 5.25} \\
&\sim N_p(0, I_p) && \text{using Equation 5.24}
\end{aligned}$$

The Cholesky decomposition or factorisation reduces a positive definite matrix $\mathbf{B}$ into the product of two upper, or lower, triangular matrices, $\mathbf{R}$ such that $\mathbf{R}^T\mathbf{R} = \mathbf{B}$.

We sample a vector $\mathbf{y}$ from this standardised MVN, and transform it such that the vector we want to sample $\mathbf{X}$ is given by $\mathbf{X} = \boldsymbol{\mu} + C^T\mathbf{y}$.

We use this approach when sampling from the full conditional distribution for $\boldsymbol{\phi}$ as given in Equation 5.17.

5.5.2 Updating $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ in two stages.

When sampling from the multivariate normal (MVN) distribution representing the full conditional distribution for $\boldsymbol{\phi}$, Equation 5.17, large differences in the variance parameters ψ_i can lead to the sampled values being inaccurate. We have defined $\Lambda = \left(0, \frac{1}{\psi_1}, \ldots, \frac{1}{\psi_p}\right)$, hence ψ_i is involved in the full conditional distribution of $\boldsymbol{\phi}$.

When we calculate the updated values for $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ computational problems arise when there are large differences between the maximum and minimum ψ_i values. One way to overcome these problems is to update $\boldsymbol{\phi}$ in two stages. The two groups of ψ_i to be updated depend on whether the ψ_i are large, i.e. above a certain threshold T_1, or small, i.e. below a different threshold T_2. In the case used in Griffin and Brown [2010], $T_1 < T_2$ therefore some ψ_i fall into both categories and actually get updated twice in each MCMC iteration.

The threshold we use here depends on $\hat{x}$, which is defined by Griffin and Brown [2010] to be 10^8. This is this limit for the ratio of the maximum ψ_i value to the minimum ψ_i value. If the ratio is greater than this then the updating is done in stages, i.e. if $\dfrac{max(\psi_i)}{min(\psi_i)} > \hat{x}$ then we update in stages. The thresholds for the stages of updating are $\psi_i > min(\psi_i)\frac{\hat{x}}{10}$ for larger values of ψ_i, and $\psi_i < min(\psi_i) \times 10\hat{x}$ for smaller values of ψ_i, giving overlapping groups.

We split the full conditional distribution into two stages and then we apply the Cholesky decomposition (see Section 5.5.1). We use the two stage updating to overcome problems with efficiency of sampling both large and small values and any corresponding accuracy issues with respect to rounding and storing of values. The Cholesky decomposition is used as it allows us to sample from an uncorrelated MVN distribution and then transform using the decomposition so that we have sampled from the desired MVN in a computationally more efficient and accurate way.

When we update in stages, this changes the full conditional distribution for ϕ_i which is stated in Equation 5.17. We assume that we are updating α and the first $(q-1)$ β_i values, the column vector of which we call $\boldsymbol{\phi}_F$. The corresponding data matrix will be referred to as $X^{*F} = X^*_{i,1}, \ldots, X^*_{i,q}$. The elements of $\boldsymbol{\phi}$ that are not updated will be referred to as $\boldsymbol{\phi}_S$, similarly we will also refer to $X^{*S} = X^*_{i,q+1}, \ldots, X^*_{i,p}$. We define the full conditional distribution as follows:

$$\begin{aligned} &f(\boldsymbol{\phi}_F|\gamma^{-2}, \sigma^{-2}, \lambda, \psi_i, \mathbf{y}) \\ &\qquad \propto \exp\left(-\frac{1}{2}\boldsymbol{\phi}_F^T\left(diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\right)\boldsymbol{\phi}_F\right) \\ &\qquad \times \exp\left(-\frac{1}{2}\sigma^{-2}\left((y - X^{*S}\boldsymbol{\phi}_S) - X^{*F}\boldsymbol{\phi}_F\right)^T\left((y - X^{*S}\boldsymbol{\phi}_S) - X^{*F}\boldsymbol{\phi}_F\right)\right), \end{aligned} \tag{5.26}$$

where $y - X^{*S}\boldsymbol{\phi}_S$ is the response adjusted for the variables not currently being updated. We now define $y^* = y - X^{*S}\boldsymbol{\phi}_S$, such that

$$\begin{aligned}
&\left((y - X^{*S}\boldsymbol{\phi}_S) - X^{*F}\boldsymbol{\phi}_F\right)^T \left((y - X^{*S}\boldsymbol{\phi}_S) - X^{*F}\boldsymbol{\phi}_F\right)\\
&= \left(y^* - X^{*F}\boldsymbol{\phi}_F\right)^T \left(y^* - X^{*F}\boldsymbol{\phi}_F\right)\\
&= \boldsymbol{\phi}_F^T(X^{*F})^T(X^{*F})\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^* + (y^*)^T y^*. \qquad (5.27)
\end{aligned}$$

Letting $\mathbf{A} = \left(\sigma^2 diag\left(\frac{1}{\boldsymbol{\psi}_F}\right) + (X^{*F})^T(X^{*F})\right)$ and substituting Equation 5.27 into Equation 5.26 gives the full conditional distribution for $\boldsymbol{\phi}_F$ as:

$$\begin{aligned}
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\phi}_F^T diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\boldsymbol{\phi}_F - \frac{1}{2\sigma^2}\left[\boldsymbol{\phi}_F^T(X^{*F})^T(X^{*F})\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^*\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\boldsymbol{\phi}_F^T diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\boldsymbol{\phi}_F + \sigma^{-2}\left(\boldsymbol{\phi}_F^T(X^{*F})^T(X^{*F})\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^*\right)\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\sigma^2\boldsymbol{\phi}_F^T diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\boldsymbol{\phi}_F + \boldsymbol{\phi}_F^T(X^{*F})^T(X^{*F})\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^*\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\boldsymbol{\phi}_F^T\left(\sigma^2 diag\left(\frac{1}{\boldsymbol{\psi}_F}\right) + (X^{*F})^T(X^{*F})\right)\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^*\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\boldsymbol{\phi}_F^T\mathbf{A}\boldsymbol{\phi}_F - 2\boldsymbol{\phi}_F^T(X^{*F})^T y^*\right]\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\phi}_F - \mathbf{A}^{-1}(X^{*F})^T y^*\right)^T\mathbf{A}\left(\boldsymbol{\phi}_F - \mathbf{A}^{-1}(X^{*F})^T y^*\right)\right). \qquad (5.28)
\end{aligned}$$

Equation 5.28 above gives the form of a multivariate normal distribution with:

$$\begin{aligned}
E[\phi_F|\sigma^2, \psi_i, \mathbf{y}] &= \mathbf{A}^{-1}(X^{*F})^T y^*\\
&= \left(\sigma^2\left(diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\right) + (X^{*F})^T(X^{*F})\right)^{-1}(X^{*F})^T y^*.\\
Var(\phi_F|\sigma^2, \psi_i, \text{data}) &= \left(\frac{1}{\sigma^2}\mathbf{A}\right)^{-1}\\
&= \sigma^2\left(\sigma^2\left(diag\left(\frac{1}{\boldsymbol{\psi}_F}\right)\right) + (X^{*F})^T(X^{*F})\right)^{-1}.
\end{aligned}$$

We use this to update the values for $\boldsymbol{\phi}$ in two stages.

5.5.3 Increasing the accuracy when sampling from the Full Conditional for $\boldsymbol{\psi}$.

When sampling directly from the Generalised Inverse Gaussian (GIG) distribution computational problems arise when the parameter estimates for β_i become very small. For computational efficiency, we use the special cases of the GIG.

If $\psi_i|\sigma^{-2}\lambda, \gamma^{-2}, \beta_i \sim GIG\left(\lambda - \frac{1}{2},\, \frac{1}{\gamma^2},\, \beta_i^2\right)$ then the probability density function is written as

$$\begin{aligned} f(\psi_i|\sigma^{-2}\lambda, \gamma^{-2}, \beta_i) &= \psi_i^{\lambda-\frac{3}{2}} \exp\left(-\frac{\beta_i^2}{2\psi_i}\right) \exp\left(-\frac{\psi_i}{2\gamma^2}\right) \\ &= \left(\frac{1}{\psi_i}\right)^{(\frac{1}{2}-\lambda)+1} \exp\left(-\frac{\beta_i^2}{2}\frac{1}{\psi_i}\right) \exp\left(-\frac{\psi_i}{2\gamma^2}\right). \qquad (5.29) \end{aligned}$$

This is the density function of the inverse gamma distribution, $IG(\frac{1}{2} - \lambda, \frac{\beta_i^2}{2})$, multiplied by $exp\left(-\frac{\psi_i}{2\gamma^2}\right)$. Hence if $\lambda < \frac{1}{2}$ (to ensure $\frac{1}{2} - \lambda > 0$) and $\gamma^2 \to \infty$ we sample from an $IG(\frac{1}{2} - \lambda, \frac{\beta_i^2}{2})$ distribution instead of the GIG.

Equation 5.29 is also in the form of a gamma distribution, $Ga(\lambda - \frac{1}{2}, \frac{1}{\gamma^2})$ multiplied by $exp\left(-\frac{\beta_i^2}{2\psi_i}\right)$. Hence if $\lambda > \frac{1}{2}$ and as $\beta_i \to 0$ we sample from a $Ga(\lambda - \frac{1}{2}, \frac{1}{\gamma^2})$ distribution instead of the GIG.

Hence we can summarise the special cases used and stated in Johnson et al. [1994] as follows:

1. As $\beta_i^2 \to 0$ and $\lambda > \frac{1}{2}$, we sample from a $Ga\left(\lambda - \frac{1}{2},\, \frac{1}{\gamma^2}\right)$.

2. As $\gamma^2 \to \infty$ (and $\lambda < \frac{1}{2}$) we sample from an $IG\left(\lambda - \frac{1}{2},\, \frac{\beta_i^2}{2}\right)$.

Special Cases of the GIG

We begin with the case where $\gamma^2 \to \infty$. The Inverse Gamma (IG) and the Generalised Inverse Gaussian (GIG) distributions are very similar for most values of ψ_i as $\exp\left(-\frac{\psi_i}{2\gamma^2}\right) \approx 1$ as $\gamma^2 \to \infty$. But, if ψ_i is large, say $\gamma^2 = \psi_i$, then $\exp\left(-\frac{\psi_i}{2\gamma^2}\right) = \exp\left(-\frac{1}{2}\right) < 1$. This gives a larger tail density for this IG compared to the GIG. Due to this we will sample a very large value for ψ_i from the IG too often (because of the larger density at high ψ_i compared to the GIG). The ratio of the densities of the GIG and IG is $\exp\left(-\frac{\psi_i}{2\gamma^2}\right)$, and so to sample with the correct probability from the GIG we have to check whether $U_{[0,1]} < \exp\left(-\frac{\psi_i}{2\gamma^2}\right)$. This gives us exactly the correct sampling probability.

The second case is where $\beta_i \to 0$. The Gamma (Ga) and GIG are very similar for most values of β_i because $\exp\left(-\frac{\beta_i^2}{2\psi_i}\right) \to 1$ as $\beta_i \to 0$. If ψ_i is also small, say $\psi_i = \beta_i$, then $\exp\left(-\frac{\beta_i^2}{2\psi_i}\right) \to \exp\left(-\frac{1}{2}\right) < 1$. This means that we will sample a very small value for ψ_i from the Gamma distribution too often (as it has a larger density at small values of ψ_i compared to the GIG). The ratio of the densities of the GIG and Gamma distribution is $\exp\left(-\frac{\beta_i^2}{2\psi_i}\right)$, and so to sample with the correct probability from the GIG, we check whether $U_{[0,1]} < \exp\left(-\frac{\beta_i^2}{2\psi_i}\right)$.

This gives the correct sampling probability from the Gamma distribution with respect to the GIG.

5.6 Conclusion

In this chapter we have recalculated all the full conditional distributions and have replicated the calculations for increasing the computational efficiency of the Normal Gamma. Where necessary we have corrected the calculations from Griffin and Brown [2010]. In all cases, we use the stated distributions here in our Normal Gamma code rather than those published in Griffin and Brown [2010].

Chapter 6

Comparing the performance of eQTL methods on non-simulated data

In this chapter we apply the statistical methods we are comparing - Spike and slab, HyperLasso, LR test, LS/MLLS, piMASS and the Normal Gamma - to the three real datasets described in Section 2.1; a yeast dataset and two human datasets, Fairfax and Hulse. We compare and contrast the statistical methods with the Normal Gamma prior to investigate whether its superior performance on simulated data, shown in Sections 4.1 and 4.2, is replicated on real data. In this chapter we do not include any functional information in the Normal Gamma; we use the basic model as specified by Griffin and Brown [2010] and described in Chapter 5.

For the NG, piMASS and Spike and slab, we use the posterior mean to rank the estimates. For the LS/MLLS we use the estimate of the regression coefficient, and we use the MAP estimate for HyperLasso. When plotting ROCs, we use the LR test statistic, but when reporting the effect size, we report the LS estimate based on the single SNP model for the SNP with the greatest LR test statistic.

6.1 Yeast data

We know yeast SNPs are identified by the gene they reside in. We have a total of 1802 SNPs that fall into 1307 unique genes/regions. We assess the success of the NG and other methods at identifying the target SNPs, which we have defined based on the hotspot location given in two or three of Lee et al. [2009], Yvert et al. [2003] and Zhu et al. [2008] and defined in Section 2.1.2, by assessing how each method ranks these target SNPs. Initially we number the SNPs 1 to 1802 rather than maintaining the gene names (as there is often more than one SNP in a gene and the gene name is given to all the SNPs in that gene).

We have removed SNPs where all of the subjects have the same genotype (when coded as 0 and 1). As such not all SNPs in the hotspot regions are in our dataset. We tabulate, in Table 6.1, the possible SNPs from the regions that we hope to be identified in each gene. We have identified the genes and SNPs using the Saccharomyces Genome Database [Cherry et al., 2012].

Gene name	Hotspot location	Genes in hotspot region	SNP numbers
TOS1 (YBR162C), AMN1(YBR158W)	Chr2: 560,000	YBR161W, YBR162W, YBR163W, YBR165W, YBR154C, YBR156C	SNP153, SNP154, SNP155, SNP156, SNP148, SNP149
ILV6 (YCL009C)	Chr3: 100,000	YCL018W (3), YCL014W, YCL009C	SNP205, SNP206, SNP207, SNP208, SNP209
MATALPHA1 (YCR040W)	Chr3: 230,000	YCR064C	SNP217
GPA1 (YHR005C), GPA1 (YHR005C-A)	Chr8: 130,000	YHR011W, YHR012W, YHR016C	SNP750, SNP751, SNP752
HAP1 (YLR256W)	Chr12: 680, 000	YLR263W, YLR265C, YLR267W, YLR269C, YLR273C	SNP1232, SNP1233, SNP1235, SNP1236, SNP1237
SIR3 (YLR442C)	Chr12: 1,070,000	YLR460C (4), YLR465C (3)	SNP1316, SNP1317, SNP1318, SNP1319, SNP1320, SNP1321, SNP1322
TOP2 (YNL088W)	Chr14: 503,000	YNL066W	SNP1513
PHM7 (YOL084W)	Chr15: 180,000	YOL081W (2)	SNP1600, SNP1601
CAT5 (YOR125C)	Chr15: 590,000	YOR135C, YOR139C (2), YOR140W (2)	SNP1678, SNP1679, SNP1680, SNP1681, SNP1682

Table 6.1: Table showing the genes, hotspot locations and SNPs in the hotspot regions that are also in our dataset. Results will be judged against these target SNPs to see how effectively each method is ranking these hotspots. The numbers in the brackets of the genes in the hotspot regions tells you the number of SNPs, if greater than 1, in the gene. These SNPs may or may not have been biologically validated as causal. We define then based on proximity to the hotspot location.

We look at multiple ways of presenting the results for this data. Initially we select only gene YOL084W/PHM7 to investigate. For these results, the Normal Gamma and Spike and slab are run with 100,000 iterations, discarding the first 5,000 as burn-in; we use the MLLS estimate instead of the LS estimate as we have more SNPs than yeast (individuals); HyperLasso is run defining 'shape' as 0.1 with all other parameters left to default; and piMASS is run with 100,000 iterations, a 10,000 iteration burn-in and thinning based on maintaining every 10th iteration, as suggested in the documentation.

Table 6.2 shows the top 10 ranked SNPs by SNP number, across each method. We see that at least one of the two SNPs in the hotspot region (SNP1600 and SNP1601) is selected in the top 10 ranked SNPs for all methods except HyperLasso (HL). We notice strong concordance across methods with the top 2 ranked SNPs (SNPs 1592 and 1591). These two SNPs are close to the hotspot location, so given a wider hotspot region, could be included in our target SNP set.

Rank	NG	piMASS	MLLS	HL	LR	SS
1	1592	1592	1592	1591	1592	1597
2	1591	1591	1591	1596	1591	1592
3	1597	1601	1595	3	1595	1591
4	1600	1600	1600	1770	1601	1770
5	1601	1597	1593	1668	1593	1771
6	1596	1595	1601	180	1597	1596
7	1771	1593	1596	94	1600	3
8	1770	1596	1597	1731	1589	9
9	1593	1589	1594	1316	1586	10
10	20	1586	1589	506	1596	1601

Table 6.2: SNP identification numbers (1-1802) for the top 10 ranked SNPs based on posterior mean effect size for each method for yeast gene YOL084W (PHM7). SNPs 1600 and 1601 are in the hotspot region, according to Lee et al. [2009], and are highlighted in red. We notice strong concordance between the different methods and the SNPs in the top 10.

Table 6.3 shows the rankings of the causal SNPs for each method compared. There are no results for the HyperLasso in this table because the hotspot SNPs are not selected by HL. We notice that the combined ranks of the two causal SNPs is lowest for piMASS and the Normal Gamma. The highest combined rank is for Spike and slab. This indicates that piMASS and the Normal Gamma are performing best in terms of the combined rank of the posterior estimate of the hotspot SNPs.

	NG	piMASS	MLLS	HL	LR	SS	MAF
SNP1600	4	4	4	.	7	11	0.440
SNP1601	5	3	6	.	4	10	0.450

Table 6.3: Table showing the ranking of the SNPs in the hotspot regions according to the different methodologies for yeast gene YOL084W. It also includes the minor allele frequency (MAF) for each of the causal SNPs. The MAF affected performance in simulation studies.

When applying the NG method to the yeast data we found there was very little variation in the raw gene expression values for some genes, leading to computational problems, see Table 6.4. We believe these problems have arisen due to the very small values for M. Recall that M is defined as $M = \frac{1}{p}\sum_{i=1}^{p}\hat{\beta}_i^2$ where $\hat{\boldsymbol{\beta}}$ is the least squares (LS) or minimum length least squares (MLLS) estimate, see Section 3.4.1. Small variability in the gene expression could lead to small variability in the MLLS estimates and therefore small values of M. The only genes that have not been problematic are YCR040W and YOL084W. To overcome the computational problems we increased M to be either 5 or 0.5 in further analyses.

YBR158W	0.6693	YCR040W	1.5631	YLR256W	0.5274	YOL084W	2.026
YBR162C	0.2723	YHR005C	0.4757	YLR442C	0.2144	YOR125C	0.3489
YCL009C	0.2670	YHR005C-A	0.3092	YNL088W	0.4725		

Table 6.4: The standard deviation of the gene expression values for each gene used in the yeast dataset.

Table 6.5 shows the rankings of the SNPs that are in the hotspot regions for all comparison methods. Note that for genes YHR005C-A and YCR040W, the indicator variable representing the difference in growth conditions for the two Yeast (glucose and ethanol) was not included as one of the posterior SNP set for HL. As such we omit the HL results for these two genes. This is one drawback of the HL. For the Yeast data, HL selects between 0.5% and 2% of the SNPs as having a non-zero effect. This is a low percentage of SNPs, but the raw numbers of SNPs (between 6 and 37) are similar to would be expected for biological validation. Note also that Yeast gene YCR040W reports the NG with M defined from the data (M data) and $M = 5$ to assess the effect of M on the posterior mean effect sizes from the NG.

	NG (M=5)	**NG (M=0.5)**	**piMASS**	**MLLS**	**HL**	**LR**	**SS**	**MAF**
YOR125C								
SNP1678	1145	1548	4	9	.	4	861	0.459
SNP1679	1638	1191	23	88	.	23	1393	0.440
SNP1680	224	226	21	52	.	19	478	0.450
SNP1682	642	81	14	21	.	12	965	0.450
YBR158W								
SNP148	3	3	2	4	1	2	5	0.404
SNP149	114	993	6	5	.	7	742	0.385
SNP153	1	2	5	3	2	4	2	0.413
SNP154	17	1270	7	7	.	6	6	0.422
SNP155	21	148	9	10	.	9	11	0.413
SNP156	5	892	10	18	.	11	15	0.404
YCL009C								
SNP205	528	1129	98	86	.	67	45	0.459
SNP206	1462	635	107	68	.	81	56	0.450
SNP207	503	1297	117	204	.	98	76	0.450
SNP208	379	992	70	28	.	45	37	0.450
SNP209	1793	1329	12	11	1	5	5	0.468
YHR005C-A								
SNP750	1006	1778	533	1717	NA	543	741	0.385
SNP751	96	18	933	1066	NA	993	655	0.367
SNP752	285	939	351	1500	NA	345	337	0.376
YLR256W								

Continued on next page

Table 6.5 – continued from previous page

	NG (M=5)	NG (M=0.5)	piMASS	MLLS	HL	LR	SS	MAF
SNP1232	107	1538	3	11	.	3	6	0.422
SNP1233	17	1542	11	10	.	11	273	0.394
SNP1235	553	579	16	18	.	16	372	0.413
SNP1236	518	1361	9	14	.	9	90	0.440
SNP1237	46	1083	13	16	.	13	73	0.450
YLR442C								
SNP1316	1185	1472	16	16	.	16	290	0.413
SNP1317	286	595	9	6	.	9	62	0.477
SNP1318	1245	532	11	7	.	11	79	0.486
SNP1319	190	37	10	13	.	10	31	0.477
SNP1320	1067	1703	101	141	.	122	1511	0.229
SNP1321	1163	963	93	180	.	110	1757	0.239
SNP1322	1760	574	135	341	.	157	1194	0.248
YNL088W								
SNP1513	16	100	3	3	.	3	27	0.394
YBR162C								
SNP148	180	868	66	737	.	74	182	0.404
SNP149	977	1531	86	391	.	102	311	0.385
SNP153	20	1681	338	1674	.	411	657	0.413
SNP154	266	179	305	1009	.	356	642	0.422
SNP155	1784	435	388	1255	.	452	977	0.413
SNP156	1796	1345	738	723	.	853	901	0.404
YHR005C								
SNP750	479	1375	7	7	.	7	430	0.385
SNP751	136	257	9	5	.	8	260	0.367
SNP752	1499	1734	12	9	.	10	552	0.376
YCR040W								
	NG (data M)	**NG (M=5)**	**piMASS**	**MLLS**	**HL**	**LR**	**SS**	**MAF**
SNP217	30	10	18	8	NA	18	166	0.477

Table 6.5: Table showing the ranking of the SNPs in the hotspot regions according to the different methodologies for Yeast data. It also includes the minor allele frequency (MAF) for each of the causal SNPs. The MAF affected performance in simulation studies. We note that for YCR040W M is approximated from the data, and also defined as 5.

We notice that in many cases the SNPs in the hotspot regions have rank far away from 1 (the top rank) in the NG, meaning that it is not detecting these target SNPs. This could be for a number of reasons. Firstly the maximum posterior effect size (mean) is very small, see Table 6.6. When the maximum posterior effect size (mean) of the SNPs is very small, the ranking of the corresponding SNP and any other lower ranked SNPs is meaningless due to the biologically negligible effect sizes. Secondly the effect sizes of the associated SNPs may be very small. The NG will not detect these due to the amount of shrinkage enforced since the prior puts a lot of mass close to 0. Thirdly, the MAF (minor allele frequency) may be affecting detection. This is unlikely given the high MAF for all SNPs - the simulation scenario where the Normal Gamma and Spike and slab significantly outperformed all other methods. Finally the ranking of SNPs becomes irrelevant at a certain point when effect sizes become too small (less than 10^{-5} say). This is something we see with the NG as it appears to be applying heavy shrinkage to the Yeast SNPs compared to the other methods.

	NG (M=5)	NG (M=0.5)	piMASS	MLLS	LR	HL	SS
YBR158W	SNP153 -0.04412	SNP150 -0.12023	SNP150 -1.05549	SNP150 -0.0362	SNP150 1.1753 (303.0085)	SNP148 -0.60792	SNP150 -0.76106
YBR162C	SNP558 0.022749	SNP1263 -0.041955	SNP1624 0.334559	SNP1716 0.011254	SNP1624 0.1369 (13.8509)	SNP1624 -0.160326	SNP1624 0.0336
YCL009C	SNP986 0.025559	SNP1081 -0.048408	SNP1273 0.339765	SNP1446 0.011681	SNP1273 0.1283 (12.89339)	SNP209 0.118669	SNP1281 -0.023065
YCR040W	SNP213 0.75828	NG M from data SNP213 1.2056	SNP213 1.07656	SNP213 0.2225	SNP213 2.9960 (556.09117)	NA	SNP213 3.060564
YHR005C	SNP1 -0.02735	SNP509 0.51195	SNP749 -0.33882	SNP746 -0.01943	SNP749 0.2285 (12.6842)	SNP749 -0.2450	SNP1716 0.1458
YHR005C-A	SNP1723 -0.01747	SNP90 0.04685	SNP1217 0.30188	SNP619 0.01474	SNP1217 0.1373 (11.0025)	NA	SNP1031 -0.001189
YLR256W	SNP1228 0.037925	SNP1229 0.11088	SNP1229 0.54866	SNP1230 0.03975	SNP1229 0.8040 (186.71)	SNP1229 0.5710	SNP1229 0.8409
YLR442C	SNP424 -0.043521	SNP6 -0.038729	SNP1309 0.11348	SNP1305 -0.010192	SNP1309 0.1658 (35.2878)	SNP1305 -0.1327	SNP1309 -0.05555
YNL088W	SNP1512 0.03611	SNP1512 0.1080	SNP1511 0.9024	SNP1511 0.06157	SNP1511 0.6829 (141.1604)	SNP1511 0.4453	SNP1511 0.4288
YOL084W	NG (M data) SNP1592 -0.4185		SNP1592 -0.8960	SNP1592 -0.1138	SNP1592 2.7363 (131.6976)	SNP1591 -1.7971	SNP1597 -1.0759
YOR125C	SNP967 -0.03128	SNP1513 -0.05735	SNP1673 -0.7381	SNP1512 -0.0223	SNP1673 0.3768 (74.9411)	SNP50 0.3288	SNP1673 -0.3600

Table 6.6: Table showing the SNP with maximum effect size for each Yeast gene across each method. We take absolute effect size based on the posterior mean for piMASS, the Normal Gamma and Spike and slab, the MAP estimate for HyperLasso, the LS estimate for the LR test single SNP model compared to the null model, and the MLLS estimate. We report the ratio of the likelihoods in brackets for the LR test.

When using the Yeast data with the extra parameter for environmental confounding, the Normal Gamma was easy to adapt to this. For the Spike and slab we had to increase the prior inclusion probability of the indicator to as close to 1 as possible. Even so the indicator was not always included at each iteration in the estimates. This may make the estimates unreliable. In the HyperLasso model, covariates can be included but they are treated similarly to the SNPs. Hence we discard the results when the indicator is not included in the final set of SNPs. In piMASS there is also no way to control for the inclusion of the indicator. We used the posterior β estimate as a summary which takes into account the inclusion probability of the indicator. We adapted the LR test to compare the single SNP plus indicator model to the background plus indicator model. In this case, we can fit our required model including α_{environ} to the MLLS, LR test and the Normal Gamma. The other methods 'select' this confounder with probability less than one. This means that we do not always take the confounder into account when estimating the β_i parameters from the other methods. This flexibility to include counfounders is an advantage of the Normal Gamma compared to the other Bayesian methods.

To conclude, the lack of variation in the gene expression values and the small value for M means that we have struggled to ensure the NG is performing as well as in simulations. In terms of detecting the target SNPs, piMASS appears to be giving the most appropriate responses within the Bayesian framework. We expected the NG not to detect small effect sizes, and with such small overall effects, we are not surprised that other methods appear to be outperforming the NG with respect to the rank of the posterior means.

6.2 Hulse data

Using the genes in Table 2.2, page 17, we analyse the results of applying all methods to the Hulse dataset. There are no causal SNPs, or any biologically validated SNPs for this dataset, hence we can only compare SNPs between methods. For this reason we will report the top 5 ranked SNPs across all methods in Table 6.7, and compare these for similarities.

For the Hulse data results, the Normal Gamma and Spike and slab are run with 100,000 iterations, discarding the first 5,000 as burn-in; we use the MLLS estimate instead of the LS estimate as we have more SNPs than individuals; HyperLasso is not run as it cannot handle imputed data which we have here; and piMASS is run with 100,000 iterations, a 10,000 iteration burn-in and thinning based on maintaining every 10th iteration, as suggested in the documentation.

We report the top 5 ranked SNPs for all methods in Table 6.7. The SNPs are listed in order of their effect sizes (largest to smallest).

Method	Top 5 ranked SNPs
ADCY1	
NG	rs6959709, rs17172584, rs6949064, rs1294903, rs11771815
piMASS	rs12532570, rs36115872, rs6939924, rs13222078, rs1294898

Continued on next page

Table 6.7 – continued from previous page

Method	Top 5 ranked SNPs
LR	rs36115872, rs1315919, rs17172499, rs1294892, rs1294903
MLLS	rs4142878, rs11761143, rs950971, rs12702161, rs1294911
Spike and slab	rs4142878, rs6975239, rs1294914, rs1294892, rs36115872
CTNNA2	
NG	rs10779960, rs13416246, rs13409348, rs7592817, rs732260
piMASS	rs17653642, rs2566554, rs11889086, rs10178923, rs10170833
LR	rs12104529, rs7570774, rs11675845, rs6741085, rs2196152
MLLS	rs1368900, rs10178923, rs1368952, rs11126702, rs11902274
Spike and slab	rs1868925, rs7605358, rs12997174, rs11679118, rs10185018
DAAM2	
NG	rs2504090, rs9394630, rs9380895, rs2504100, rs7750130
piMASS	rs3003947, rs3004070, rs3004060, rs3004071, rs882559
LR	rs3003947, rs3004060, rs3004070, rs3003931, rs3004062
MLLS	rs9380890, rs7776096, rs303649, rs11759168, rs3008804
Spike and slab	rs2504803, rs9380892, rs6458124, rs6917582, rs2504805
IL6	
NG	rs12700386, rs17302823, rs1476483, rs2961310, rs2905324
piMASS	rs12700386, rs2067074, rs2905316, rs2961310, rs17302823

Continued on next page

Table 6.7 – continued from previous page

Method	Top 5 ranked SNPs
LR	rs12700386, rs17302823, rs1476483, rs2067074, rs2961310
MLLS	rs2905345, rs12535797, rs6969927, rs12536091, rs17778126
Spike and slab	rs10499563, rs4310110, rs10251555, rs11981074, rs1524098
PLOD2	
NG	rs6440269, rs11926970, rs9289711, rs1731398, rs11707136
piMASS	rs4561830, rs6779715, rs6770862, rs4611822, rs10513260
LR	rs36025939, rs16857603, rs1967207, rs1398776, rs2055989
MLLS	rs9846710, rs11926420, rs12491824, rs7611641, rs12487322
Spike and slab	rs11719883, rs13088646, rs1880902, rs962823, r7432214
SNX7	
NG	rs11166113, rs12756402, rs4908110, rs7416451, rs571344
piMASS	rs4402170, rs4469760, rs17386441, rs11166011, rs10875151
LR	rs1482139, rs11811184, rs9725840, rs12738955, rs12757095
MLLS	rs766204, rs1384167, rs1434362, rs1328310, rs7548566
Spike and slab	rs1384167, rs1482163, rs12401685, rs1482139, rs12757095
TNFRSF11B	
NG	rs4319131, rs3103991, rs4532625, rs10107202, rs4372031
piMASS	rs2875845, rs13279492, rs1564860, rs4532619, rs3103989

Continued on next page

Table 6.7 – continued from previous page

Method	Top 5 ranked SNPs
LR	rs17830456, rs3103992, rs3103991, rs11985044, rs6996974
MLLS	rs17683937, rs7007167, rs11989516, rs1385511, rs3103982
Spike and slab	rs1485302, rs1485303, rs7464496, rs4319131, rs2073618

Table 6.7: A list of the top 5 SNPs from the different methods run on the 7 selected genes from the Hulse dataset. SNP rs12700386 in IL6 is highlighted as it appears in the NG as the top ranked SNP, and is in the top 10 ranked SNPs for all other methods.

In Table 6.7, we notice that SNP rs12700386 in IL6 (highlighted) is the top ranked SNP for the NG and appears in the top 5 ranked SNPs for piMASS and the LR test, and is rank 6 for MLLS and 9 in Spike and slab. This SNP is reported as a Caucasian maternal and African-American fetal SNP by Velez et al. [2008], which looks into the haplotypes in IL6 and IL6-R that are associated with amniotic fluid protein concentrations in pre-term birth. Also, in Wang et al. [2013] which discusses rheumatoid arthritis, rs12700386 was found to be statistically significant at the 5% level but had a small estimated effect size (estimated in the two studies within the paper as 0.22 and 0.16). The paper studies associations between the number of DNA, RNA and protein biomarkers that are directly related to IL6 signalling. The authors use gene expression and genotyping data in their statistical analysis which consists of correlations and standard linear regression for defining associations. The small effect size of rs12700386 means that it is not investigated further in the paper. IL6 is an interleukin gene which is involved in immunity and inflammation. Given the annotation of the SNP, it seems plausible for it to be truly associated to a change in gene expression in IL6 given the reported function of the SNP and gene. SNP rs2961310 also in IL6, ranks in the top 5 SNPs for the NG, piMASS and the LR test but has no reported annotations. There is very little other concordance across the top 5 reported SNPs in these genes across the methods used. It is interesting to note that the lack of concordance between the methods suggests that there is no clearly causal SNP(s) with a moderate to large effect size within these genes. Given that we have no validated SNPs for this dataset, there is limited extra information that can be obtained from these results.

6.3 Fairfax data

We choose the genes and SNPs to analyse here based on Figure 6(b) in Fairfax et al. [2012]. These 6 genes and their respective gene expression source (bcells or monocytes) lead to 8 scenarios, see Table 2.3, page 19 for details of the number of individuals, SNPs, causal (validated) SNPs and their minor allele frequencies (MAFs). These are examples of genes containing eQTLs involving SNPs associated with multiple traits in GWAS for particular diseases.

For the results in this section, the Normal Gamma and Spike and slab are run with 100,000 iterations, discarding the first 5,000 as burn-in; we use the MLLS estimate instead of the LS estimate as we have more SNPs than individuals; HyperLasso is omitted as it cannot be run using imputed genotypes; and piMASS is run with 100,000 iterations, a 10,000 iteration burn-in and thinning based on maintaining every 10th iteration, as suggested in the documentation.

We tabulate the results for the Fairfax data in Table 6.8. The rank of the causal SNPs from the literature is given.

Method	Maximum Effect Size	Causal SNP rank
ERAP2 bcell (792 SNPs)		
NG	0.11424	1
LS	0.166	63
LR	0.4521 (31.7396)	1
SS	0.172275	1
piMASS	0.48618	1
ERAP2 mono (792 SNPs)		
NG	0.67332	1
LS	1.1155 (0.28443)	1
LR	183.36	1
SS	1.102613	1
piMASS	1.29736	1
CARD9 mono (511 SNPs)		
NG	0.0393	3
LS	0.0875	119
LR	0.1435 (60.219)	2
SS	0.065754	2
piMASS	0.86126	5
FADS1 bcell (1076 SNPs)		

Continued on next page

Table 6.8 – continued from previous page

Method	Maximum Effect Size	Causal SNP rank
NG	0.09312	639
LS	0.10731	123
LR	1.0445 (13.03626)	102
SS	0.235892	272
piMASS	0.10593	128
RBM6 bcell (932 SNPs)		
NG	0.047661	39
LS	0.054905	450
LR	0.08120 (14.20943)	5
SS	0.054853	27
piMASS	1.08936	30
RBM6 mono (932 SNPs)		
NG	0.05562	191
LS	0.039585	186
LR	0.1625 (56.87475)	4
SS	0.076683	240
piMASS	1.15543	147
CD40 mono (468 SNPs)		
NG	0.07733	1
LS	0.22978	18
LR	0.1594 (16.01593)	1
SS	0.0943	2
piMASS	0.40448	2
FAM167A bcell (551 SNPs)		
NG	0.01552	528
LS	0.10914	226
LR	0.06782 (8.741665)	469
SS	0.017145	523
piMASS	1.01494	488

Table 6.8: Table showing the results from the analysis of the Fairfax data. Each gene has one causal SNP reported in the literature, we state the rank of this causal SNP. The maximum posterior effect size is also reported, with the maximum effect size for the LR test calculated using the LS estimate of the single SNP model. We report the ratio of the likelihoods in brackets for the likelihood ratio test.

We notice that the LR test, the only univariate methods we compare, always has the (joint) lowest rank (rank closest to the top rank of 1) even when other methods are performing poorly such as in RBM6 bcell. The LR test does not have (joint) lowest rank in FAM167A bcell where all methods perform very badly.

The good performance of the univariate method could be explained by the fact that the associated SNPs were identified using univariate analysis. Techniques such as correlation analysis, using Z-scores to compare the difference in gradient of lines representing the effect of the number of minor alleles on the monocyte and bcell expression (ANCOVA) were used. Linear and Spearman rank models were also used in some of the analysis. It is not clear if these are used for multiple SNPs or simply to compare across bcell and monocyte expression - the latter is more likely given the results and aims of the work. Given that the SNPs were initially selected using a univariate method, it is not surprising that the univariate method performs well here.

The NG assigns the causal SNP a rank of 1 for 3 of the cases, rank 3 for 1 case, rank 39 for 1 case, and with rank worse than 100 in the remaining cases. This means that out of the 8 scenarios for the Fairfax data, the NG puts the causal SNP in the top 5 ranked SNPs 50% of the time. This is the same as the number of times the causal SNP is ranked in the top 5 for Spike and slab and piMASS. Only the LR test outperforms this with the causal SNP being identified in the top 5 SNPs for 75% of the genes. The LS performs worst, only identifying the causal SNP in the top 5 ranked SNPs in 1 of the 8 scenarios.

We conclude here that there are genes where the causal SNPs within them, such as ERAP2 mono, are very obvious and may have a large effect size relative to all other SNP as all methods detect the causal SNP with rank 1. Other genes such as FAM167A have a causal SNP which we hypothesise has a smaller effect size that is similar to other SNPs as all methods fail to detect the causal SNP with rank greater than 226 (out of a total of 551 SNPs, hence it is ranked in the bottom 60% of the SNPs by all methods). Genes such as RBM6 bcell have a causal SNP that is more easily detected by the LR test than other methods, but that has an effect size that means the NG, piMASS and Spike and slab detect it with rank between $27 - 39$ (out of 932 SNPs, meaning it is ranked in the top 4.5% of SNPs).

6.4 Which statistical method is most appropriate?

Even though we cannot conclude based only on the results from the Hulse dataset and conclusions from Fairfax and Yeast data are limited to comparisons from the literature, we can try to infer more about these results using the results from applying these methods to simulated data. We know the LR test detects SNPs with high marginal effects. It is therefore good at detecting SNPs with large effect sizes.

Due to the heavy shrinkage enforced by the NG, any SNPs with very small effect sizes are often missed as they have posterior mean estimates shrunk to 0. Simulations imply that rarer SNPs need larger effect sizes for detection by the NG.

From our simulations, we conclude that piMASS performs better for detecting rarer SNPs than for more common SNPs. The definition of piMASS to prioritise SNPs in close proximity also means that if we are assessing SNPs where the causal/associated SNPs are in close proximity/high LD, piMASS will perform better. In the Hulse dataset, we take all exonic SNPs on the same chromosome as the gene. From this we hypothesize that there may be groups of SNPs in high LD, and hence piMASS will detect these.

Spike and slab performs section of SNPs. As with the Normal Gamma, Spike and slab performs better on more common SNPs, therefore rarer SNPs with small effect sizes maybe missed.

In conclusion, none of the statistical methods compared are certain to identify the true causal SNP in every dataset. Having studied these methods carefully, the method for retaining the causal SNP in the majority of cases involves using the likelihood ratio test to initially filter the SNPs before using this filtered set of SNPs in the NG framework. Ensuring a relaxed threshold for the LR test will mean that we do not discard too many potentially causal SNPs. Using the NG framework means we can investigate the full posterior distribution for possible smaller, causal/associated effect sizes in some SNPs, while controlling for the effect sizes of other SNPs.

6.5 Convergence and Computational Time for the Normal Gamma

When comparing methods that use MCMC on new datasets, it is important to assess the convergence, as well as the computational time. Methods that have been published contain information to enable these factors to be assessed. These are not readily available for the Normal Gamma. To check the convergence, we use the same procedure as we defined in Section 4.3.

6.5.1 Checking convergence

We check convergence of the Markov chains using the R-hat statistic, see Brooks and Gelman [1998]. This investigates the scaled, weighted difference between the within and between chain variances, see Section 4.3 for more details. We test this once for each dataset, using 10 runs of the NG, the minimum suggested number of runs for the test. Brooks and Gelman [1998] suggests an R-hat value of < 1.1 is advisable for convergence. We tabulate the maximum R-hat values for all parameters in the Normal Gamma for each dataset in Table 6.9. This shows that for all datasets, 50,000 iterations is sufficient for convergence, although 30,000 is sufficient for the Fairfax dataset.

Dataset	Maximum R-hat value
Yeast data (YOL084W)	1.003331
Hulse data (DAAM2)	1.001987
Fairfax data (CD40 mono)	1.067799

Table 6.9: Table showing the maximum R-hat values [Brooks and Gelman, 1998] for all the Normal Gamma parameters for Yeast and Hulse datasets run with 50,000 iterations and a 5,000 iteration burn-in. The Fairfax data achieved convergence with 30,000 iterations and a 5,000 iteration burn-in. The R-hat statistic was calculated using 10 datasets.

For the Hulse and Yeast datasets, convergence has been achieved with 50,000 iterations of the Normal Gamma and a 5,000 iteration burn-in. For the Fairfax dataset, convergence has been achieved with 30,000 iterations and a 5,000 iteration burn-in. The convergence diagnostic shows that the between chain variance is very small which implies that the within chain variances must be similar across chains. This leads us to believe that the stationary distribution has been reached at each run of the chain. For the Yeast, Hulse and Fairfax results in this chapter, the Normal Gamma was run with 100,000 iterations and a 5,000 iteration burn-in.

6.5.2 Computational time

When checking the convergence we also monitor the time taken for the Normal Gamma to run. The time taken, and the dimension of the problem in terms of number of individuals (n) and number of SNPs (p) are tabulated in Table 6.10.

Dataset	n	p	Computational time (mean (range))
Yeast data (YOL084W)	218	1803	30 hours (14 hours - 54 hours)
Hulse data (DAAM2)	39	149	3.2 hours (1 hour - 5 hours)
Fairfax data (CD40 mono)	243	467	3.6 hours (3 hours - 4 hours)

Table 6.10: The computational time taken for each of the datasets used to calculate the R-hat convergence statistic. We include n, the number of individuals and p the number of SNPs included in each dataset. Note that Yeast and Hulse datasets are for 50,000 iterations with a 5,000 iteration burn-in, and the Fairfax data is for 30,000 iterations with a 5,00 iteration burn-in.

The 3 datasets tabulated here cover a wide range of dimensions of the data used in the Normal Gamma. The large range in computational time for the same dataset can be explained for the same reasons as in Section 4.3.2, page 64, based on sampling from different distributions within the MCMC algorithm and other processes running on the same nodes of the HPC facility.

Similarly with simulated data, the Normal Gamma is computationally much slower than other methods. In the case of CTNNA2 for the Hulse data, the Spike and slab is similarly slow to the NG but HL, LS, piMASS and LR test all run in less than 1 hour for all datasets, including CTNNA2. Importantly, on the real data, the time taken does not appear to be prohibitive to using the NG.

6.6 Conclusion

When comparing all methods on real data it is very difficult to define how well each method detects causal/associated SNPs. The ranking works well when the effect sizes are sufficiently different to one another but once the effect sizes become small then ranking is no longer meaningful.

The Normal Gamma enforces very harsh shrinkage on SNPs with small estimated effect sizes. As such, the Normal Gamma does not find many associated SNPs. This is good in the sense that it reduces the number of false positives, but it could also reduce the number of true positives.

The ability of the Normal Gamma to perform equally well as any other method is noted here. We have used SNPs coded as $\{0, 1\}$ and as $\{0, 1, 2\}$, and SNP data with and without imputation. We have also used SNP and expression data from many different platforms. This shows the versatility of the Normal Gamma which is not reflected in all other methods.

The ability to adapt the Normal Gamma to include confounding factors such as an environment confounding effect as with the Yeast data is essential. The LS, LR test and NG are the only methods to be able to do this accurately. The prior inclusion probability of a SNP in the Spike and slab cannot be defined as exactly 1. Hence we cannot enforce the inclusion of the SNP in the model with complete certainty. This is better than other methods but does not reflect the inclusion of the confounder with certainty. The Normal Gamma and all methods except HyperLasso, are able to handle imputed data. The ability to include imputed data is essential when using real data, as we often have missing data. HyperLasso can handle missing data, but not imputed data. We choose not to run these datasets through HyperLasso with missing values instead of imputed values because we maintained only SNPs and individuals whose data was complete post imputation. This meant removing individuals and SNPs where the imputed values did not have a high enough info score. We could have used only these SNPs and individuals with the imputed values removed but we did not believe that this could lead to a direct comparison of effect sizes across methods, especially if the handling of the missing values in HyperLasso led to a very different representation of the imputed values to we used.

We believe this to be the first time that the Normal Gamma has been applied to eQTL data and its performance compared on competing methods. Testing these models on real data available allows us to see the versatility the Normal Gamma and its ability to handle imputed data as well as other forms of SNP coding. The harsh shrinkage enforced by the Normal Gamma is limiting its success at detecting any true or false positive results.

The Normal Gamma is still well placed to identify causal SNPs in these datasets. By adapting the amount of shrinkage it enforces according to prior information about SNPs, we may be able to improve the ranking of SNPs validated in the literature. This is the focus of Chapter 7.

Chapter 7

Including Functional Information in the Normal Gamma prior

In this chapter we adapt the Normal Gamma prior structure to include functional information with the aim of making it more effective at detecting causal sequence variants. A directed acyclic graph (DAG) representing this can be found in Figure 7.1. We are making the assumption a priori that synonymous SNPs are less likely to be causal, based on the functional significance score, described in Section 2.7. Here we assess which functional information to use and how best to incorporate it. We test our method on the simulated data from HapGen2 described in Sections 4.1.1 and 4.2.1, and suggest improvements to the method based on these results.

We begin with a simple linear transformation of the FS score which we assess via simulation. This simple approach has the computational advantage of yielding Gibbs updates. We also consider a more complex transformation of the FS score which gives much clearer differential shrinkage between the synonymous and non-synonymous SNPs, but requires more computationally intensive Metropolis Hastings updates.

7.1 Which functional information should we include?

Even with an extensive biological knowledge, it is difficult to fully understand which functional information should be included. Lirnet [Lee et al., 2009] uses regulatory features, or general functional information that is available for specific organisms. For a given disease, we could use an expert in the field to

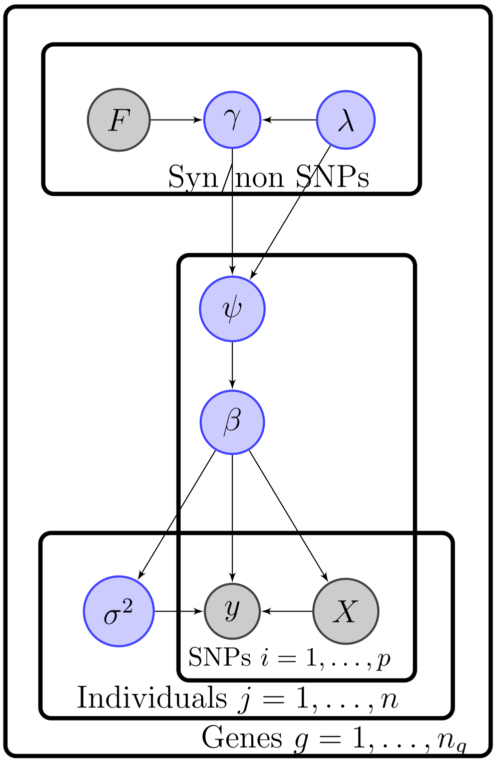


Figure 7.1: DAG representing the hierarchical relationships between the variables in the NG model with functional information, F (synonymous (syn) and non-synonymous (non) SNPs). Details of the parameters and the basic NG prior structure can be found in Chapter 5. A DAG represents the relationships between the parameters in a model using the arrows between nodes. The nodes are shaded grey when the variable is observed and nodes within a plate are iterated over by the feature stated on the bottom right corner of the plate.

identify and/or group the key features for that particular disease from a list of functional information categories. This would tailor the functional information and hence the model to the disease. Unfortunately most functional information data contains a lot of missing or highly correlated values, for example the ENCODE database, which makes it difficult to use.

For consistency across human data and disease neutrality, we will use the Functional Significance (FS) score [Lee and Shatkay, 2009] which combines functional information features into one score per SNP. It combines the functional features, protein coding, splicing regulation, transcriptional regulation and post-transcriptional regulation into a score. The score focuses on the deleterious effect of individual SNPs from multiple on-line resources. The functional effects of each SNP are then combined using weights which reflect the importance of the feature and reliability of the on-line resource. The resultant score is constrained to $[0, 1]$ where 0 represents no deleterious effect, 1 represents a highly deleterious effect and 0.5 represents no knowledge. More details of this can be found in Section 2.7, page 24.

7.2 How to incorporate the functional information

To enhance the ability of the NG to detect causal SNPs, we allow the variance of $\beta|\lambda, \gamma$ to depend on the FS score through a parameter B. In the standard Normal Gamma, $\pi(\text{var}(\beta|\lambda, \gamma)) = 2\lambda\gamma^2$ is given an $IG(2, M)$ prior distribution with expectation M, where M is defined as $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$ for $n > p$ and $M = \frac{1}{n}\sum_{i=1}^{p} \hat{\beta}_i^2$ for $p > n - 1$ and $\hat{\beta}_i$ represents the least squares (LS) estimate of β if $n > p$ and the minimum length least squares (MLLS) estimate otherwise. M provides an approximate estimate of the variance of the LS/MLLS estimates. We modify this prior to become $IG(2, MB)$ so that $E[\pi(var(\beta|\lambda, \gamma, B))] = MB$ and $\pi(\gamma^{-2}|\lambda, B) \sim Ga\left(2, \frac{MB}{2\lambda}\right)$. Larger values of B support, a priori, larger values of β. By relating B to the FS score and allowing different priors for the FS score for synonymous and non-synonymous SNPs, we allow non-synonymous SNPs to a priori have larger effect sizes. Figure 7.2 shows how changing B affects the marginal prior variance of $\beta|\lambda, \gamma, B$.

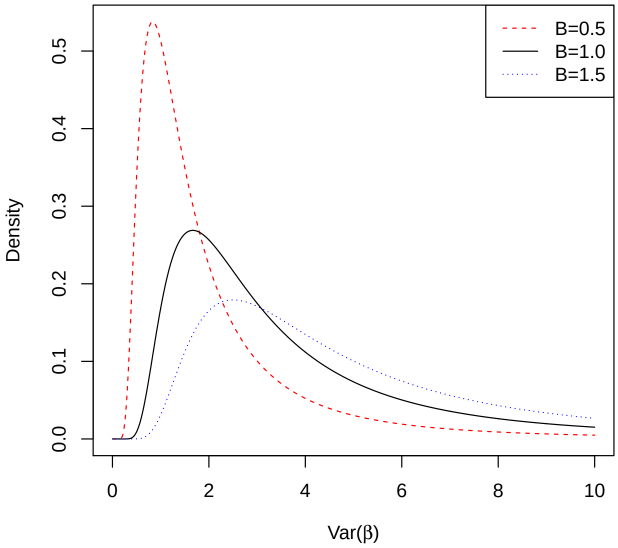


Figure 7.2: The effect on the variance of the marginal prior distribution of β, when changing B, where $\text{var}(\beta|\lambda, \gamma, B) \sim IG(2, MB)$. In this case we fix $M = 5$ and vary B.

7.3 Prior distributions for the Functional Significance (FS) Scores of Synonymous and Non-synonymous SNPs

Using approximately 6500 FS scores for non-synonymous SNPs and 4500 FS scores for synonymous SNPs from the FS score database [Lee and Shatkay, 2009] we create the distribution of FS scores. The histograms of the FS scores for synonymous and non-synonymous SNPs are shown in Figure 7.3.

7.3.1 Obtaining the prior distributions

We fitted a truncated mixture of Gamma distributions to the FS scores for synonymous SNPs and used a Uniform distribution on $[0, 1]$, $(U_{[0,1]})$, for the prior distribution of FS scores for non-synonymous SNPs. When considering which distributions would be appropriate to fit, we considered the biological effect of the SNPs. We also noted that the mean of the FS scores for synonymous SNPs was very different from non-synonymous SNPs, 0.36 and 0.48 respectively, showing that the non-synonymous SNPs have a more deleterious effect on average. Figure 7.3a (top) shows the histogram for non-synonymous SNPs. Due to the biological factors associated with non-synonymous SNPs we chose the

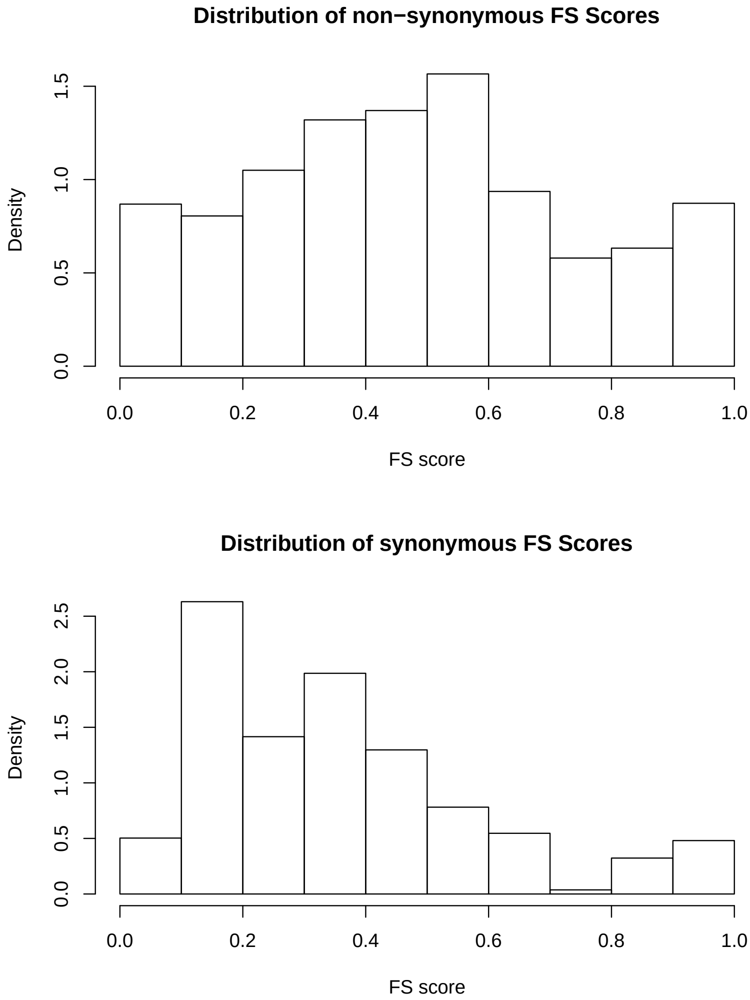


Figure 7.3: The distribution of functional significance (FS) scores [Lee and Shatkay, 2009]. (a) FS scores for non-synonymous SNPs (top) (b) FS score for synonymous SNPs (bottom).

Uniform distribution to represent this. Non-synonymous SNPs, by definition, change the amino acid and therefore the protein they encode. Depending on the particular type of codon the mutation effects, the effect can be very large or sometimes very small. For example, if the mutation leads to a premature stop codon, the protein will be truncated. This may prevent it from functioning correctly. However, if the mutation leads to a small chemical change in an amino acid in the centre of the protein the effect might be negligible due to the rest of the protein being correct. The synonymous SNPs have two distinct groups of effects - the less common being highly deleterious (FS score close to 1) and the most common having a small deleterious effect (FS score close to 0). Biologically, Hunt et al. [2009] explain that synonymous SNPs, also referred to as silent mutations due to their lack of change of the amino acid, can affect messenger

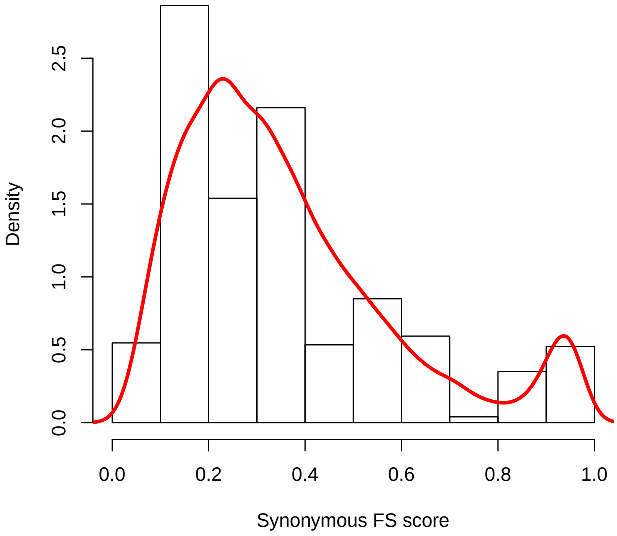


Figure 7.4: The distribution of functional significance (FS) scores [Lee and Shatkay, 2009] for synonymous SNPs. The histogram represents the actual FS score values, while the red line is the fitted mixture distribution $\pi(FS_{syn}) \sim \mathbb{1}_{FS_{syn} \in [0,1]} \left\{0.946Ga\left(2.929, \frac{1}{0.113}\right) + 0.054Ga\left(640.5, \frac{1}{0.0015}\right)\right\}$. There is small mass beyond 1, but we truncate this using the indicator variable.

RNA, as well as stability and structure of proteins and protein folding. These can adversely affect the function of a protein. The two groups of FS scores for synonymous SNPs fit well with a mixture distribution, see Figure 7.4. The Gamma distribution provides the skewness and heavy-tails required to create a well fitting mixture distribution. The prior we use is defined as

$$\pi(FS_{syn}) \sim \mathbb{1}_{FS_{syn} \in [0,1]} \left\{0.946Ga\left(2.929, \frac{1}{0.113}\right) + 0.054Ga\left(640.5, \frac{1}{0.0015}\right)\right\}. \tag{7.1}$$

This prior distribution for FS_{syn}, without the indicator function, is not constrained to be within $[0, 1]$. During the updating of FS_{syn}, we discard any values that are sampled outside of $[0, 1]$. This is an inefficient method, but we only discard around 8% of sampled values and so we believe, given the good fit of the mixture distribution to the FS score data, that this is sufficient. We tried to fit distributions that are naturally constrained to $[0, 1]$ such as the Beta distribution but we couldn't find any that provided a good fit.

7.3.2 Transforming the FS score

To use the FS score, constrained to $[0, 1]$, to enforce a change in the mean of the prior variance of β using B we require the interval of values for B to include 1 (the "standard" shrinkage of the Normal Gamma). This can be achieved by many transformations. We choose simply to translate the FS score region by 0.5, such that $B = FS + 0.5$ and $B \in [0.5, 1.5]$. Figure 7.2 shows the effect of B in this range on the prior variance of β.

7.4 Computational changes to the Normal Gamma

When including functional information we update the parameters of the Normal Gamma in two stages for the synonymous and non-synonymous SNPs, meaning that we have group specific parameters for λ, γ^{-2} and FS/B. The full conditional distributions and acceptance probability for λ that are used can be found in Section 5.5.2. In this case, the two stages of updating correspond to updating synonymous SNPs and then non-synonymous SNPs.

To show the changes to the standard Normal Gamma code, which we call the NG splitting function, we include pseudocode in Appendix D.2, page 185. This is useful for understanding the extra computational effect of the inclusion of functional information.

When calculating the full conditional distributions for FS_{syn} and FS_{non} we only need to include the prior distributions that contain either an F or B term. This means we only include the prior distribution for FS and the prior distribution for $\pi(\gamma^{-2}|\lambda)$. The probability density function (pdf) for $\pi(\gamma^{-2}|\lambda, B)$ is given by

$$\begin{aligned}\pi(\gamma^{-2}|\lambda, B) &\sim Ga\left(2, \frac{MB}{2\lambda}\right)\\ &= \frac{\left(\frac{MB}{2\lambda}\right)^2}{\Gamma(2)} \left(\gamma^{-2}\right)^{2-1} exp\left(\frac{-MB}{2\lambda}\gamma^{-2}\right)\\ &= \frac{1}{4\lambda^2\gamma^2}(MB)^2 \exp\left(\frac{-MB}{2\lambda\gamma^2}\right) \qquad (7.2)\end{aligned}$$

7.4.1 Full conditional distribution for FS_{non}

The full conditional distributions for B or FS change dependent on whether the SNP is synonymous or non-synonymous. For non-synonymous SNPs the full conditional distribution is

$$
\begin{aligned}
f(FS_{non}|\lambda,\gamma) \propto & \pi(FS_{non}) f\left(\gamma^{-2}|B = FS_{non} + 0.5, \lambda\right) \\
\propto & (FS_{non} + 0.5)^2 \exp\left(-\frac{M}{2\lambda\gamma^2}(FS_{non} + 0.5)\right) \mathbb{1}_{FS_{non}\in[0,1]} \\
\propto & (FS_{non}^2 + FS_{non} + \frac{1}{4}) \exp\left(-\frac{M}{2\lambda\gamma^2} FS_{non}\right) \mathbb{1}_{FS_{non}\in[0,1]} \\
= & \left[FS_{non}^2 \exp\left(-\frac{M}{2\lambda\gamma^2} FS_{non}\right) + FS_{non} \exp\left(-\frac{M}{2\lambda\gamma^2} FS_{non}\right)\right. \\
& \left. + \frac{1}{4} \exp\left(-\frac{M}{2\lambda\gamma^2} FS_{non}\right)\right] \mathbb{1}_{FS_{non}\in[0,1]} \qquad (7.3)
\end{aligned}
$$

We recognise this as a mixture of three Gamma distributions on a restricted support with the following parameters.

$$
\begin{aligned}
& f(FS_{non}|\lambda,\gamma) \\
& \propto \left[w_1 Ga\left(3, \frac{M}{2\lambda\gamma^2}\right) + w_2 Ga\left(2, \frac{M}{2\lambda\gamma^2}\right) + w_3 Ga\left(1, \frac{M}{2\lambda\gamma^2}\right)\right] \mathbb{1}_{FS_{non}\in[0,1]}, \qquad (7.4)
\end{aligned}
$$

where the weights are defined as follows:

$$
\begin{aligned}
w_{1A} = & \frac{16\lambda^3\gamma^6}{M^3} \times \gamma_f\left(1, 3, \frac{M}{2\lambda\gamma^2}\right) \\
w_{2A} = & \frac{4\lambda^2\gamma^4}{M^2} \times \gamma_f\left(1, 2, \frac{M}{2\lambda\gamma^2}\right) \\
w_{3A} = & \frac{\lambda\gamma^2}{2M} \times \gamma_f\left(1, 1, \frac{M}{2\lambda\gamma^2}\right),
\end{aligned}
$$

where $\gamma_f(a, b, c)$ represents the lower incomplete CDF (cumulative distribution function) of the gamma distribution between $(0, a]$ with parameters b and c respectively, and $w_1 = \frac{w_{1A}}{w_{1A} + w_{2A} + w_{3A}}$, $w_2 = \frac{w_{2A}}{w_{1A} + w_{2A} + w_{3A}}$ and $w_3 = \frac{w_{3A}}{w_{1A} + w_{2A} + w_{3A}}$.

When applying the MCMC updating, we simulate the value for FS_{non} by sampling a uniform value, u, which determines which of the three Gamma distributions we sample from. This normalisation and scaling ensures that the relative weights of the three Gamma distributions are sampled with the correct weights.

7.4.2 Full conditional distribution for FS_{syn}

The full conditional distribution for synonymous SNPs, FS_{syn} is defined as follows. For brevity, we omit the indicator function but the following is only non-zero when $FS_{syn} \in [0,1]$ or $B \in \left[\frac{1}{2}, \frac{3}{2}\right]$, we also write FS for FS_{syn}.

$$\begin{aligned} f(FS|\lambda,\gamma) =& \pi(FS) f\left(\gamma^{-2}|\lambda, B = FS + 0.5\right) \\ =& \left(\frac{M(FS+0.5)}{2\lambda}\right)^2 \gamma^{-2} \exp\left\{-\frac{M}{2\lambda\gamma^2}(FS+0.5)\right\} \\ &\times \left\{0.946 \frac{\left(\frac{1}{0.113}\right)^{2.929}}{\Gamma(2.929)} FS^{1.929} \exp\left(-\frac{1}{0.113} FS\right)\right\} \\ &+ \left(\frac{M(FS+0.5)}{2\lambda}\right)^2 \gamma^{-2} \exp\left\{-\frac{M}{2\lambda\gamma^2}(FS+0.5)\right\} \\ &\times \left\{0.054 \frac{\left(\frac{1}{0.0015}\right)^{640.5}}{\Gamma(640.5)} FS^{639.5} \exp\left(-\frac{1}{0.0015} FS\right)\right\} \end{aligned} \tag{7.5}$$

After some algebra, we notice this is a mixture of 6 Gamma distributions with shape and rate parameters as follows:

$$\begin{aligned} f(FS_{syn}|\lambda,\gamma) \propto & \sum_{i=1}^{3} \left\{ w_i Ga\left(1.929 + i, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right)\right\} \\ &+ \sum_{i=4}^{6} \left\{ w_i Ga\left(636.5 + i, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right)\right\} \end{aligned} \tag{7.6}$$

The relative weights $w_1, \ldots, w_6$ are defined as follows.

$$\begin{aligned} w_{1B} =& \frac{0.946}{4} \left(\frac{\left(\frac{1}{0.113}\right)^{2.929}}{\Gamma(2.929)}\right) \left(\frac{\Gamma(2.929)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right)^{2.929}}\right) G\left(1, 2.929, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right) \\ w_{2B} =& 0.946 \left(\frac{\left(\frac{1}{0.113}\right)^{2.929}}{\Gamma(2.929)}\right) \left(\frac{\Gamma(3.929)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right)^{3.929}}\right) G\left(1, 3.929, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right) \\ w_{3B} =& 0.946 \left(\frac{\left(\frac{1}{0.113}\right)^{2.929}}{\Gamma(2.929)}\right) \left(\frac{\Gamma(4.929)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right)^{4.929}}\right) G\left(1, 4.929, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.113}\right) \\ w_{4B} =& \frac{0.054}{4} \left(\frac{\left(\frac{1}{0.0015}\right)^{640.5}}{\Gamma(640.5)}\right) \left(\frac{\Gamma(640.5)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right)^{640.5}}\right) G\left(1, 640.5, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right) \\ w_{5B} =& 0.054 \left(\frac{\left(\frac{1}{0.0015}\right)^{640.5}}{\Gamma(640.5)}\right) \left(\frac{\Gamma(641.5)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right)^{641.5}}\right) G\left(1, 641.5, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right) \\ w_{6B} =& 0.054 \left(\frac{\left(\frac{1}{0.0015}\right)^{640.5}}{\Gamma(640.5)}\right) \left(\frac{\Gamma(642.5)}{\left(\frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right)^{642.5}}\right) G\left(1, 642.5, \frac{M}{2\lambda\gamma^2} + \frac{1}{0.0015}\right), \end{aligned}$$

with $w_i = \frac{w_{iB}}{\sum_{j=1}^{6} w_{jB}}$, and where $G(a, b, c)$ represents the lower incomplete Gamma function up to a for shape parameter b and rate parameter c.

When applying the MCMC updating, we simulate the value for FS_{syn} by sampling a uniform value, u, which determines which of the six Gamma distributions we sample from. This normalisation and scaling ensures that the relative weights of the Gamma distributions are sampled with the correct weights.

7.5 Simulation Results, HapGen dataset 2

The NG splitting function is the name we give to the NG function with functional information for synonymous and non-synonymous SNPs included. The pseudocode can be found in Appendix D.2. The NG splitting function was first tested on HapGen datasets 1A and 1B but the results were similar to those that we present for HapGen datasets 2A and 2B, and so are omitted. HapGen dataset 2A and 2B are both simulated using HapGen2 [Su et al., 2011] with 6 causal SNPs in each of the 9 sub-datasets of dataset 2A and 2B. The effect sizes are simulated to be 0.4. The causal SNPs in dataset 2A have approximate population MAF 0.2 while in dataset 2B they have approximate population MAF 0.02. We simulate 9 sub-datasets of dataset 2A and another 9 sub-datasets of dataset 2B, each with a total of 631 SNPs and 300 individuals. This means that in our analysis via ROC curves, we have $9 \times 6 = 54$ causal SNPs and $9 \times (631 - 6) = 5625$ non-causal SNPs. More details can be found in Section 4.2.1.

We test the NG splitting function on the HapGen simulated dataset 2A and 2B using two scenarios. Firstly the 'best case' where all simulated causal SNPs are treated as non-synonymous and all non-causal SNPs are treated as synonymous. We denote this case 'true causal'. The second case is the 'worst case' where all simulated causal SNPs are treated as synonymous and all non-causal SNPs are treated as non-synonymous. We denote this case 'false causal'. This is the worst case as the causal SNPs are all in the group where, a priori, there will be more shrinkage on this set of SNPs. The results for datasets 2A and 2B can be seen in Figure 7.5 top and bottom respectively.

Looking at the ROC curve for dataset 2A (MAF 0.2), Figure 7.5 (top), the AUCs for NG, NG true causal and NG false causal are 0.968, 0.975 and 0.971 respectively. Using DeLong's test the NG is not statistically different to either NG true causal or false causal (p-values 0.735 and 0.997 respectively), neither is the AUC for NG true causal and false causal, p-value 0.366. We notice that at a false positive rate (FPR) < 0.1, the splitting function is superior to the standard

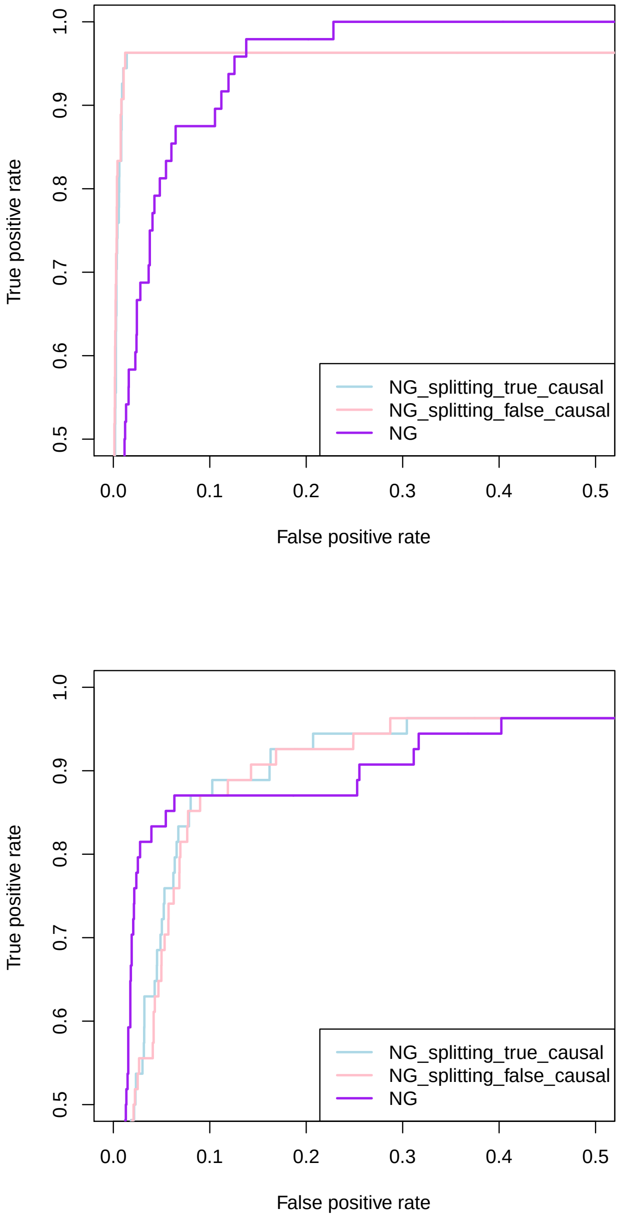


Figure 7.5: ROC curve assessing the difference between the Normal Gamma with the same shrinkage across all SNPs, to the NG with different shrinkage across SNPs. True causal represents the 'best case' scenario where all simulated causal SNPs are treated as non-synonymous and false causal represents the 'worst case' scenario where all simulated causal SNPs are treated as synonymous. The data is simulated using HapGen2 [Su et al., 2011] to include 6 causal SNPs with effect size 0.4 and 625 non-causal SNPs for each of 9 sub-datasets. **Top:** Dataset 2A with a population MAF (minor allele frequency) of approximately 0.2. **Bottom:** Dataset 2B with a population MAF (minor allele frequency) of approximately 0.02.

Normal Gamma, even when the causal SNPs are allocated to the synonymous group. This shows that the NG splitting is outperforming the standard Normal Gamma with respect to ranking of the causal SNPs in the most relevant FPR range.

For the ROC curve for dataset 2B (MAF 0.02), Figure 7.5 (bottom), the AUCs for NG, NG true causal and NG false causal are 0.932, 0.931 and 0.929 respectively. Using DeLong's test the NG is not statistically different to either NG true causal or false causal (p-values 0.894 and 0.666 respectively). The NG true causal and false causal scenarios are not statistically different at the 5% level, but are at the 10% level, p-value 0.089. We notice that the standard Normal Gamma is performing better with respect to the ranking of the causal SNPs up to a false positive rate (FPR) of approximately 0.05 in comparison to the NG splitting, even when the causal SNPs are allocated to the non-synonymous SNP group. This is the opposite effect to we see in the case of MAF 0.2. We believe that the prior is having less of an effect in the MAF 0.2 case (dataset 2A) because of the relative information in the likelihood.

7.5.1 Investigating the differential shrinkage from the NG splitting function

Given the similarity of the ROC curves and hence the posterior ranks of the SNPs, we need to investigate the posterior distribution of B for synonymous and non-synonymous SNPs, to ensure that different levels of shrinkage are being enforced in each group. Figure 7.6 shows an example of the prior and posterior distributions for B_{syn} and B_{non} for one of the sub-datasets of dataset 2B. These histograms are representative of all posterior distributions in dataset 2A and 2B. We show the histograms for the NG false causal (causal SNPs are identified as synonymous) in Figure 7.6 (top) and for the NG true causal (causal SNPs defined as non-synonymous) in Figure 7.6 (bottom).

For the false causal case (top histogram in Figure 7.6), the prior and the posterior for B_{syn} (causal SNPs) are similar, although the posterior appears to have a higher mean. There is clearly more information in the likelihood, hence increasing the posterior mean in comparison to the prior. This is in contrast to the true causal case, where the prior mean is higher than the posterior mean. This is what we would expect in this case as the 6 causal SNPs are in the synonymous group. Hence, reducing the shrinkage (increasing B) on their posterior effect sizes is our aim. We need to be aware that there are only 6 SNPs (the causal SNPs) in this group, and so there is not much difference between the prior and the posterior.

In the true causal case (bottom histogram in Figure 7.6) we notice there is a large difference between the prior and the posterior for B_{syn}. B_{syn} represents the non-causal SNPs, of which there are 625, and so the likelihood is very informative for this group.

For B_{non} there does not appear to be much difference between the true and false causal cases. This may be because the prior is uninformative and therefore the posterior is based only on the likelihood. It may also be due to the upper bound of 1.5 on B. This may not be sufficient for enforcing a reduction in the shrinkage. We also note that in the false causal case, the SNPs in the non-synonymous group (the non-causal SNPs) may be being given large effect sizes because those in the synonymous group (the causal SNPs) are not, due to a lack of information in the likelihood.

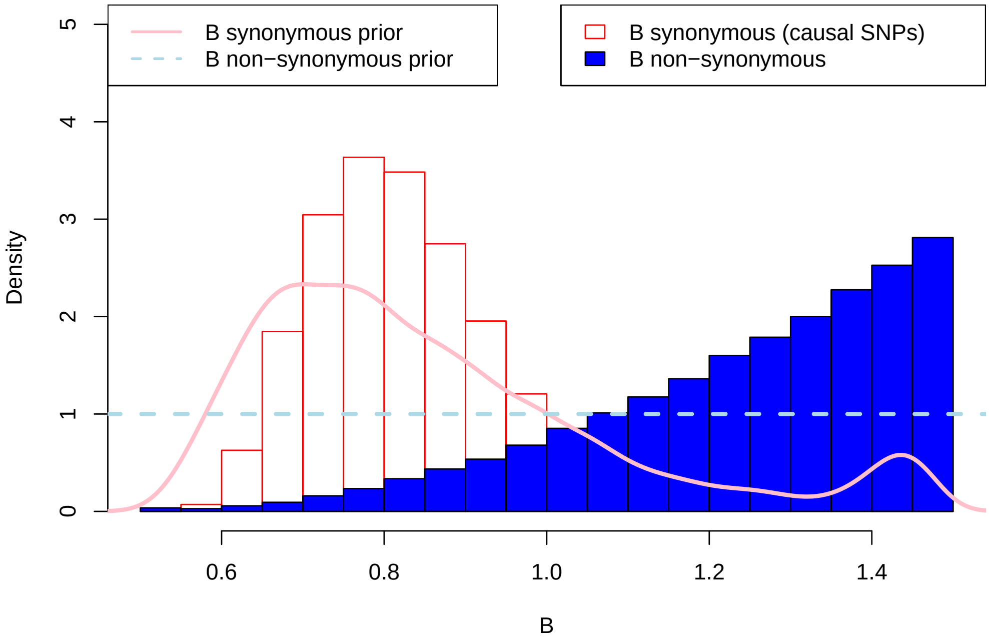
Table 7.1 shows the range of posterior means for B over all 9 datasets within each of datasets 2A and 2B. We notice that the posterior means for B_{non} are similar across both MAFs and across both scenarios (true causal and false causal). In both scenarios it is noticeable higher than the prior mean of 0.98.

The posterior means for B_{syn} are very different to B_{non}, and to each other for the true causal and false causal scenario. As expected, the posterior means for B_{syn} are higher in the false causal case, where the causal SNPs are defined to be synonymous. This is as we would have expected. In the true causal case where the non-causal SNPs are synonymous, the posterior mean range is smaller than the prior mean of 0.86. The prior mean is included in the posterior mean range for the false causal case, where the causal SNPs are defined as synonymous.

	B_{syn} (0.86)	B_{non} (0.98)
NG false causal MAF 0.02 (dataset 2B)	0.8-0.95	1.22-1.24
NG false causal MAF 0.2 (dataset 2A)	0.8-0.89	1.22-1.24
NG true causal MAF 0.02 (dataset 2B)	0.57-0.63	1.17-1.23
NG true causal MAF 0.2 (dataset 2A)	0.57-0.59	1.21-1.23

Table 7.1: The range of posterior means across all 9 datasets within datasets 2A and 2B for B_{syn} and B_{non} for both NG true causal (causal SNPs defined as non-synonymous, non-causal SNPs defined as synonymous) and NG false causal (non-causal SNPs defined as non-synonymous, causal SNPs defined as synonymous). The values that relate to the group where the causal SNPs are found is highlighted in red. The values in brackets represent the prior means for B_{syn} and B_{non}.

We investigate the magnitude of the posterior effect sizes in the NG, NG splitting true causal and NG splitting false causal cases. Figure 7.7 shows the effect sizes for the 54 causal SNPs after a single run of the NG, NG splitting false causal and NG splitting true causal for dataset 2B. All causal SNPs in



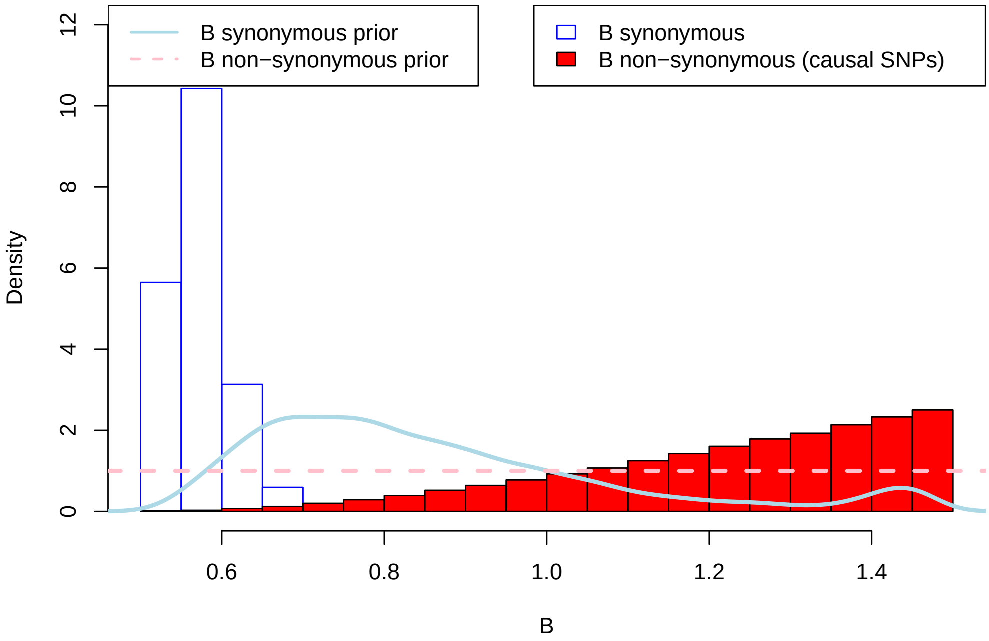


Figure 7.6: Histograms showing the posterior distribution of B for synonymous and non-synonymous SNPs. The results are based on one sub-dataset of HapGen simulated dataset 2B with causal SNPs effect size 0.4 and MAF (minor allele frequency) of approximately 0.02 in the population. These distributions are based on 100,000 iterations with a 5,000 iteration burn-in. The B that represents the group containing only the causal SNPs are coloured red in these histograms. **Top:** The false causal case where causal SNPs are all denoted as synonymous (worst case of NG splitting). **Bottom:** The true causal case, where causal SNPs are all denoted as non-synonymous (best case of NG splitting).

dataset 2B and in Figure 7.7 are simulated to have effect size 0.4. We notice that the posterior magnitudes of the effect sizes are different.

The top plot in Figure 7.7, for dataset 2A, MAF 0.2, shows that the causal SNP posterior mean estimates are highest for the NG true causal, next the NG false causal and finally for the NG in most cases. We expect the NG true causal to enforce less shrinkage on the causal SNPs as they are in the non-synonymous functional group. We see this reflected in the posterior mean effect sizes (simulated to be 0.4) in comparison to the NG false causal where the causal SNPs are in the synonymous functional group, and have a priori more shrinkage. We believe that the extra flexibility in the shrinkage is allowing larger posterior effect sizes compared to the standard NG.

The bottom plot in Figure 7.7, for MAF 0.02, dataset 2B, does not show the clear separation of posterior effect sizes between the NG, NG true causal and NG false causal cases. We notice that there are two causal SNPs which have large effect sizes in the NG but not as large effect sizes in either of the NG splitting cases. Remember that all causal SNPs have a simulated effect size of 0.4. Even though the effect sizes are not as large, these are still the top two ranked SNPs according to all three scenarios (standard NG, NG true causal and NG false causal).

The investigative results imply that the change in shrinkage based on the FS score is performing as expected and changing the amount of shrinkage on the SNPs in the synonymous and non-synonymous groups. However, the results are not being reflected in the posterior estimates for β. We conclude that the range of $B \in [0.5, 1.5]$ enforced by our simple transformation $B = FS + \frac{1}{2}$ is insufficient. We could consider other linear transformations to increase the support of B which would allow Gibbs updates, but to avoid data-specific supports, we consider non-linear transformations.

7.6 An alternative transformation from FS to B

To overcome the restriction on the values for B enforced by our previous transformation from FS to B, we propose an alternative non-linear transformation.

We define our new monotonic transformation of the FS score as

$$B = \tan\left(\frac{FS\pi}{2}\epsilon\right), \tag{7.7}$$

where $\epsilon = 0.99$, and prevents $B \to \infty$ as $FS \to 1$. This transformation was

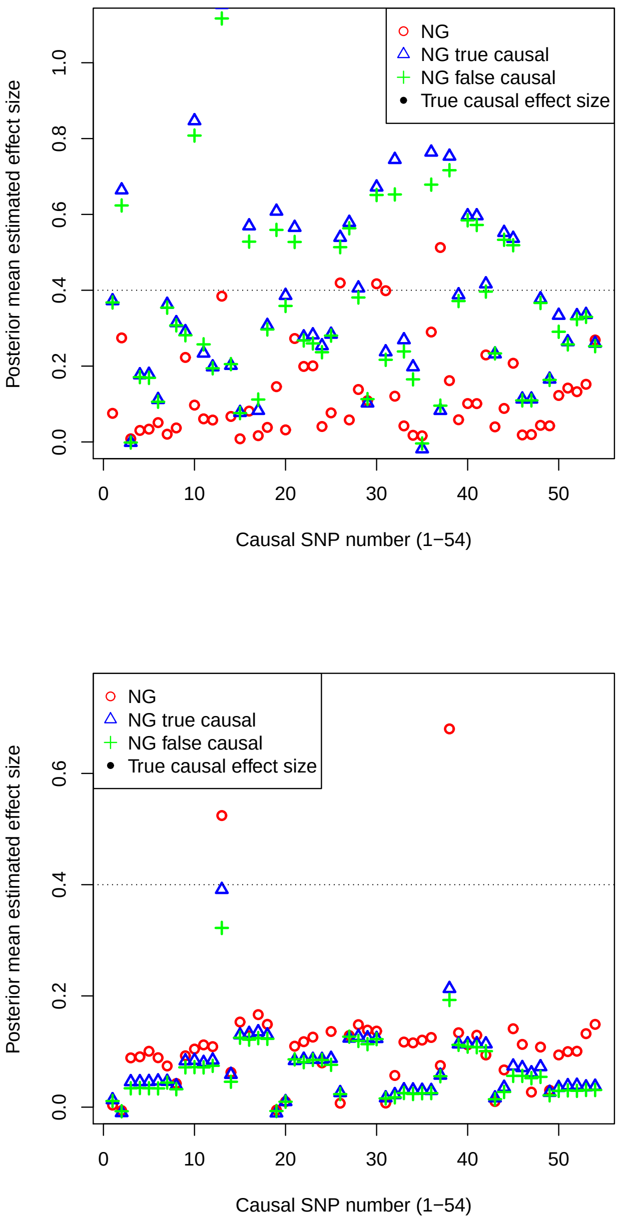


Figure 7.7: The posterior effect size estimates for all 54 causal SNPs in the HapGen datasets 2A and 2B for the NG, NG true causal and NG false causal. The false causal case is the case where only the causal SNPs are defined as synonymous (worst case of NG splitting). The true causal case is the case where only the causal SNPs are all defined as non-synonymous (best case of NG splitting). The non-causal SNPs are all defined to be in the other group, non-synonymous in the false causal case, and synonymous in the true causal case. **Top:** The results are based on HapGen simulated dataset 2A which has causal SNPs effect size 0.4 (shown on the plot) and MAF (minor allele frequency) of approximately 0.2 in the population. **Bottom:** The results are based on HapGen simulated dataset 2B which has causal SNPs effect size 0.4 (shown on the plot) and MAF (minor allele frequency) of approximately 0.02 in the population.

chosen as it maps $FS \in [0,1]$ to $B \in [0,\infty)$. B enters the prior hierarchical structure through $\text{var}(\pi(\beta|\lambda,\gamma^2)) = 2\lambda\gamma^2 \sim IG(2,B)$. We remove M from the scale of the Inverse Gamma distribution here, which allows direct comparison of B across datasets.

The calculations for the full conditional distributions for F_{non} and F_{syn} using the new FS to B transformation, see Equation 7.7, can be found in Section 8.3.2.

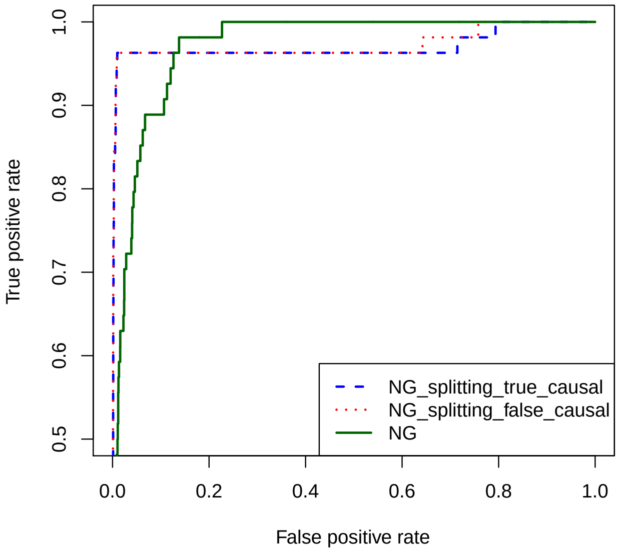
The Normal Gamma function that we have written to include the new transformation of B to FS for the two groups of SNPs, synonymous and non-synonymous, is the NG splitting function. We will refer to this throughout the remainder of this chapter as NG splitting. When comparing to the standard NG function, we will refer to the standard NG function as the NG.

7.7 Simulation Results, HapGen dataset 2

We will now investigate the results of applying the NG splitting function with the new transformation to HapGen datasets 2A and 2B. To recap, these datasets have 9 sub-datasets each with 300 individuals and 631 SNPs, of which 6 are simulated to be causal with effect size 0.4. When analysing dataset 2A or 2B, we have 54 causal SNPs in total and 5625 non-causal SNPs.

We begin by investigating the ROC curves for dataset 2A (MAF 0.2) and 2B (MAF 0.02), Figure 7.8, top and bottom respectively. These ROC curves show little separation. Precision-Recall (PR) curves can be used when there are different numbers of observations in each group. This is the case here, but having plotted the PR curve, there is still little to no separation between the NG splitting true and false causal, and so we omit the plot, preferring the ROC curves for consistency.

The AUCs of the ROC curves for dataset 2A (top ROC in Figure 7.8) are very similar although the shape of the ROC curves are quite different for the NG and the NG splitting. We notice that the shape of the ROCs are similar to Figure 7.5, although the AUCs are larger here. The AUCs for the NG, NG splitting true causal and NG splitting false causal are 0.970, 0.970 and 0.973 respectively. For the two NG splitting cases the difference between these two ROC curves may be due to inter chain MCMC variability. However it is unlikely, given the large difference in shapes of the ROC curves, that the difference between the NG and the two NG splitting cases is due to MCMC variability. In this case, the NG splitting detects the majority of causal SNPs at a much lower false positive rate ($< 5\%$) than the standard NG, although the standard NG detects the entire set of causal SNPs at a much lower false positive rate (less than 30%) compared to the NG splitting (around $70\% - 80\%$). This



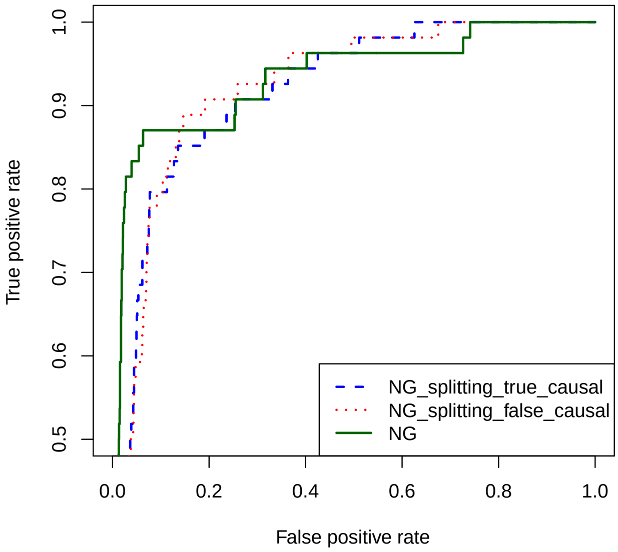


Figure 7.8: ROC curve comparing the difference between the standard Normal Gamma with the same shrinkage across all SNPs, to the NG with different shrinkage across the two SNP groups. True causal represents the 'best case' scenario where all simulated causal SNPs are treated as non-synonymous and false causal represents the 'worst case' scenario where all simulated causal SNPs are treated as synonymous. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a specified minor allele frequency (MAF). **Top:** The ROC is for dataset 2A with causal SNP MAF of approximately 0.2 in the population. **Bottom:** The ROC is for dataset 2B with causal SNP MAF of approximately 0.02 in the population.

is very different to the case where the causal SNPs have MAF (minor allele frequency) 0.02 in dataset 2B.

The AUC's for dataset 2B (bottom ROC in Figure 7.8) are, as expected, less than for dataset 2A. The AUC for the standard NG is 0.932, and for the NG splitting true causal and NG splitting false causal are 0.914 and 0.919 respectively. Again we notice a difference in shape between the NG and the two NG splitting ROC curves. In this case, the NG detects approximately 90% of causal SNPs at a much lower false positive rate (approximately 10%) compared to the two NG splitting cases which have a false positive rate of approximately 20% to achieve a 90% true positive rate. It is possible in this case that the difference between the two NG splitting cases is based on MCMC variation alone and not due to differential shrinkage.

The ROC curves in Figure 7.8 appear to show that the difference in ranking of the posterior mean effect sizes for the SNPs in the NG splitting true cause and false causal cases are similar, regardless of whether the SNPs are in a group with a priori more or less shrinkage. The ROC curves are taking into account relative effect sizes of the posterior mean estimates through the rankings, but they do not assess the actual effect sizes.

To assess whether there is a difference between the two groups, synonymous and non-synonymous, and whether differential shrinkage is being enforced, we begin by assessing the shrinkage parameters B_{syn} and B_{non} for the two NG splitting cases.

In the sub-dataset assessed here, M=0.2724. The sub-dataset used is sub-dataset 2 from dataset 2A, where the MAF of the causal SNPs are approximately 0.02. Table 7.2 shows the comparison between the posterior B and M used in the standard NG. Where $B > M$, there is less shrinkage being enforced in the NG splitting compared to the standard NG.

Investigating only the posterior summary statistics for this sub-dataset, which is representative of the other sub-datasets, has shown that the differential shrinkage using B is performing well, and as we would expect for the two different groups of SNPs. We note that in the NG false causal, a few large values for B_{non} are skewing the posterior mean and 95% posterior credible interval.

Given that we are certain that the shrinkage being applied to each group is different, and is tailored to the data, we now investigate the posterior mean effect sizes for the β estimates. We maintain the estimates from the standard Normal Gamma that have previously been used for comparison.

Figure 7.9 shows the comparison between the posterior mean effect sizes of the causal SNPs for the standard NG, NG splitting true causal and NG splitting false causal. The top plot shows the comparison for dataset 2A, MAF 0.2. As

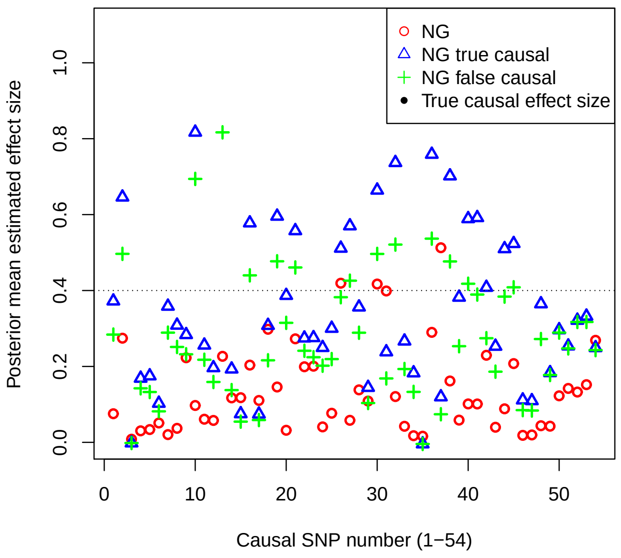
	Posterior $B > M$	**Posterior mean**	**Posterior median**	95% **posterior credible interval**
NG splitting false causal				
B_{syn}	29%	0.2996	0.1801	(0.0379,0.8030)
B_{non}	8%	24.72	0.003	(0.000112,146.54)
NG splitting true causal				
B_{syn}	0% max is 0.047	0.00768	0.0064	(0.00159,0.0206)
B_{non}	12%	0.786	0.659	(0.00383,2.659)

Table 7.2: A summary of the difference between the amount of shrinkage enforced by the NG splitting through B and the standard NG through fixed $M = 0.2724$. These results are for one sub-dataset, sub-dataset 2 from dataset 2A, where the MAF of causal SNPs is approximately 0.02. This is representative of all other sub-datasets. We highlight in red the row of the true causal and false causal results to indicate the group in which the 6 causal SNPs are located.

expected, the NG true causal SNPs have higher posterior mean effect sizes than the false causal and the standard NG. For dataset 2B, MAF 0.02, the bottom plot of Figure 7.9, shows that in comparison to the bottom plot in Figure 7.7, the increased range of B appears to be increasing the shrinkage on the posterior mean effect size estimates for the NG splitting true causal and false causal cases, although it is differentiating more between the effect size estimates for the NG false causal and NG true causal cases.

Figures 7.10 and 7.11 show the posterior mean effect sizes of all SNPs for MAF 0.2 and 0.02 respectively for the standard NG, the NG true causal and the NG false causal cases. The posterior means are much smaller for the NG splitting in the case of dataset 2B (MAF 0.02) compared to the standard NG. However, for dataset 2A (MAF 0.2), the posterior mean effect sizes for the NG splitting are in general larger than for the NG, in particular for the causal SNPs. In most cases the posterior effect sizes for the NG splitting true causal (where the causal SNPs are defined to be non-synonymous) are higher than the NG splitting false causal (where the causal SNPs are defined to be synonymous), which is consistent with our prior distributions. We note that there are only 6 SNPs in the group containing the causal SNPs, for each of the 9 sub-datasets.

We have seen that there is clearly differential shrinkage being enforced on the posterior mean estimates of effect size β through changes in B. The differences in the effect size estimates between the NG false causal and true causal are quite small. We hypothesize this may be due to amount of information in the likelihood and that we are maintaining all causal SNPs together in one of the two groups at all times. We will, for dataset 2A only (population MAF 0.2 for



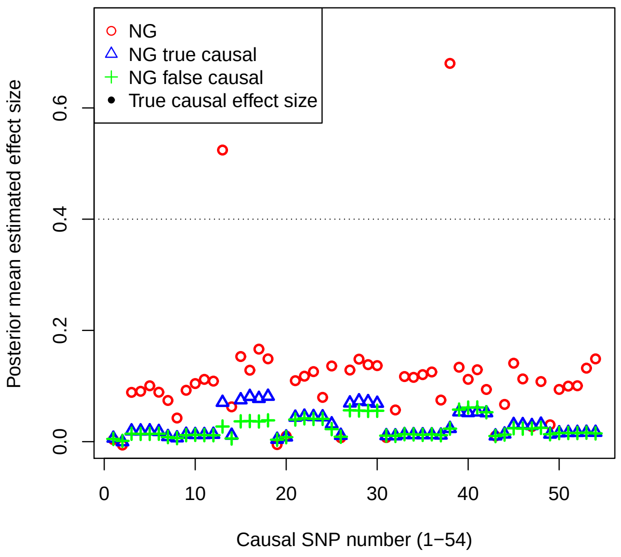


Figure 7.9: The posterior effect size estimates for all 54 causal SNPs in the HapGen datasets 2A and 2B for the NG, NG true causal and NG false causal with the larger range transformation from $F \rightarrow B$. The false causal case is the case where only the causal SNPs are defined as synonymous (worst case of NG splitting). The true causal case, is the case where only the causal SNPs are defined as non-synonymous (best case of NG splitting). The non-causal SNPs are all defined to be in the other group, non-synonymous in the false causal case, and synonymous in the true causal case. **Top:** The results are based on HapGen simulated dataset 2A which has causal SNPs effect size 0.4 (shown on the plot) and MAF (minor allele frequency) of approximately 0.2 in the population. **Bottom:** The results are based on HapGen simulated dataset 2B which has causal SNPs effect size 0.4 (shown on the plot) and MAF (minor allele frequency) of approximately 0.02 in the population.

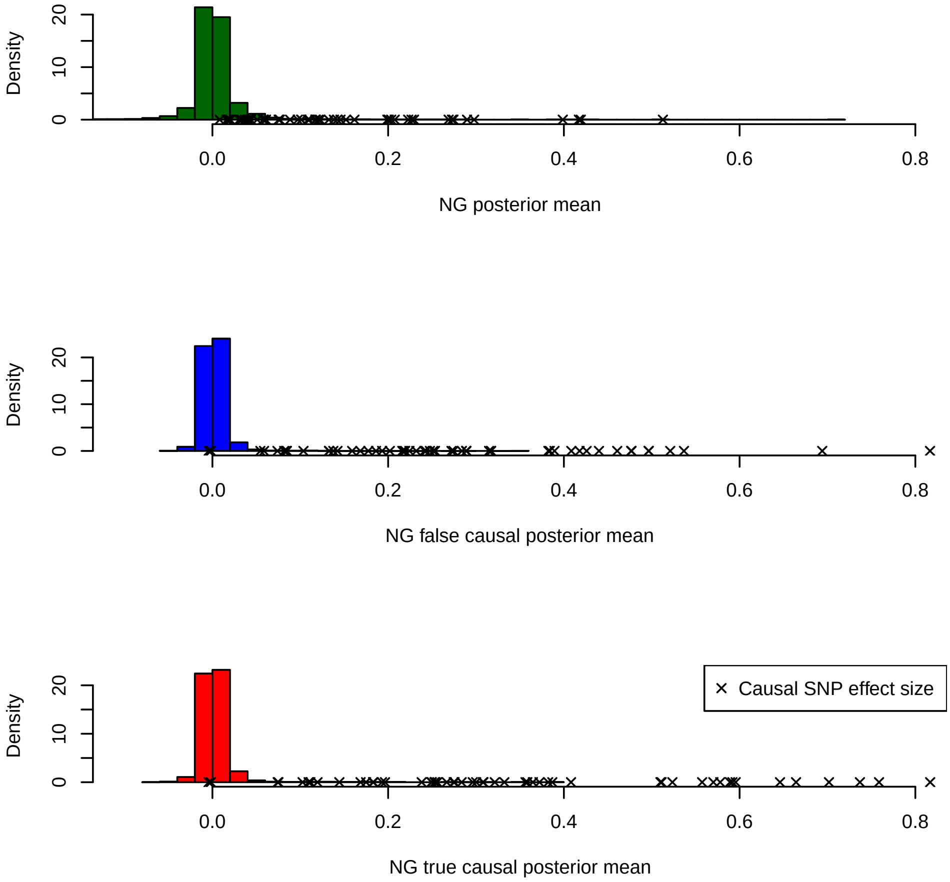


Figure 7.10: Histograms comparing the posterior mean effect size for all SNPs in dataset 2A. The 54 causal SNPs from the NG, NG splitting true and NG splitting false causal scenarios are marked with '×' on the histograms. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) in the population of approximately 0.2 (dataset 2A). The NG splitting false causal case is where all 54 causal SNPs are defined as synonymous and all 625 non-causal SNPs are defined as non-synonymous. The NG splitting true causal case is where all 54 causal SNPs are defined as non-synonymous and the remaining 625 non-causal SNPs are defined as non-synonymous.

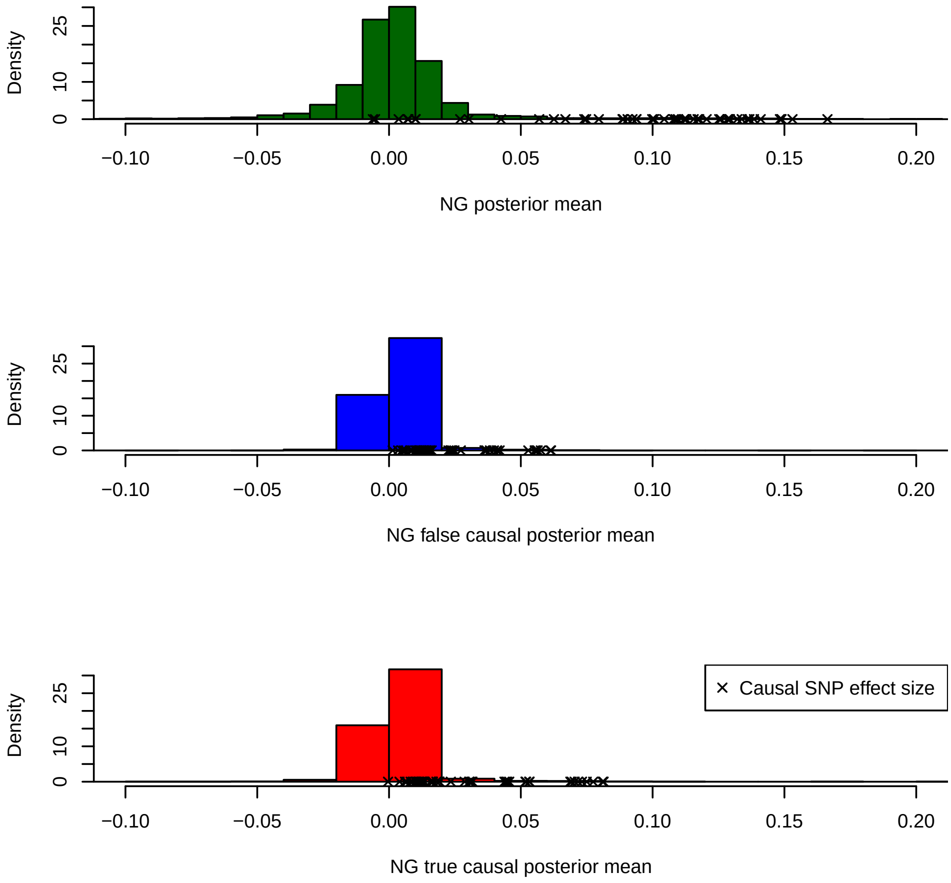


Figure 7.11: Histograms comparing the posterior mean effect size for all SNPs in dataset 2B. The 54 causal SNPs from the NG, NG splitting true and NG splitting false causal scenarios are marked with '×' on the histograms. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) in the population of approximately 0.02 (dataset 2B). The NG splitting false causal case is where all 54 causal SNPs are defined as synonymous and all 625 non-causal SNPs are defined as non-synonymous. The NG splitting true causal case is where all 54 causal SNPs are defined as non-synonymous and the remaining 625 non-causal SNPs are defined as non-synonymous.

the causal SNPs), reduce n to increase the influence of the prior distribution and create two groups of SNPs; one containing all causal SNPs plus extra non-causal SNPs, and the other containing only non-causal SNPs. We will also split the 6 causal SNPs equally between the two groups, synonymous and non-synonymous as this is a more realistic scenario.

7.7.1 Reducing n

In this section we reduce the number of individuals in dataset 2A to 50 and 100 with the aim of assessing the relative influence of the prior distribution on the posterior. The individuals to keep were chosen randomly, and checks have been made to ensure the MAF remains similar to the full datasets with $n = 300$.

We assess, using a ROC curve, the difference between $n = 50$ and $n = 100$ and the true and false causal cases. We compare this to the standard NG case.

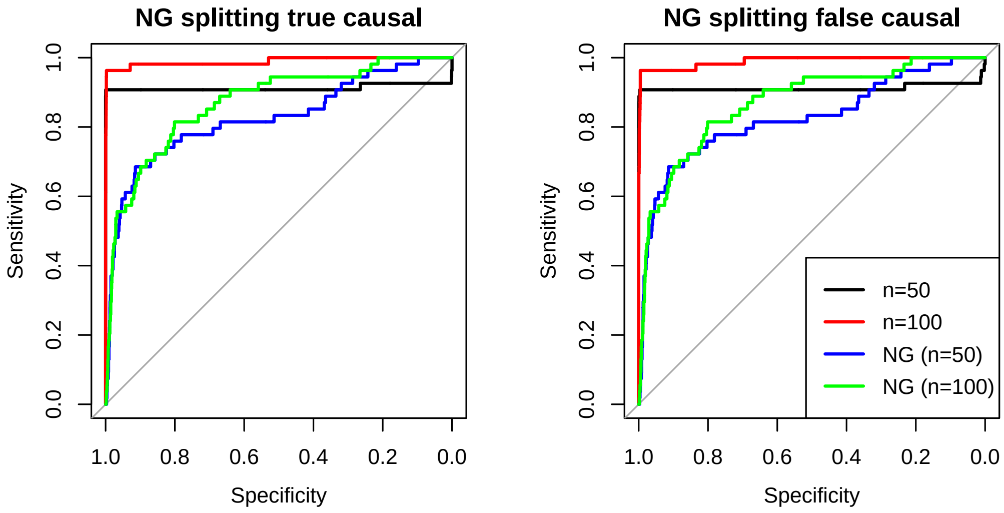


Figure 7.12: ROC curve comparing the difference between the standard Normal Gamma with the same shrinkage across all SNPs to the NG with different shrinkage across the two SNP groups for different values of n. True causal represents the ‘best case’ scenario where all simulated causal SNPs are treated as non-synonymous and false causal the ‘worst case’ scenario where all simulated causal SNPs are treated as synonymous. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a specified minor allele frequency (MAF). The ROC is for MAF of approximately 0.2 in the population (dataset 2A) with reduced numbers of individuals.

Figure 7.12 shows clear differences between the NG splitting results for $n = 50$ and $n = 100$. As expected, when $n = 100$ the AUC is larger, see Table 7.3. We note that the NG splitting true and false causal with $n = 50$ both perform better with respect to the AUC than the NG with $n = 100$. This reinforces the hypothesis that the extra flexibility on the model by not stating a fixed M for the prior on $2\lambda\gamma^2$ enhances the model performance. The differences between the ROCs for NG splitting true and false causal appear most likely to be due to MCMC variation. However, when assessing the posterior effect size estimates for the causal SNPs only, we notice those for the false causal are in general closer to 0 than for the true causal, see Figure 7.13 top and bottom for $n = 50$ and $n = 100$ respectively. This reflects what we expect given our prior distributions.

NG ($n = 50$)	0.8263
NG splitting true causal ($n = 50$)	0.9124
NG splitting false causal ($n = 50$)	0.9121
NG($n = 100$)	0.8704
NG splitting true causal ($n = 100$)	0.9896
NG splitting false causal ($n = 100$)	0.9904

Table 7.3: The AUC for the ROC curves in Figure 7.12.

Having reduced n to 50 and 100 we clearly see a difference in the posterior mean effect size estimates for the causal SNPs. The comparison with the standard NG reinforces the advantage of the extra level of complexity we have introduced into the model in the form of a prior on B based on functional information.

7.7.2 Re-grouping the SNPs

To assess the NG splitting model in a more realistic setting, we regroup the SNPs such that there are the 6 causal SNPs plus 218 randomly selected non-causal SNPs in one group, and the remaining 407 non-causal SNPs in the second group. For the ‘causal plus non’ scenario, the 6 causal SNPs plus 218 non-causal SNPs are defined as non-synonymous and the remaining 407 SNPs are defined as synonymous. This is equivalent to the true causal case with extra, non-causal SNPs also defined as non-synonymous. The ‘causal plus syn’ scenario has the 6 causal SNPs plus 218 non-causal SNPs defined as synonymous and the remaining 407 SNPs defined as non-synonymous. This is equivalent to the false causal case with extra non-causal SNPs also defined as synonymous.

We begin by assessing the AUC of the ROC curves, see Figure 7.14. Note that there are 30 causal SNPs and 3155 total SNPs as we are only using 5 sub-datasets. We do this as the results from 5 or 9 sub-datasets remain invariant. The AUC for the NG is 0.932, for the NG causal plus non is 0.989 and the NG causal plus syn is 0.987. Again, assessing only the ROC curves, we would hypothesize that the difference between the NG causal plus non and NG causal plus syn are only differing due to MCMC variation. When we assess the

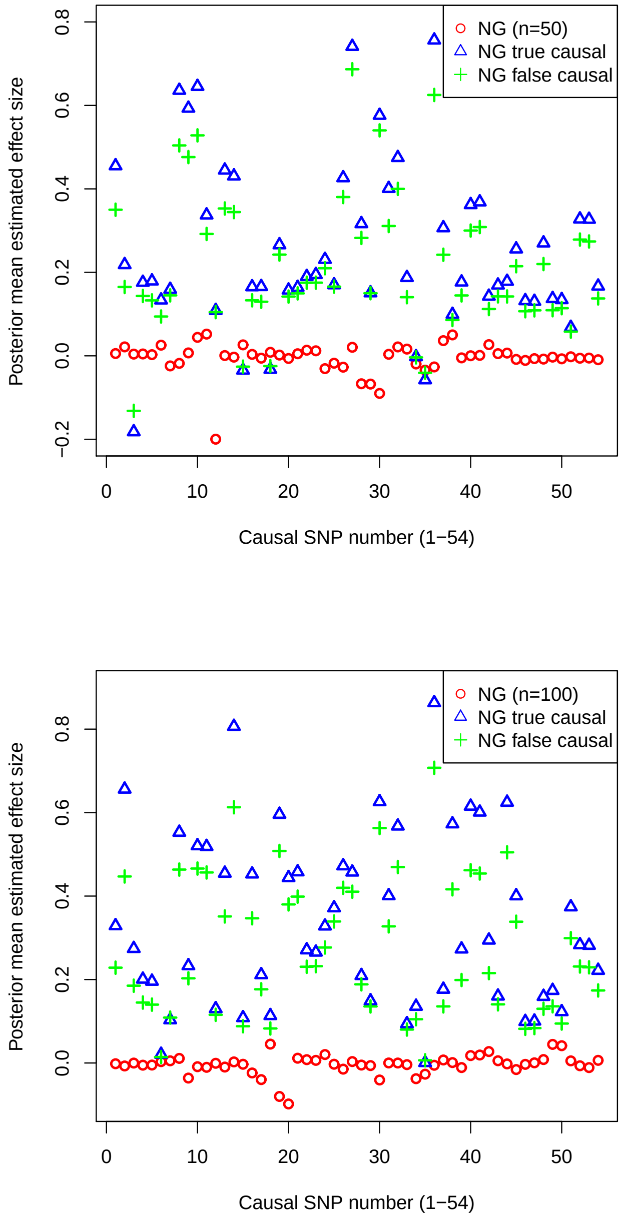


Figure 7.13: Plots comparing the posterior mean effect size for the 30 causal SNPs from the NG, NG splitting true and NG splitting false causal scenarios. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) in the population of approximately 0.2 (dataset 2A). The NG splitting false causal case is where all 6 causal SNPs are defined as synonymous and all 625 non-causal SNPs are defined as non-synonymous. The NG splitting true causal case is where all 6 causal SNPs are defined as non-synonymous and the remaining 625 non-causal SNPs are defined as non-synonymous for each of the 5 sub-datasets used. **Top:** Dataset contains $n = 50$ individuals. **Bottom:** Dataset contain $n = 100$ individuals.

posterior mean effect sizes for the causal SNPs, see Figure 7.15, we notice that the NG causal plus non, which has a priori larger effect sizes, also has larger posterior effect sizes in most cases.

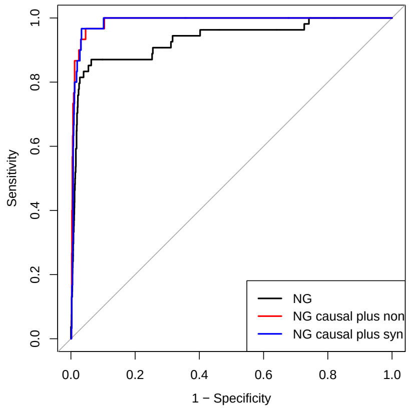


Figure 7.14: ROC curve comparing the difference between the standard Normal Gamma with the same shrinkage across all SNPs to the NG with different shrinkage across the two SNP groups. For the 'causal plus non' group, the 6 causal SNPs plus 218 non-causal SNPs are defined as non-synonymous and the remaining 407 SNPs are defined as synonymous. The 'causal plus syn' group has the 6 causal SNPs plus 218 non-causal SNPs defined as synonymous and the remaining 407 SNPs defined as non-synonymous. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a specified minor allele frequency (MAF). The ROC is for data with MAF of approximately 0.2 in the population (dataset 2A).

We conclude that even when disguising the causal SNPs in a group with other, non-causal SNPs the NG splitting still out-performs the standard NG. Within the NG splitting, we still see from the ROC curves that the causal SNPs are being selected amongst those SNPs with the highest posterior mean effect sizes even when categorised as synonymous (having a priori smaller effect sizes) with many other non-causal SNPs.

7.7.3 Splitting the causal SNPs between groups

In this section we split the causal SNPs such that there are 307 SNPs in the synonymous group, of which 3 are causal, and 324 SNPs in the non-synonymous group, 3 of which are also causal. This is a more realistic setting for the distribution of causal SNPs across the synonymous and non-synonymous functional information groups.

We assess, using a ROC curve, the AUC and compare this to the AUC for the NG, NG true causal and NG false causal when n is reduced, see Table 7.3.

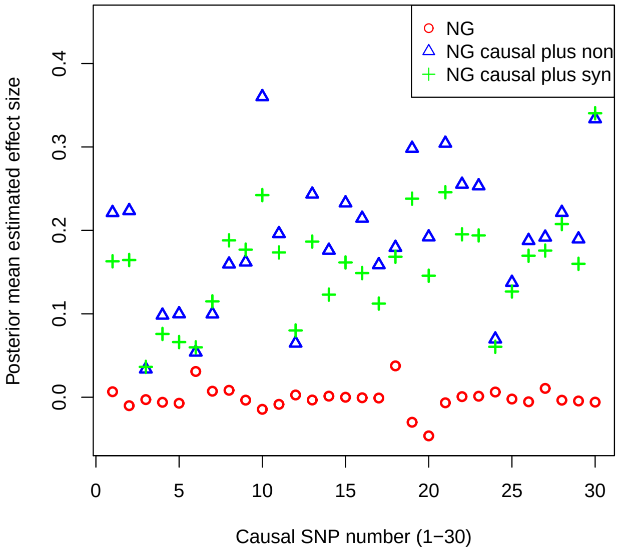


Figure 7.15: Plots comparing the posterior mean effect size for the 30 causal SNPs from the NG, NG causal plus non and NG causal plus syn scenarios. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) in the population of approximately 0.2 (dataset 2A). For the 'causal plus non' group, the 6 causal SNPs plus 218 non-causal SNPs are defined as non-synonymous and the remaining 407 SNPs are defined as synonymous. The 'causal plus syn' group has the 6 causal SNPs plus 218 non-causal SNPs defined as synonymous and the remaining 407 SNPs defined as non-synonymous for each of the 5 sub-datasets used.

We expect that, given n is greater in this case, the AUC will be comparable to the NG splitting cases for $n = 100$ in this table. We find that the AUC for the NG with causal SNPs split across synonymous and non-synonymous SNPs is 0.985. This is almost as high as for the NG splitting true causal and false causal cases. With such a high AUC we conclude that even when splitting the causal SNPs across the two groups, we do not decrease the detection of causal SNPs with respect to the rank.

We also assess the posterior mean effect sizes for the causal SNPs in comparison with the standard NG, see Figure 7.16. We have coloured the SNPs to represent which were in the synonymous and which were in the non-synonymous group. We notice that there are causal SNPs with larger posterior effect sizes than the true causal effect size of 0.4. There are more of these causal SNPs with large effect sizes from the synonymous group than the non-synonymous group. This may be due to which non-causal SNPs are allocated to which group, and the LD between causal and non-causal SNPs within a shrinkage group. We notice also that there is less shrinkage of the largest causal effect sizes in comparison to the standard NG.

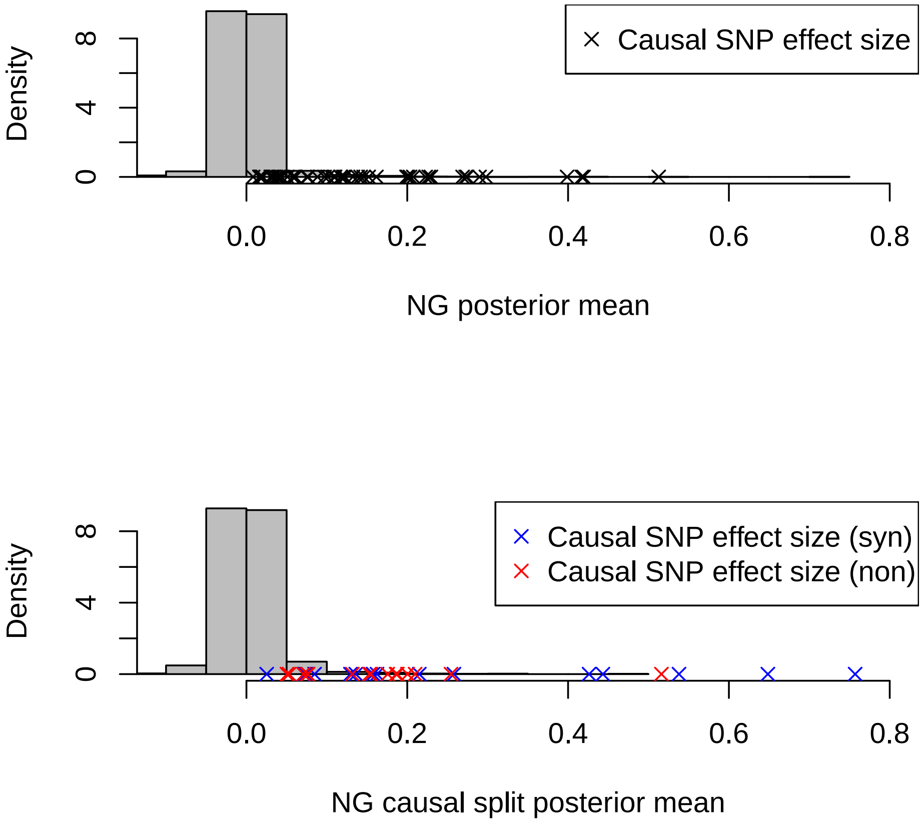


Figure 7.16: Plots comparing the posterior mean effect size for the 30 causal SNPs from the NG and the NG causal split scenarios. The causal split scenario is when the 6 causal SNPs have been split between the two groups, 3 causal SNPs plus 307 non-causal SNPs in the synonymous (with a priori more shrinkage) group and 3 causal SNPs plus 324 non-causal SNPs in the non-synonymous (a priori less shrinkage) group. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) in the population of approximately 0.2 (dataset 2A).

We can conclude here that even when the causal SNPs are split across groups, the ranking of the causal SNPs still leads to a very high AUC of the ROC, therefore the causal SNPs are ranked highly compared to the non-causal SNPs. We can also see that the causal SNPs are still detected regardless of the functional group they are in.

7.8 Conclusion

In this chapter we have assessed two different methods for including functional information into the NG framework. Using the more complex transformation, we have shown that, given the type of SNP, synonymous or non-synonymous, we can enforce differential shrinkage on the posterior mean estimates of effect size β through changes in B. Although we can change the posterior effect size estimates, the posterior ranking remains similar based on the AUC of the ROC curves. By assessing the effect sizes when the number of individuals is reduced, and by combining causal and non-causal SNPs into the same groups we have shown that in a realistic setting, the NG splitting performs very well. By splitting the causal SNPs across groups we have also shown that it is possible to detect causal SNPs even if they are in a group with a priori greater shrinkage and when there are causal SNPs in another group with a priori less shrinkage.

In Chapter 8 we extend the inclusion of synonymous and non-synonymous SNP categories in the Normal Gamma to include more categories for the SNPs beyond exonic mutations. For the Fairfax and Hulse datasets we have used throughout this thesis, SNPs can be found in intronic, intergenic, upstream, downstream, exonic (synonymous and non-synonymous), UTR3, UTR5, splicing, non-coding regions etc. Therefore now we have a model that appears to be successfully enforcing differential shrinkage on the simulated data, we are going to adapt the model to include the extra functional information groups that we/our expert believes are necessary in this type of analysis.

Chapter 8

Extending the Normal Gamma prior with functional information

Having chosen to target only exome sequence data initially, we only required synonymous and non-synonymous SNPs in our NG function with functional information, see Chapter 7. The data we are using to test the Normal Gamma on includes causal SNPs that reside in many different regions of the genome. For this reason, we have decided to extend our NG function with functional information in Chapter 7 to include other SNP groups in our "NG super function".

To decide which extra groups to include, we discussed the possible annotations from ANNOVAR with an expert who understands the biological features affecting gene expression. Using our experts knowledge, we compiled a list of 6 important groups of SNPs including the synonymous and non-synonymous SNPs from before. The 6 groups we now include are synonymous, non-synonymous, UTR3 (also referred to as $3'$ UTR (un-translated region)), splicing, intronic and intergenic. For completeness, we include a 7^{th} group representing the 'other' category to take into account SNPs with no known location in ANNOVAR or SNPs from other ANNOVAR annotations that we are not including. Note that at this stage we group SNPs with upstream, downstream and intergenic labels together as upstream and downstream regions are outside the genes, often between genes. Hence we treat these all as intergenic regions.

A DAG (directed acyclic graph) representing the inclusion of these extra functional information groups can be seen in Figure 8.1, and the corresponding pseudo-code can be found in Appendix D.3.

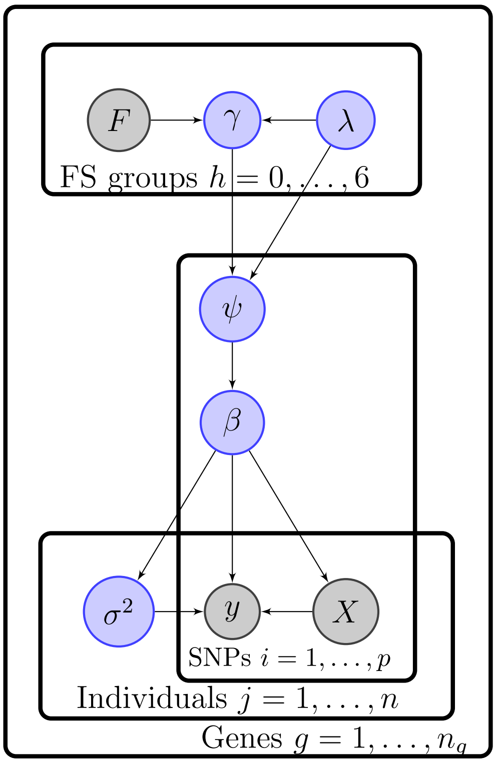

F
γ
λ
FS groups $h = 0, \ldots, 6$
ψ
β
σ^2
y
X
SNPs $i = 1, \ldots, p$
Individuals $j = 1, \ldots, n$
Genes $g = 1, \ldots, n_g$

Figure 8.1: DAG representing the relationships between the variables in the NG super function; the standard NG with seven groups of functional information model. A DAG represents the relationships between parameters in a model. The plates represents the parameters that are specific to groups e.g. individuals, genes etc. The grey nodes represent the observed variables while the arrows show the dependences between parameters.

8.1 Prior distributions

Using the complete FS score database from Lee and Shatkay [2009], omitting all FS score values of 0.5 as these represent no knowledge of the deleterious effect of the SNP, we plot histograms for all functional information groups, see Figure 8.2.

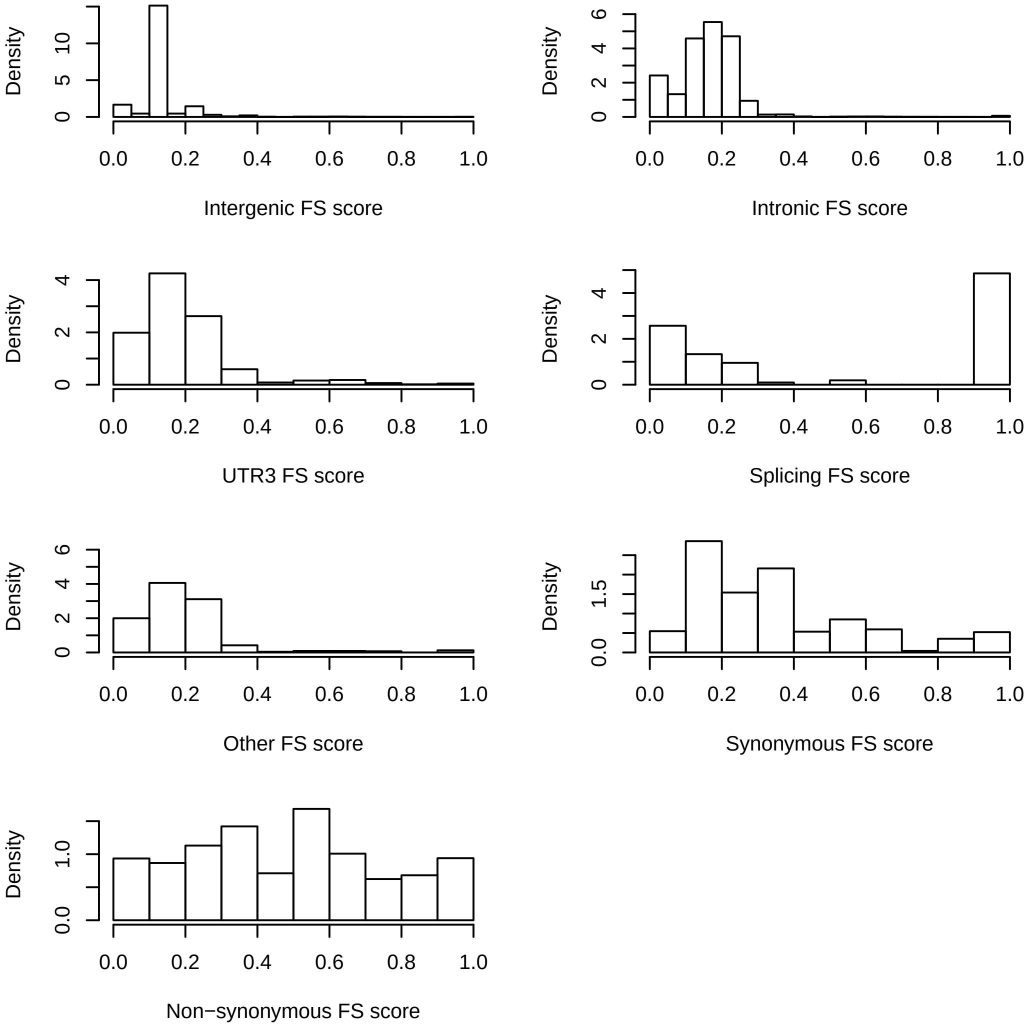


Figure 8.2: Histograms for the 7 functional information groups we are using. The histograms plot the raw FS score values from Lee and Shatkay [2009] excluding 0.5 as this value represents no knowledge of the deleterious effect or functional role of the SNP.

Using the histograms and the raw FS score values, we explore the data to decide which prior distributions to fit to the data. Discussing the biological effect of our mutation types with our expert, mutations in the splicing region either have an effect (which is highly deleterious) or they do not. Due to the shape of the histogram for splicing SNPs, see Figure 8.2 we see a noisier version of this binary effect. We therefore choose to fit a Bernoulli distribution in the proportions seen in the data for $FS < 0.5$ and $FS > 0.5$. This gives a Bernoulli($\frac{53}{105}$) prior distribution for splicing SNPs, such that $P(F_{splicing} = 1) = \frac{53}{105}$. All other distributions are fitted using continuous mixture distributions where appropriate, including point masses where these are consistent with the functional data.

The best fit distributions, selected for their ability to replicate the shape of the histograms, for our extended set of functional information priors can be seen below in Equations 8.1-8.7. The fitted densities have been plotted on top of the histograms in Figure 8.3.

$$\begin{aligned}
\pi(F_{\text{Intergenic}}) &\sim \mathbb{1}_{F\in[0,1]}\left\{0.789\delta_{[0.101866]} + 0.211Ga(1.296, 6.365)\right\} && (8.1)\\
\pi(F_{\text{Intronic}}) &\sim \mathbb{1}_{F\in[0,1]}\left\{0.121\delta_{[0]} + 0.879Ga\left(7.359, \frac{1}{0.0235}\right)\right\} && (8.2)\\
\pi(F_{\text{UTR3}}) &\sim \mathbb{1}_{F\in[0,1]}\left\{Ga(1.45, 8.08)\right\} && (8.3)\\
\pi(F_{\text{Splicing}}) &\sim \text{Bernoulli}\left(\frac{53}{105}\right) && (8.4)\\
\pi(F_{\text{Other}}) &\sim \mathbb{1}_{F\in[0,1]}\left\{0.085\delta_{[0]} + 0.915Ga\left(4.349, \frac{1}{0.0340}\right)\right\} && (8.5)\\
\pi(F_{\text{Syn}}) &\sim \mathbb{1}_{F\in[0,1]}\left\{0.946Ga\left(2.929, \frac{1}{0.113}\right) + 0.054Ga\left(640.5, \frac{1}{0.0015}\right)\right\} && (8.6)\\
\pi(F_{\text{Non-syn}}) &\sim \text{Uniform}[0,1]. && (8.7)
\end{aligned}$$

There is small mass of approximately 8% outside the $[0, 1]$ range for F_{syn} which we truncate using the indicator function. This means that the prior, in its current form, is not a proper prior as it does not integrate to 1, and the value of the density based on the non-truncated Gamma will be too small. To make this integrate to 1, we rescale the prior for F_{syn} by multiplying it by $\frac{1}{0.92} = 1.09$ so that it integrates to 1 over $[0, 1]$. Hence the prior for F_{syn} is defined as follows:

$$1.09 \times \mathbb{1}_{F\in[0,1]}\left\{0.946Ga\left(2.929, \frac{1}{0.113}\right) + 0.054Ga\left(640.5, \frac{1}{0.0015}\right)\right\}. \quad (8.8)$$

For all other densities, the mass outside $[0, 1]$ which we truncate with the indicator function is negligible.

Using the prior distributions in Equations 8.1-8.8 and the following transformation for B,

$$B = \tan\left(F\frac{\pi}{2}\epsilon\right), \quad (8.9)$$

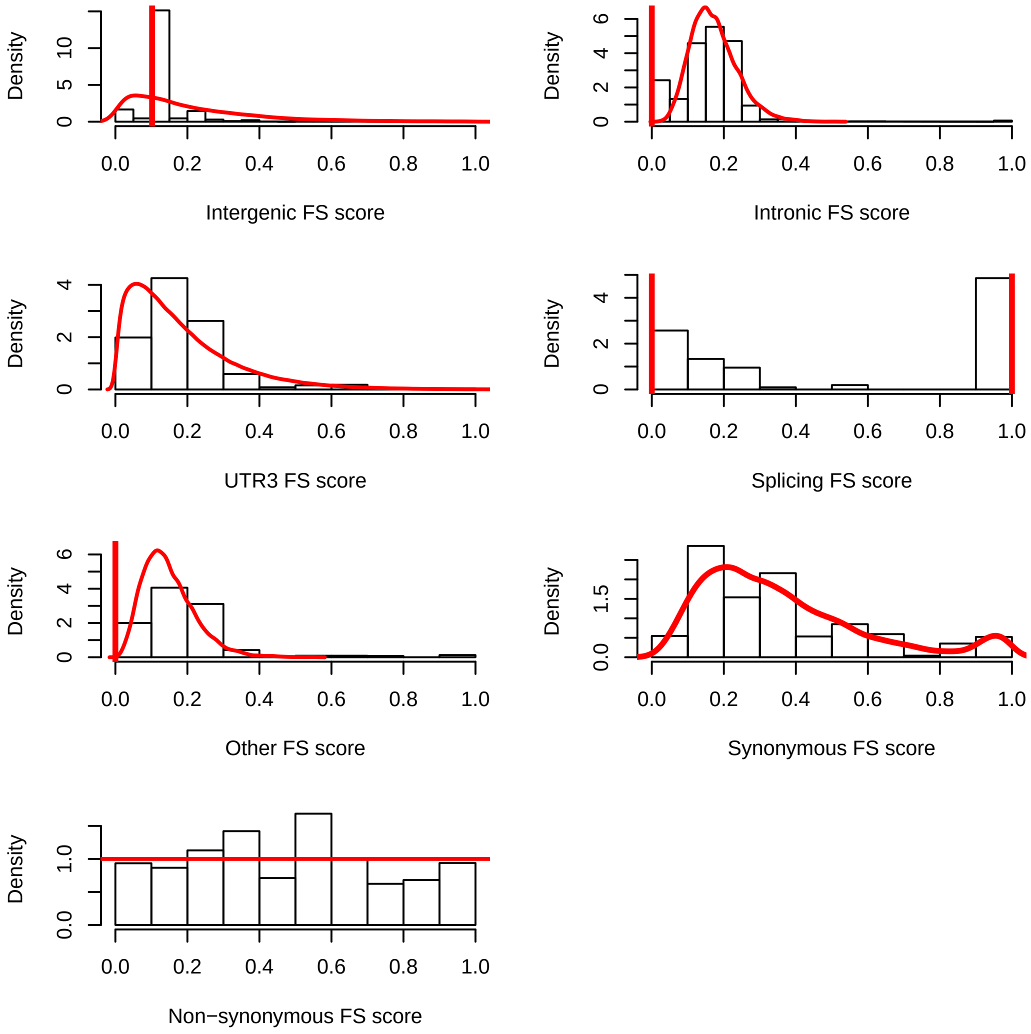


Figure 8.3: An array showing the histograms for the 7 functional information groups we are using with the red lines representing the prior distributions we are placing on each group. The histograms plot the raw FS score values from Lee and Shatkay [2009] excluding 0.5 as this value represents no knowledge.

with $\epsilon = 0.99$ to avoid $B \to \infty$ as $F \to 1$, we calculate the full conditional distributions for updating F for the new SNP groups. Given that many of the prior distributions are mixture distributions consisting of one continuous density and one point mass, we use the technique described in Gottardo and Raftery [2004] to update our F parameters. In all cases we have to use Metropolis Hastings updating as the $B \to F$ transformation prevents us from obtaining the form of a standard distribution for the full conditional distribution for F.

8.2 Transformation F to B

Our transformation from Chapter 7 defined in Equation 7.7 as

$$B = \tan\left(\frac{FS\pi}{2}\epsilon\right), \tag{8.10}$$

needs to be edited slightly to prevent $B = 0$ which causes some computational issues based on the point masses in certain prior distributions. As such, we edit the transformation as follows:

$$B = \tan\left(\frac{FS\pi}{2}\epsilon\right) + (1-\epsilon), \tag{8.11}$$

where $\epsilon = 0.99$ as before.

We do not need to include the Jacobian of the transformation when we calculate the full conditional distribution as both ur prior distribution and our full conditional distribution are calculated in terms of F. We use B as shorthand notation for the transformation applied to a single value from the distribution of F in the prior distribution (and subsequent full conditional distributions) for γ^{-2}.

8.3 Full conditional distributions

When calculating the full conditional distributions for $F|\lambda, \gamma^2$ we only need to include the prior distributions that contain either an F or B term. This means we only include the prior distribution for F for the SNP group we are updating, and the prior distribution for $\pi(\gamma^{-2}|\lambda, B)$. The probability density function (pdf) for $\pi(\gamma^{-2}|\lambda, B)$ is given by

$$\begin{aligned}\pi(\gamma^{-2}|\lambda, B) &\sim Ga\left(2, \frac{B}{2\lambda}\right)\\ &= \frac{\left(\frac{B}{2\lambda}\right)^2}{\Gamma(2)}\left(\gamma^{-2}\right)^{2-1} exp\left(\frac{-B}{2\lambda}\gamma^{-2}\right)\\ &= \frac{1}{4\lambda^2\gamma^2}B^2\text{exp}\left(\frac{-B}{2\lambda\gamma^2}\right)\end{aligned}$$

Substituting $B = \tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)$, Equation 8.9, gives:

$$\pi(\gamma^{-2}|\lambda, F) = \frac{1}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right) - (1-\epsilon)}{2\lambda\gamma^2}\right). \tag{8.12}$$

All full conditional distributions are calculated using:

$$f(F|\lambda, \gamma^{-2}) \propto \pi(F)\pi(\gamma^{-2}|\lambda, F) \tag{8.13}$$

and acceptance probabilities are calculated from:

$$\min\left\{1, \frac{f(F'|\lambda, \gamma^{-2})\pi(F')}{f(F|\lambda, \gamma^{-2})\pi(F)}\frac{q(F', F)}{q(F, F')}\right\}, \tag{8.14}$$

where $f(F'|\lambda, \gamma^{-2})$ is the full conditional distribution for F', $\pi(F')$ is the prior distribution on F' and $q(F', F)$ is the density transition kernal from the current value F to the proposed value F'.

8.3.1 Proposal value for F

In all cases where a proposal value is required, we either propose the value of the point mass, F^*, or we propose a value of $F \in [0, 1]\backslash\{F^*\}$ as detailed below:

- Simulate β as $\beta = \sigma^2 z$ where $z \sim N(0, 1)$ and σ^2 is calibrated to allow us to move around the sample space efficiently, preventing too many rejections of proposed moves, but with enough large jumps to explore the sample space (we initially use $\sigma^2 = 1$ and tune it).

- Given the current value of F, propose $F' = F + (1-F)\Phi(\beta)$ if $z > 0$ and $F' = F - F(1 - \Phi(\beta))$ if $z < 0$ where Φ is the standardized normal distribution function.

8.3.2 Updating the single continuous priors

In this section we state the updating for F_{syn}, F_{non} and F_{UTR}. These are grouped as they are updated using the standard Metropolis Hastings updating.

We propose $F' = F + (1 - F)\Phi(\beta)$ if $z > 0$ and $F' = F - F(1 - \Phi(\beta))$ if $z < 0$ as defined in Section 8.3.1. As the proposal values are not symmetric, we define the density transition kernel as follows:

$$q(F, F') = \phi\left(\frac{F' - F}{1 - F}\right). \tag{8.15}$$

This is the same for F_{syn}, F_{non} and F_{UTR}.

We state the acceptance probabilities for F_{syn}, F_{non} and F_{UTR3} as follows. Let $\tan\left(F'\frac{\pi}{2}\epsilon\right) + (1 - \epsilon) = T'$, $\tan\left(F\frac{\pi}{2}\epsilon\right) + (1 - \epsilon) = T$ and $Q = \frac{\phi\left(\frac{F'-F}{1-F}\right)}{\phi\left(\frac{F-F'}{1-F'}\right)},.$

$$AP_{F_{syn}} =$$
$$\min\left\{1, Q\frac{AF'^{1.929}\left(T'\right)^2\exp\left(-\frac{F'}{0.113} - \frac{1}{2\lambda\gamma^2}\left(T'\right)\right) + CF'^{639.5}\left(T'\right)^2\exp\left(-\frac{F}{0.0015} - \frac{1}{2\lambda\gamma^2}\left(T'\right)\right)}{AF^{1.929}\left(T\right)^2\exp\left(-\frac{F}{0.113} - \frac{1}{2\lambda\gamma^2}\left(T\right)\right) + CF^{639.5}\left(T\right)^2\exp\left(-\frac{F}{0.0015} - \frac{1}{2\lambda\gamma^2}\left(T\right)\right)}\right\}, \tag{8.16}$$

where $A = 0.946 \times \frac{\frac{1}{0.113}^{2.929}}{\Gamma(2.929)}$ and $C = 0.054 \times \frac{\frac{1}{0.0015}^{640.5}}{\Gamma(640.5)}$.

$$AP_{F_{non}} = \min\left\{1, Q\left(\frac{T'}{T}\right)^2\exp\left(-\frac{1}{2\lambda\gamma^2}\left[T' - T\right]\right)\right\}. \tag{8.17}$$

$$AP_{F_{UTR3}} = \min\left\{1, Q\left(\frac{F'}{F}\right)^{0.45}\left(\frac{T'}{T}\right)^2\exp\left(-8.08\left(F' - F\right)\right)\exp\left(-\frac{1}{2\lambda\gamma^2}\left[T' - T\right]\right)\right\}. \tag{8.18}$$

8.3.3 Updating one continuous and one point mass prior

In this section we state the acceptance probabilities for the three prior distributions $F_{intronic}$, $F_{intergenic}$ and F_{other} whose prior distributions are a mixture of one point mass and one continuous prior distribution.

To calculate the the full conditional distributions in these cases, we use the technique described in Gottardo and Raftery [2004]. This requires that, for the two measures ν_1 and ν_2, there exists a measurable set A such that $\nu_1(A) = 0$ and $\nu_2(A^C) = 0$, where A^C defines the complement of the set A, i.e. the two supports do not overlap.

The Gottardo and Raftery [2004] technique can be summarised as follows for cases where there is a mixture of one continuous distribution and one point mass, i.e. of the form $\pi(D) \sim (1-w)\delta_{D^*}+wg(D)$, where D^* is the value of D for which there exists a point mass and w is the mixture proportion. Formally, we have to exclude the value of the point mass from the support of the continuous distribution to ensure that $\nu_1(A) = 0$ and $\nu_2(A^C) = 0$. Hence the prior must be of the form $\pi(D) \sim (1-w)\delta_{D^*} + wg(D)\mathbb{1}_{\mathbb{R}\setminus\{D^*\}}(D)$.

1. Let part 1 represent the point mass ($(1-w)\delta_{D*}$) and part 2 represent the continuous density ($wg(D)$).

2. Calculate the component-wise full conditional distributions for $F|\lambda, \gamma$ for part 1 and part 2, which depend on the mixing proportion (w for part 2, $1-w$ for part 1). We cannot calculate the normalising constant in this case.

3. As we cannot calculate the normalising constant due to our $F \to B$ transformation, we sample a random uniform value u to define which part we sample our proposal value from. If $u > p_1$ we propose the value of the point mass, if $u \leq p_1$ we sample a proposal value from our proposal distribution defined in Section 8.3.1.

4. Calculate the values of the proposal density transition kernel for the four possible scenarios of $q(F|F')$.

5. Calculate the acceptance probabilities for each scenario of starting and ending in either part 1 or part 2 of the mixture prior.

The proposal distribution can be summarised for $F_{intronic}$, $F_{intergenic}$ and F_{other} in terms of moving in and out of the point mass and the continuous parts of the mixture prior. We propose the value of the point mass with probability p_1. We propose $F' = F + (1-F)\Phi(\beta)$ if $z > 0$ and $F' = F - F(1-\Phi(\beta))$ if $z < 0$ as defined in Section 8.3.1 with probability $1-p_1$.

The proposal density transition kernel is defined as follows, where F^* is the value of the point mass:

$$\begin{aligned} q(F, F' = F^*) &= p_1 \\ q(F, F' \in \mathbb{1}_{[0,1]\setminus\{F^*\}}) &= (1-p_1)\phi\left(\frac{F'-F}{1-F}\right) \end{aligned} \tag{8.19}$$

Updating $F_{intronic}$

We calculate part 1 of the full conditional distribution for the point mass as follows:

$$\begin{aligned} f(F_{\text{part 1}}|\lambda,\gamma) =& \frac{0.121}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right) \\ =& \frac{0.121}{4\lambda^2\gamma^2}T^2\text{exp}\left(\frac{-T}{2\lambda\gamma^2}\right), \end{aligned} \tag{8.20}$$

where $T = \tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)$.

We now calculate part 2 of the component-wise full conditional distribution for the Gamma part of the mixture distribution prior.

$$\begin{aligned} f(F_{\text{part 2}}|\lambda,\gamma) =& \frac{0.879}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right) \\ & \times\frac{1}{\Gamma(7.359)\times 0.0235^{7.359}}F^{7.359-1}\text{exp}\left(-\frac{F}{0.0235}\right) \\ =& \frac{0.879}{0.0235^{7.359}4\lambda^2\gamma^2\Gamma(7.359)}T^2F^{7.359-1}\text{exp}\left(\frac{-T}{2\lambda\gamma^2}-\frac{F}{0.0235}\right). \end{aligned} \tag{8.21}$$

We now state the acceptance probabilities. Let $\tan\left(F'\frac{\pi}{2}\epsilon\right) + (1-\epsilon) = T'$, $\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon) = T$ and $\dfrac{\phi\left(\dfrac{F'-F}{1-F}\right)}{\phi\left(\dfrac{F-F'}{1-F'}\right)} = Q$.

$$AP = \begin{cases} 1 & \text{if } F = F' = 0 \\ \min\left\{1, \dfrac{p_1 \times 0.879F'^{6.359}\left(T'\right)^2\text{exp}\left(\dfrac{-T'}{2\lambda\gamma^2}\right)\text{exp}\left(-\dfrac{F'}{0.0235}\right)}{(1-p_1)\phi\left(\dfrac{F'-F}{1-F}\right)\times 0.121T^2\text{exp}\left(\dfrac{-T}{2\lambda\gamma^2}\right)}\right\} & \text{if } F = 0 \text{ and } F' \neq 0 \\ \min\left\{1, \dfrac{(1-p_1)\phi\left(\dfrac{F'-F}{1-F}\right)\times 0.121\left(T'\right)^2\text{exp}\left(\dfrac{-T'}{2\lambda\gamma^2}\right)}{p_1\times 0.879F^{6.359}T^2\text{exp}\left(\dfrac{-T}{2\lambda\gamma^2}\right)\text{exp}\left(-\frac{F}{0.0235}\right)}\right\} & \text{if } F \neq 0 \text{ and } F' = 0 \\ \min\left\{1, Q\dfrac{F'^{6.359}\left(T'\right)^2\text{exp}\left(\dfrac{-T'}{2\lambda\gamma^2}\right)\text{exp}\left(-\frac{F'}{0.0235}\right)}{F^{6.359}T^2\text{exp}\left(\dfrac{-T}{2\lambda\gamma^2}\right)\text{exp}\left(-\frac{F}{0.0235}\right)}\right\} & \text{if } F \neq 0 \text{ and } F' \neq 0. \end{cases} \tag{8.22}$$

Updating $F_{intergenic}$

We calculate part 1 of the component-wise full conditional distribution for the point mass as follows, where $F = F^*$:

$$\begin{aligned} f(F_{\text{part 1}}|\lambda,\gamma) =& \frac{0.789}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right) - (1-\epsilon)}{2\lambda\gamma^2}\right) \\ =& \frac{0.789}{4\lambda^2\gamma^2}T^2\exp\left(\frac{-T}{2\lambda\gamma^2}\right). \end{aligned} \tag{8.23}$$

We now calculate part 2 of the component-wise full conditional distribution for the Gamma part of the mixture distribution prior.

$$\begin{aligned} f(F_{\text{part 2}}|\lambda,\gamma) =& \frac{0.211}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right) - (1-\epsilon)}{2\lambda\gamma^2}\right) \\ & \times \frac{6.365^{1.296}}{\Gamma(1.296)}F^{1.296-1}\exp(-6.365F) \\ =& \frac{0.211 \times 6.365^{1.296}}{4\lambda^2\gamma^2\Gamma(1.296)}\left(\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)\right)^2 F^{0.296} \\ & \times \exp\left(-6.365F - \frac{1}{2\lambda\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right) + (1-\epsilon)\right)\right) \\ =& \frac{0.211 \times 6.365^{1.296}}{4\lambda^2\gamma^2\Gamma(1.296)}T^2F^{0.296}\exp\left(-6.365F - \frac{T}{2\lambda\gamma^2}\right). \end{aligned} \tag{8.24}$$

We now state the acceptance probabilities, where the point mass is $F^* = 0.101866$.

$$AP = \begin{cases} 1 & \text{if } F = F' = F^* \\ \min\left\{1, \dfrac{p_1\frac{0.211\times 6.365^{1.296}}{\Gamma(1.296)}\left(T'\right)^2 F'^{0.296}\exp\left(-6.365F' - \frac{1}{2\lambda\gamma^2}T'\right)}{(1-p_1)\phi\left(\dfrac{F'-F}{1-F}\right)0.789T^2\exp\left(\dfrac{-T}{2\lambda\gamma^2}\right)}\right\} & \text{if } F = F^* \text{ and } F' \neq F^* \\ \min\left\{1, \dfrac{(1-p_1)\phi\left(\frac{F'-F}{1-F}\right)0.789\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}\right)}{p_1\frac{0.211\times 6.365^{1.296}}{\Gamma(1.296)}T^2F^{0.296}\exp\left(-6.365F - \frac{T}{2\lambda\gamma^2}\right)}\right\} & \text{if } F \neq F^* \text{ and } F' = F^* \\ \min\left\{1, Q\dfrac{\left(T'\right)^2 F'^{0.296}\exp\left(-6.365F' - \frac{1}{2\lambda\gamma^2}\left(T'\right)\right)}{\left(T\right)^2 F^{0.296}\exp\left(-6.365F - \frac{T}{2\lambda\gamma^2}\right)}\right\} & \text{if } F \neq F^* \text{ and } F' \neq F^*. \end{cases} \tag{8.25}$$

Updating F_{other}

We calculate part 1 of the full conditional distribution for the point mass as follows:

$$\begin{aligned} f(F_{\text{part 1}}|\lambda,\gamma) =& \frac{0.085}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right)\\ =& \frac{0.085}{4\lambda^2\gamma^2}T^2\exp\left(\frac{-T}{2\lambda\gamma^2}\right). \end{aligned} \tag{8.26}$$

We now calculate part 2 of the component-wise full conditional distribution for the Gamma part of the mixture distribution prior.

$$\begin{aligned} f(F_{\text{part 2}}|\lambda,\gamma,B) =& \frac{0.915}{4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right)\\ &\times \frac{1}{\Gamma(4.349)\times 0.0340^{4.349}}F^{4.349-1}\exp\left(-\frac{F}{0.0340}\right)\\ =& \frac{0.915}{0.0340^{4.349}\times 4\lambda^2\gamma^2\Gamma(4.349)}T^2F^{3.349}\exp\left(\frac{-T}{2\lambda\gamma^2}-\frac{F}{0.0340}\right). \end{aligned} \tag{8.27}$$

We now state the acceptance probabilities.

$$AP = \begin{cases} 1 & \text{if } F=F'=0\\ \min\left\{1, \dfrac{\dfrac{0.915p_1}{0.0340^{4.349}\Gamma(4.349)}F'^{3.349}\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}-\dfrac{F'}{0.0340}\right)}{(1-p_1)\phi\left(\dfrac{F-F'}{1-F'}\right)\times 0.085\left(T\right)^2\exp\left(\dfrac{-T}{2\lambda\gamma^2}\right)}\right\} & \text{if } F=0 \text{ and } F'\neq 0\\ \min\left\{1, \dfrac{(1-p_1)\phi\left(\dfrac{F'-F}{1-F}\right)\times 0.085\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}\right)}{\dfrac{0.915p_1}{0.0340^{4.349}\Gamma(4.349)}F^{3.349}T^2\exp\left(\dfrac{-T}{2\lambda\gamma^2}-\dfrac{F}{0.0340}\right)}\right\} & \text{if } F\neq 0 \text{ and } F'=0\\ \min\left\{1, Q\dfrac{F'^{3.349}\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}-\dfrac{F'}{0.0340}\right)}{F^{3.349}T^2\exp\left(\dfrac{-T}{2\lambda\gamma^2}-\dfrac{F}{0.0340}\right)}\right\} & \text{if } F\neq 0 \text{ and } F'\neq 0. \end{cases} \tag{8.28}$$

8.3.4 Updating two point mass priors

In this section we assess the updating for $F_{splicing}$. This is a Bernoulli prior which we treat as point masses at 0 and 1. In this case we use the technique described in Gottardo and Raftery [2004] but treat our sample space to be only $\{0,1\}$.

We calculate part 1 of the component-wise full conditional distribution as the component relating to $\delta_{[0]}$.

$$\begin{aligned} f(F_{\mathrm{F}=0}|\lambda,\gamma) =& \frac{52}{105\times 4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right) \\ =& \frac{52}{105\times 4\lambda^2\gamma^2}\left(T\right)^2 \exp\left(\frac{-T}{2\lambda\gamma^2}\right). \end{aligned} \tag{8.29}$$

We now calculate part 2 of the component-wise full conditional distribution as the component relating to $\delta_{[1]}$.

$$\begin{aligned} f(F_{\mathrm{F}=1}|\lambda,\gamma) =& \frac{53}{105\times 4\lambda^2\gamma^2}\left(\tan\left(F\frac{\pi}{2}\epsilon\right)+(1-\epsilon)\right)^2 \exp\left(\frac{-\tan\left(F\frac{\pi}{2}\epsilon\right)-(1-\epsilon)}{2\lambda\gamma^2}\right) \\ =& \frac{53}{105\times 4\lambda^2\gamma^2}T^2\mathrm{exp}\left(\frac{-T}{2\lambda\gamma^2}\right). \end{aligned} \tag{8.30}$$

We now state the acceptance probabilities.

$$AP = \begin{cases} 1 & \text{if } F = F' = 0 \\ \min\left\{1, \dfrac{53\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}\right)}{52T^2\mathrm{exp}\left(\dfrac{-T}{2\lambda\gamma^2}\right)}\times\dfrac{p_1}{(1-p_1)}\right\} & \text{if } F = 0 \text{ and } F' = 1 \\ \min\left\{1, \dfrac{52\left(T'\right)^2\exp\left(\dfrac{-T'}{2\lambda\gamma^2}\right)}{53T^2\mathrm{exp}\left(\dfrac{-T}{2\lambda\gamma^2}\right)}\times\dfrac{1-p_1}{p_1}\right\} & \text{if } F = 1 \text{ and } F' = 0 \\ 1 & \text{if } F = F' = 1. \end{cases} \tag{8.31}$$

8.3.5 Parameter choice

We initially specified $\sigma^2 = 1$ in our proposal value. This value works well in our simulations and so we fix it to be 1 in all further work. We also initially specified $p_1 = 0.5$. Again this works well in the simulations so we remain using this value. We maintain these values as they provide an acceptance rate similar to the acceptance rate of $20\% - 30\%$ for λ'.

8.4 Conclusion

In this chapter we have created extra groups within the Normal Gamma to cover SNPs in all the different regions of the genome that our expert believes are important for our type of model (SNPs affecting gene expression). Calculating the full conditional distributions and the acceptance probabilities has allowed us to create a function that we can test. In the next chapter we will carry out simulation studies to show that the function is performing as expected, and we will show how this affects the posterior results. We will then test the function on a subset of the Hulse and Fairfax datasets in Chapter 10.

Chapter 9

Simulation results of the Normal Gamma prior with seven functional information groups

In this chapter we use HapGen simulated dataset 2A to test the NG super function described in Chapter 8. Recall that HapGen dataset 2A is simulated using HapGen2 [Su et al., 2011] with 6 causal SNPs in each of the 9 sub-datasets within dataset 2A. In this case we only use 5 of the 9 sub-datasets. The causal effect size is 0.4 for all 6 causal SNPs with a population MAF 0.2. Each of the 5 datasets has a total of 631 SNPs and 300 individuals. This means that in our analysis via ROC curves, we have $5 \times 6 = 30$ causal SNPs and $5 \times (631 - 6) = 3125$ non-causal SNPs. We maintain the same error structure of the model. More details of the dataset can be found in Section 4.2.1.

To test the Normal Gamma super function from Chapter 8, we split the 631 SNPs into the 7 different functional information groups. We define the proportion of SNPs in each category based on up-weighting and down-weighting the percentages from the FS score data and the Fairfax data. We define the minimim percentage in a group to be 1% to ensure that there are enough SNPs in each group when running the NG super function. We do not use the Hulse data as we selected only exonic SNPs (excluding the causal SNPs which are not exonic in all cases). We split the data such that the percentages are as stated in Table 9.1.

We test the NG super function in two cases. The best case scenario is when all 6 causal SNPs are together with no other SNPs in a group where the FS score is a priori high, therefore enforcing less shrinkage. For this case, which we call splicing causal, we allocated all 6 causal SNPs in the splicing category alone. All other SNPs are allocated to the remaining SNP groups as shown

	Fairfax	FS Score	NG super function	Splicing causal	UTR causal
Synonymous	0.82%	2.5%	7% (44)	44	44
Non-synonymous	0.327%	3.44%	3% (19)	19	19
Splicing	0.003%	0.06%	1% (7)	6	7
Intergenic	55.7%	7.5%	40% (252)	253	252
Intronic	40.7%	81.6%	40% (252)	252	252
UTR3	0.99%	4.2%	5% (32)	32	6 + 25
Other	1.46%	0.7%	4% (25)	25	25

Table 9.1: The percentages of each type of SNP found in the Fairfax data and in the SNPs used in the FS score database. These have been assessed and used to inform the estimates of the percentages of each type of SNP used in the NG super function. In many cases we increase the percentage for groups with small numbers of SNPs in and decrease the percentage for intronic and intergenic SNPs to ensure there are sufficient SNPs in each group to apply the NG splitting model successfully. The number in brackets is the number of each of the 631 SNPs that will be allocated to each category of SNPs. The numbers in red represent the causal SNPs within the two NG splitting scenarios (splicing causal and UTR causal).

in Table 9.1. We define this as the best case because the FS score prior mean for splicing is 0.545 which is the highest prior mean (the second highest is for non-synonymous SNPs at 0.48) and we believe that having only causal SNPs in one group will increase detection as the group will receive less shrinkage overall than if the SNPs were in a group with other, non-causal SNPs.

The second scenario is the worst case scenario when the causal SNPs are in a group with many other non-causal SNPs, and with a priori much larger shrinkage. We call this case utr causal. We allocate the SNPs as in Table 9.1 with the 6 causal SNPs (highlighted in red) in the UTR3 group. We define this as the worst case scenario for detecting causal SNPs because UTR3 has prior mean 0.1797 which enforces extra shrinkage on these SNPs, and there is a mixture of causal and non-causal SNPs. As the shrinkage on all SNPs in the group is the same, extra shrinkage will be enforced equally on both causal and non-causal SNPs. When only the causal SNPs are in a functional information group, we believe that the shrinkage will be less due to the information in the likelihood regarding the causality of the SNPs.

Given there are many other scenarios we could have chosen, we choose to only investigate these two as the best and worst case, but to vary n, the number of individuals. We will begin by comparing the standard NG, the NG splicing causal and the NG utr causal; we will then reduce n to 100 and 50 by choosing 100 and 50 individuals at random from the 300 in dataset 2A. We will compare the three scenarios for each n to assess the weight given to the prior in the best and worst case scenarios. We will then compare the NG splicing causal and NG utr causal across the three n to assess the effect of n. Will will also assess which, if any SNPs have a 90% credible interval that does not contain 0. We can use this for formal statistical association testing. We have not reported this previously because the results with respect to the ranking and detection of causal SNPs were poor.

9.1 Results

9.1.1 Comparing the NG super function across methods with the same n

As before, we compare the results from the NG super function when we change the group that the causal SNPs are allocated to, using a ROC curve. The results of comparing the NG and NG super function across different n are given in Figure 9.1. We notice that the NG causal splicing is detecting the most causal SNPs at the lowest false positive rate (FPR), until a FPR of approximately 0.5. The NG causal UTR is worse than NG causal splicing except in the case where the FPR > 0.6 in the $n = 50$ case. These results are as expected in that when the causal SNPs are alone in a group with a priori less shrinkage, the posterior mean effect size estimates rank the causal SNPs higher than when the causal SNPs are in a group with non-causal SNPs (26 non-causal and 6 causal in the UTR group) which has a priori more shrinkage on the effect size estimates.

To assess the comparison of methods we use the AUC of the ROC. These are shown in Table 9.2. The NG causal splicing has the largest AUC within each n (row). For $n = 100$ the NG has the lowest AUC, but for the other two n, the AUC of NG causal UTR is lowest.

9.1.2 Comparing the NG super function as sample size varies

Here we assess the effect of n on the posterior effect size ranks from the NG causal splicing, NG and NG causal UTR, see Figure 9.2 and Table 9.2. We expect that as the number of individuals n decreases, the AUC will also decrease as there is less information in the likelihood. This is reflected in the NG and NG causal UTR but not in the NG causal splicing. In this case the $n = 100$ ROC curve has greater AUC than the $n = 300$ ROC curve. The difference between $n = 300$ and $n = 100$ for NG causal splicing may simply be due to MCMC variation.

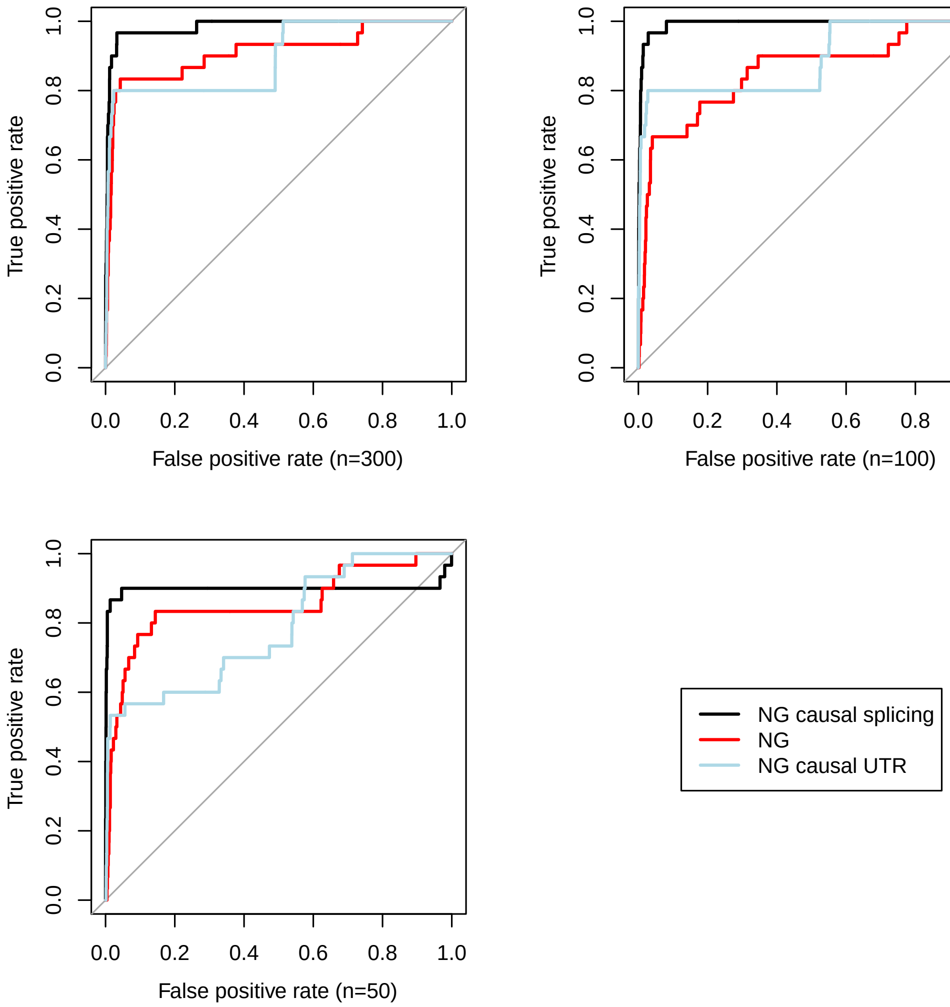


Figure 9.1: ROCs comparing the posterior mean rank of the effect size for 30 causal SNPs for the NG causal splicing, NG and NG causal UTR scenarios with $n = 300$, $n = 100$ and $n = 50$. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets of dataset 2A for these results.

The results in Figure 9.2 show that when the causal SNPs and other non-causal SNPs are allocated to a functional information group with a priori much greater shrinkage, the detection of these causal SNPs is poor. We may conclude that for the NG super function, the difference between $n = 100$ and $n = 300$

	NG	NG causal splicing	NG causal UTR
$n = 300$	0.9103	0.9848	0.8945
$n = 100$	0.8544	0.9931	0.8873
$n = 50$	0.8526	0.8985	0.7825

Table 9.2: The AUC of the ROC curves in Figures 9.1 and 9.2. The ROC curves compare the posterior mean rank of the effect size for 30 causal from the NG causal splicing, NG and NG causal UTR scenarios with $n = 300$, $n = 100$ and $n = 50$. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets of dataset 2A for these results.

is insignificant given the prior distributions and the proportion of causal SNPs based on these ROC curves. This effect is not see in the standard NG, an increase in n from 300 to 100 increases the AUC of the ROC curve.

In both the NG causal splicing and NG causal UTR cases for the $n = 50$ case, the AUC of the ROC is approximately 0.1 lower than the AUC for the ROC curves with $n = 100$ and $n = 300$. We hypothesize that with $n = 50$ it may be more difficult in the case of the NG super function to accurately estimate the posterior effect sizes with such small numbers of individuals and SNPs in each functional information group. The variance of the likelihood is affected by the minor allele frequency (MAF), and also by the number of individuals. Hence we need to ensure a balance between n and MAF to ensure causal SNPs are detected. It may be the case here that $n = 50$ is not large enough when the causal SNPs are given a priori less shrinkage, and are in functional information groups with many other, non-causal SNPs.

9.1.3 Assessing the gradient of the likelihood function

To understand the difference between the gradient of the likelihood with the change in minor allele frequency (MAF) from 0.2 to 0.02, we investigate the case where there is only one SNP. This situation is unrealistic but it allows a simple comparison between the likelihood in the case of both MAFs.

To calculate the variance of the least squares estimate we use $\text{var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2$. For simplicity we treat $\sigma^2 = 1$.

In the case where $MAF = 0.2$ with the design matrix X having two

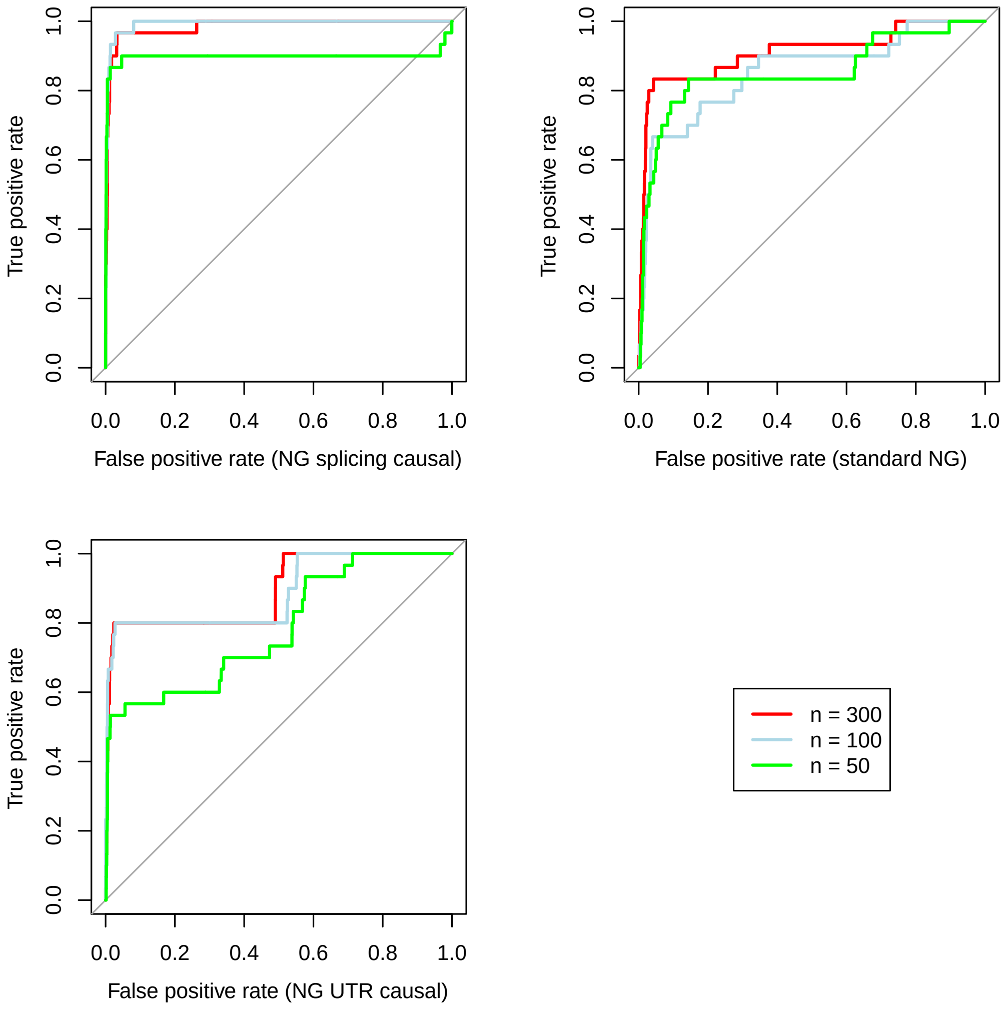


Figure 9.2: ROCs comparing the posterior mean rank of the effect size for 30 causal SNPs for the NG causal splicing, NG and NG causal UTR scenarios with $n = 300$, $n = 100$ and $n = 50$. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets of dataset 2A for these results.

columns, the first contains all 1's, the second contains 60 1's and the rest 0 - this represents MAF 0.2. Hence $X^TX = \begin{pmatrix} 300 & 60 \\ 60 & 60 \end{pmatrix} = 60 \begin{pmatrix} 5 & 1 \\ 1 & 1 \end{pmatrix}$, and

$$(X^TX)^{-1} = \frac{1}{60 \times (5-1)} \begin{pmatrix} 1 & -1 \\ -1 & 5 \end{pmatrix} \implies \text{var}(\hat{\beta}) = \frac{5}{60 \times 4} = \frac{1}{48}.$$

In the case where $MAF = 0.02$, the design matrix X has two columns, the first column contains all 1's, the second column contains 6 1's and the rest 0's - this represents MAF 0.02. Hence $X^TX = \begin{pmatrix} 300 & 6 \\ 6 & 6 \end{pmatrix} = 6\begin{pmatrix} 50 & 1 \\ 1 & 1 \end{pmatrix}$, and $(X^TX)^{-1} = \frac{1}{6 \times (50-1)}\begin{pmatrix} 1 & -1 \\ -1 & 50 \end{pmatrix} \implies \text{var}(\hat{\beta}) = \frac{50}{6 \times 49} = \frac{50}{294}$. We notice that $\text{var}(\hat{\beta}_{MAF=0.02}) > \text{var}(\hat{\beta}_{MAF=0.2})$. Hence we conclude that the gradient of the likelihood is much larger for rarer variants, meaning that the likelihood plane is steeper for more common variants. From this, we can understand why the Bayesian methods such as the Spike and slab and the Normal Gamma struggle to detect rare variants compared with more common variants.

9.1.4 Formal Statistical Association testing

In this section we state the SNPs whose 90% posterior credible interval does not contain 0. We have omitted this in previous chapters as many of the 90% posterior credible intervals for the causal SNPs contain 0. This makes this a poor method for detecting those causal SNPs. In Table 9.3 we state the number of SNPs whose 90% posterior credible interval does not contain 0, according to whether these SNPs are truly causal or not.

As with the standard Normal Gamma, the number of causal SNPs whose 90% posterior credible interval does not contain 0 is very poor, however the number of false positives is also low. In only 7 of the 30 sub-datasets do we detect one or more non-causal SNP with a 90% posterior credible interval that does not contain 0 (FPR = 0.064%). The TPR is 3.6%. If we were to define only these SNPs as associated, the true positive rate (TPR) is not high enough, although the FPR is sufficiently low. Increasing the width of the credible interval increases both the TPR and the FPR. Using only this method (based on the 90% credible interval) for summarising the NG super function results prevents detection of too many truly causal SNPs.

We consider alternative summary statistics to the mean for the NG super function, based on the credible intervals assessed above. We assess whether the AUC of the ROC curve increases when using either the median, the 5^{th} or 95^{th} percentiles compared to the mean.

Having assessed the ROC curves, we omit these from the thesis as the plots, in most cases, show the difference between using the mean, median and 95^{th} percentile as a summary statistic to be very similar. Instead of including the ROC curves, we state the AUC of each ROC in Table 9.4.

The results in Table 9.4 show that the only case where the AUC for the

n	Sub-dataset	Number of causal SNPs detected	Number of non-causal SNPs detected
Splicing causal			
300	4	1	2
	5	2	1
	6	2	2
100	5	1	1
	6	1	0
50	3	1	0
UTR causal			
300	4	0	1
	5	3	3
	6	2	2

Table 9.3: The number of SNPs (total 631, of which 6 are causal) whose 90% posterior credible interval does not contain 0. The sub-datasets that are omitted have 0 SNPs whose posterior credible interval does not contain 0. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets (numbered 2-6) of dataset 2A for these results.

posterior mean is not maximum (or within 0.01 of the maximum) is in the NG splicing causal $n = 50$ case. In the case of the NG UTR causal, the 95^{th} percentile performs poorly as a summary statistic, with AUC lower than for the posterior mean and 50^{th} percentile which are very similar. In the NG splicing causal case, there is very little difference between the AUC for the 50^{th} and 95^{th} percentiles and the posterior mean (except in the $n = 50$ case as stated). Given these results and previous results that led to using the posterior mean, there is no strong evidence here that another summary statistic would be more appropriate, therefore as the posterior mean gives large AUC of the ROC curves, see Table 9.2 and 9.4, we continue to use the posterior mean as our summary statistic.

9.1.5 Comparing posterior mean effect sizes

We now assess the values of the posterior mean effect sizes for all SNPs for NG splicing causal and NG UTR causal. We plot histograms, identifying the causal

	n = 300	**n = 100**	**n = 50**
NG splicing causal			
5^{th} percentile	0.6856	0.8234	0.9078
50^{th} percentile	0.986	0.9958	0.9045
95^{th} percentile	0.99	0.9926	0.9875
Posterior mean	0.9848	0.9931	0.8985
NG UTR causal			
5^{th} percentile	0.6701	0.5678	0.5902
50^{th} percentile	0.8927	0.8714	0.7336
95^{th} percentile	0.7971	0.7991	0.6109
Posterior mean	0.8945	0.8873	0.7825

Table 9.4: The AUC of the ROC curves for the NG super function, comparing the posterior mean summary statistic to three other percentiles, the 5^{th}, 50^{th} and 95^{th} percentiles, of the posterior distribution as the summary statistic. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing.

SNPs with ×'s on the axis in Figure 9.3.

We notice that the causal SNPs in the NG splicing causal case (dark green histograms in Figure 9.3) take larger values than for the NG UTR causal case (blue histograms). We also notice that the maximum posterior effect size for a causal SNP in the NG splicing causal case is much higher for $n = 300$ than either $n = 100$ or $n = 50$.

We now compare the cases where $n = 100$ and $n = 300$ for the NG UTR causal and NG splicing causal to the standard NG. The results can be seen in Figure 9.4.

Figure 9.4 shows how the causal SNPs and all SNPs in general, have a much smaller posterior mean effect size for the standard NG (red) compared to either of the NG splicing causal or NG UTR causal cases. In particular we notice that in the NG UTR causal case, there appears to be less shrinkage on the posterior mean effect size estimates compares to the standard NG. This, as hypothesized earlier, may be due to the more flexible prior structure we are placing on the prior variance of β, see Chapter 7 for details and examples. We also clearly see that there is less shrinkage applied to the causal SNPs in the NG splicing causal case compared to the NG UTR causal case. This implies that not only is the NG super function improving the detection of causal SNPs by increasing

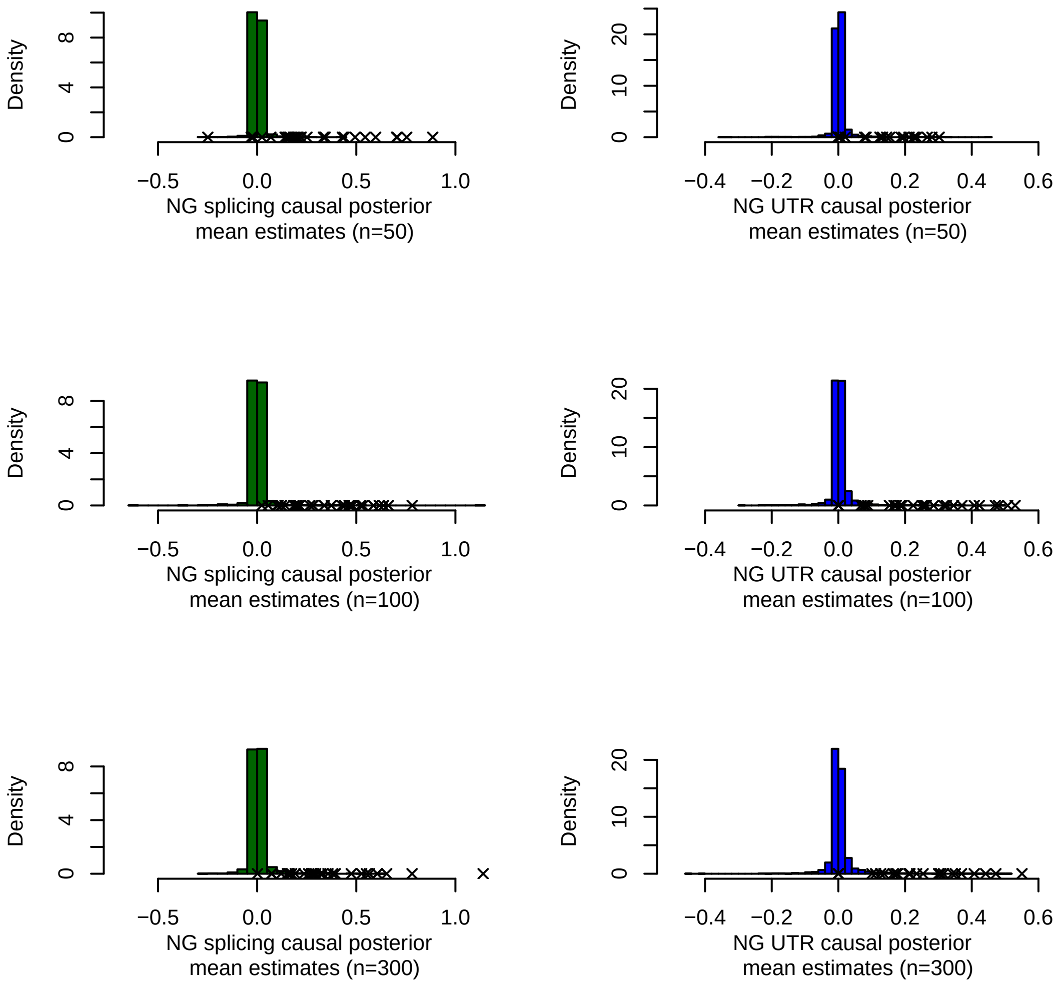


Figure 9.3: Histograms comparing the posterior mean effect sizes for 30 causal SNPs (marked with a ×) and the 3125 non-causal SNPs for the NG causal splicing (dark green) and the NG causal UTR (blue) scenarios with $n = 300$, $n = 100$ and $n = 50$. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets of dataset 2A for these results.

the flexibility of the prior, it is also differentially reducing shrinkage on SNPs with an a priori higher chance of being causal.

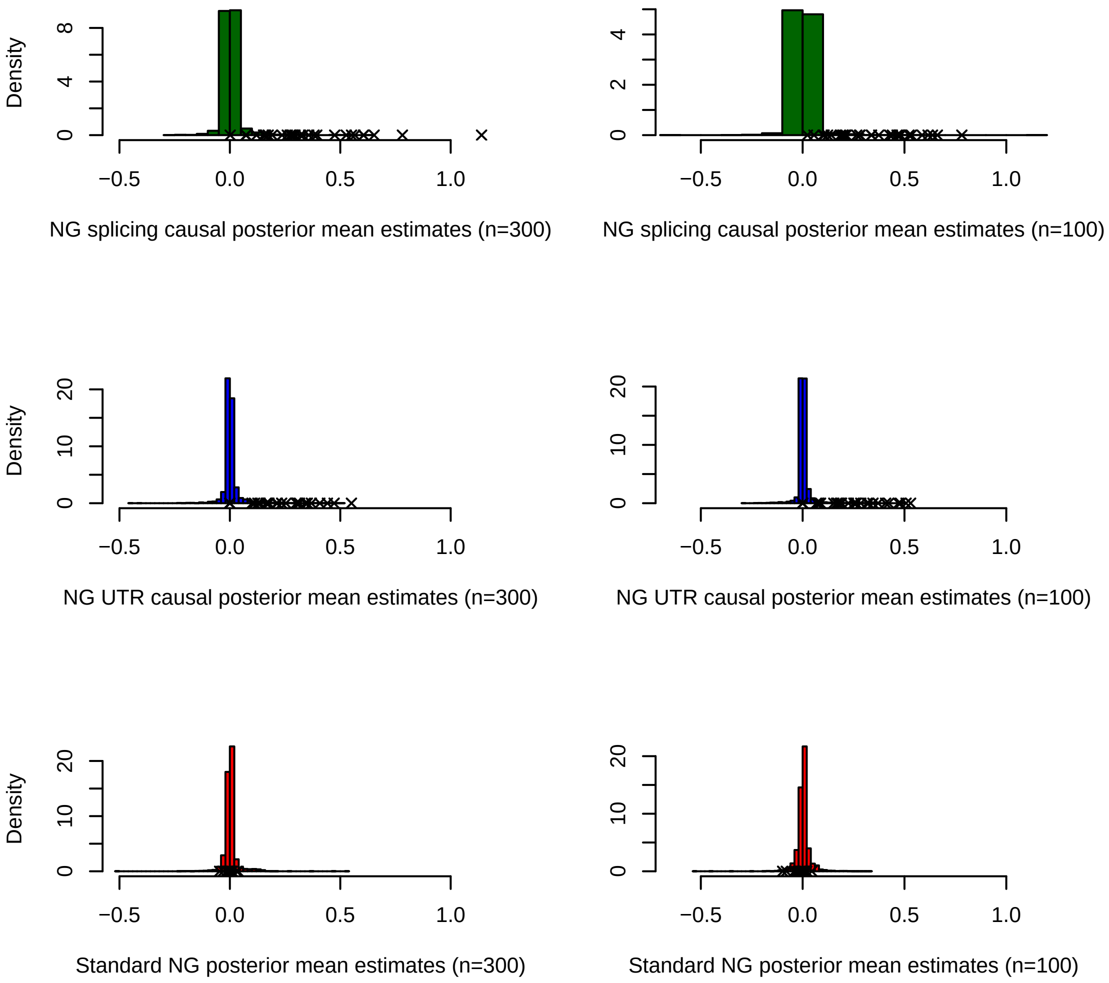


Figure 9.4: Histograms comparing the posterior mean effect sizes for 30 causal SNPs (marked with a ×) and the 3125 non-causal SNPs for the NG causal splicing (dark green), the NG causal UTR (blue) and the standard NG (red) scenarios with $n = 300$ and $n = 100$. The data is simulated using HapGen2 [Su et al., 2011] to include causal SNPs with a MAF (minor allele frequency) of approximately 0.2 in the population (dataset 2A). The NG causal splicing case is where all 6 causal SNPs are defined as splicing and all 625 non-causal SNPs are defined to be from the other 6 functional information groups that we have defined in proportions that resemble the population proportion of each type of SNP. The NG causal UTR case is similar to the NG causal splicing case but with all 6 causal and 25 non-causal SNPs defined as UTR rather than splicing. Splicing SNPs are a priori not shrunk as much as UTR SNPs. We use 5 sub-datasets of dataset 2A for these results.

9.2 Conclusion

In this chapter, we have assessed two cases of the NG super function - the first or best case is where the causal SNPs are in a group on their own with a priori less shrinkage (NG splicing causal), compared to the worst case scenario where the causal SNPs are in a group with other SNPs and have a priori greater shrinkage (NG UTR causal). As would be expected, the NG super function performs much better in the first case. In the worst case, the Normal Gamma still performs well when $n > 100$, based on an AUC of 0.8873 for the ROC curve.

The plots of the posterior mean effect sizes for both the NG splicing causal and the NG UTR causal cases shows that there is a difference in the shrinkage of the causal SNPs based on the prior distribution applied to the SNPs. This is what we expected from using the NG super function. Comparing to the standard Normal Gamma, we see that the flexibility of the NG super function hierarchy is improving the causal SNP detection as well.

Given the results in this chapter, which show that the NG super function clearly outperforms the standard NG when the SNPs are allocated to the splicing group, and gives larger posterior mean estimates for the causal SNPs in both the UTR causal and splicing causal cases, we will now assess the results of the NG super function on a subset of Fairfax and Hulse datasets.

Chapter 10

Application of the Normal Gamma super function

In this chapter we apply the NG super function to a subset of the Hulse and Fairfax data. Details on the NG super function can be found in Chapter 8 and details of the Hulse and Fairfax data can be found in Sections 2.1.3, page 15 and 2.1.4, page 18 respectively. We compare the results in this chapter with the results using the standard Normal Gamma function which can be found in Chapter 6. We do not use the Yeast data for the NG super function as the functional information groups and their prior distributions are based on human data/SNPs. We check for convergence using the R-hat statistic of Brooks and Gelman [1998] discussed in Section 4.3. Convergence is achieved with 50,000 iterations and a 5,000 iteration burn-in.

10.1 Hulse data

The Hulse data contains no validated causal or any associated SNPs, therefore it is very difficult to assess the ability of the NG to detect causal or associated SNPs. As such we will select only a subset of genes to run through the NG super function. We select CTNNA2, DAAM2 and IL6 to test the NG super function on as these represent a spread of maximum posterior effect size for the standard NG and a spread in the number of SNPs within each gene. IL6 contained the top ranked SNP for the NG that was also in the top 10 ranked SNPs for all other methods. We include this to assess where this SNP will be ranked using the NG super function. We will then compare the maximum posterior mean effect sizes for each gene in the standard NG and the NG super function. We will also assess how the ranks of the top 5 SNPs compare across the NG and the NG super function. These results are tabulated in Table 10.1. The SNPs are listed in order of their effect sizes (largest to smallest). We report the maximum effect size for the NG and NG super function to show how small the effect sizes are in the standard Normal Gamma compared to the NG super function.

Method	**Max Effect Size**	**Top 5 ranked SNPs**
CTNNA2 (2919 SNPs, 38 individuals)		
NG	1.0064	rs10779960 (1204), rs13416246 (2664), rs13409348 (205), rs7592817 (812), rs732260 (1570)
NG super function	6.665	rs1437353 (intergenic,1157), rs1427638 (intergenic,1531), rs960601 (intergenic,2800), rs993607 (intergenic,1513), rs6728409 (intergenic,349)
DAAM2 (149 SNPs, 39 individuals)		
NG	0.00007956	rs2504090 (24), rs9394630 (105), rs9380895 (101), rs2504100 (56), rs7750130 (84)
NG super function	0.4632	rs3004070 (UTR3,36), rs3004071 (UTR3,117), rs3793137 (UTR3,70), rs3003929 (syn,93), rs3004067 (syn,133)
IL6 (189 SNPs, 39 individuals)		
NG	0.007528	rs12700386 (88), rs17302823 (31), rs1476483 (128), rs2961310 (63), rs2905324 (137)
NG super function	0.03037	rs2069833 (intronic,45), rs2069832 (intronic,43), rs2066992 (intronic,69), rs1524107 (intronic,48), rs1474347 (intronic,66)

Table 10.1: A comparison of the top 5 ranked from the NG and NG super function on the selected genes from the Hulse dataset. The maximum posterior effect size is stated for the NG and the NG super function. We state in brackets, for the top 5 ranked SNPs in the NG, their rank in the NG super function and for the NG super function, we state the rank of the SNPs in the NG. For the NG super function, we state in brackets the functional group to which the top 5 ranked SNPs belongs.

In Table 10.1, we notice that the maximum effect sizes are all larger in the NG super function case compared to the standard NG case. This reflects the decrease on the shrinkage of at least some of the SNP groups with respect to the standard Normal Gamma. For CTNNA2, the effect sizes are much larger than in any other gene, and are much larger in the NG super function compared to the standard NG. We hypothesize that this is due to the much larger number of SNPs than in any other gene, giving greater uncertainty in the posterior estimates. None of the top 5 ranked NG SNPs are in the top 5 ranked SNPs for the NG super function for any of these three genes. The top 5 causal SNPs are not all from the same functional information group for DAAM2, and although the top 5 ranked SNPs for CTNNA2 and IL6 are all from intergenic or intronic functional groups respectively, these are not the groups with a priori less shrinkage. The one SNP, rs12700386 in IL6 that was top ranked by the NG and top 10 ranked by other methods in Table 6.7 is not highly ranked by the NG super function. In this case it has rank 88 with a tiny posterior mean effect size (1.6×10^{-8}).

We compare the mean rank of the SNPs in each the functional information groups across the three genes, CTNNA2, DAAM2 and IL6, in Table 10.2.

	Non-syn	**Syn**	**Intronic**	**Intergenic**	**UTR3**
Mean rank					
CTNNA2 (2919 SNPs)	NA	1499	2261.9	936.1	NA
DAAM2 (149 SNPs)	NA	14	55.9	121.2	2
IL6 (189 SNPs)	187	10	5	99.5	NA
Mean posterior mean effect size					
CTNNA2	NA	0.33	0.094	1.36	NA
DAAM2	NA	0.0066	0.0019	0.00015	0.29
IL6	6.7×10^{-11}	0.0059	0.021	1.3×10^{-8}	NA

Table 10.2: Mean rank and the mean of the posterior mean effect sizes for the three Hulse genes run on the NG super function. Splicing and 'other' functional information groups are omitted as no SNPs were in these groups for these three Hulse genes.

It is difficult to conclude based on the results in Table 10.2, how much of an effect the prior shrinkage is having on the posterior estimates of the SNPs as the two groups with a priori less shrinkage, splicing and non-synonymous, are not well used by the three selected Hulse genes. We observe that for CTNNA2 and DAAM2, the intergenic and intronic SNPs are reversed in which has the higher posterior mean rank and mean posterior mean effect size. This shows that the information in the likelihood is affecting the posterior ranking of these SNPs.

Again we conclude that it is difficult to draw any conclusions using this dataset as there are no validated causal SNPs. Comparing the results from the NG and the NG super function has highlighted clear differences between the top 5 ranked SNPs by posterior mean effect size, and assessing the mean rank of the SNPs in each functional information group has shown that the likelihood is affecting the mean ranks across the functional information groups.

10.2 Fairfax data

The Fairfax data was selected to only include SNPs in exonic regions on the same chromosome as the gene. This did not include all validated causal SNPs, and so for 5 of the 8 gene/gene expression source (bcell or monocytes) combinations, the causal SNP is not exonic, and is the only SNP in its category. Therefore we will only run the NG super function on one of these genes. We choose FADS1 bcell to represent the 5 genes where the causal SNP is not located in an exonic region. The three genes where the causal SNP is exonic are CARD9 mono, RBM6 mono and RBM6 bcell. We will run all three through the NG super function based on using only the synonymous, non-synonymous and 'other' categories of the function as the SNPs should all be either synonymous or non-synonymous if they are known, or 'other' if there is only an exonic functional annotation available.

We tabulate the results for the Fairfax data in Table 10.3, showing the rank of the causal SNP and the maximum posterior effect size for both the NG and the NG super function. We state the functional information group that the causal SNP belongs to in brackets with its rank in the NG super function.

Method	**Max Effect Size**	**Causal SNP rank**
CARD9 mono (511 SNPs, 243 individuals)		
NG	0.0393	3
NG super function	0.06911	2 (non)
FADS1 bcell (1076 SNPs, 243 individuals)		
NG	0.09312	639
NG super function	0.06997	2 (intronic)
RBM6 bcell (932 SNPs, 243 individuals)		
NG	0.047661	39
NG super function	0.02698	16 (syn)
RBM6 mono (932 SNPs, 243 individuals)		
NG	0.05562	191
NG super function	0.0602	8 (syn)

Table 10.3: Table showing the results of the Fairfax data on the selected genes for the NG and the NG super function. Each gene has one causal SNP identified in the literature, which we state the rank of for the NG and the NG super function. For the NG super function, we state in brackets the functional group to which the causal SNP belongs, with its ranking.

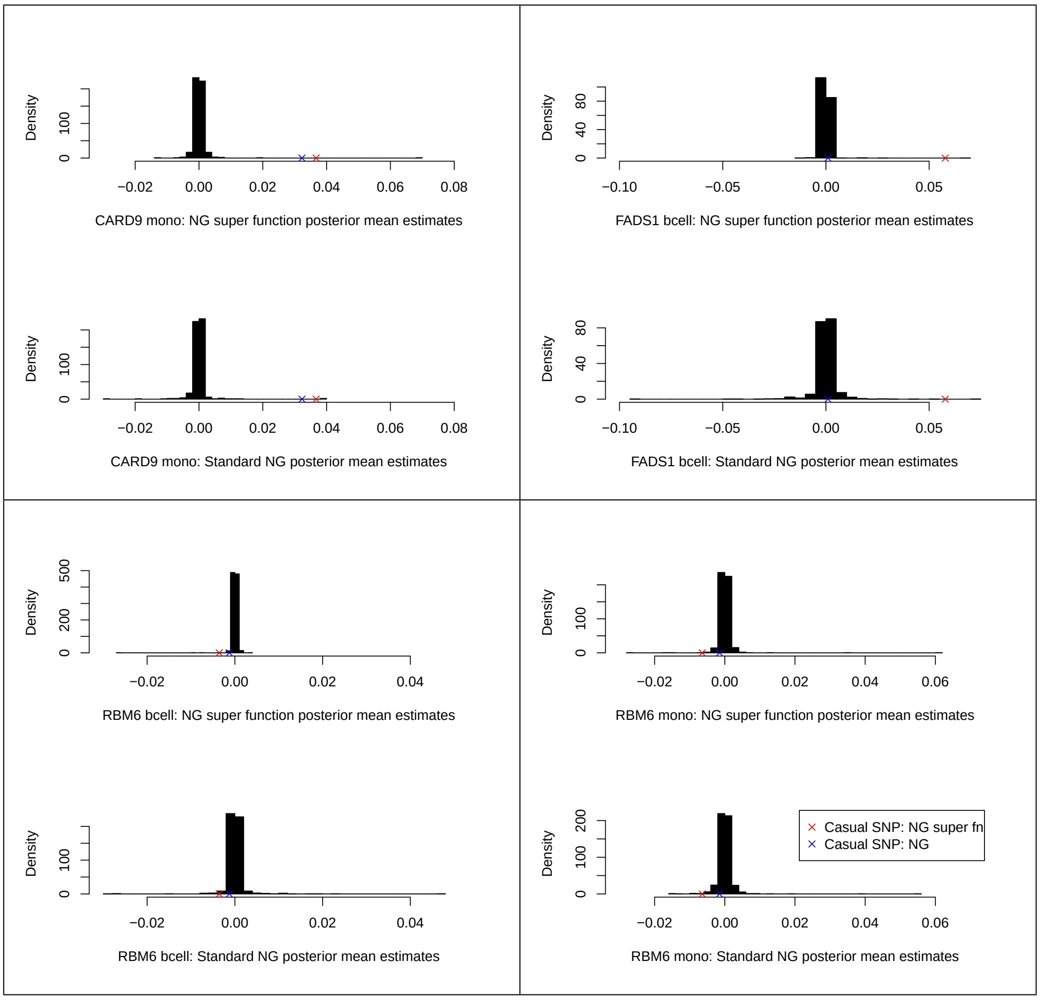


Figure 10.1: Comparison of the posterior mean estimates for the NG and the NG super function for the 4 genes in the Fairfax dataset. The causal SNPs are marked on both plots for the NG and the NG super function.

The results in Table 10.3 show that in all cases, including for FADS1 where the causal SNP is on its own in the intronic group, the rank of the causal SNP has improved, even though the causal SNPs are, in general, not in the groups with a priori less shrinkage (splicing and non-synonymous (non)). This shows that the differential shrinkage based on the location of the SNP is improving the causal SNP detection for these Fairfax genes.

We plot the posterior mean estimates for the standard NG and the NG super function, highlighting the magnitude of the causal SNP for the NG and the NG super function in both cases in Figure 10.1. This allows us to compare the posterior means relative to the causal SNP effect size.

Figure 10.1 shows that the absolute posterior mean effect size of the causal SNP is always larger in the NG super function than in the standard NG. The non-causal SNP effect sizes are, as expected, very close to 0. There are only a few non-causal SNPs which have absolute posterior mean effect size larger than the causal SNP. For FADS1, where the rank of the causal SNP is increasing from 632 to 2 when using the NG super function, we notice there is a large difference in the absolute value of the posterior mean effect size. This can only be due to the differential shrinkage of the NG super function.

The maximum posterior mean is not always larger for the NG super function compared to the standard NG as we saw in the Hulse data. The maximum posterior effect sizes remain fairly constant across both methods. This shows that, even with a priori more or less shrinkage, the posterior mean estimates can increase or decrease based on a combination of the likelihood and the prior.

10.3 Conclusion

We recall that for the Fairfax data we had 243 individuals and for the Hulse data we had either 38 or 39 individuals. This difference in the numbers of individuals (n) may be affecting the posterior estimates as the likelihood may not contain as much information in the Hulse dataset as in the Fairfax dataset. Hence the prior will be more influential and therefore the change in prior would lead to different posterior results. As we have no validated causal SNPs we cannot verify this using the Hulse dataset.

In conclusion, it is very difficult to assess the differences in the Hulse dataset as we have no reportedly causal SNPs to compare directly. Comparing across the top 5 ranked SNPs by the NG and the NG super function, we see very little agreement. With the Fairfax data, we see an increase in the rank of the causal SNP in all cases. This shows that the prioritisation of the SNPs by employing different function information groups is successfully increasing identification of causal SNPs. Even in the case when the causal SNP is put in a group with a priori lower mean FS score, the rank of the causal SNP still increases (often dramatically).

Chapter 11

Discussion

11.1 Conclusion

In Chapter 3 we investigated several statistical methods that could be applied to eQTL data. In Chapter 4 we applied these methods to simulated eQTL data. The results from the Normal Gamma prior were such that we choose to develop this model to include functional information with the aim of prioritising the most likely causal SNPs. The inclusion of seven functional information groups in Chapter 8 and the application of this NG super function in Chapter 10 increased the rank of the causal SNP in all four Fairfax genes tested, and increased the effect sizes of all simulated causal SNPs with respect to the standard Normal Gamma.

Within the field of genetics, the ability of this NG super function to take in human data with any genotype coding ($\{0,1\}$, $\{0,1,2\}$, imputed genotypes) and a vast range of n and p, and then to prioritise SNPs represents a great advancement in selecting SNPs to biologically test for disease association.

No statistical model is perfect, to maximise the information from a statistical model, a combination of model results and biological knowledge are needed. If this combination can be achieved, the NG model has the scope to aid understanding of complex biological diseases.

11.2 Further developments to the Normal Gamma

Initial work to understand gene expression data highlighted that the amount of background gene expression (α) present in most samples was not constant. In the NG we centre y to remove the majority of the background expression and then give α an uniformative prior distribution. To develop the NG model further, a proper prior on α, and updating it separately to $\boldsymbol{\beta}$ could produce a more realistic model.

At present, the NG model is coded in Matlab. If this model is to become more widely used, a quicker and more memory efficient processing tool would need to be used such as C++ within R.

The computational requirements of the NG model mean that there are limitations on the numbers of individuals and SNPs that can be included in the model if it is to be run in a reasonable time frame (less than a week) and without enormous memory requirements. The maximum number of SNPs we have included in the NG and NG super function is approximately 3000, and the largest number of individuals is 300. This cannot be run on a standard PC due to the increased memory requirements. This limitation on the size of the data that can be processed means that many GWAS could not benefit from the model without prior reduction in the SNPs and/or individuals to include.

We summarise the MCMC output using the posterior mean. As a result, MAP estimation or variational Bayes may be a computationally quicker approach. We could also consider using a similar prioritisation approach based on the FS scores applied to a MAP estimation approach such as HyperLasso which would lead to summarising the posterior using the mode rather than the mean.

When summarising the Normal Gamma posterior distribution, we currently use only the posterior mean. As discussed in Section 4.1.2, we could have used formal statistical testing. Using the histograms in Figure 4.3, page 58, we notice one bimodal posterior distribution. This is for a truly causal SNP (simulated to be so). The maximum mode represents the prior distribution which surrounds 0. Using a statistical test such as the t-statistic that compares for a statistically significant difference between the prior and the posterior distributions, having taken into account the prior. This may allow us to detect causal/associated SNPs more robustly than using the posterior mean as we have done throughout this thesis.

Throughout this thesis we use four simulated datasets, HapGen datasets 1A, 1B, 2A and 2B. In these datasets we keep the MAF of the causal SNPs constant. This does not allow us to investigate the effect of different MAFs on the posterior mean effect size estimates. Further simulation studies on datasets where different causal SNPs have different MAFs would allow us to investigate more about the interaction between the ability to detect a causal/associated SNP and the MAF of the SNP. This is beyond the aims of this thesis but the results of such a study would increase understanding of the cases in which the NG performs optimally.

For the NG super function, we use a $tan()$ transformation from the FS score to B. We could consider a linear transformation that would allow Gibbs updating and would also provide a wide range of values for B. This would be computationally much simpler, and may therefore increase computational speed.

If the computational speed of the NG could be increased and the computational demands could be reduced, there would be scope to model the gene expression jointly across genes rather than applying a gene-by-gene approach. This would allow the complex regulatory interactions between genes that are caused by genetic mutations to be taken into account.

The error structure of the Normal Gamma is very simple, using only a scalar σ^2. We suggest that the inclusion of a covariance matrix Σ with an Inverse Wishart prior might better represent the complex error structure of genetics data more effectively. Using this type of error structure could also enable incorporation of gene expression technical variability from PUMA [Liu, 2006]. Details of preliminary work on this can be found in Appendix C, page 181.

The minor allele frequency of the causal SNPs on the simulated data had a large effect on the ability of the NG to detect causal SNPs. This means that on rare SNPs the prior seemed to be dominating the posterior and enforcing large amounts of shrinkage. Developing the model to take into account the minor allele frequency as well as the functional information group may increase detection of rare SNPs that are found in functionally important groups such as splicing SNPs, or even in groups that have lower functional effect based on the FS score, such as intronic or intergenic SNPs. We could also develop the grouping of SNPs to include other functional information that is not related to the FS scores. The ENCODE database is being increasingly populated with functional information that we could use for this purpose.

The functional information is specific to humans, and therefore the NG super function cannot be directly applied to data from other organisms. We have seen within this thesis, that the NG is improved by increasing the flexibility of the model by placing a prior distribution on $E[\pi(\text{var}(\beta|\lambda,\gamma))]$. This expectation is fixed in Griffin and Brown [2010] to be M, but we adapt this when including functional information to be either MB or B in the NG super function. We could therefore develop a general, non-specific version of the NG super function that could be used with any organism.

We rely on the functional information score being accurate to inform our prior distributions. The database of FS scores that we used was produced in 2008, and contains only a small proportion of the total number of SNPs with annotations in the different databases. The information that is in the database we believe to have been correct at the time of calculating the FS score. This information may have been updated, and there may be more SNPs that could be used, hence leading to possible changes the prior distributions that we have fitted. To maintain the complete accuracy of the prior distributions, we would need to re-implement the FS score calculations such that the priors were automatically updated with any adjustments to the set of all FS scores. This would be computationally very demanding.

We initially focused the model on eQTL data consisting of information on genotypes and gene expression. There is no reason that the model could not be extended to GWAS data where the response variable is binary based on a disease phenotype. In its current form, the NG super function would be readily able to handle any type of sequence and gene expression data where there were two groups of individuals such as two extreme phenotype groups, or case-control data. The NG can easily be adapted to include indicator variables to identify the case/control status or extreme phenotype group.

The NG currently only includes SNPs in the model. If we could define a code to represent other genetic mutations then there is scope for easy incorporation of other genetic mutations. The NG super function would also need to be developed to include categories for representing the deleterious effect of other mutations such as indels and CNVs. This would be possible with more time, and the data to support the definition of the prior distributions.

The NG super function in its current form has much scope for development. However, compared to the standard NG model it is much improved in terms of the ranking of causal SNPs. Using the top number or percentage of SNPs defined by the NG super function for biological validation would lead to a higher chance of detecting the truly causal SNP than using any other statistical method compared in this thesis.

Appendix A

Standard Distributions

There are some standard distributions that are used when stating the prior distributions and deriving the full conditional distributions from the Normal Gamma prior. For clarity, the probability density functions and the shorthand notation for these distributions are defined below.

A.1 The Gamma distribution.

Throughout this work we will denote the Gamma distribution by $Ga(\alpha, \beta)$, for shape parameter α and rate parameter $\beta = \frac{1}{\theta}$, where θ is the scale parameter. This has mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$. The probability density function of the $Ga(\alpha, \beta)$ density is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x).$$

A.2 The Generalised Inverse Gaussian distribution.

The Generalised Inverse Gaussian distribution is denoted GIG(a,b,c) and has the probability density function as below:

$$f(x) = \frac{(\frac{c}{b})^{\frac{a}{2}} x^{a-1} exp\left(-\frac{1}{2}\left(\frac{b}{x} + cx\right)\right)}{2K_a(\sqrt{bc})},$$

where $K_a(.)$ represents the modified Bessel function of the third kind.

We use the special cases of the GIG stated by Johnson et al. [1994] when calculating the full conditional distributions. The special cases are as follows:

1. Gamma distribution occurs when $b = 0$ and $a > 0$.

2. Inverse Gamma occurs when $c = 0$ and $a < 0$.

Appendix B

Simple dataset simulation results

B.1 Basic simulated data

We begin testing the statistical methods by using a simple dataset that contains no information on correlation of variables, and has a wide range of effect sizes, some of which may not be realistic. We use this to initially assess the methods and their implementation.

B.1.1 Simulating the data

We simulate the SNP data by generating standard Normal random vairables. We use the threshold of 0 to define SNPs (1) or not SNPs (0) for each individual. We simulated data with 50 individuals and 100 SNPs. We simulate effect sizes for 8 causal SNPs to be 2, 2, 2, 0.6, 0.6, 0.5, 0.4 and 0.4. The non-causal SNPs have effect sizes generated from $N(0, 0.01^2)$. We then calculate the gene expression value $y_j = \sum_{i=1}^{p} X_{j,i} + \epsilon$, for individual j with SNPs $i = 1, \ldots, p$ and $\epsilon \sim N(0, 1)$.

B.1.2 Results

We use this dataset to initially assess different methods of summarising the Normal Gamma prior, posterior mean and median, as well as comparing the methods suggested in Chapter 3; piMASS (posterior beta estimate), HyperLasso (HL), Minimum Length Least Squares (MLLS/LS) and likelihood ratio (LR) test. The results of the initial comparison are shown in the ROC curve, see Figure B.1.

All methods are performing very well with the smallest AUC of 0.8246 for the LR test. The maximum is HyperLasso with an AUC of 0.9901. The NG posterior mean performs better than the posterior median with AUCs 0.9779

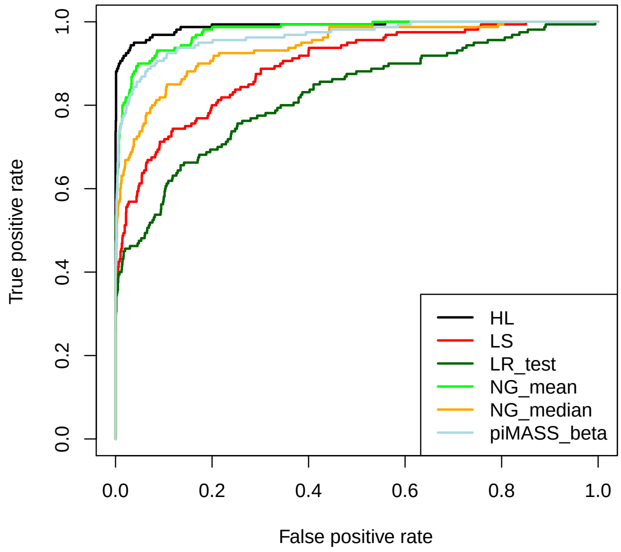


Figure B.1: ROC curve comparing the different methods for detecting causal SNPs. The data is simulated to have 8 causal SNPs with effect sizes ranging from 2-0.4 (basic simulated data). We compare the difference here between the posterior mean and median for the Normal Gamma.

and 0.9390 respectively. The LS has an AUC of 0.8943 and piMASS has an AUC of 0.9659. These are very high, but we expect this as there are 8 causal SNPs in each of the 20 datasets (160 causal SNPs out of 2000) of which 3 in each dataset (60 in total) have effect size of 2. This is a very large effect, much larger than we would expect in real gene expression data. We included such a large effect initially to ensure the methods were able to perform efficiently with data of the format $n < p$.

The Normal Gamma is an MCMC method which provides the full updating steps as output. We use this output to investigate possible summary statistics for the posterior distributions and assess their effectiveness. Initially, from Figure B.1 we notice that the posterior mean is superior to the posterior median with respect to the AUC.

We assess the posterior means of the SNPs whose credible intervals do not include 0. We vary the percentage credible interval. Those credible intervals that include 0 have $\beta_i := 0$ for the ROC, those that do not include 0 have the posterior mean as the estimate of β_i. We compare the 99%, 95%, 90%, 80%, 70%, 60% and 50% credible intervals in Figure B.2. The widest credible interval leads to the largest AUC, as we maintain the largest proportion of SNPs. Even in the case of the 50% credible interval, we still only obtain around 80% of the true causal SNPs. This is not as good as we hoped for this method and so we will not use any of the credible intervals to define causal or associated SNPs.

We could have used formal association testing to define whether SNPs are associated for the NG. To show why we didn't do this, we initially decided to define only those SNPs whose 90% credible interval did not contain 0. Upon investigation of simulated data, we find that although the proportion of SNPs retained decreases, there are still a large number of SNPs retained. More importantly, the proportion of causal SNPs retained is quite low, see Table B.1.

% credible interval	% SNPs with credible interval not containing 0	% causal SNPs with credible interval not containing 0
99	5.5	54.5
95	7	59
90	8	60.5
80	10.5	69
70	13.5	72
60	17	75.5
50	22	82

Table B.1: The percentages of SNPs and causal SNPs retained when reporting only those SNPs whose credible interval does not contain 0 in the Normal Gamma.

The proportions of SNPs retained are small, however when there are more than approximately 500 SNPs say, the 90% credible interval not containing 0 will still retain around 40 SNPs. This is a large number to propose for biological validation, especially when we expect only around 60% of these (approximately 24 SNPs) to be truly causal.

We also assess difference scalings of the posterior β_i estimates. We use $\frac{\text{mean}}{\text{SD}}$; $\frac{\text{mean}}{(\text{width of 90\% credible interval})}$; and $\frac{\text{mean}}{\sqrt{(\text{width of 90\% credible interval})}}$. The results, compared with the posterior mean and median can be seen in Figure B.3.

Tables B.2 and B.3 shows the AUC for each method in Figures B.2 and B.3 respectively. Of all these methods for summarising the posterior distribution of the Normal Gamma, the posterior mean appears to be the best method with an AUC of 0.9487, compared to the next best method of the $\frac{\text{mean}}{\text{SD}}$ with AUC 0.9456, which also includes the posterior mean. The posterior median has AUC 0.9390, considerably smaller than for the posterior mean.

B.1.3 Conclusion

Comparing all summary methods for the Normal Gamma we see that the AUC for the posterior mean is the largest. Using the mean relies on the posterior

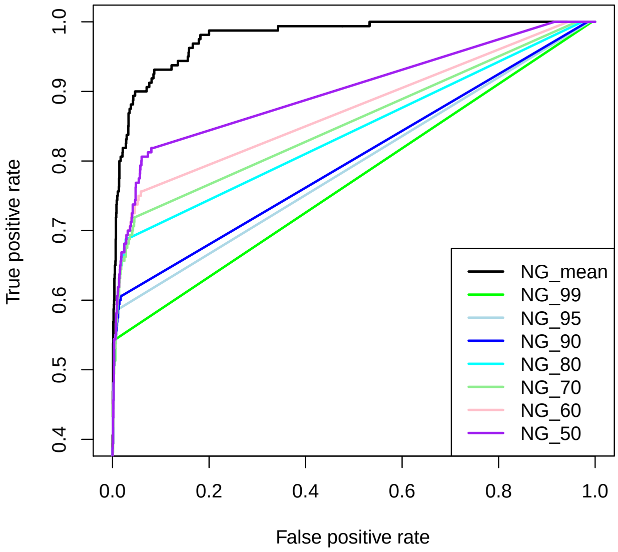


Figure B.2: ROC curve comparing the different credible intervals containing 0 as a method for detecting causal SNPs. The data is simulated with 8 causal SNPs with effect sizes ranging from 2-0.4 (basic simulated data). We compare the difference here between the posterior mean for the Normal Gamma and the credible intervals for each posterior β.

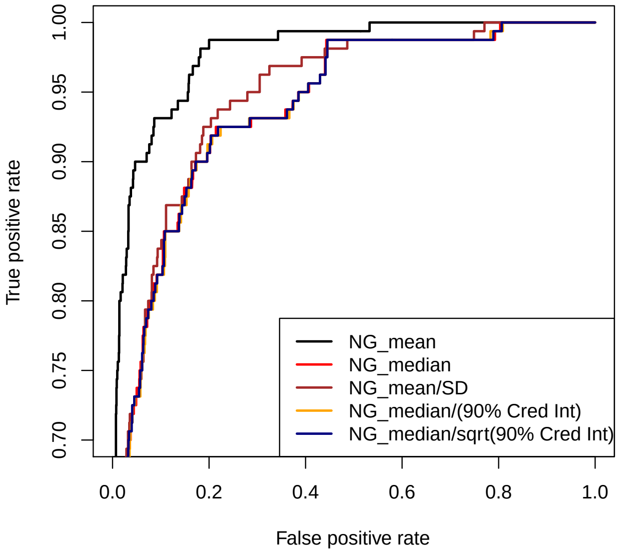


Figure B.3: ROC curve comparing the different summary statistics for the Normal Gamma for detecting causal SNPs. The data is simulated with 8 causal SNPs with effect sizes ranging from 2-0.4 (basic simulated data).

99% CI	95% CI	90% CI	80% CI	70 % CI	60 % CI	50 % CI
0.7716	0.7926	0.8015	0.8411	0.8544	0.8724	0.9006

Table B.2: The AUCs of the ROC curves in Figure B.2 based on only maintaining SNPs whose posterior credible does not contain 0. Where this is the case, the ranks are based on the posterior mean for the SNP. We vary the posterior credible interval from $99\% - 50\%$ for the Normal Gamma on the basic simulated data.

Mean	Median	$\frac{\text{mean}}{\text{SD}}$	$\frac{\text{mean}}{\text{(width of 90\% credible interval)}}$	$\frac{\text{mean}}{\sqrt{\text{(width of 90\% credible interval)}}}$
0.9487	0.9390	0.9456	0.9374	0.9385

Table B.3: The AUCs for the ROC curves in Figure B.3, where we test different posterior summary statistics for the Normal Gamma on the basic simulated data.

distribution not having too many outlying values. We think this is unlikely and therefore that the mean the best summary statistic for the Normal Gamma, hence we will use this in all future work.

The Normal Gamma is not the best method on this dataset, Hyper Lasso performs exceptionally well, although piMASS and the Normal Gamma are very good. All methods, even the univariate LR test and the MLLS estimates, provide good results. To be able to appropriately compare these methods, we feel it would be better to compare on a more realistic dataset with more appropriate effect sizes and a correlation structure that reflects human data.

Appendix C

Future developments to the Normal Gamma model

C.1 Including Gene Expression Uncertainty

Our initial method for including this extra uncertainty in the Normal Gamma prior framework is to scale the variance of the likelihood to be $\sigma^2\mathbf{\Sigma}$. σ^2, the uncertainty parameter, remains the same as in the standard NG. $\mathbf{\Sigma}$ is now a weight matrix with the diagonal elements corresponding to the squared standard error estimates for each gene, for each subject in the model. This allows us to take into account the gene expression technical variability from PUMA [Liu, 2006] when modelling the parameter estimates.

The effect of this change on the full conditional distributions is minor. The only full conditional distributions to change are the full conditional distributions for σ^2 and $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$. This is due to the change in the likelihood, which becomes:

$$\begin{aligned} f(data|\lambda, \gamma^{-2}, \psi_i, \alpha, \beta_i, \sigma^{-2}) &= N_n(\mathbf{y} - \alpha - X\boldsymbol{\beta}, \sigma^2\mathbf{\Sigma}) \\ &\propto |\sigma^2\mathbf{\Sigma}|^{\frac{1}{2}} exp\left(-\frac{1}{2}(\mathbf{y} - \alpha - X\boldsymbol{\beta})^T(\sigma^2\mathbf{\Sigma})^{-1}(\mathbf{y} - \alpha - X\boldsymbol{\beta})\right) \\ &\propto \left(\sigma^{-2}\right)^{\frac{n}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}} exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha - X\boldsymbol{\beta})^T\mathbf{\Sigma}^{-1}\mathbf{y} - \alpha - X\boldsymbol{\beta})\right). \end{aligned}$$

The new full conditional distribution for σ^2 is as follows.

$$\begin{aligned} f(\sigma^{-2}|\Sigma, \mathbf{y}, \alpha, \boldsymbol{\beta}) &\propto \left(\sigma^{-2}\right)^{\frac{n}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}} exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha - X\boldsymbol{\beta})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \alpha - X\boldsymbol{\beta})\right) \\ &\propto \left(\sigma^{-2}\right)^{\frac{n}{2}} exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{y} - \alpha - X\boldsymbol{\beta})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \alpha - X\boldsymbol{\beta})\right). \end{aligned}$$

Comparing this to the standard distributions, we notice that this is proportional to a Gamma distribution of the form:

$$Ga\left(\frac{n}{2}+1,\ \frac{(\mathbf{y}-\alpha-X\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\alpha-X\boldsymbol{\beta})}{2}\right).$$

Hence, from the updated likelihood, the new full conditional distribution for $\boldsymbol{\phi}=(\alpha,\boldsymbol{\beta})^T$ is:

$$N\left(\left(X^{*T}\Sigma^{-1}X^{*}+\sigma^2\Lambda\right)^{-1}X^{*T}\mathbf{y},\sigma^2\left(X^{*T}\Sigma^{-1}X^{*}+\sigma^2\Lambda\right)^{-1}\right).$$

C.1.1 Discussion of inclusion of gene expression uncertainty

Considering this, it appears that this may not be the most appropriate way to include this extra variability. Including the variability in this way overcomes any identifiability problems, but it does not necessarily make sense that the overall variance of each gene will be weighted by the technical variance in the gene expression values. The gene expression variability may contain biological as well as technical variability. This variability is not scaled based on the technical variability. Other methods that allow us to take into account the extra sources of variability may encounter problems based on identifiability of the error.

It is important to develop a method for incorporating this extra knowledge of the uncertainty in gene expression calls. We have considered modelling the error in the form $\sigma^2+\rho\Sigma$, where Σ is the diagonal matrix of standard errors. We fear this will lead to identifiability problems with ρ and σ.

Another suggestion was to define a covariance matrix K where the diagonal elements are defined as the scaled technical variability estimates. Again, we believe this would lead to the problems discussed above.

Appendix D

Pseudocode

Pseudo-code is used to give the flow of code without giving all the commands. It can be though of as almost just the annotations/comments that accompany any well written code. We use it here to convey the stages of the updating of the NG prior parameters to achieve convergence.

D.1 Pseudocode for the standard NG function

```
function [outputs] = NG_function(Inputs: SNPs, gene expr and nbrIters)

% Define the nbr of obs, n
% Define the nbr of parameters p
% Define sigmaSqLambda

% Calculate M
if ( p > n )
    % Use the MLLS estimate
else
    % Use the LS estimates
end

% Define the initial parameter values
% Create the locations to store the values from the MCMC

for i=1:nbrIterations
    %%% Updating alpha and beta %%%
    % Define xhat
                      xhat = 10^8
```

```
  % Calculating the updating of alpha and beta in stages
% based on the size of the associated psi values
  if ( max(psi) / min(psi) > xhat )
% Define an indicator vector to show which elements are to be
% included at which stage
indicator = psi > min(psi)*(xhat/10);
% Calculate the residuals, mean and var of excluded elements
    residual = y - X(:, indicator==0) * beta(indicator==0);

% Use the Cholesky decomposition and propose values from Normal dist

% Redefine the indicator vector to show which psi are smaller than
% the min psi multiplied by xhat*10
    indicatorVector = psi < min(psi)*(xhat*10);

 % Calculate the new residuals, mean and var of excluded elements
    residual = y - alpha  - X(:, indicator=0) * beta(indicator=0);

% Use the Cholesky decomposition and propose values from Normal dist.
  else
    % Calculate the expectation and variance
% Use the Cholesky decomposition and propose values from Normal dist.
  end

  %%%% Updating psi %%%
  for j=1:p
    if (beta approx 0)
      if (first parameter is negative) % Case where the first
                       % parameter is less than zero
        % Use the inverse gamma distribution to update
      else % Case where parameter 2 is approx 0 and parameter one
                        % is non-negative
        % Use the gamma distribution to update
      end
    else % No special case
      % Use the GIG distribution for updating
    end
  end
```

```
    %%% Updating sigmaSq %%%
    % Update using the Inverse Gamma distribution

    %%% Updating lambda %%%
    % Define the proposed  lambda value and the
    % corresponding proposed gamma value

    % Calculate the acceptance probability
% Update sigmaSqLambda
    % Accept if the acceptance prob is greater than a uniform
    % random value

    %%% Updating gamma %%%
% Update using the inverse gamma distribution

end

% Generate summary statistics
```

D.2 Pseudocode for NG splitting function

```
function [outputs] = NG_splitting_function(Inputs:
   SNPs, gene expression, nbrIterations, vector of syn/non status)

% Ensure the function looks in the correct folder for the GIG code,
% randraw script

% Define the number of observations, n and number of parameters p
% Define sigmaSqLambda

% Calculate M
if ( p > n )
    % Use the MLLS estimate
else
    % Use the LS estimates
end
```

```
% Define the initial values and the syn/non elements
% Create the initial locations for the values to be stored in

for i=1:nbrIterations
    %%% Updating alpha and beta SYNONYMOUS %%%
% Define xhat
xhat = 10^8

    % Calculating the updating of alpha and beta_syn in stages
% based on the size of the associated psi_syn values
    if ( max(psi) / min(psi) > xhat )

% Define an indicator vector to show which elements are to be
% included at which stage
indicator_syn = psi > min(psi)*(xhat/10);

% Calculate the residual, partial variance and partial expectation
% based on those elements not included
  residual = y - X(:, indicator_syn=0) * beta(indicator_syn=0);

% Use Cholesky decomposition to check the values are appropriate

% Redefine the indicator vector to show which psi are smaller than
% the min psi multipled by xhat*10
        indicator = psi < min(psi)*(xhat*10);

% Calculate the residual, partial variance and partial expectation
% based on those elements not included
  residual = y - alpha  - X(:, indicator_syn=0) * beta(indicator_syn=0);

% Use the Cholesky decomposition to check the values are appropriate
    else
        % Calculate the expectation and variance
% Use the Cholesky decomposition to check the values are appropriate
    end

    %%% Updating psi SYNONYMOUS %%%
for j=1:p_syn
        if (beta approx 0)
```

```
      if (first parameter is negative) % Case where the first
   % parameter is less than zero
        % Use the inverse gamma distribution to update
      else % Case where parameter 2 is 0 and parameter one is
   % non-negative
        % Use the gamma distribution to update
      end
    else % No special case
      % Use the GIG distribution for updating
    end
  end

  %%% Updating sigmaSq %%%
  % Update using the Inverse Gamma distribution

  %%% Updating lambda SYMONYMOUS %%%
  % Define the proposed lambda value and the
  % corresponding proposed gamma value

  % Calculate the acceptance probability
% Update sigmaSqLambda
  % Accept if the acceptance prob is greater than a uniform random
  % value

  %%% Updating gamma SYNONYMOUS %%%
% Update using the inverse gamma distribution

  %%% Updating B/FS SYNONYMOUS %%%
  % Update using sum of gamma distributions
  % Define B-syn from FS
  B_syn = FS_syn + 0.5;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%% Updating alpha and beta NON-SYNONYMOUS %%%
  % Define xhat
xhat = 10^8

  % Calculating the updating of alpha and beta_non in stages
% based on the size of the associated psi_non values
    if ( max(psi) / min(psi) > xhat )

% Define an indicator vector to show which elements are to be
% included at which stage
indicator_syn = psi > min(psi)*(xhat/10);

% Calculate the residual, partial variance and partial expectation
% based on those elements not included
        residual = y - X(:, indicator_non=0)*beta(indicator_non=0);

% Use the Cholesky decomposition to check the values are appropriate

% Redefine the indicator vector to show which psi are smaller than
% the min psi multipled by xhat*10
        indicator = psi < min(psi)*(xhat*10);

% Calculate the residual, partial variance and partial expectation
% based on those elements not included
        residual=y-alpha-X(:, indicator_non=0)*beta(indicator_non=0);

% Use the Cholesky decomposition to check the values are appropriate
    else
        % Calculate the expectation and variance
% Use the Cholesky decomposition to check the values are appropriate
    end

%%% Updating psi NON-SYNONYMOUS %%%
for j=1:p_syn
        if (beta approx 0)
            if (first parameter is negative) % Case where the first
     % parameter is less than zero
                % Use the inverse gamma distribution to update
            else % Case where parameter 2 is 0 and parameter one
     %is non-negative
                % Use the gamma distribution to update
            end
        else % No special case
            % Use the GIG distribution for updating
        end
    end

    %%% Updating sigmaSq %%%
    % Update using the Inverse Gamma distribution

    %%% Updating lambda NON-SYMONYMOUS %%%
    % Define the mean of psi, the proposed lambda value and the
    % corresponding proposed gamma value

    % Calculate the acceptance probability
% Update sigmaSqLambda
    % Accept if the acceptance prob is greater than a uniform random
    % value

    %%% Updating gamma SYNONYMOUS %%%
% Update using the inverse gamma distribution

    %%% Updating B/FS NON-SYNONYMOUS %%%
    % Update using the gamma distributions
    % Define B-non from FS
    B_non = FS_non + 0.5;

end

% Generate summary statistics
```

D.3 Pseudocode for NG super function

We use functions based on the standard Normal Gamma updating for the subsets generated by each of the functional information group.

Firstly we provide the pseudocode for the Master function which calls each individual function. The individual functions have pseudocode similar to the pseudocode in Section D.2, excluding σ^2 which we only update once per iteration.

```
function [outputs] = NG_master_function(Inputs:
   SNPs (X), gene expression (y), nbrIterations, groupingVector)
% Note: groupingVector has 0 for non-syn, 1 for syn,
% 2 for intronic, 3 for intergenic, 4 for splicing, 5 for utr3,
% and 6 for other

% Define the number of observations, n and number of parameters p

% Define
sigmaSqLambda = 0.01;
sigmaSqFS = 1;
burnin = 5000;

% Defining the initial values for all parameters

% Determine which functional groups are represented in the data

% Create the initial locations for the values to be stored in

% Create the functional group specific initial values

for iter=1:nbrIterations

%%% Non-synonymous updating
if (number_of_non_SNPs > 0)
                      % Update all non_syn SNP parameters
end

%%% Synonymous updating
if (number_of_syn_SNPs > 0)
                      % Update all syn SNP parameters
end

%%% Intronic updating
if (number_of_intronic_SNPs > 0)
                      % Update all intronic SNP parameters
end

%%% Intergenic updating
if (number_of_intergenic_SNPs > 0)
                      % Update all intergenic SNP parameters
end
```

```
%%% Splicing updating
if (number_of_splicing_SNPs > 0)
                      % Update all splicing SNP parameters
end

%%% UTR3 updating
if (number_of_utr_SNPs > 0)
                      % Update all UTR3 SNP parameters
end

%%% Other updating
if (number_of_other_SNPs > 0)
                      % Update all other SNP parameters
end

    % Updating sigmaSq

    % Save the updated values of the parameters
end

% Generate summary statistics
```

Bibliography

B. Alberts, J. H. Wilson, and T. Hunt. Molecular biology of the cell. Garland Science, 5 edition, 2008.

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

J. H. Barrett, M. M. Iles, M. Harland, J. C. Taylor, J. F. Aitken, P. A. Andresen, L. A. Akslen, B. K. Armstrong, Marie-Francoise Avril, E. Azizi, B. Bakker, W. Bergman, G. Bianchi-Scarr, B. Bressac de Paillerets, D. Calista, L. A. Cannon-Albright, E. Corda, A. E. Cust, T. Debniak, D. Duffy, A. M. Dunning, D. F. Easton, E. Friedman, P. Galan, P. Ghiorzo, G. G. Giles, J. Hansson, M. Hocevar, V. Hiom, J. L. Hopper, C. Ingvar, B. Janssen, M. A. Jenkins, G. Jnsson, R. F. Kefford, G. Landi, M. T. Landi, J. Lang, J. Lubinski, R. Mackie, J. Malvehy, N. G. Martin, A. Molven, G. W. Montgomery, F. A. van Nieuwpoort, S. Novakovic, H. Olsson, L. Pastorino, S. Puig, J. A.

Puig-Butille, J. Randerson-Moor, H. Snowden, R. Tuominen, P. Van Belle, N. van der Stoep, D. C. Whiteman, D. Zelenika, J. Han, S. Fang, J. E. Lee, Q. Wei, G. M. Lathrop, E. M. Gillanders, K. M. Brown, A. M. Goldstein, P. A. Kanetsky, G. J. Mann, S. MacGregor, D. E. Elder, C. I. Amos, N. K. Hayward, N. A. Gruis, F. Demenais, J. A. Newton Bishop, and D. T. Bishop. Genome-wide association study identifies three new melanoma susceptibility loci. *Nature Genetics*, 43:1108–1113, 2011.

A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Springer-Verlag, New York, 2nd edition, 2003.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7 (4):434–455, 1998. doi: 10.1080/10618600.1998.10474787.

W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, December 2012. doi: 10.1371/journal.pcbi.1002822.

N. J. Camp, M. Parry, S. Knight, R. Abo, G. Elliott, S. H. Rigas, S. P. Balasubramanian, M. W. R. Reed, H. McBurney, A. Latif, W. G. Newman, L. A. Cannon-Albright, D. G. Evans, and A. Cox. Fine-mapping casp8 risk variants in breast cancer. *Cancer Epidemiology Biomarkers and Prevention*, 21 (1):176–181, 2012.

R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. 86:6–22+, 2010. ISSN 0002-9297. doi: 10.1016/j.ajhg.2009.11.017.

G. Casella and C. P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, March 1996. doi: 10.1093/biomet/83.1.81.

J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, 40, 2012. doi: 10.1093/nar/gkr1029.

S. C. T. Choi. *Iterative Methods For Singular Linear Equations And Least-Squares Problems.* PhD thesis, Stanford University, December 2006.

E. Choy, R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, and D. Altshuler. Genetic analysis of human traits in vitro: Drug response and gene expression in lymphoblastoid cell lines. *PLoS Genetics*, 4(11):e1000287, November 2008.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–45, September 1988.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998. ISSN 1091-6490. doi: 10.1073/pnas.95.25.14863.

F. Emmert-Streib and G. V. Glazko. Pathway analysis of expression data: Deciphering functional building blocks of complex diseases. *PLoS Computational Biology*, 7(5), 2011.

B. P. Fairfax, S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. Genetics of gene expression in primary immune cells identifies cell typespecific master regulators and roles of HLA alleles. *Nature Genetics*, 44:502510, 2012.

P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Khri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernndez-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012. doi: 10.1093/nar/gkr991.

L. Franke and R. C. Jansen. eqtl analysis in humans. In Keith DiPetrillo, editor, *Cardiovascular Genomics*, volume 573 of *Methods in Molecular Biology*, pages 311–328. Humana Press, 2009. ISBN 978-1-60761-246-9. doi: 10.1007/978-1-60761-247-6_17.

W. J. Fu. Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998. ISSN 10618600. doi: 10.2307/1390712. URL `http://dx.doi.org/10.2307/1390712`.

R. Gottardo and A. E. Raftery. Markov chain monte carlo with mixtures of singular distributions. *Journal of Computational and Graphical Statistics*, 17 (4):949–975, 2004.

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

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

C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7):e1000130, 2008.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, June 2009. doi: 10.1371/journal.pgen.1000529.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008. ISSN 1750-2799. doi: 10.1038/nprot.2008.211.

A. M. Hulse and J. J. Cai. Genetic variants contribute to gene expression variability in humans. *Genetics*, 193(1):95–108, 2013.

R. Hunt, Z. E. Sauna, S. V. Ambudkar, M. M. Gottesman, and C. Kimchi-Sarfaty. Silent (synonymous) snps: Should we care about them? In A. A. Komar, editor, *Single Nucleotide Polymorphisms*, volume 578 of *Methods in Molecular Biology*, pages 23–39. Humana Press, 2009. ISBN 978-1-60327-410-4. doi: 10.1007/978-1-60327-411-1_2.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, April 2005.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions.*, volume 1 of *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons, New York, 2nd edition, 1994.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

G. Kang, D. Lin, H. Hakonarson, and J. Chen. Two-Stage Extreme Phenotype Sequencing Design for Discovering and Testing Common and Rare Genetic Variants: Efficiency and Power. *Human Heredity*, 73:139–147, 2012.

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

S. I. Lee, A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan, D. Peér, and D. Koller. Learning a prior on regulatory potential from eqtl data. *PLoS Genetics*, 5(1):e1000358, January 2009. doi: 10.1371/journal.pgen.1000358.

C. Leng, M. N. Tran, and D. Nott. Bayesian Adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 66:221–244, April 2014.

W. A. Link and M. J. Eaton. On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115, 2012. doi: 10.1111/j.2041-210x.2011.00131.x.

X. Liu. *Microarray Data Analysis Using Probabilistic Methods.* PhD thesis, Univeristy of Manchester, November 2006.

A. Lourdusamy, S. Newhouse, K. Lunnon, P. Proitsi, J. Powell, A. Hodges, S. K. Nelson, A. Stewart, S. Williams, and I. Kloszewska. Identification of cis-regulatory variation influencing protein abundance levels in human plasma, 2012.

S. Ma, X. Song, and J. Huang. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8(1):60, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-60.

N. Malo, O. Libiger, and N. J. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375 – 385, 2008. ISSN 0002-9297. doi: http://dx.doi.org/10.1016/j.ajhg.2007.10.012.

MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, September 2006. ISSN 1087-0156. doi: 10.1038/nbt1239. URL `http://dx.doi.org/10.1038/nbt1239`.

I. W. McKeague and W. Wefelmeyer. Markov chain monte carlo and raoblackwellization. *Journal of Statistical Planning and Inference*, 85(12):171 – 182, 2000. doi: http://dx.doi.org/10.1016/S0378-3758(99)00079-8.

J. Miller, C. Cai, P. Langfelder, D. Geschwind, S. Kurian, D. Salomon, and S. Horvath. Strategies for aggregating gene expression data: The collapserows r function. *BMC Bioinformatics*, 12(1):322, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-322.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. With comments by James Berger and C. L. Mallows and with a reply by the authors. *Journal of the American Statistical Association*, 83 (404):1023–1032, 1988.

A. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. http://ai.stanford.edu/ãng/papers/icml04-l1l2.pdf, 2004.

A. F. Palazzo and T. R. Gregory. The case for junk dna. *PLoS Genetics*, 10 (5):e1004351, May 2014.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

L. Parts, O. Stegle, J. Winn, and R. Durbin. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genetics*, 7(1): e1001276, January 2011. doi: 10.1371/journal.pgen.1001276.

A. Petersen, C. Alvarez, S. DeClaire, and N. L. Tintle. Assessing methods for assigning snps to genes in gene-based tests of association using common variants. *PLoS ONE*, 8(5):e62161, May 2013.

H. P. Piepho. Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49:1165–1176, July 2009. doi: 10.2135/cropsci2008.10.0595.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Shi, J. Wielaard, R. T. Smith, and P. Sajda. Perceptual decision making through the eyes of a large-scale neural model of v1. *Frontiers in Psychology*, 4(161), 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00161.

B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007. ISSN 1095-9203. doi: 10.1126/science.1136678.

Z. Su, J. Marchini, and P. Donnelly. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 2011. doi: 10.1093/bioinformatics/btr341.

S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker. eqed: an efficient method for interpreting eqtl associations using protein networks. *Molecular Systems Biology*, 4(1), 2008. ISSN 1744-4292. doi: 10.1038/msb.2008.4. URL `http://dx.doi.org/10.1038/msb.2008.4`.

M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. Sa-Correia. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*, 34(suppl. 1): D446–D451, 2006. doi: 10.1093/nar/gkj013.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 0028-0836. doi: 10.1038/nature11247. URL `http://dx.doi.org/10.1038/nature11247`.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, October 2005. doi: 10.1038/nature04226.

The MathWorks Inc. MATLAB. Natick, MA, 2000.

The University of Sheffield. Iceberg High Performance Computing (HPC) Facility. http://www.sheffield.ac.uk/wrgrid/iceberg.

P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–41, 2003.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

D. R. Velez, S. J. Fortunato, S. M. Williams, and R. Menon. Interleukin-6 (IL-6) and receptor (IL6-R) gene haplotypes associate with amniotic fluid protein concentrations in preterm birth. *Human Molecular Genetics*, 17(11):1619–30, 2008. ISSN 1460-2083.

J. Wang, A. Platt, R. Upmanyu, S. Germer, G. Lei, C. Rabe, R. Benayed, A. Kenwright, A. Hemmings, M. Martin, and O. Harari. IL-6 pathway-driven investigation of response to IL-6 receptor inhibition in rheumatoid arthritis. *BMJ Open*, 3(8), 2013. doi: 10.1136/bmjopen-2013-003199.

J. C. Whittaker, R. Thompson, and M. C. Denham. Marker-assisted selection using ridge regression. *Genetical Research*, 75:249–252, April 2000. ISSN 1469-5073.

E. A. Winzeler, D. R. Richards, A. R. Conway, A. L. Goldstein, S. Kalman, M. J. McCullough, J. H. McCusker, D. A. Stevens, L. Wodicka, D. J. Lockhart, and R. W. Davis. Direct allelic variation scanning of the yeast genome. *Science*, 281(5380):1194–1197, 1998. doi: 10.1126/science.281.5380.1194.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, 89:82–93, 2011.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25 (6):714–721, 2009. doi: 10.1093/bioinformatics/btp041.

N. Yi and S. Xu. Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*, 179(2):1045–1055, 2008. doi: 10.1534/genetics.107.085589.

G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, and L. Kruglyak R. Mackelprang. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, 35: 57–64, September 2003.

J. Zhu, B. Zhang, E. Smith, B. Drees, R. Brem, L. Kruglyak, R. Bumgarner, and E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854–861, July 2008. doi: 10.1038/ng.167.

J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40 (7):854–861, 2008.