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-E□ects model. The approach is particularly

Table of Contents

List of Tables

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 ne-mapping can be thought of as the step following a GWAS, where regions identi ed 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, ne-mapping can be understood as taking the GWAS data that shows these complex associations and essentially, `untangling' it to nd the causal genes. The purpose of nding 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

to emulate the requirements of a genetic ne-mapping attempt using a simulation of the genomic information for chromosome 1.

In the literature, ne-mapping is formulated as a variable selection problem in regression. A typical analysis of this sort is treated in [10]. What di erentiates genetic ne-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 di erent loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their di erent 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 ne-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 identi cation of causal variants at risk loci in ne-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 E ects (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

$$y = Xb + ;$$

where y is the vector of phenotypes for n individuals, X is the genotype matrix of dimension $n \quad p$, b is a vector of regression coefficients with exactly one non-zero entry and other entries of zero, and $\quad N(0; \ ^2 I_n)$. Let $b = b$ where b is a scalar corresponding to the effect size of a single single-nucleotide variant (SNV) and $= [\ _1; :::; \ _p]$ 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 $b \quad N(0; \ ^2_0)$ and the vector $\quad$ Multinomial$(m; \ _p)$ where $m = 1$ is the number of resamples and $_p$ 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 as

the resampling probability, prior knowledge of which SNV is more likely to be causal can be incorporated through .

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

$$jX;y; \ ^2; \ ^2_0 \quad \text{Multinomial}(1; \ );$$

where is

parameters, consider their marginal likelihood under the SER model:

$$`_{SER}(y; \ _0^2) = p_0(y j \ ^2)$$

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 $b_1; :::b_L$. The idea is to find an approximation $q(b_1; :::b_L)$ to the posterior $p_{post} = p(b_1; :::b_L jX; y; \ ^2; \ ^2_0)$. This can be done by minimizing $D_{KL}(q; p_{post})$ where D_{KL} is the Kullbeck-Leibler (KL) Divergence between

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

$$\ell(\gamma, \alpha; y) = \log \int p(y|b; \alpha)$$

as

$$F(q; g; ; y) = E_q \left[\frac{\log p(y|b; g;)}{q(b)} \right]$$
$$= E_q[\log p(y|b; g;)] + E_q \left[\log \frac{g(b)}{q(b)} \right]$$

Now consider an additive model:

$$y = \sum_{l=1}^{L} {}_l +$$
$$N(0; \ ^2 I_n)$$
$${}_l \quad g_l :$$

The Susie model is an example, where $q = Xb_l$ for some $l \ 2 \ f 1; :::Lg$ and each g_l is the prior distribution for b_l. De ne a simple model M_l which is derived from the original model M by setting ${}_{l^0} = 0$ for all $l^0 \ 6= \ l$. Therefore 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 E ects Regression (SER). The Susie method hinges on the idea that the model M can be t if each of the simpler models M_l can be t. In order to t each of the simpler models, allow the class of distributions on $Xb_1; :::Xb_L$ to factorize over $Xb_1; :::Xb_L$. This leads to:

$$q(Xb_1; :::Xb_L) = \prod_{l=1}^{L} q_l(Xb_l):$$

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

$$F(q; g; ; y) = E_q[\log p(y|b;)] + E_q \left[\log \frac{g(b)}{q(b)} \right]$$

$$= \frac{n}{2} \log(2 \quad ^2) \quad \frac{1}{}$$

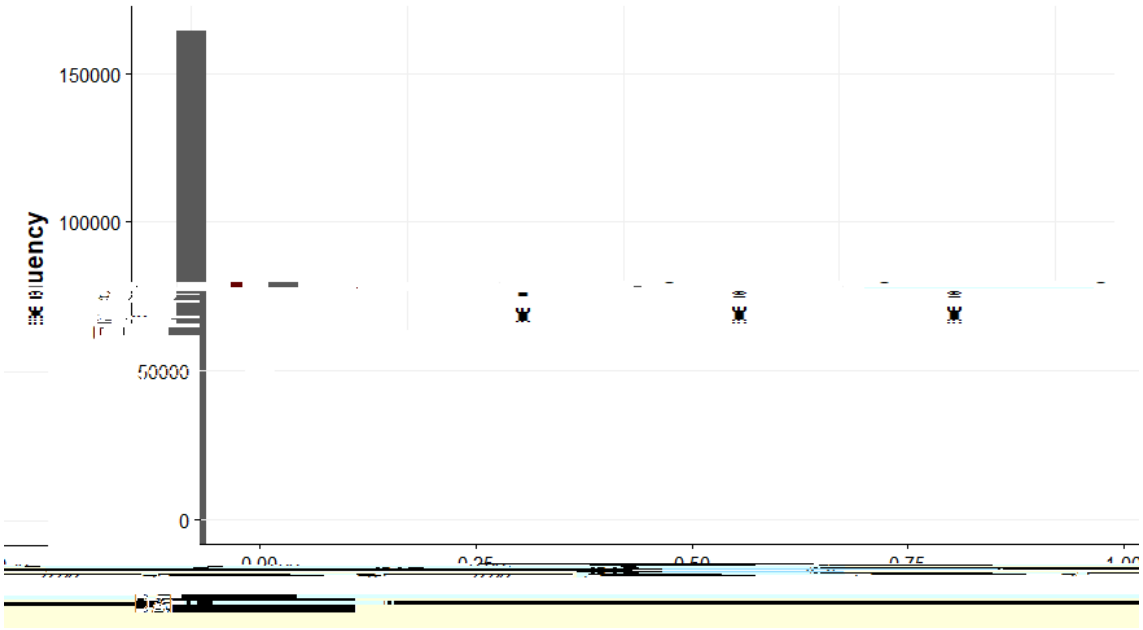
Chapter 3

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 $1 \quad 10^{-8}$ 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 $(249000000 \div 26288)$ 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 $3100 \quad 287668$ Only possible entries for


150000
100000
50000
0

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

$$^2 = \frac{b^2}{f(1 \quad f)};$$

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:

$$\begin{aligned} g &= 2 \ ^2 f(1 \quad f) \\ &= 2 \quad \frac{b^2}{f(1 \quad f)} \quad f(1 \quad f) \\ &= 2b^2: \end{aligned}$$

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 80% 6%. That is, the heritability of height, Y,

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

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 $Var(E[Y|X])$

In order to work with these data e□ciently in R, the information in these four files was combined into a BGDataobject from the BGDatapackage [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 di□erence in proportions of SNVs between the causal gene, non-causal gene and non-gene regions in Chromosome 1.

The first quality control stepty/rtolved firlery/out-239(NV)82(s)-38(y/r)2239(he)-3480non-gene egi

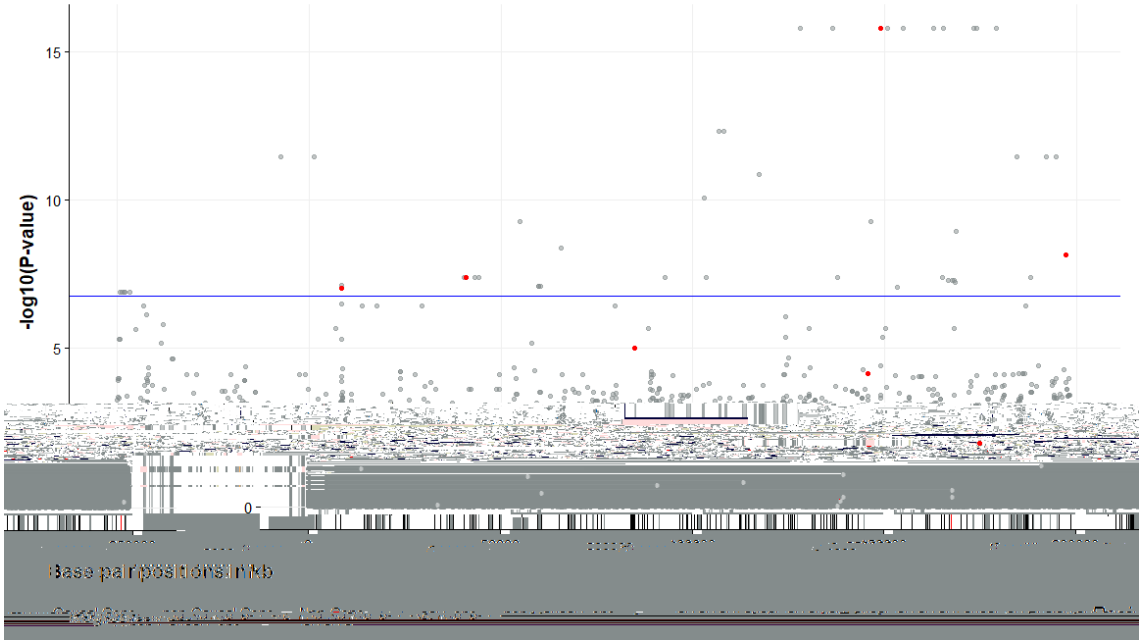


Figure 3.3: Manhattan plot of an initial Chromosome-wide scan of association

```
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	$\hat{}$	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

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.

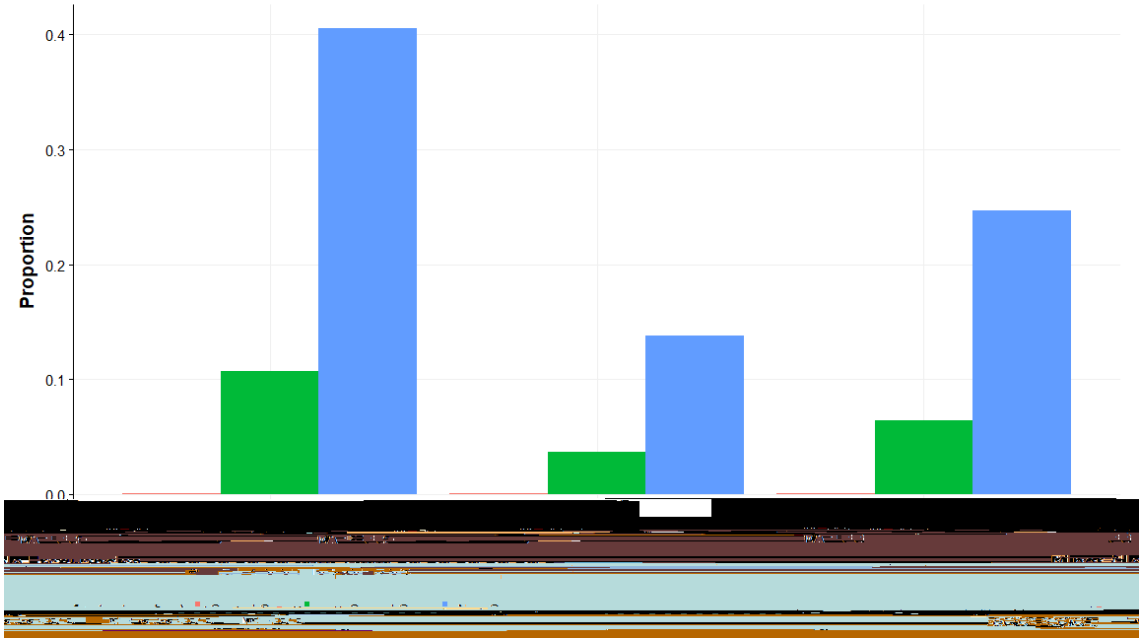


Figure 3.4: Proportions of SNVs on the chromosome that are causal, non-causal and non-gene along with their minor allele frequencies grouped into 3 categories

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

```
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 newgenosmatrix and the corresponding SNP.support dataframe, a new BGDataobject 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.

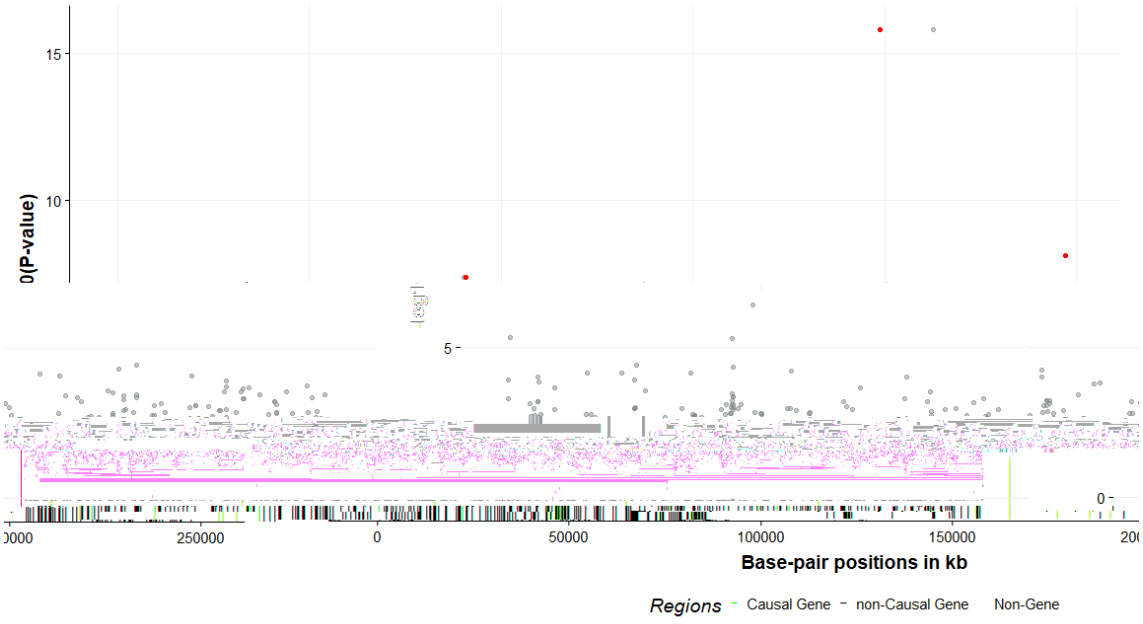


Figure 4.1: Manhattan plot of Chromosome 1 after Pre-processing and quality control steps

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

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.

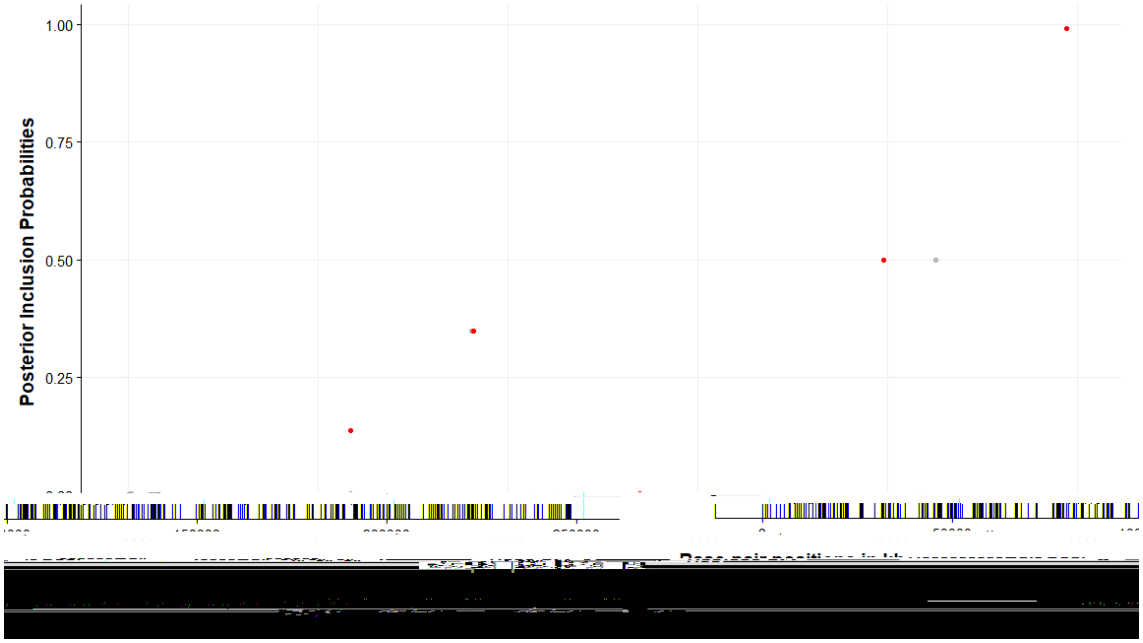


Figure 4.2: Plot of posterior inclusion probabilities after the first run of the Susie method

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 e□ect SNV.

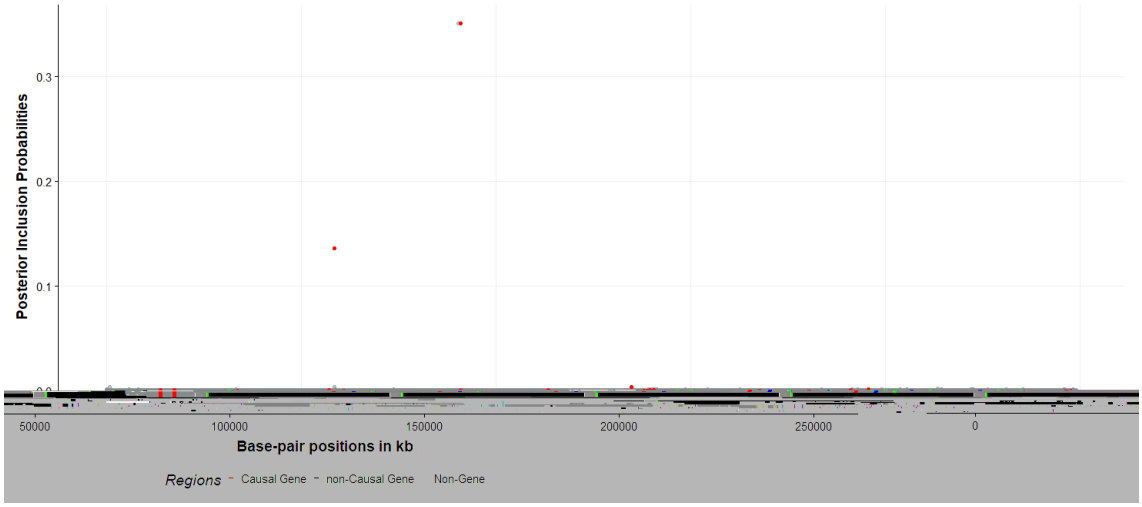


Figure 4.3: Plot of PIPs after the third run of the Susie method highlighting the reduction in the highest scoring SNV.

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%

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 su□cient 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.

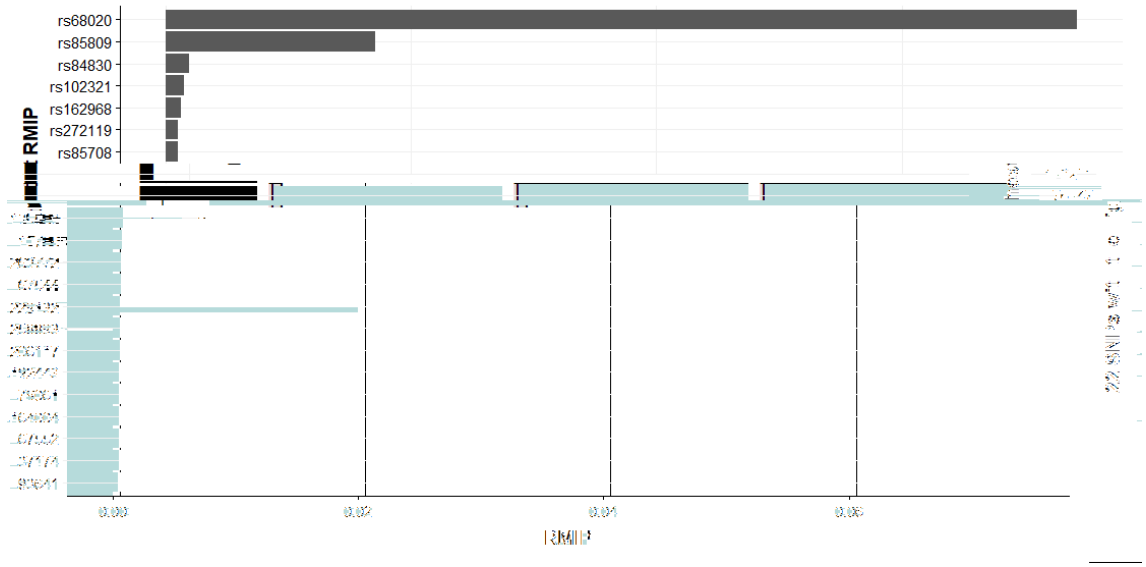


Figure 5.2: RMIP values for SNVs from the LLARRMA method; only rs68020 was correctly identified

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 e□ect model. If dominance e□ects or other locus-e□ect architectures (usually some combination of additive and dominance e□ects) 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 di□erent priors and it could be

possible to simulate data where causal gene e□ects 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

[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 Je□rey 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.