Bayesian variable selection in regression with genetics application

by

Zayed Shahjahan

B.Sc., Dickinson College, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name: **Zayed Shahjahan**

Degree: **Master of Science**

Thesis title: **Bayesian variable selection in regression with genetics application**

Committee: **Chair:** Joan Hu
Professor, Statistics and Actuarial Science

Jinko Graham
Supervisor
Professor, Statistics and Actuarial Science

Lloyd Elliott
Committee Member
Assistant Professor, Statistics and Actuarial Science

Brad McNeney
Examiner
Associate Professor, Statistics and Actuarial Science

Abstract

In this project, we consider a simple new approach to variable selection in linear regression based on the Sum-of-Single-Effects model. The approach is particularly well-suited to big-data settings where variables are highly correlated and effects are sparse. The approach shares the computational simplicity and speed of traditional stepwise methods of variable selection in regression, but instead of selecting a single variable at each step, computes a distribution on variables that captures uncertainty in which variable to select. This uncertainty in variable selection is summarized conveniently by credible sets of variables with an attached probability for the entire set. To illustrate the approach, we apply it to a big-data problem in genetics.

Keywords: variable selection; Bayesian regression; uncertainty quantification; genomic data science

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since the completion of the Human Genome Project and the International HapMap Project in 2003 and 2005 respectively, geneticists have established thousands of associative relationships between genetic variants (usually Single-Nucleotide Variants, or SNVs) and disease traits using Genome-wide Association Studies (GWAS) [3]. What these association studies cannot do is establish a causal link between the variants and the traits they are associated with. In order to establish causality, further study is required. Genetic fine-mapping can be thought of as the step following a GWAS, where regions identified by a GWAS (after exceeding some Bonferroni-esque threshold) are analyzed to ascertain whether or not a causal variant exists in that region.

As such, fine-mapping can be understood as taking the GWAS data that shows these complex associations and essentially, 'untangling' it to find the causal genes. The purpose of finding these causal genes is to home-in on the precise gene mechanisms that are involved in driving the causation and potentially alter the mechanisms to change the trait. Fine-mapping requires three essential components: (1) all the single-nucleotide variants in the region need to be genotyped or imputed with high confidence, (2) stringent quality control of data and (3) large sample sizes to provide enough power to differentiate between variants in high LD [12]. Even with these requirements in place, it is still not possible to precisely identify, with perfect confidence, what the causal variants are. The procedures described in this project attempt to emulate the requirements of a genetic fine-mapping attempt using a simulation of the genomic information for chromosome 1.

In the literature, fine-mapping is formulated as a variable selection problem in regression. A typical analysis of this sort is treated in [10]. What differentiates genetic fine-mapping from a standard variable selection problem is the fact that genetic variants tend to be very highly correlated due to a phenomenon called linkage disequilibrium. This is the non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly [11]. There may be instances where the correlation between variants is as high as 0.99 or even 1 [15].

The most rudimentary procedures in fine-mapping assume only one causal variant in each locus [3]. The assumption that there is only one causal variant in the selected locus is not the most realistic given the direction of the most recent research. Linkage disequilibrium hinders the identification of causal variants at risk loci in fine-mapping studies as at each locus, there are often tens to hundreds of variants tightly linked [2]. The local LD structure can also induce higher association statistics for neighboring variants rather than the causal variants [13]. As such, there exists a need to develop methodologies that can be used to model multiple SNVs simultaneously. Using these methodologies, it should be possible to obtain more robust measures of which SNV is likely to be causal. The Sum of Single Effects (Susie) regression method [15] is one such methodology which aims to compute a posterior distribution on multiple variants.

This study focused on using simulated SNV data for chromosome 1 on 3100 diploid individuals to assess the Susie regression method's performance on a dense genomic region. This was done to answer the question of whether or not variable selection methods utilizing Variational Bayes techniques can be scaled up to detect variants in larger regions than those typically seen in a fine-mapping study. Since chromosome 1 is the largest chromosome in the human genome, the results from this study can be extended to other chromosomes as well.

The remainder of this thesis report is organized as follows: The Methods section will walk the reader through the underlying mathematics of the Susie method and its accompanying fitting procedure, the Iterative Bayesian Step-wise Selection (IBSS). The Data section will dive deeper into the simulation used to generate the variant data for Chromosome 1 as well as the pre-processing, quality control and study design considerations before the Susie method was used. The workflow and the results of the analysis are discussed in the Application section while the limitations of this project and recommendations for future research in this area are addressed in the Discussion section. Also included in the Discussion section is a comparison of the Susie method with a penalized-likelihood based approach that utilizes re-sampling to quantify uncertainty in the variable selection procedure [14].

Chapter 2

Methods

The Susie method is founded on the idea of a single-effects regression (SER) [9]. Essentially, the SER model assumes there is exactly one non-zero regression coefficient. In the genomics setting, this would be akin to assuming only one causal variant. The single-effects regression model is formulated as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon,$$

where $\mathbf{y}$ is the vector of phenotypes for n individuals, $\mathbf{X}$ is the genotype matrix of dimension $n \times p$, $\mathbf{b}$ is a vector of regression coefficients with exactly one non-zero entry and other entries of zero, and $\epsilon \sim N(0, \sigma^2 I_n)$. Let $\mathbf{b} = b\gamma$ where b is a scalar corresponding to the effect size of a single single-nucleotide variant (SNV) and $\gamma = [\gamma_1, ..., \gamma_p]$ is a vector with elements indicating inclusion of a single SNV among the p SNVs i.e. the vector can only take on values 0 or 1.

Assume that the scalar $b \sim N(0, \sigma_0^2)$ and the vector $\gamma \sim \text{Multinomial}(m, \pi_p)$ where $m = 1$ is the number of resamples and π_p is the p-vector of resampling probabilities. These can be thought of as the prior distributions for the Susie method. From this it can be seen that the single SNV with non-zero effect is obtained from a resampling procedure with $m = 1$ draw. Since the sampling distribution is multinomial with π as the resampling probability, prior knowledge of which SNV is more likely to be causal can be incorporated through π.

Under the SER model, the posterior distribution of an SNV being included is given by

$$\gamma|\mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2 \sim \text{Multinomial}(1, \alpha),$$

where α is the p-vector of posterior inclusion probabilities (PIP). In order to construct the vector α, we use Bayes' Factor (BF). The BF of a simple linear-regression model, with slope β, quantifies the support (from the data) that the model has over a competing model. In this setting, the proposed model assumes $\beta \neq 0$ and the null model assumes $\beta = 0$. The BF is given by

$$\text{BF}(\mathbf{x}, \mathbf{y}, \sigma_0^2) = \sqrt{\frac{s^2}{\sigma_0 + s^2}} \exp\left\{\frac{z^2}{2} \times \frac{\sigma_0^2}{\sigma_0^2 + s^2}\right\},$$

where $s^2 = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$ and $z = \hat{\beta}/s = [(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T y]/s$.

Notice that the BF is calculated from summary statistics of a univariate regression. This fact hints at the connection between the single-SNV procedures of a genome-wide association study (GWAS) and the Susie Method. From the BF of the jth SNV we can compute its corresponding PIP with the formula

$$\alpha_j = \frac{\pi_j \text{BF}(\mathbf{x_j}, \mathbf{y}, \boldsymbol{\sigma_0^2})}{\sum_{j'=1}^{p} \pi_{j'} \text{BF}(\mathbf{x_j}, \mathbf{y}, \boldsymbol{\sigma_0^2})},$$

where π_j is the prior probability of the j^{th} SNP. In order to construct credible sets from the PIPs, sort the PIPs in descending order. Then, include SNPs with the highest PIPs in the credible set until the cumulative PIP in the set is below some level ρ.

In order to compute the PIPs for each SNP and the resulting credible sets, estimates of the hyper-parameters σ^2 and σ_0^2 are also required. To estimate these hyper-parameters, consider their marginal likelihood under the SER model:

$$\ell_{\mathrm{SER}}(\mathbf{y}, \sigma_0^2) = p_0(\mathbf{y}|\sigma^2) \sum_{j=1}^{p} \mathrm{BF}(\mathbf{x}_j, \mathbf{y}, \sigma_0^2),$$

where $p_0(\mathbf{y}|\sigma^2)$ is the distribution of $\mathbf{y}$ under the null hypothesis. This likelihood can be maximised over both σ^2 and σ_0^2 numerically.

All of the above expressions can be obtained in the single effects regression (SER) framework. The Susie method is extending the Single-Effects framework to multiple effects. This entails introducing multiple single effects vectors $\mathbf{b}_1, ..., \mathbf{b}_L$. The total number of single effects vectors is L and this is chosen *a priori*. Therefore L is a hyperparameter. The overall effect vector $\mathbf{b}$ is defined as the sum of these single-effects vectors. Just like in the SER framework, each single-effect vector $\mathbf{b}_l = \gamma_l b_l$ where b_l is the scalar effect size for the l^{th} single-effect vector and γ_l is the vector of indicator variables that is distributed Multinomial $(1, \pi)$. Notice here that the number of resamples for the vector of indicator variables is set at 1, just like in the SER method. Everything remains the same as the previously described SER method. To fit the Susie model, Wang et al. [15] propose a new fitting procedure called Iterative Bayesian Step-wise Selection (IBSS).

Given single-effect vectors $\mathbf{b}_1, ...\mathbf{b}_{L-1}$, estimating $\mathbf{b}_L$ involves fitting a SER model. Within each iteration of the IBSS algorithm, the SER model is used to estimate b_l given the current estimates of $b_{l'}$ for $l \neq l'$. In doing so, the IBSS computes a posterior distribution on which variable should be selected. This computation is based on the results of p univariate regressions (like in a GWAS). This posterior distribution captures the uncertainty in the model selection process and this uncertainty is taken into account when computing residuals using a model-averaged estimate of the regression coefficient. The residuals are again fed back into the IBSS algorithm as the new response variable and the process repeats itself until a stopping criterion is reached.

The IBSS is a hill-climbing algorithm that optimizes a variational approximation to the posterior distribution for $\mathbf{b}_1, ...\mathbf{b}_L$. The idea is to find an approximation $q(\mathbf{b}_1, ...\mathbf{b}_L)$ to the posterior $p_{post} = p(\mathbf{b}_1, ...\mathbf{b}_L|X, y, \sigma^2, \sigma_0^2)$. This can be done by minimizing $D_{KL}(q, p_{post})$ where D_{KL} is the Kullbeck-Leibler (KL) Divergence between the posterior distribution and the approximation to the posterior distribution. KL-Divergence is a statistical measure of how one probability distribution is different from another. In this case, the idea is to reduce the distance between the posterior distribution and the IBSS approximation to the posterior.

However, in practice, $D_{KL}(q, p_{post})$ is difficult to calculate. As a result, Wang et al (2020), reformulate this in terms of an Evidence Lower Bound (ELBO) function, $F(q, \sigma^2, \sigma_0^2)$, as

$$D_{KL}(q, p_{post}) = \log p(\mathbf{y}, \sigma^2, \sigma_0^2) - F(q, \sigma^2, \sigma_0^2).$$

Since q appears only in the ELBO function $F(q, \sigma^2, \sigma_0^2)$, minimizing D_{KL} over q amounts to maximizing F over q. Under the variational approximation, the posterior $q \approx \prod_{l=1}^{L} q_l(b_l)$ so that optimizing over q becomes finding $\max_{q_1,...,q_L} F(q_1, ..., q_L; \sigma^2, \sigma_0^2)$. Where $1, ...L$ indexes the single effects. Note that maximizing over an individual q_l reduces to fitting an SER model where the phenotype vector is replaced by the l^{th} residual vector.

To be precise, consider the following set-up:

$$\mathbf{y} \sim p(\mathbf{y}|\mathbf{b}, \theta)$$
$$\mathbf{b} \sim g(\mathbf{b}),$$

where $\mathbf{b}$ is a vector of latent variables, θ denotes other additional parameters that need to be estimated, $p(\cdot)$ represents the likelihood for $\mathbf{b}$ and θ and $g(\cdot)$ the prior distribution on the latent variable vector $\mathbf{b}$. Obtain estimates $\hat{g}$ and $\hat{\theta}$ via maximum likelihood where

$$\ell(g, \theta; \mathbf{y}) = \log \int p(\mathbf{y}|\mathbf{b}, \theta) g(\mathbf{b}) d\mathbf{b}.$$

Using the estimates $\hat{g}$ and $\hat{\theta}$, compute the posterior distribution of $\mathbf{b}$ as $\hat{p}_{post} = p(\mathbf{b}|\mathbf{y}, g, \theta) \propto p(\mathbf{y}|\mathbf{b}, \theta) g(\mathbf{b})$. The two steps of (i) estimating $\hat{g}$ and $\hat{\theta}$ and (ii) computing the posterior distribution for $\mathbf{b}$ can be expressed as

$$\begin{aligned}(\hat{p}_{post}, \hat{g}, \hat{\theta}) &= \underset{g \in \mathcal{G}, \theta \in \Theta, q}{\operatorname{argmax}} F(q, g, \theta, \mathbf{y}) \\ &= \underset{g \in \mathcal{G}, \theta \in \Theta, q}{\operatorname{argmax}} [\ell(g, \theta; \mathbf{y}) - D_{KL}(q, \hat{p}_{post})] \\ &= \underset{g \in \mathcal{G}, \theta \in \Theta, q}{\operatorname{argmax}} \left[\ell(g, \theta; \mathbf{y}) - \int q(\mathbf{b}) \log \left(\frac{q(\mathbf{b})}{\hat{p}_{post}(\mathbf{b})} \right) d\mathbf{b} \right],\end{aligned}$$

where ℓ is the marginal log-likelihood, $D_{KL}(q, \hat{p}_{post})$ is the Kullback-Leibler (KL) divergence from q to $\hat{p}_{post}$ and the optimization of q in the first line of the equation is over all possible distributions on $\mathbf{b}$. The function F is called the evidence lower bound or ELBO. To make the optimization of the ELBO function tractable, Wang et al. [15] proposed restricting the optimization of the posterior over all possible distributions on $\mathbf{b}$ to $q \in \mathcal{Q}$ where $\mathcal{Q}$ is a suitably chosen set of distributions. This procedure is called variational approximation since it can be interpreted as finding the best possible approximation to the posterior distribution from a set of distributions $\mathcal{Q}$. Thus, the original optimization problem becomes $\operatorname{argmax}_{g \in \mathcal{G}, \theta \in \Theta, q \in \mathcal{Q}} F(q, g, \theta; \mathbf{y})$, with the additional constraint that $q \in \mathcal{Q}$. The ELBO function can be decomposed as

$$
\begin{aligned}
F(q, g, \theta; \mathbf{y}) &= \mathbb{E}_q\left[\frac{\log p(\mathbf{y}|\mathbf{b}, g, \theta)}{q(\mathbf{b})}\right] \\
&= \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{b}, g, \theta)] + \mathbb{E}_q\left[log\frac{g(\mathbf{b})}{q(\mathbf{b})}\right]
\end{aligned}
$$

Now consider an additive model:

$$
\begin{aligned}
&y = \sum_{l=1}^{L} \mu_l + \epsilon \\
&\epsilon \sim N(0, \sigma^2 I_n) \\
&\mu_l \sim g_l.
\end{aligned}
$$

The Susie model is an example, where $\mu_l = \mathbf{X}\mathbf{b}_l$ for some $l \in \{1, ...L\}$ and each g_l is the prior distribution for $\mathbf{b}_l$. Define a simple model $\mathcal{M}_l$ which is derived from the original model $\mathcal{M}$ by setting $\mu_{l'} = 0$ for all $l' \neq l$. Therefore $\mathcal{M}_l$ is the model that includes only the l^{th} additive term. With respect to the Susie model, this simpler model corresponds to Single Effects Regression (SER). The Susie method hinges on the idea that the model $\mathcal{M}$ can be fit if each of the simpler models $\mathcal{M}_l$ can be fit. In order to fit each of the simpler models, allow the class of distributions on $\mathbf{X}\mathbf{b}_1, ...\mathbf{X}\mathbf{b}_L$ to factorize over $\mathbf{X}\mathbf{b}_1, ...\mathbf{X}\mathbf{b}_L$. This leads to:

$$
q(\mathbf{X}\mathbf{b}_1, ...\mathbf{X}\mathbf{b}_L) = \prod_{l=1}^{L} q_l(\mathbf{X}\mathbf{b}_l).
$$

Combining the above result with the decomposition of the ELBO function from earlier, the following expression for the ELBO function is obtained:

$$
\begin{aligned}
&F(q,g,\theta;\mathbf{y}) = \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{b},\theta)] + \mathbb{E}_q\left[\log \frac{g(\mathbf{b})}{q(\mathbf{b})}\right] \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_q[||\mathbf{y} - \sum_{l=1}^{L}\mathbf{X}\mathbf{b}_l||^2] + \sum_{l=1}^{L}\mathbb{E}_{q_l}\left[\log \frac{g_l(\mathbf{X}\mathbf{b}_l)}{q_l(\mathbf{X}\mathbf{b}_l)}\right]
\end{aligned}
$$

Notice that the expected residual sum of squares is the term $\mathbb{E}_q[||\mathbf{y}-\sum_{l=1}^{L}\mathbf{X}\mathbf{b}_l||^2]$ where $||.||$ denotes the Euclidean distance between the phenotype vector and the predicted phenotype vector obtained from $\mathbf{X}\mathbf{b}_l$. The q_l and g_l that maximizes the ELBO function for the Susie method can be found by maximizing the ELBO function for the SER model when the phenotype vector $\mathbf{y}$ is replaced by its expected residuals. The proof of this is provided in Proposition A1 of Appendix B.4 in the original paper by Wang et al. [15]. The Proposition A1 can be thought of as the crux of the Susie method and the reason it tends to be computationally efficient.

In sum, the Susie method uses a variational approximation within a Bayesian variable selection frame-work to provide a measure of uncertainty regarding which SNP is likely to be causal. This measure of uncertainty (Posterior inclusion probability) is obtained via summary statistics of a univariate regression and passing the summary statistics into a formulation of the Bayes Factor. Using PIPs, the Susie method constructs Credible sets. These sets are superficially similar to confidence sets but have a probabilistic interpretation regarding the causal SNPs. The method relies on a simpler and easier to fit single effects regression (SER) model as a building block for the fitting procedure. The fitting procedure (IBSS) relies on the fact that fitting the Susie model implies fitting L simpler SER models on the residuals derived from the original phenotype vector.

To test the Susie method in a high dimensional setting, a simulated dataset of sequencing data was used. These data will be described in the next section.

Chapter 3

Data

Simulated sequencing data were used to assess the Susie method's performance. The python library `msprime` was used to simulate sequencing data for Chromosome 1 [4]. The SNV data were simulated using a backwards Wright-Fisher model, followed by a coalescent model to approximate the ancestry further back in time [5]. The Wright-Fisher reproductive model assumes that generations do not overlap and that each copy of the gene found in the new generation is drawn independently at random from all copies of the gene in the old generation. The population consisted of 3100 diploid individuals corresponding to 6200 sequences. The mutation and recombination rates were set to be 1×10^{-8} per base-pair per generation.

Chromosome 1 consists of 249 million base-pairs and human genes tend to consist of a median of 26288 base-pairs. Consequently, the sequence data for chromosome 1 was divided into 9427 (24900000/26288) non-overlapping regions. Of these 9427 regions, 2000 were randomly selected to be genes. To better summarize distribution of polymorphisms within this population, consider Figure 3.1 which shows the distribution of the derived (mutated) allele frequency of the population. The allele frequency can be thought of as the amount of genetic variation within a locus expressed as a percentage.

The sequence data was then randomly paired with 3100 individuals. The subsequent genotype matrix X had dimensions 3100×287668. Only possible entries for

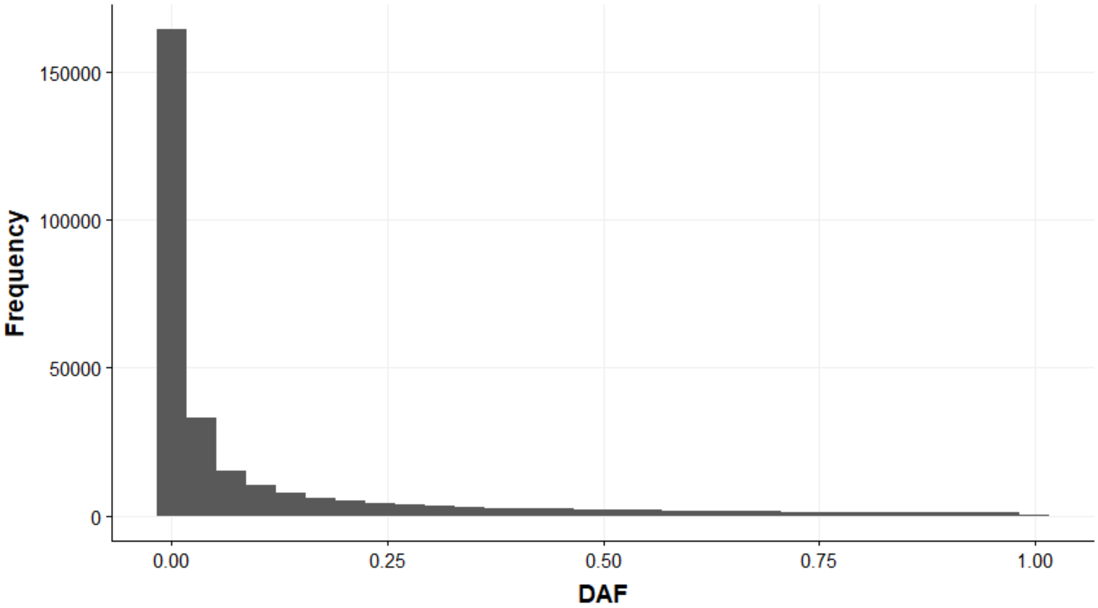

Figure 3.1: Histogram of derived allele frequencies in the genotype matrix

the matrix X were $0, 1$ or 2. These corresponded to having no, one or two copies of the derived allele.

The phenotype of interest was human height in centimetres. The expected value of the phenotype (human height in centimetres) is modelled using:

$$\mathbb{E}[y|X] = \beta_0 + \sum_{j=1}^{m} \beta_j X_j,$$

where $\mathbb{E}[y|X]$ denotes the expected height of an adult given the vector $\mathbf{X}$, β_0 is the baseline height and β_j is the effect size of the j^{th} causal SNV. The error terms of the linear model were normally distributed with mean 0 and variance σ^2. Based on the height distribution of Canadians, σ was set at 10.5 cm.

According to Park et al. [6], the squared scalar effect size (β^2) and the population derived allele frequency are inversely related as:

$$\beta^2 = \frac{b^2}{f(1-f)},$$

where b^2 is a random error term obtained from a Laplace distribution with location parameter 0 and shape parameter 1. Further, the contribution of each causal SNV to the total genetic variation in the phenotype is:

$$\begin{aligned} g &= 2\beta^2 f(1-f) \\ &= 2 \times \frac{b^2}{f(1-f)} \times f(1-f) \\ &= 2b^2. \end{aligned}$$

From [16], height was assumed to have a heritability of 80%. That is, 80% of the variation in height was due to genetic factors. We are assuming an omnigenetic relationship between our SNVs and phenotype. Since Chromosome 1 contains about 7% of the genes in the human genome, the contribution of its genes to the variability of height was calculated as $0.07 \times 80\% \approx 6\%$. That is, the heritability of height, Y,

$$H = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)},$$

where the random variable X encodes the genetic information on chromosome 1, is $H = 0.06$. Given $H = 0.06$ and $Var(Y) = 10.5^2$, the genetic variation $Var(\mathbb{E}[Y|X])$ of height was calculated to be 6.6 cm^2. Among the 2000 simulated genes on chromosome 1, 150 were randomly chosen as candidates to contain causal SNVs. Causal SNVs were randomly selected from these 150 candidate genes until the sum of their contribution to the genetic variance was 6.6 cm^2. As a result, 22 SNVs in 19 genes were randomly selected to be causal. Figure 3.2 represents the proportion of genetic variation in height attributable to each of the 22 causal SNVs.

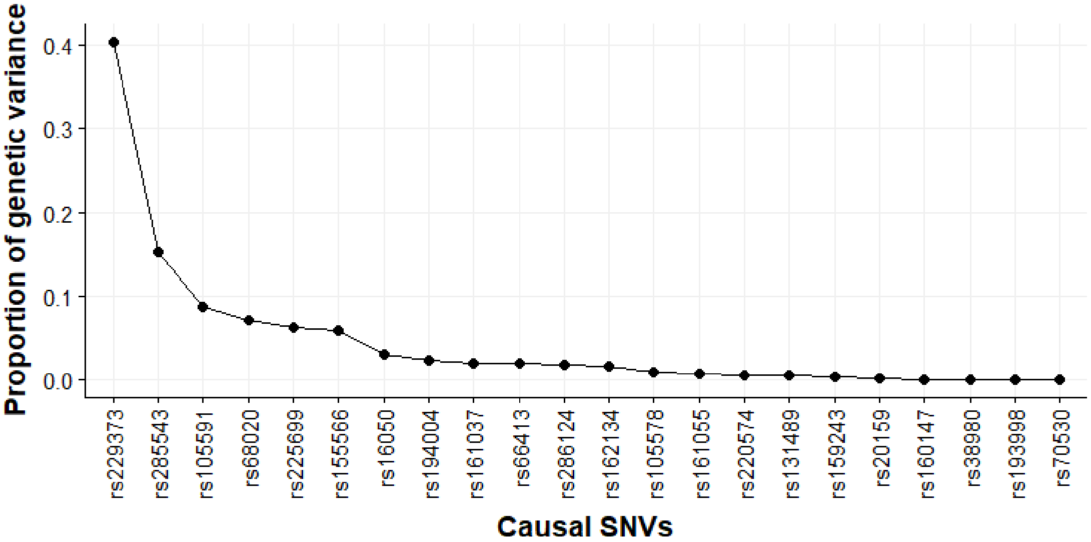

Figure 3.2: Proportion of genetic variation attributable to the Causal SNVs

After attributing the genetic variation to the 22 causal SNVs, β_0 in the height model was determined to be the difference between the average height of a Canadian male (173.19 cm) and the genetically attributable height calculated from the expression $\sum_{j=1}^{22} \beta_j X_j$.

The resulting genotype and height data was grouped into four different datasets stored as text files. The `gene.matrix.txt` file contained the genotype-dosage matrix $\mathbf{X}$. The gene support matrix contained the gene IDs for the 2000 genes as well as their physical start and end positions in base pairs. The phenotype dataset contained the height information for the 3100 individuals in centimetres and feet, along with the error term and related quantities in the linear model for height. The SNV support dataset contained information regarding the SNV positions in base-pairs, whether or not the SNV was causal and their derived-allele frequencies.

In order to work with these data efficiently in R, the information in these four files was combined into a BGData object from the BGData package [1]. Before the data could be analyzed using the Susie method, pre-processing and data quality control steps were performed. These are described below. The quality control, pre-processing and the assessment of the Susie method were all performed in R [7]. All the scripts for this analysis are freely available at https://github.com/SFUStatgen/ZJ/tree/main/Thesis/DataScripts.

The results of a chromosome-wide scan of association before performing the Quality control steps are provided in Figure 3.3. The striations above 15 on the y-axis involves a number of SNVs that are in perfect LD. Notice that non-gene regions of the Chromosome are significantly denser than the gene regions. The causal genes in the Chromosome account for an even smaller proportion than the gene regions. Despite this, 4 causal SNVs that account for most of the genetic variation in the phenotype are detected after adjusting for multiple testing (blue horizontal line). However, these are detected along with a host of other non-causal SNVs. In a real-world study the causal SNVs would be indistinguishable from the non-causal. Figure 3.4 highlights the glaring difference in proportions of SNVs between the causal gene, non-causal gene and non-gene regions in Chromosome 1.

The first quality control step involved filtering out SNVs in the non-gene regions that had minor allele frequencies (MAF) less than 0.01. This is done in the code chunk below.

```
#perform a chromosome-wide scan of association
res1 <- GWAS(formula = height.cm ~ 1,
             data = DATA,
             method = "lsfit")
#res1 contains results of a chromosome-wide scan of association without quality control
gene_summaries <- BGData::summarize(DATA@geno)
snvs <- as.data.frame(rownames(gene_summaries))
```

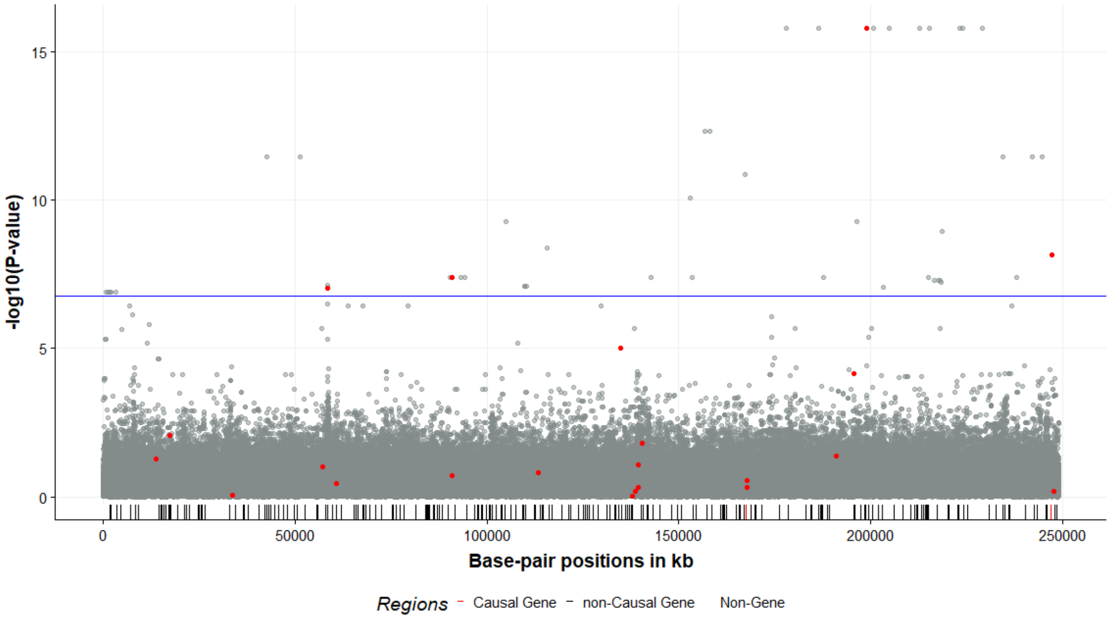

Figure 3.3: Manhattan plot of an initial Chromosome-wide scan of association

```
colnames(snvs)[1] <- "snp"

gene_summaries <- cbind(snvs, gene_summaries)

#Merge Allele frequency information with the results of univariate regressions for later

gwasdata1 <- merge(res1, gene_summaries, by = "snp")

#Obtain minor allele frequencies

gwasdata1 <- gwasdata1 %>%

  mutate(maf = ifelse(allele_freq > 0.5,

                      1 - allele_freq, allele_freq))

#Filter the SNPs in the non-exome regions for MAF > 0.01

exome <- gwasdata1 %>%

  filter(region == "Causal Gene" | region == "non-Causal Gene")

nongene <- gwasdata1 %>% filter(region == "Non-Gene")

filtered.nongene <- nongene %>% filter(maf > 0.01)

filtered.SNVs <- rbind(exome,filtered.nongene)

filtered.SNVs %>% arrange(snp)

filsnps <- filtered.SNVs %>% select(snp)

#Making the necessary adjustments in the BGData object
```

Table 3.1: Information on Causal SNVs

SNV	gene.ID	position (bp)	MAF	$\hat{\beta}$	gv^a	prop.gv^b
rs16050	103	13812836	0.4885	-0.63	0.199	0.030
rs20159	134	17300226	0.0134	0.70	0.013	0.002
rs38980	253	33573167	0.3663	0.02	0.000	0.000
rs66413	458	57093607	0.0013	7.10	0.130	0.020
rs68020	474	58512510	0.3029	1.05	0.469	0.071
rs70530	491	60751870	0.2479	-0.01	0.000	0.000
rs105578	740	90888746	0.0021	3.68	0.057	0.009
rs105591	740	90900340	0.0002	42.51	0.583	0.088
rs131489	932	113319820	0.0619	0.56	0.036	0.005
rs155566	1110	134739717	0.0010	14.22	0.391	0.059
rs159243	1132	137934895	0.2800	-0.23	0.022	0.003
rs160147	1141	138722190	0.0032	0.54	0.002	0.000
rs161037	1147	139542705	0.0002	-20.57	0.137	0.021
rs161055	1147	139557027	0.1937	-0.39	0.048	0.007
rs162134	1156	140483427	0.3469	0.49	0.110	0.017
rs193998	1366	167670283	0.0002	-0.67	0.000	0.000
rs194004	1366	167676775	0.2245	0.68	0.161	0.024
rs220574	1538	191142554	0.0115	-1.32	0.040	0.006
rs225699	1577	195651125	0.0437	2.22	0.410	0.062
rs229373	1606	198898287	0.0006	-45.62	2.684	0.405
rs285543	1983	247120904	0.0060	-9.26	1.017	0.153
rs286124	1985	247641972	0.0865	0.89	0.125	0.019

[a] gv=genetic variance contribution.

[b] prop.gv=proportion of contribution to total genetic variance.

```
filtered.Geno <- DATA@geno[,unlist(filsnps)]

DATA2 <- DATA

DATA2@geno <- filtered.Geno

#The code used to adjust the map file was not included
```

From Figure 3.4, it can also be seen that most of the SNVs in non-gene regions tend to have very low values for MAF (see table 3.1). At the same time 12 of the 22 causal SNVs also have very low MAF values. To avoid losing information from the gene regions, only SNVs in non-gene regions were filtered based on MAF values.

The filtering resulted in a reduced genotype matrix with dimensions 3100×170551, corresponding to a 40 percent decrease in the number of SNVs available for analysis.

After the filtering process, a second chromosome-wide scan of association was performed and the results of this test are shown in Figure 4.1. Notice that now there are far fewer SNVs that have p-values above the Bonferroni-adjusted threshold. In addition to the analysis of the fully processed and quality-control adjusted genotype matrix, the Susie regression method was performed on this partially quality-controlled dataset. The results of this analysis are presented in the Results section for comparison.

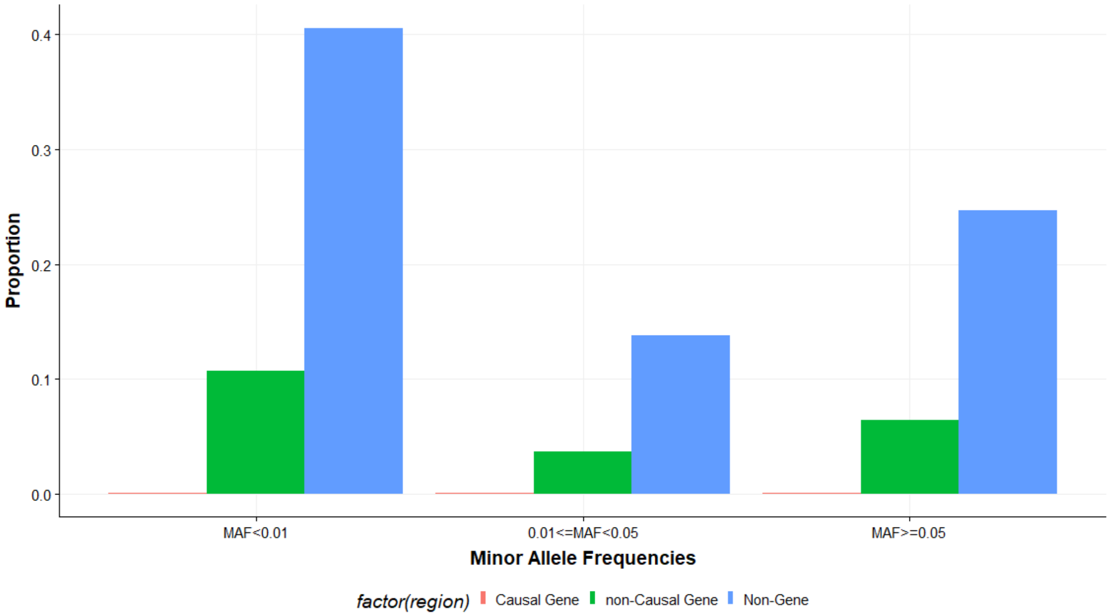

Figure 3.4: Proportions of SNVs on the chromosome that are causal, non-causal and non-gene along with their minor allele frequencies grouped into 3 categories

The second quality control step involved further reducing the number SNVs to be consistent with a study that uses exome-sequencing in combination with array genotyping of common SNVs or single-nucleotide polmorphisms (SNPs). In exome-sequencing only the protein-coding regions (exomes) are sequenced. The process of spacing non-gene SNPs in the chromosome can be intuitively thought of as a form of systematic sampling of exomes SNPs spaced roughly 105 kilobases apart. This corresponds to the average spacing on a SNP-array chip. The code to wrangle the genotype matrix to conform to this study design is shown below.

```
#Create an "array" of nongenic SNPs spaced roughly 105kb.
#Get the common SNVs with MAF>1%.
MAFs <- pmin(SNV_support$DAF,
           1-SNV_support$DAF)
SNPs <- SNV_support[which(MAFs>.01),]
#Determine the number of SNPs on the array with spacing 105 kb
numarray<-ceiling((SNPs[nrow(SNPs),
                                   "SNV.posn"]-SNPs[1,
                                                    "SNV.posn"])/105000)
#Third, set up positions of hypothetical SNPs
#on an idealized grid with spacing exactly 105 kb
targetgrid <- as.numeric(SNPs[1,"SNV.posn"])+105000*(0:(numarray-1))
targetgrid <- c(targetgrid, as.numeric(SNPs[nrow(SNPs),"SNV.posn"]))
#Write a function that returns the index of the SNP that is closest
#to the position specified by an index on the idealized grid.
#Make copy of SNV positions and temporarily change position of
#genic SNPs to huge values so that they’re never the closest to
#a grid point
snv.posn <- unlist(SNPs[,"SNV.posn"])
snv.posn[!is.na(SNPs[,"gene.ID"])] <- Inf
findSNP<-function(index){
  which.min(abs((targetgrid[index]-snv.posn)))
  }
#Apply the function to find the non-genic SNPs
#closest to those on the idealized grid.
arraySNPs<-sapply(1:numarray,findSNP)
# Sixth, set up the information for the array.
array_support<-SNPs[arraySNPs,]
keep<-which(is.na(array_support[,"gene.ID"]))
array_support<-array_support[keep,]
```

```
spacings<-diff(unlist(array_support[,"SNV.posn"]) )
#For the exome-sequencing step,
# First, pare down SNV_support into newSNV_support
arraySNV.IDs<-SNPs[arraySNPs, "SNV.ID"]
include<-((unlist(SNV_support[, "SNV.ID"]) %in%
             unlist(arraySNV.IDs)) |
            !is.na(SNV_support[,"gene.ID"]))
newSNV_support<-SNV_support[which(include),]
# Second, pare down genos into newgenos
cols <- which(include)+1
newgenos<-subset(genos,,cols)
```

Using the `newgenos` matrix and the corresponding SNP.support dataframe, a new `BGData` object was created. The code for that is available in the Github repository for this project.

After these steps were completed, the resulting genotype matrix had dimensions 3100×62534. The Susie regression method was used on this matrix to identify SNVs associated with the phenotype. The results of this analysis are available in the next section.

Chapter 4

Application

Figure 4.1 shows the results of a second Chromosome-wide scan of univariate association on the reduced data obtain after the pre-processing, quality-control and exome-sequencing design steps described in the previous section. Now, the plot of the negative log-10 p-values allows for a better understanding of which loci could potentially be causal.

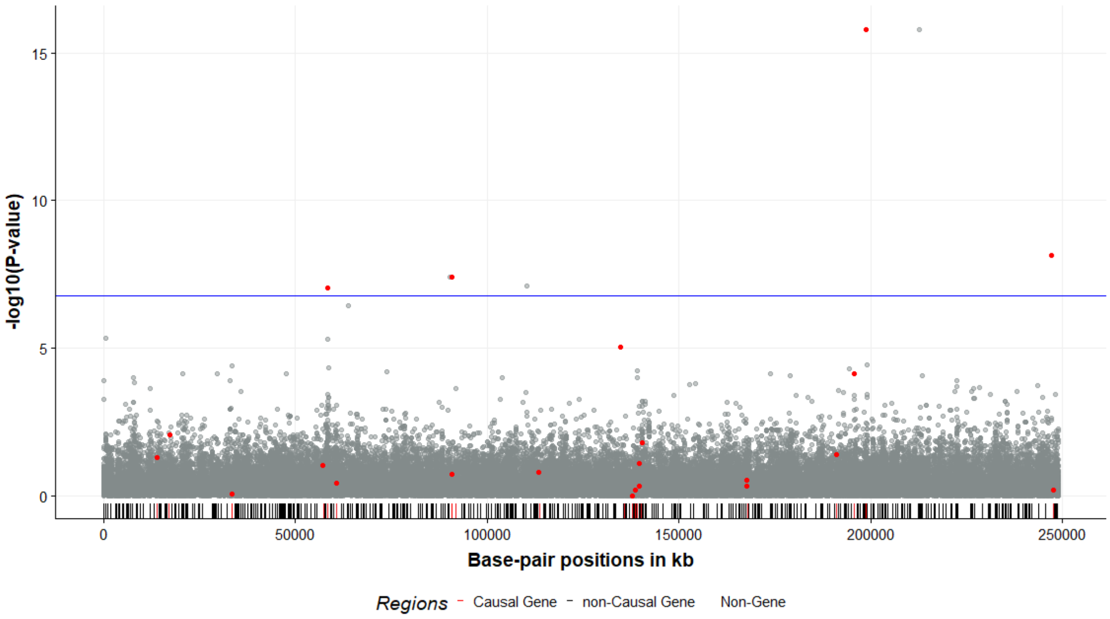

Figure 4.1: Manhattan plot of Chromosome 1 after Pre-processing and quality control steps

However, the results of the chromosome-wide scan only establish which SNVs are associated with the phenotype and the strength of the univariate association via p-values. These univariate p-values cannot be used to make statements regarding which SNVs are jointly associated with the variation in the phenotype. Statements of that sort require a quantification of the uncertainty in the variable selection process. The credible sets obtained from the Susie regression provide this kind of statement. The Susie Regression method [15] was therefore applied to the reduced data. The interpretation of its credible sets is unrelated to univariate tests of association. Rather the interpretation is probabilistic, and focuses on the likelihood that at least one of the SNVs in the credible set is an effect variable in the multiple regression for the phenotype. We applied the following procedure to obtain credible sets:

1. Fit a Susie regression model with the genotype dosages as potential effect variables and the phenotype as the dependent variable.

2. If more than one credible set exists, choose the set with the smallest number of SNVs and record its posterior inclusion probability or PIP.

3. Fit a univariate regression to the phenotype using a representative SNV from the credible set. For example, if the set contains more than one SNV, choose one SNV at random to represent the set.

4. Obtain residuals from the univariate regression.

5. Remove the SNVs in the credible set and any associated SNVs from consideration in the genotype dosage matrix. Here, associated SNVs are defined as those outside the credible set but from the same genes as those in the credible set.

6. Fit a Susie regression using the new genotype dosage matrix and the residuals as the new dependent variable.

7. Continue this process until no credible sets are reported.

8. If no credible sets are reported, lower the requested coverage probability and try again.

All scripts for the analysis in this section can be found at `https://github.com/SFUStatgen/ZJ/tree/main/Thesis` in the folder `AnalysisScripts`

Initially, we fit a Susie regression with the default settings. The resulting PIPs are plotted in Figure 4.2. This first run yielded two credible sets. The information for each of the SNVs in these credible sets is summarized in Table 4.1, along with the rest of the SNVs detected by the Susie method from the start to the finish of the project. The first credible set (CS1) contained one causal SNV, rs229373 and one non-causal SNV, rs245524. This is due to extremely high LD between these two SNVs. The correlation between rs229373 and rs245524 was calculated to be 1 using Pearson's r. Both of these SNVs have a PIP of 0.5001592. The second credible set (CS2) contained only the causal SNV rs285543. The posterior coverage probabilities for CS1 and CS2 were 1 and 0.99, respectively. The coverage probability was requested to be at least 0.95. Notice that while CS1 had a higher coverage probability than CS2, this higher coverage probability was due to both SNVs in the set contributing equally to the coverage. This indicates some degree of ambiguity in the selection process. However, even though CS2 had a slightly lower coverage probability, only one SNV, rs285543, contributed to all of it. Since the identity of the effect SNV was unambiguous, CS2 and specifically rs285543 was chosen for further analysis.

For the next step, we fit a univariate regression with height as the dependent variable and the dosage of rs285543 as the independent variable. The residuals from this univariate regression were then used as a new phenotype for a Susie regression. Before fitting this regression involving the residuals, all the SNVs in the gene with ID 1983 (containing rs285543) were removed.

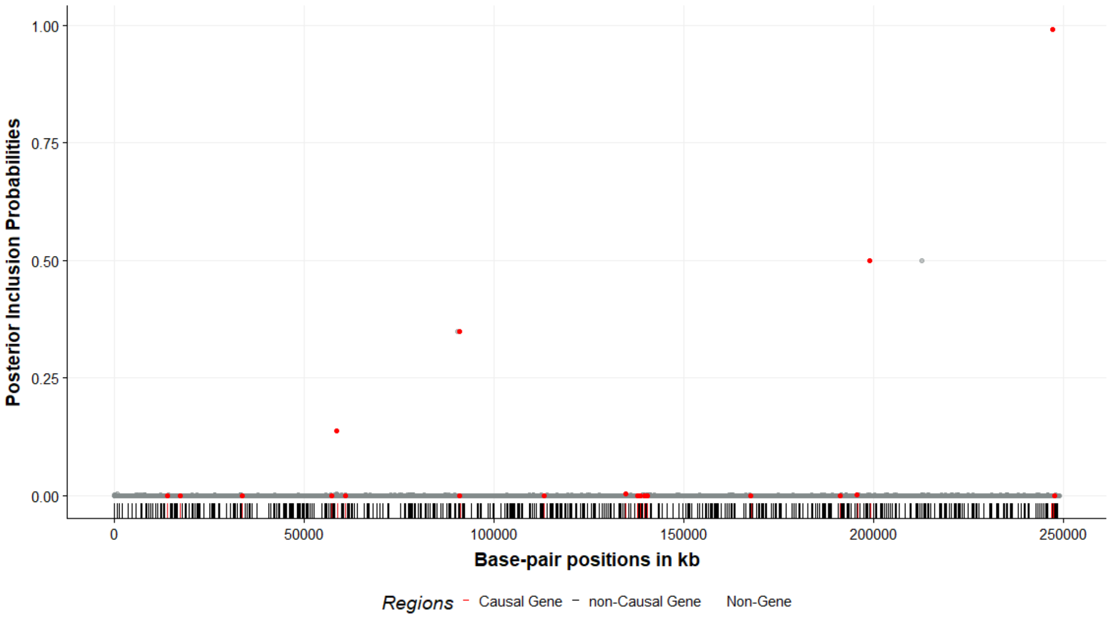

Figure 4.2: Plot of posterior inclusion probabilities after the first run of the Susie method

From the second run of the Susie method, only one credible set was obtained. As expected this contained rs229373 and rs245524, the SNVs reported in the first run. Again, both SNVs had PIPs of approximately 0.5, indicating that the ambiguity in the selection process was not improved by the adjustment. This time, a univariate regression was performed with the residuals of rs285543 as the dependent variable and the dosage of rs245524. Note that rs245524 is a non-causal SNV in linkage disequilibrium (LD) with the actual causal SNV rs229373. However, in order to simulate a real-world investigation, the non-causal SNV was chosen at random from the two SNVs in the credible set. When adjusting for this SNV in the genotype matrix, all SNVs from both genes (IDs 1606 and 1725) represented by the credible set were removed from the genotype dosage matrix to adjust for the LD.

For the third run of the Susie method, the dependent variable was the residual vector of rs245524. After the third run of the Susie method, no credible sets were reported at the 95% coverage level. Coverage was then reduced to 85% but to no avail. The reason for this is evident in the plot of the posterior inclusion probabilities after the third run, as shown in Figure 4.3. Since the top SNVs in this third run have a combined coverage that is less than 95%, a much lower requested coverage probability would yield credible sets in this setting. Subsequently, coverage was reduced to 50%. This action yielded one credible set with coverage probability of approximately 70%. The new credible set contained two SNVs, rs105591 which was causal and rs104853 which was not. As was the case in the previous run, both of these SNVs had PIPs equal to 0.3510689, indicating the ambiguity in identifying which was an effect SNV.

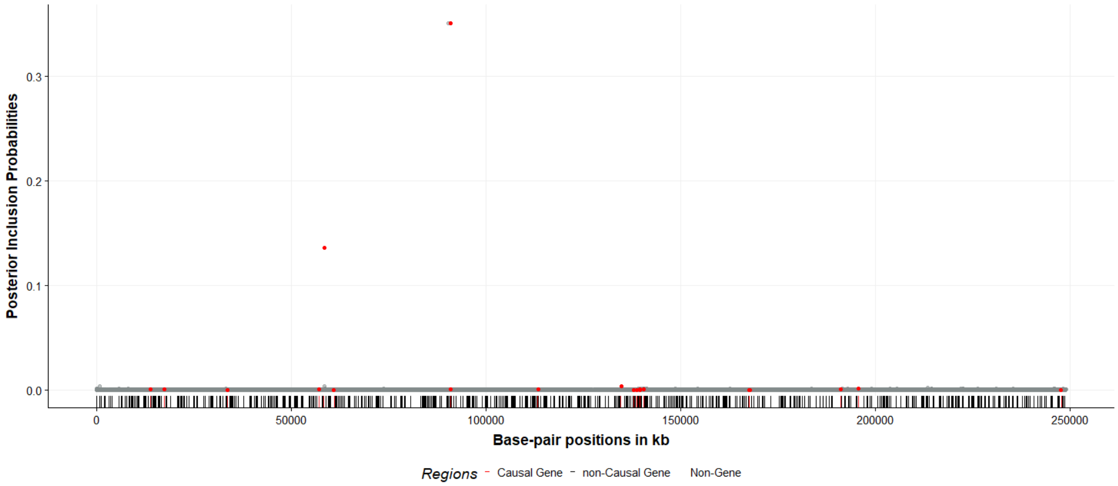

Figure 4.3: Plot of PIPs after the third run of the Susie method highlighting the reduction in the highest scoring SNV.

For the fourth run of the Susie method, the residuals of rs245524 were now the dependent variable in a univariate regression with rs105591 as the independent variable. The residuals of this regression would then be re-fed into the Susie function with the genotype dosage matrix adjusted appropriately. For this run, requested coverage was further reduced to 10% since no credible sets were reported at the 50% coverage level. Only one credible set was reported after the change in requested coverage and this contained causal SNV rs68020. The SNV had a PIP of 0.4.

On subsequent runs, using the residuals from regressing the residuals of rs105591 with the dosage of rs68020 as the phenotype, the Susie method failed to report any new credible sets. This was the case even after reducing the requested coverage probability to a ridiculously low 0.05. Upon further inspection, after adjusting for rs68020 all PIPs had been reduced to 0. This indicated that the method could not find the remaining 18 causal SNVs after adjusting for the first 4. Therefore, as summarized in Table 4.1, the causal SNVs identified by Susie in this study were: rs229373, rs285543, rs105591 and rs68020. At first it may seem that Susie was only 20% successful. However, note that the four identified SNVs were also four of the top five SNPs in terms of proportion of genetic variation.

After no more credible sets could be obtained, two different strategies were implemented to recover the remaining causal SNVs. These are highlighted below.

The first strategy entailed using a multivariable regression model to simultaneously adjust for the discovered SNPs, instead of the earlier strategy of making adjustments sequentially. The residuals from a multivariable regression on height using rs285543, rs245524, rs105591, rs68020 were fed to the Susie method as the dependent variable. However, this method also failed to produce a credible set at the 10% level.

The second strategy entailed removing the genes that were associated with the previously discovered SNVs (both causal and non-causal) and then applying Susie regression to the height phenotype instead of the residuals. This procedure yielded a credible set at the 10% level. This credible set contained the SNV rs128017, which was non-causal. However, this was still treated as a discovery and the residuals from this SNV were re-fed into the Susie function. Upon re-fitting Susie with these new residuals, no new credible sets were reported at the 5% level. The interpretations, implications, limitations and recommendations in light of these findings are discussed in the Discussion section.

Table 4.1: Information on SNVs detected by Susie

SNV	causal	gene.ID	base-pair	prop.gv	CS[a]	RC[b]	Susie run	PIP[c,d]
rs229373	Yes	1606	198898287	0.405	L1	95%	first, second	0.50
rs245524	No	1725	212777762	0	L1	95%	first, second	0.50
rs285543	Yes	1983	247120904	0.153	L2	95%	first	0.99
rs104853	No	738	90313575	0	L1	50%	third	0.35
rs105591	Yes	740	90900339	0.088	L1	50%	third	0.35
rs68020	Yes	474	58512510	0.071	L1	10%	fourth	0.44

[a] CS=Credible Set.

[b] RC=Requested coverage probability.

[c] PIP=Posterior Inclusion Probability.

[d] first and second runs had same PIPs for rs229373 and rs245524

Chapter 5

Discussion

A Bayesian variable-selection method, Susie regression (Wang et al. 2020), was applied to chromosome-wide data on SNV genotypes. The goal was to assess the method's performance in finding causal variants in a dense genomic region. After quality control and pre-processing steps, the region consisted of 62,534 SNVs for 3100 diploid individuals. The phenotype of interest mimicked human height. Sex and ethnicity were not incorporated in the data simulation procedure. The Susie method was able to correctly identify four of the twenty-two causal variants in the data. These four SNVs contributed disproportionately to the total genetic variation in the phenotype. The genotype distributions for the top SNVs are provided in Table 5.1.

After accounting for the first four causal SNVs, the Susie method failed to report more credible sets. The Posterior Inclusion Probabilities (PIPs) of the remaining SNVs were reduced to 0. The locations of the detected SNVs are highlighted in Figure 5.1. Two different strategies were then implemented to coax out more SNVs, to no avail.

Table 5.1: Gene dosage breakdown of the top five causal SNVs

SNP	0	1	2	gv
rs229373	3096	4	0	40.00%
rs285543	3063	37	0	15.00%
rs105591	3099	1	0	8.80%
rs68020	292	1294	1514	7.10%
rs225699	2836	257	7	6.20%

However, given the rarity of the discovered SNVs and their lower contributions to genetic variation, the results suggest that the Susie method can be used in settings where traditional fine-mapping procedures come up short. It also seems to be the case that the Susie method can be extended to larger regions than previously possible in genetic fine-mapping applications while still yielding acceptable credible sets.

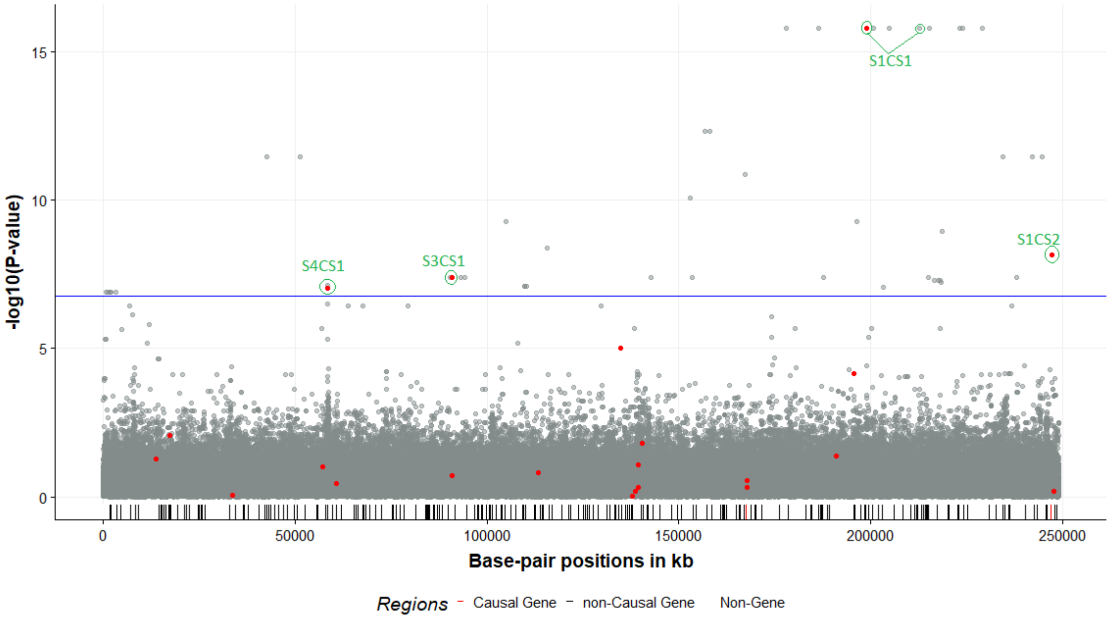

Figure 5.1: Plot showing the positions of the Susie-discovered SNVs

Genetic fine-mapping procedures rely on genome- and/or chromosome-wide tests of association to identify regions that are strongly associated with the trait under investigation. These regions of high association are 'isolated' for further investigation. Bear in mind that true causality can only be established using a controlled experiment. In the process of dividing regions in a chromosome into subregions for further analysis, linkage disequilibrium (LD) between distant variants cannot be taken into consideration [8]. LD structures in human populations tend to be fairly complex [11]. This complexity can inhibit a fine-mapping procedure's ability to accurately find causal variants. It could be the case that top SNVs (according to negative log-10 p-values) were simply in high LD with actual causal SNVs. This problem highlights the need for fine-mapping procedures that can be used on larger and denser genomic regions while yielding accurate results.

The results from this investigation indicate that it is possible extend fine-mapping procedures to (at least) a chromosome-wide region. These results also indicate that variable selection methods based on variational approximations can be highly successful in identifying causal variants when these variants make a large contribution to the total genetic variation. For comparison purposes, a penalized-likelihood based fine-mapping procedure was also used on this data.

The method, LASSO local automatic regularization resample model averaging (LLARRMA) developed by Valdar et al. [14] combines LASSO shrinkage with resample model-averaging in order to estimate, for each SNP, the probability of being included in a multivariable model in alternate realizations (subsamples) of the data. The sparse matrix of LLARRMA Re-sampled Model Inclusion probabilities (RMIP) contained 164 SNPs. Out of the twenty-two causal SNVs, only rs68020 was included. The RMIP values of the top 22 SNVs are shown in Fig 5.2.

It seems that the LLARRMA has been able to correctly identify rs68020 because it had a sufficient contribution to genetic variation while also being relatively common in the population. This is in contrast to the SNVs obtained by Susie. These were reported in credible sets which usually contained one or two SNVs at a time. In the first 4 iterations of the procedure described in the Analysis section, all of the reported credible sets contained causal SNVs. The LLARRMA-identified SNV rs68020 was among the reported causal SNVs identified by the Susie method.

While this project demonstrated that it is possible to extend Bayesian fine-mapping procedures to larger and denser genomic regions and still obtain fairly accurate results, the study had certain limitations which must be taken into consideration.

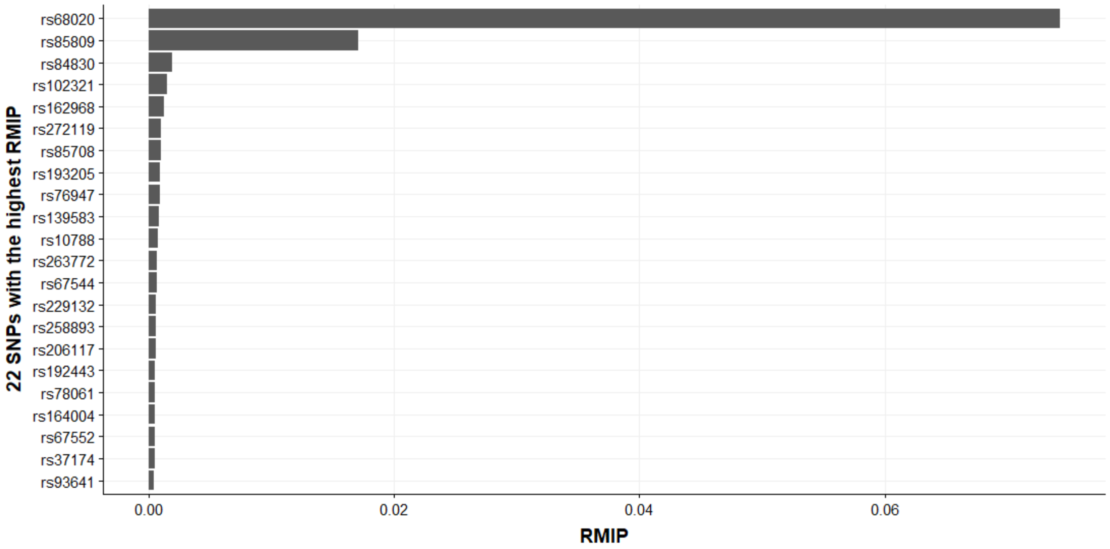

Figure 5.2: RMIP values for SNVs from the LLARRMA method; only rs68020 was correctly identified

The data used to assess the Susie methods performance was obtained using a simulation method. Therefore, weaknesses in the assumptions underlying the simulation model can make the results of this study less useful. For instance, the simulation model (and the Susie method) are both based on an additive effect model. If dominance effects or other locus-effect architectures (usually some combination of additive and dominance effects) are present, this may make the detection of the causal variants more challenging or may require a shift in the methodology.

The Susie method was used on a phenotype that was Normally distributed. Future endeavors could also focus on how the Susie method performs in genetic case-control studies that have binary variables as the phenotype.

At the same time, initial results from performing Susie on a chromosomal dataset that was partially pre-processed seemed to yield promising results. This indicates that pre-processing may not be necessary. On the methodology side, there is scope to refine the search for causal variants using more information on the gene regions. The Susie method has the capacity to incorporate different priors and it could be possible to simulate data where causal gene effects are emulated using known gene functions (knowledge of genes coding for certain proteins). This information could then be passed on to the Susie method as a functionally informed prior and the results compared to a baseline model with flat priors.

Bibliography

[1] Alexander Grueneberg and Gustavo de los Campos. BGData - A Suite of R Packages for Genomic Analysis with Big Data. *G3 Genes/Genomes/Genetics*, 9(5):1377–1383, 05 2019.

[2] Farhad Hormozdiari, Emrah Kostem, Eun Y Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[3] Anna Hutchinson, Jennifer Asimit, and Chris Wallace. Fine-mapping genetic associations. *Human Molecular Genetics*, 29(R1):82–89, 2020.

[4] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[5] Dominic Nelson, Jerome Kelleher, Aaron P Ragsdale, Claudia Moreau, Gil McVean, and Simon Gravel. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS genetics*, 16(5):e1008619, 2020.

[6] Ju-Hyun Park, Mitchell H. Gail, Clarice R. Weinberg, Raymond J. Carroll, Charles C. Chung, Zhaoming Wang, Stephen J. Chanock, Joseph F. Fraumeni, and N Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

[7] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[8] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome- wide associations to candidate causal variants by statistical fine- mapping. *Nature Reviews Genetics*, 19:491–501, 2018.

[9] Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics*, 3(7):1–13, 07 2007.

[10] Mikko J Sillanpaa and Maddhucchanda Bhattacharjee. Bayesian association-based fine mapping in small chromosomal segments. *Genetics*, 169:427–439, 2005.

[11] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9:477–485, 2008.

[12] Sarah L Spain and Jeffrey C Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):111–119, 2015.

[13] Miriam S. Udler, Kerstin B. Meyer, and Karen A. Pooley. Fgfr2 variants and breast cancer risk: fine-scale mapping using african american studies and analysis of chromatin conformation. *Human Molecular Genetics*, 18(9):1692–1703, 2009.

[14] William Valdar, Jeremy Sabourin, Andrew Nobel, and Christopher C Holmes. Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genetic Epidemiology*, 36:451–462, 2012.

[15] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[16] Andrew R. Wood, Tonu Esko, and Jian Yang. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, 2014.

Appendix A

Code

The code used to summarize, explore and analyse the data in this thesis is available from https://github.com/SFUStatgen/ZJ.