A BAYESIAN MODELING FOR PAIRED DATA IN GENOME-WIDE ASSOCIATION STUDIES WITH APPLICATION TO BREAST CANCER

by

Yashi Bu

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Min Chen, Chair

Dr. Swati Biswas

Dr. Sunyoung Shin

Dr. Zhenyu Xuan

Dedicated to my family.

A BAYESIAN MODELING FOR PAIRED DATA IN GENOME-WIDE ASSOCIATION STUDIES WITH APPLICATION TO BREAST CANCER

by

YASHI BU, BS, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

STATISTICS

THE UNIVERSITY OF TEXAS AT DALLAS

December 2020

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor Professor Min Chen for his constant guidance throughout the PhD study and research. He has been extremely supportive and patient to students, and offers valuable advice with his professional expertise and knowledge, which is essential in helping me improve academic work and overcome challenges.

I would also like to extend my thanks to my dissertation committee: Professor Swati Biswas, Professor Sunyoung Shin, and Professor Zhenyu Xuan, who have provided valuable suggestions in completing this work. I would like to thank all Professors and staff in the Department of Mathematical Sciences for their impressive knowledge and sincere support during my research.

I am deeply grateful to my parents, who have accompanied me and offered unconditional love and emotional support throughout this long journey.

November 2020

A BAYESIAN MODELING FOR PAIRED DATA IN GENOME-WIDE ASSOCIATION STUDIES WITH APPLICATION TO BREAST CANCER

Yashi Bu, PhD
The University of Texas at Dallas, 2020

Supervising Professor: Dr. Min Chen, Chair

Many complex human diseases are associated with genetic factors. Identifying genetic markers is the key step to account for disease heritability, and develop disease diagnosis, risk prediction, prevention and therapeutic strategies. Genome-wide association study (GWAS) has emerged as a powerful tool to identify genetic variants that are associated with various cancers. The common statistical methodologies in GWAS focus on case-control data where cases and controls are sampled independently from the populations. Despite the success of GWAS in finding a number of genetic variants that are associated with cancers, the power of conventional GWAS is limited. Extensive research has shown that many tumors develop as a consequence of the progressive accumulation of somatic mutations over time. We focus on GWAS data from tumor and paired normal tissues to unravel the genetic association of somatic mutations. To address the limitation that conventional GWAS methods are not applicable to matched-paired data, we propose in this dissertation a framework that incorporates allelic relative risk, frequency and mutation rate to accommodate the structure of paired data.

We first apply the penalized maximum likelihood estimation (MLE) to perform single marker analysis based on the framework. Simulation studies are carried out to assess the performance of penalized MLE. To further improve the estimation accuracy and power of single

marker analysis, we develop a Bayesian hierarchical model that takes advantage of applying Bayesian shrinkage and making inferences based on the posterior distributions. The hierarchical Bayesian model has the flexibility to take into account the prior knowledge and extend to multiple marker analysis. We find that the single-marker Bayesian model has improved the estimation and power performance in most simulation scenarios. To identify DNA segments and SNP sets, rather than single genetic variants that are associated with the disease, we develop a multiple-SNP Bayesian model which considers SNP sets that are grouped together in a biologically meaningful way, such as genes or pathways. The multiple-SNP analysis considers the joint effects of the SNP set, which improves the power to identify SNPs that have moderate marginal effects by themselves. Simulation studies show that the multi-marker Bayesian model has higher power to identify associated SNPs and lower type I error rates. Next, we apply the proposed methods to a breast cancer data set from The Cancer Genome Atlas (TCGA).We compare the most significant genes identified by single marker analysis and multiple marker analysis to external resources on somatic mutations of breast cancer. We find that both methods identify genes associated with breast cancer, and multiple marker analysis provides more consistent results with external resources.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

1.1.1 Review of Genome-wide association studies with traditional case-control data

Many human complex diseases, including diabetes, heart disease, hypertension and various types of cancers, are associated with genetic factors. Identifying the genetic factors of complex diseases is essential to the understanding of disease heritability, which may lead to better strategies of the diagnosis and the treatment of disease. A single nucleotide polymorphism (SNP) is the single base substitution in genome, which contributes to nearly 90% of the genetic variations (Smith et al., 2009). In recent years, Genome-wide association study (GWAS) has emerged as a powerful tool to identify common genetic variations that are associated with complex diseases, such as various types of cancers, type I and II diabetes, Crohn's disease, bipolar disorder, and hypertension (Stadler et al., 2010; Marees et al., 2018; Kim et al., 2013). A typical GWAS design is a case-control study, which independently samples controls from the healthy population and cases from the diseased population. The aim of GWAS is to identify SNPs that are linked to the disease susceptibility from hundreds of thousands or even millions of genetic loci across the entire genome. Despite the great success of traditional GWAS, the identified SNPs only explains a small portion of disease heritability (Manolio et al., 2009). In this dissertation, we focus on several alternative strategies to explore the missing heritability in cancer and develop statistical techniques to improve the power to detect contributory genetic variants.

The typical GWAS strategy is to examine the individual association status of large amount of biomarkers across the entire genome. Nowadays over one million SNPs can be genotyped as genetic markers among cases and controls. The common statistical methods

in GWAS often requires independent case-control samples. The contingency table method, often known as chi-square test or Fisher's exact test, is widely used in GWAS with dichotomous traits. The null hypothesis of the test is that there is no association between the genotypes and the phenotype. The test p-values are used to determined statistical significance in GWAS (Timpson et al., 2009; Bush et al., 2008). However, these approaches suffer from the disadvantage of lacking estimation for allelic effect size. Another common method is the logistic regression, which is able to include clinical covariates to account for additional heritability and provide odds ratios as a proxy for effect size. The generalized linear model (GLM) methods are generally applied for quantitative traits. With these common approaches, the conventional GWAS has identified numerous loci that are associated with a variety of human cancers over the recent years (Zeng et al., 2015).

Although it has been successful to identify significant genetic loci from case-control study, most GWAS methods require the assumption of independent samples. Under typical GWAS using independent case-control samples, a common practice is that the confounding factors such as population stratification need to be adjusted. The population stratification, which represents the systematic allele frequency differences between cases and controls due to ancestry differences, can cause spurious discoveries in GWAS. The principal components analysis (PCA) is used to infer the axes of ancestry differences and correct for population stratification (Price et al., 2006). The mixed models are developed to correct for population structure using random effects (Kang et al., 2010; Yu et al., 2005). These corrections generally have demonstrated to be useful for GWAS but their adjustment performances subject to the experimental design of choosing cases and controls (François et al., 2016).

1.1.2 Genome-wide association studies with matched pair design

Unlike traditional case-control design, another type of samples in cancer studies is the paired tumor-normal data, where cases are genotyped from tumor tissues and controls are from normal tissues of the same patients. The genotypic differences between normal tissue and tumor

tissue are due to somatic mutations. The somatic mutations accumulate in any somatic cells throughout lifetime. Many of mutations do not have noticeable effects, however some somatic mutations are linked to cellular dysfunction and cancerogenesis (Martincorena and Campbell, 2015). It has been reported that the majority of cancer genes are somatically mutated in cancers (Futreal et al., 2004). Many somatic mutations are closely associated with heritable genetic risk factors and they can be treated as phenotypes that have direct links with carcinogenesis. Study of somatic mutations in cancer may unravel previously unknown genetic markers. The Cancer Genome Atlas (TCGA) project has collected genotypic sequencing data from over 22,000 tumor and match normal samples for 33 types of cancers. Among all cancers, the breast cancer is the most common invasive cancer in women worldwide (Anastasiadi et al., 2017). To address the limitation of conventional GWAS, we propose a framework to accommodate the structure of paired data and develop statistical methods to measure the association of somatic mutations. Most traditional GWAS evaluates effects of germline variations among independent samples, where cases and controls are independent. In spite of the combination of GWAS and independent samples have successfully discovered germline variants with high or moderate effect sizes, the effects of somatic mutations associated to tumor development are largely understudied. To study the effects of somatic mutations, the matched tumor-normal data are collected In summary, there is an increasing need of studying matched tumor-normal genetic data to better understand the associations of genetic factors and enhance explanations of heritability.

In case-control studies, GWAS typically focuses on individual SNP susceptibility detection. Existing methods such as logistic regression, association test are implemented to derived an estimation for each genetic marker. Then a decision threshold, such as p-value, is determined to conclude the association status for each marker. In conventional GWAS, more than 1 million genetic markers are commonly sequenced and evaluated. In order to control the overall type I error, a very stringent threshold must be used to adjust for multiple hypothesis testing to reach a genome-wide significance level. The threshold is usually

too stringent that many genetic markers with moderate effects fail to reach beyond the this boundary and are concluded as not associated. In fact, The underlying basis of GWAS is the Common Disease Common Variant (CDCV) hypothesis, which states that complex diseases are mainly attributed to common variants with moderate effect size (Reich and Lander, 2001; Pritchard, 2002). With a threshold at genome-wide significant level, the moderate genetic markers can hardly be identified in hypothesis testings. Complex diseases are linked to many genetic factors. While standard GWAS has identified numerous significant variants, many traditional GWAS only focuses on single SNP analysis. As a matter of fact, it is widely known that interactions between genetic factors are commonly presented in many diseases (Moore, 2003; Millstein et al., 2006; Zhang and Liu, 2007; Kooperberg and LeBlanc, 2008). Traditional GWAS only studies individual SNP and fails to discover the joint effects of these genetic factors.

1.1.3 Single-marker and multi-marker analysis

Multiple marker analysis methods in GWAS such as gene-based association have become an important tool. In addition, thanks to the blossom of high throughput sequencing technologies, there is an increasing demand for appropriate statistical methodologies that can be applied to multi-locus association studies. Researchers have developed many SNP-set analysis recently (Tibshirani, 1996; Gayán et al., 2008; Borecki and Province, 2008; Bush et al., 2008; Mukhopadhyay et al., 2009). A SNP-set analysis is proposed to combine a set of SNPs that are proximate in terms of genomic characteristics such as genes or haplotypes (Wu et al., 2010; Larson and Schaid, 2013) . The SNP-set is tested for the joint association status without concluding the association status for each SNP. In this dissertation, a Bayesian model in the novel framework to analyze joint effects of variants on the gene level.

1.2 Overview of the dissertation

The remaining of this dissertation is organized as follows: In Chapter 2 we introduce the proposed framework for tumor and normal paired data. The penalized maximum likelihood estimation (MLE) is developed to measure single allelic effect size. The Wald test, Score test and likelihood ratio test are used in hypothesis testings in simulation studies. To improve the model performance, a hierarchical Bayesian model is developed to incorporate the Bayesian inferential framework for single-marker analysis. We conduct simulation studies under different settings to compared the Bayesian and penalized MLE methods. To further increase power to detect associated variants, Chapter 3 extends the hierarchical Bayesian model to group biologically related SNPs and perform multi-marker analysis. We use the simulated data to evaluate performance of the proposed methods under different scenarios. In Chapter 4 the single- and multi- marker models are applied to the TCGA breast cancer data. We use external databases of breast cancer genes to compared with the genes identified by the hierarchical Bayesian models. Finally, Chapter 5 summarizes the proposed methods and discusses potential future work.

CHAPTER 2

SINGLE MARKER ANALYSIS

2.1 Introduction

The widely available high throughput sequencing has greatly facilitated the understanding and studies of genetic factors in complex diseases. Conventional genome-wide association studies (GWAS) have reported meaningful findings, where a large number of single nucleotide polymorphisms (SNPs) are genotyped to test the individual association between these SNPs and the disease. One of the common statistical methodologies to test for association in case-control studies is the chi-square test, which has considerable power but comes with a drawback that covariates cannot be added (Cantor et al., 2010). Another common approach, logistic regression, allows inclusion of additional factors to take into account confounding effects. In these methods, the p-values or effect size of SNPs are evaluated and thresholds are set up to determine significant association. They have proven useful in GWAS to detect genetic variants associated with complex disease susceptibility, such as type 2 diabetes (Voight et al., 2010) , Crohn's disease and rheumatoid arthritis (The Wellcome Trust Case Control Consortium, 2007). However, the common approaches are not able to address the impact of somatic mutations. Numerous studies have suggested that the accumulation of somatic mutations over the lifetime can gradually lead to changes in protein functions and tumor progression (Greenman et al., 2007; Martincorena and Campbell, 2015). The Cancer Genome Atlas (TCGA) has provided abundant tumor and normal paired data, where genotypes are assayed from normal and tumor tissues of the same individual, for many kinds of cancers. However, the common statistical GWAS methods requires that cases and controls are independent. There is a growing need of appropriate methods for matched tumor-normal data to investigate the effects of somatic variants on tumor development.

We present novel approaches to developing single-marker analysis using frequentist and Bayesian methods, along with the comparisons of these two methods in this chapter. The

approaches utilize a framework suited for the matched-pair data. Point estimation, interval estimation and hypothesis testing can be derived to make inferences about the loci associated with cancer risk. In this chapter, the estimation and inference will focus on single marker analysis, where only one marker is considered. The penalized maximum likelihood estimation and Bayesian methods will be used to estimate and infer the association of each locus.

The remainder of this chapter is organized as follows. Section 2.2 describes the framework for tumor and normal paired data. Section 2.3 describes the penalized maximum likelihood estimation with results of the simulation study under different situations. More specifically, the penalized term, or regularization is accomplished through Ridge Regression to keep parameters away from the boundaries. In addition, the performance of penalized maximum likelihood estimation will be evaluated through mean squared error (MSE), hypothesis testing and power curve. Then the Bayesian method is introduced to further improve the analysis. Section 2.4 incorporates Bayesian model and shows the model formulation and posterior distribution for parameters. Markov Chain Monte Carlo (MCMC) algorithm is used to sample from the posterior distributions. Simulation studies are conducted to compare the two methods and the results are reported in Section 2.5.

2.2 Framework for matched pair data

A growing number of research has suggested that various cancers are associated with somatic mutations (Alexandrov and Stratton, 2014). Often the development of cancer is a progressive process in which multiple mutations are accumulated in a normal cell that eventually evolves to a cancerous cell, which can evade the immune mechanisms and start to proliferate. Understanding the role of somatic mutations in carcinogenesis can be critical in risk prediction, continuous monitoring and early detection of cancer, and may lead to individualized prevention and therapeutic strategies. The Cancer Genome Atlas (TCGA) has been collecting tumor tissues along with their matched normal samples for different types of

cancers since 2005. More than 30 cancer types are involved in genomic characterization and sequence analysis in the TCGA project by 2014. A number of studies focusing on different types of cancer have discovered somatic variants that are associated with the susceptibility based on the TCGA database. One advantage of association study involving tumor DNA and the matched normal DNA is that the two samples share identical genetic and environmental background. The study design of matched data intrinsically controls many confounding risk factors, and therefore has the potential to greatly improve the power of detecting true signals coming from genetic variations. The proposed methods are designed for matched data in which the cases and controls are no longer independent.

Many existing methods for GWAS fail to provide valid results for matched data, since the assumption of independence is violated. To perform GWAS on matched pairs, we propose a sampling scheme to illustrate the difference between germline variations and somatic mutations. Specifically, this sampling scheme characterizes the allelic relative risks for somatic mutations. Using the following sampling scheme, the relative risk, allele frequency and somatic mutation rate can be estimated and tested.

2.2.1 Sampling framework

In the matched-pair design specimens are only sampled from the disease population. The cases are tumor tissues while controls are tumor-adjacent normal tissues of the same patient. Thus it can provide a reliable detection of somatic mutations. A general matched data have the same structure as the case-control data in GWAS. Assume the sample size is 1000 and the number of genetic marker is 1,000,000. For a certain sample, the genotype can be 0, 1 or 2 at each SNP. The raw data structure is as follow:

The sampling framework is introduced to characterize the relation of relative risk, allele frequency and mutation rate. Given that the tumor and normal tissues are from the same patient, the alteration of SNPs at the same locus is due to somatic mutation. The somatic mutation rate acts as a bridge between dependent cases and controls.

Table 2.1: Tumor tissue data structures

Tumor tissue genotypes

Samples	**SNP 1**	**SNP 2**	...	**SNP 1000000**
1	0	0	...	1
2	1	0	...	1
⋮	⋮	⋮		⋮
1000	0	2	...	1

Table 2.2: Matched normal data structures

Matched Normal tissue genotypes

Samples	**SNP 1**	**SNP 2**	...	**SNP 1000000**
1	0	0	...	1
2	1	0	...	0
⋮	⋮	⋮		⋮
1000	0	1	...	0

Let N be the whole population size, A be the risk allele frequency, M be the somatic mutation rate. Genotype is the genetic makeup of alleles at a locus. Under genetic equilibrium, the probability of carrying genotype 0, 1, 2 are $(1-A)^2, 2A(1-A), A^2$, respectively. Hence the expected frequencies of genotypes in whole population are as follows:

Table 2.3: Expected genotype frequency

Normal Genotype	**0**	**1**	**2**
Expected Frequency	$(1-A)^2$	$2A(1-A)$	A^2

Considering the somatic mutations accumulate over time, the mutation rate M characterizes the probability of the alternation at a locus. Given normal genotype is 0, the mutant genotypes in tissue can be 0, 1 and 2 with probabilities of $(1-M)^2, 2M(1-M), M^2$, respectively. Given normal genotype is 1, the mutant genotypes can be 0, 1 and 2 with probabilities of $M(1-M), M^2+(1-M)^2, M(1-M)$. Given normal genotype is 2, the mutant genotypes can be 0, 1 and 2 with probabilities of $M^2, 2M(1-M), (1-M)^2$. The Table 2.4 shows the expected frequencies of genotypes in whole population given normal genotypes and mutant

genotypes. The total of the expected frequencies is population size N. The population can be divided into 9 categories based on the genotypes in normal and mutant tissues.

Table 2.4: Expected frequency of genotypes in whole population

Freq	**Tissue with mutant genotypes**		
Normal genotype	0	1	2
0	$(1-A)^2(1-M)^2N$	$2(1-A)^2M(1-M)N$	$(1-A)^2M^2N$
1	$2A(1-A)M(1-M)N$	$2A(1-A)[M^2+(1-M)^2]N$	$2A(1-A)M(1-M)N$
2	A^2M^2N	$2A^2M(1-M)N$	$A^2(1-M)^2N$

The penetrance is defined as the probability of having cancer given a specific genotype. In general, if the SNP is unassociated with disease, $\pi_2 = \pi_1 = \pi_0$. If the SNP is associated with disease, $\pi_2 \geq \pi_1 \geq \pi_0$.

$$\pi_t = P(Having\ disease|Mutated\ genotype = t)\ wheret = 0, 1, 2 \tag{2.1}$$

Let N_p be the total number of patients in the population. The Table 2.5 shows the expected frequencies of genotypes in the patient population. The sum of the expected frequencies is the patient population size N_p.

Table 2.5: Expected frequency of genotypes in the patient population

Freq	**Tumor genotype**		
Normal genotype	0	1	2
0	$(1-A)^2(1-M)^2N\pi_0$	$2(1-A)^2M(1-M)N\pi_1$	$(1-A)^2M^2N\pi_2$
1	$2A(1-A)M(1-M)N\pi_0$	$2A(1-A)[M^2+(1-M)^2]N\pi_1$	$2A(1-A)M(1-M)N\pi_2$
2	$A^2M^2N\pi_0$	$2A^2M(1-M)N\pi_1$	$A^2(1-M)^2N\pi_2$

Given the expected genotype frequencies in patient population, we assume a simple random sample is drawn from all patients. Let n be the sample size and $P_{sampled}$ be the sampling rate. The Table 2.6 shows the expected frequencies of genotypes in the sample. The sum of the expected frequencies in this table is sample size n.

Let R_t be the allelic relative risk (RR) given a specific genotype: $R_1 = \pi_1/\pi_0, R_2 = \pi_2/\pi_0$. Under different types of genetic risk models, there are 4 representations of R_1 and R_2 using a

Table 2.6: Expected frequency of genotypes in the sample

Freq	**Tumor genotype**		
Normal genotype	0	1	2
0	$(1-A)^2(1-M)^2N\pi_0P_{sampled}$	$2(1-A)^2M(1-M)N\pi_1P_{sampled}$	$(1-A)^2M^2N\pi_2P_{sampled}$
1	$2A(1-A)M(1-M)N\pi_0P_{sampled}$	$2A(1-A)[M^2+(1-M)^2]N\pi_1P_{sampled}$	$2A(1-A)M(1-M)N\pi_2P_{sampled}$
2	$A^2M^2N\pi_0P_{sampled}$	$2A^2M(1-M)N\pi_1P_{sampled}$	$A^2(1-M)^2N\pi_2P_{sampled}$

unified parameter. If the somatic mutation are not associated with cancer, the corresponding relative risk should be 1. The additive genetic risk model is used in this dissertation. Other genetic risk models, such as dominant, recessive and multiplicative can be considered as well.

Table 2.7: Genetic models

Genetic Model	**Allelic Penetrance**			**Relative Risk**	
	aa	**Aa**	**AA**	$\mathbf{R_1}$	$\mathbf{R_2}$
Dominant	π_0	π_1	π_2	R	R
Recessive	π_0	π_1	π_2	1	R
Additive	π_0	π_1	π_2	$\frac{R+1}{2}$	R
Multiplicative	π_0	π_1	π_2	R	R^2

Given the expected genotype frequencies in the sample, a normalization of sample size n is performed. During the normalization process, the population size N, sampling rate $P_{sampled}$ are cancelled out. The penetrance π_2, π_1, π_0 can be represented by relative risk parameter R. Thus, the expected probability of genotypes in the sample can be derived as in Table 2.8, where q_{sum} is the normalization constant such that probabilities add up to 1 and q_{sum} is a function of R, A, M.

Therefore, for each patient, the joint distribution of normal-tumor genotypes of a SNP follows Multinoulli distribution, which is also known as the generalized Bernoulli distribution. The normal-tumor genotypes of the SNP among all patients in the sample can be summarized into one table. Let $P_{i,j}$ be the probability in each category, where $i = 0, 1, 2$ is the normal tissue genotype and $j = 0, 1, 2$ is the tumor tissue genotype. The sum of $P_{i,j}$ is 1 and each $P_{i,j} = f_{i,j}(R, A, M)$ is a function of R, A, M. The Table 2.9 shows Multinoulli distribution with probabilities $P_{i,j}$ for all patients.

Table 2.8: Expected probability of genotypes in the sample

Prob	**Tumor genotype**		
Normal genotype	0	1	2
0	$\frac{(1-A)^2(1-M)^2}{q_{sum}}$	$\frac{2(1-A)^2M(1-M)}{q_{sum}}\frac{R+1}{2}$	$\frac{(1-A)^2M^2}{q_{sum}}R$
1	$\frac{2A(1-A)M(1-M)}{q_{sum}}$	$\frac{2A(1-A)[M^2+(1-M)^2]}{q_{sum}}\frac{R+1}{2}$	$\frac{2A(1-A)M(1-M)}{q_{sum}}R$
2	$\frac{A^2M^2}{q_{sum}}$	$\frac{2A^2M(1-M)}{q_{sum}}\frac{R+1}{2}$	$\frac{A^2(1-M)^2}{q_{sum}}R$

Table 2.9: Paired tumor-normal data follows Multinoulli distribution

Probabilities	**Tumor genotypes**		
Normal genotypes	**0**	**1**	**2**
0	P_{00}	P_{01}	P_{02}
1	P_{10}	P_{11}	P_{12}
2	P_{20}	P_{21}	P_{22}

where

$$
\begin{aligned}
P_{00} &= \frac{(1-A)^2(1-M)^2}{q_{sum}} \\
P_{01} &= \frac{2(1-A)^2M(1-M)}{q_{sum}} \cdot \frac{R+1}{2} \\
P_{02} &= \frac{(1-A)^2M^2}{q_{sum}}R \\
P_{10} &= \frac{2A(1-A)M(1-M)}{q_{sum}} \\
P_{11} &= \frac{2A(1-A)[M^2+(1-M)^2]}{q_{sum}} \cdot \frac{R+1}{2} \\
P_{12} &= \frac{2A(1-A)M(1-M)}{q_{sum}}R \\
P_{20} &= \frac{A^2M^2}{q_{sum}} \\
P_{21} &= \frac{2A^2M(1-M)}{q_{sum}} \cdot \frac{R+1}{2} \\
P_{22} &= \frac{A^2(1-M)^2}{q_{sum}}R
\end{aligned}
\tag{2.2}
$$

For each SNP, the joint genotype of the tumor and its matched normal tissue follows the Multinoulli distribution with 9 categories, of which the probabilities are functions of the parameters (R, A, M) as defined previously. In the next section, we will derive the estimation of the parameters.

2.3 Maximum Likelihood Method

The frequentist inference of parameters is based on the penalized maximum likelihood estimation for each parameter. The likelihood function is the probability mass function of the observed value of the Multinoulli distribution, which is derived under the sampling framework for matched data described in Section 2.2. In the sampling scheme framework, three parameters: relative risk, allele frequency and mutation rate are defined to describe the relation between normal data and tumor data. Under Hardy-Weinberg equilibrium assumptions, there are expected genotypic frequencies given the allele frequency for each marker. With mutation rate, the expected genotypic frequencies of germline with and without somatic mutation can be derived as a 3 by 3 table, where each column represent a genotype. In reality, the paired data set is sampled from patients that have already developed cancer. The above expected genotypic frequencies incorporates penetrance to filter out samples that carry the disease . Thus the risk alleles tend to be enriched in the samples with disease, which causes allele frequencies to shift higher in the sample. The final stage of the sampling scheme is normalize genotypic frequencies by dividing the sum of all frequencies, such that the data follows a Multinoulli distribution. And relative risk is incorporated as the ratio of penetrance given one or two risk alleles and the penetrance given homozygous major alleles. Therefore, the framework purposed for paired data allows to handle dependent genotype data and derive estimation and inference about relative risk, the link between genetics and disease susceptibility.

The sampling framework transforms the paired data into 9 proportions according to the corresponding genotypes, as shown in the following table and forms a Multinoulli distribution with sample size n and probabilities $\mathbf{P} = (P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21}, P_{22})$, where each $P_{i,j}$ is a function of parameters, i.e., $P_{i,j} = f_{i,j}(R, A, M)$. Then the likelihood function can be derived using data as a function of these parameters.

Table 2.10: Multinoulli Distribution on each genetic marker

Counts		Tumor genotypes		
		0	1	2
Normal genotypes	0	n_{00}	n_{01}	n_{02}
	1	n_{10}	n_{11}	n_{12}
	2	n_{20}	n_{21}	n_{22}

Thus, on each genetic marker, the sample $\mathbf{n} = (n_{00}, n_{01}, n_{02}, n_{10}, n_{11}, n_{12}, n_{20}, n_{21}, n_{22})$ follows Multinoulli$\big(n, (P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21}, P_{22})\big)$, where n is the summation of the counts in all components. The likelihood function is written as follows:

$$\begin{aligned} & L(P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21}, P_{22}|\mathbf{n}) \\ & = f\big(\mathbf{n}|P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21}, p_{22}\big) \\ & = \prod_{i=0,1,2} \prod_{j=0,1,2} (P_{i,j})^{n_{i,j}} \end{aligned} \tag{2.3}$$

Replace the $P_{i,j}$ with the functions of parameters $P_{i,j} = f_{i,j}(R, A, M)$. The likelihood function can be represented as follows:

$$L(R, A, M|\mathbf{n}) = \prod_{i=0,1,2} \prod_{j=0,1,2} (f_{i,j}(R, A, M))^{n_{i,j}} \tag{2.4}$$

Correspondingly likelihood function is as follows:

$$l(R, A, M|\mathbf{n}) = \sum_{i=0,1,2} \sum_{j=0,1,2} n_{i,j} \cdot log(f_{i,j}(R, A, M)) \tag{2.5}$$

The log-likelihood function can be represented as a function of the three parameters.

2.3.1 Likelihood Function

The model uses penalized maximum likelihood estimation method to estimate the value of parameters.The penalized term is added to prevent parameter estimations on the boundary. The boundaries of relative risk (R), allele frequency (A) and mutation rate (R) are $(0, \infty)$, $(0, 1)$ *and* $(0, 1)$ respectively. The penalty term is introduced as the L2 (Ridge) norm to penalize extreme allele frequency (A) and mutation rate (R):

$$\begin{aligned} l_p(R, A, M|\mathbf{n}) = \sum_{i=0,1,2} \sum_{j=0,1,2} n_{i,j}\ log(f_{i,j}(R, A, M)) - \lambda\ log^2(A)\ -\ \lambda\ log^2(1-A) \\ - \lambda\ log^2(M) - \lambda\ log^2(1-M) - \alpha\ log^2(R) \end{aligned} \tag{2.6}$$

where α and λ are the tuning parameters. When the estimations of parameters A and M are close to the their boundaries, at least one of the terms A, 1-A, M, 1-M are close to 0. Thus, at least one of the terms $log^2(A)$, $log^2(1-A)$, $log^2(M)$, $log^2(1-M)$, will be fairly large enough to add penalty through tuning parameters to the likelihood function, and to pull parameter estimations away from the boundaries. In the following simulation study, the value of tuning parameter λ will be 0.05, which is not too large to introduce much bias, and still carries the influence of penalty.

The penalized maximum likelihood is derived as:

$$\begin{aligned} \underset{R,A,M}{\operatorname{argmax}}\, l_p(R, A, M|\mathbf{n}) = \underset{R,A,M}{\operatorname{argmax}} \sum_{i=0,1,2} \sum_{j=0,1,2} n_{i,j}\ log(f_{i,j}(R, A, M)) - \lambda log^2(1-A) \\ - \lambda log^2(A) - \lambda log^2(M) - \lambda log^2(1-M) - \alpha\ log^2(R) \end{aligned} \tag{2.7}$$

2.3.2 Hypothesis Testing

Hypothesis testing for each SNP is conducted to decide if relative risk is statistically significant. We are interesting in testing null hypothesis $H_0 : R = R_0 = 1$, which states that the genetic marker is not associated with cancer, versus alternative hypothesis $H_1 : R \neq R_0 = 1$,

which states that the genetic marker is associated with the disease. The Wald test is one approach to determine statistical significance. It measures the squared difference $\hat{R}_{MLE} - R_0$ weighted by the curvature of log-likelihood function. The Wald statistic is calculated as follows:

$$W = \frac{(\hat{R}_{MLE} - R_0)^2}{var(\hat{R}_{MLE})} \tag{2.8}$$

where $var(\hat{R}_{MLE})$ is the variance of penalized maximum likelihood estimator. The variance is estimated by the inverse of the expected information matrix evaluated at the maximum likelihood estimate. Under the null hypothesis, the Wald test statistic W follows an asymptotic χ^2-distribution with one degree of freedom.

Score test assesses the statistical significance of parameter based on the gradient of the likelihood function. The value of score function is evaluated at R_0 and equaled to 0 when $R_0 = \hat{R}_{MLE}$. When the score function at R_0 deviates far from 0, the alternative hypothesis is more plausible than H_0. The Score test statistic is calculated as follows:

$$S = \frac{(u(R_0))^2}{I(R_0)} \tag{2.9}$$

where $u(R_0)$ is the score function evaluated at R_0 and other parameters are replaced by the MLE. $I(R_0)$ is the fisher information evaluated by R_0 and other parameters are fixed at their MLE. The test statistic asymptotically follows a χ^2 distribution with one degree of freedom.

The likelihood-ratio test is another method to assess statistical significance based on comparison of log-likelihood function evaluated at MLE and R_0. The likelihood-ratio test statistic is calculated as follows:

$$LR = 2l(\hat{R}_{MLE}) - 2l(R_0) \tag{2.10}$$

where $l(\hat{R}_{MLE})$ and $l(R_0)$ is the log-likelihood function evaluated at $\hat{R}_{MLE}$ and R_0, respectively. Under large samples the likelihood-ratio test statistic asymptotically follows a χ^2 distribution with one degree of freedom.

2.3.3 Simulations

Simulation studies are carried out with these test statistics in different settings. Figure 2.1 shows the p-value distribution of Wald test under settings (1) $n = 1000$, $M = 0.001$ and (2) $n = 3000$, $M = 0.005$. Figure 2.2 shows the p-value distribution of Score test under settings where $n = 3000$ and $M = 0.005$. Figure 2.3 shows the p-value distribution of likelihood-ratio test under the same setting. The full simulation results are listed in Appendix A. Statistical theory has stated that p-value should be uniformly distributed under null hypothesis $R = 1$. As can be seen in these figures, the p-value in Wald test is approximately uniformly distributed under the null hypothesis with large sample sizes, but this is not true for small sample sizes. For Score test and likelihood-ratio test, the performance is similar. In addition, the type I errors are inflated in settings with small sample size, low allele frequency. The powers under these scenarios may not be reliable due to the corresponding type I errors are not well controlled. This suggests that the performance of penalized MLE method needs to be improved in some settings. These results motivate us to develop the Bayesian method which can incorporate prior knowledge to the model.

2.4 Bayesian Hierarchical Models

2.4.1 Motivation

The performance of penalized Maximum likelihood estimation has relatively low power in the testing. We propose Bayesian hierarchical model that incorporates prior distributions on the SNP association status and other model parameters. The advantage of this method is that it

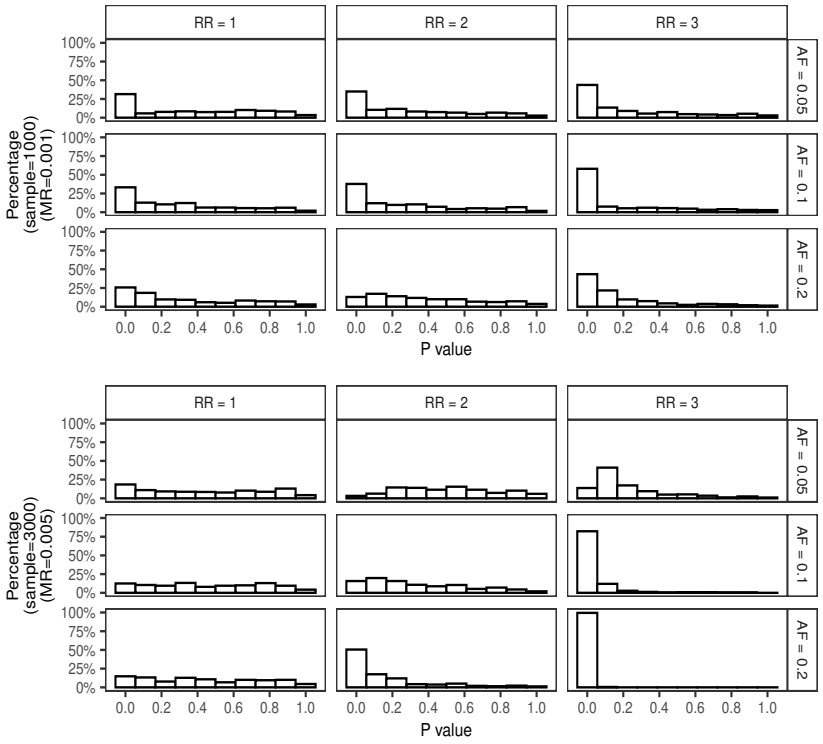

Figure 2.1: The p-value distribution of Wald test in settings (1) sample = 1000, MR = 0.001 and (2) sample = 3000, MR = 0.005.

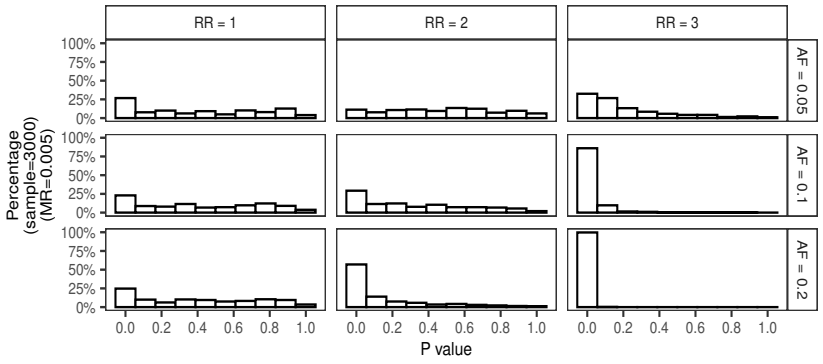

Figure 2.2: The p-value distribution of Score test in settings sample = 3000 and MR = 0.005.

allows us to perform Bayesian regularization through prior distributions and make Bayesian inferences based on the posterior distributions. In this section, we describe the Bayesian hierarchical model on single markers. Simulation studies are carried out to compare the Bayesian model with the Maximum likelihood estimation method.

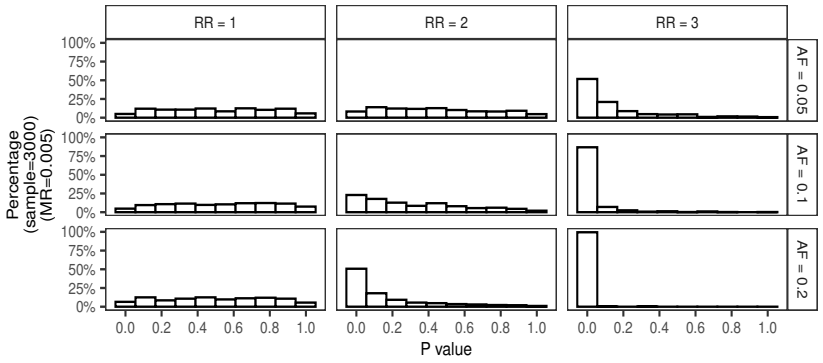

Figure 2.3: The p-value distribution of likelihood-ratio test in settings sample = 3000 and MR = 0.005.

2.4.2 Prior Distribution

Let H represent the SNP association status. If a SNP is associated with cancer, then $H = 1$, otherwise, $H = 0$. The SNP status H follows a Bernoulli distribution with probability b, where b is a hyper-parameter with a hyper prior. The hyper-parameter b has a range between 0 and 1, and follows a Beta distribution with mean at 0.2.

The prior distribution of SNP association status and hyper-parameter can be expressed as follows:

$$\begin{aligned} H &\sim Bernoulli(b) \\ b &\sim Beta(\alpha_b, \beta_b) \end{aligned} \tag{2.11}$$

The prior distribution of effect size depends on the association status of each SNP. If the SNP status is associated, the effect size is expected to be greater than 1. Then conditional effect size $R|(H = 1)$ is assigned a prior $Gamma(k_1, \theta_1)$ with a mean greater than 1 when it is associated. On the other hand, when the SNP status is not associated, the effect size should be 1. Then conditional effect size $R|(H = 0)$ is assigned a prior $Gamma(k_0, \theta_0)$ with the mean at 1 when it is irrelevant. Thus, the relative risk distribution is as follows:

$$
\begin{aligned}
R|(H=1) &\sim Gamma(k_1, \theta_1) \\
R|(H=0) &\sim Gamma(k_0, \theta_0) \\
R|H &\sim Gamma(k_1, \theta_1)^H \cdot Gamma(k_0, \theta_0)^{1-H}
\end{aligned}
\tag{2.12}
$$

Let A denote the allele frequency (AF). The prior distribution of A depends on the association status of the SNP. When the SNP is irrelevant to the tumor, the variant is neither enriched nor diminished in disease population. The AF in disease population is expected to be the same as the AF in normal population. Thus for non-associated SNP, the conditional AF $A|(H=0)$ is assigned a prior $Beta(\alpha_0, \beta_0)$. The mean of $Beta(\alpha_0, \beta_0)$ is at A_{obs}.

On the other hand, if a SNP is associated with tumor development, the SNP frequency will be enriched in disease population. Therefore the AF observed in disease population is higher than the AF in the normal population. For $A|(H=1)$, we assign a prior $Beta(\alpha_{A1}, \beta_{A1})$. From above, the distribution of AF is as follows:

$$
\begin{aligned}
A|(H=1) &\sim Beta(\alpha_{A1}, \beta_{A1}) \\
A|(H=0) &\sim Beta(\alpha_{A0}, \beta_{A0}) \\
A|H &\sim Beta(\alpha_{A1}, \beta_{A1})^H \cdot Beta(\alpha_{A0}, \beta_{A0})^{1-H}
\end{aligned}
\tag{2.13}
$$

The mean of $Beta(\alpha_{A1}, \beta_{A1})$ depends on the effect size. For a higher effect size, more allele enrichment is expected in disease population. To determine the mean of $Beta(\alpha_{A1}, \beta_{A1})$, let A_{obs} be the AF observed in samples. The observed AF in disease population has the expectation and estimation as follows:

$$
\begin{aligned}
A_{obs} &= \frac{n_{10}+n_{11}+n_{12}+2n_{20}+2n_{21}+2n_{22}}{2n} \\
\mu_{A_{obs}} = E(A_{obs}) &= E\left(\frac{n_{10}+n_{11}+n_{12}+2n_{20}+2n_{21}+2n_{22}}{2n}\right) \\
&= \frac{A(R-A+RA+2AM-2RAM+1)}{2(RA-M-A+RM+2AM-2RAM+1)}
\end{aligned} \tag{2.14}
$$

When $\mu_R = 1$, there is $\mu_{A_{obs}} = A$. When $R \neq 1$, the equation 2.14 is a quadratic function of A when $R \neq 1$ with coefficients:

$$
\begin{aligned}
a &= (R-1)(1-2_M) \\
b &= (R+1) - (R-1)2\mu_{A_{obs}}(1-2M) \\
c &= -2A_{obs}(RM-M+1)
\end{aligned} \tag{2.15}
$$

For a risky SNP, the effect size $R > 1$. The parameter A can be represented by A_{obs} and other parameters (R, M). The relation can be used to determine the AF prior mean $Beta(\alpha_{A1}, \beta_{A1})$ when the SNP is associated.

$$
\hat{A} = A_{obs} + \frac{R+1-\sqrt{[R+A_{obs}2(2M-1)(R-1)+1]^2 - 8A_{obs}(2M-1)(R-1)(RM-M+1)}}{2(2M-1)(R-1)} \tag{2.16}
$$

The relative risk parameter R is set to 2.5 for a risky SNP based on the empirical observation. Let MR parameter M be estimated by the observed MR from samples since it has little impact. The expect A_{obs} can be estimated from samples. Thus the prior mean of A can be determined from the below equation:

$$
\hat{A} = A_{obs} + \frac{3.5-\sqrt{(3A_{obs}(2M-1)+3.5)^2 - 12A_{obs}(2M-1)(1.5M+1)}}{3(2M-1)} \tag{2.17}
$$

Let M denote the mutation rate (MR). The prior distribution of M depends on the SNP association status. When the SNP is non-associated, the mutation rate is assigned a

prior $Beta(\alpha_{M0}, \beta_{M0})$. When the SNP is associated, the mutation rate is assigned a prior $Beta(\alpha_{M1}, \beta_{M1})$.

$$\begin{aligned} M|(H=1) &\sim Beta(\alpha_{M1}, \beta_{M1}) \\ M|(H=0) &\sim Beta(\alpha_{M0}, \beta_{M0}) \\ M|H &\sim Beta(\alpha_{M1}, \beta_{M1})^{H} \cdot Beta(\alpha_{M0}, \beta_{M0})^{1-H} \end{aligned} \tag{2.18}$$

Let M_{obs} be the observed mutation from tumor samples. The observed MR has the expectation and estimation as follows:

$$\begin{aligned} M_{obs} &= \frac{n_{01} + 2n_{02} + n_{10} + n_{11mut} + n_{12} + 2n_{20} + n_{21}}{2n} \\ \mu_{M_{obs}} = E(M_{obs}) &= E\left(\frac{n_{01} + 2n_{02} + n_{10} + n_{11mut} + n_{12} + 2n_{20} + n_{21}}{2n}\right) \\ &= \frac{M(R - M + RM + 2AM - 2RAM + 1)}{2(RA - M - A + RM + 2AM - 2RAM + 1)} \end{aligned} \tag{2.19}$$

When $R = 1$, there is $\mu_{M_{obs}} = M$. The equation 2.19 is a quadratic function of M when $R \neq 1$. The quadratic equation has coefficients:

$$\begin{aligned} a &= (R-1)(1-2A) \\ b &= (R+1) - (R-1)2\mu_{M_{obs}}(1-2A) \\ c &= -2M_{obs}(RA - A + 1) \end{aligned} \tag{2.20}$$

For a risky SNP, the effect size $R > 1$. The parameter M can be represented by M_{obs} and other parameters R, A. The relation can be used to determine the MR prior mean of $Beta(\alpha_{M1}, \beta_{M1})$ when SNP is associated.

$$\hat{M} = M_{obs} + \frac{-(R+1) + \sqrt{[R + 1 - 2M_{obs}(R-1)(1-2A)]^2 + 8M_{obs}(1-2A)(R-1)(RA - A + 1)}}{2(R-1)(1-2A)} \tag{2.21}$$

In the above equation, the relative risk parameter R is estimated to be 2.5 for a risky SNP. The AF parameter A is estimated by equation 2.17. Then the prior mean 2.18 of MR can be estimated.

2.4.3 Joint Posterior Distribution

Let $\mathbf{\Theta} = \{R, A, M\}$ denote the set of parameters for each SNP, with prior distribution depending on the SNP status H. The joint distribution of these parameters can be derived as follows:

$$\begin{aligned} f(\mathbf{\Theta}, H, b) &= f(\mathbf{\Theta}|H, b) \cdot f(H, b) \\ &= f(\mathbf{\Theta}|H, b) \cdot f(H|b) \cdot f(b) \end{aligned} \tag{2.22}$$

The parameters (R, A, M) have prior distribution conditional on the SNP status. Assume that relative risk, allele frequency and mutation rate are conditionally independent given the SNP status, there is

$$\begin{aligned} f(\mathbf{\Theta}, H, b) &= f(\mathbf{\Theta}|H, b) f(H|b) f(b) \\ &= f(R, A, M|H, b) f(H|b) f(b) \\ &= f(R, A, M|H) f(H|b) f(b) \\ &= f(R|H) f(A|H) f(M|H) f(H|b) f(b) \end{aligned} \tag{2.23}$$

Based on the proposed framework, samples are collected from both tumor and controlled normal tissues. For each SNP, the corresponding samples follows a Multinoulli distribution with 9 categories, where the expected probability of each category is a function of relative risk, AF and mutation rate, each P_{ij} can be presented as $P_{i,i'} = f_{i,i'}(\mathbf{\Theta})$ and $i, i' = 0, 1, 2$. Let $\mathbf{n} = (n_{00}, n_{01}, n_{02}, n_{10}, n_{11}, n_{12}, n_{20}, n_{21}, n_{22})$ denote the samples, then there is

$$f(\mathbf{n}|\mathbf{\Theta}) = \prod_{i,i'}^{0,1,2} (P_{i,i'})^{n_{i,i'}} = \prod_{i,i'}^{0,1,2} f_{i,i'}(R, A, M)^{n_{i,i'}} \tag{2.24}$$

The distribution of sample $\mathbf{n}$ is dependent on SNP status. Therefore, from previous equations 2.23 the joint probability distribution of all parameters is as follows:

$$\begin{aligned} f(\mathbf{n}, \boldsymbol{\Theta}, H, b) &= f(\mathbf{n}|\boldsymbol{\Theta}, H, b) \cdot f(\boldsymbol{\Theta}, H, b) \\ &= f(\mathbf{n}|\boldsymbol{\Theta}, H, b) f(\boldsymbol{\Theta}|H, b) f(H, b) \\ &= f(\mathbf{n}|\boldsymbol{\Theta}, H, b) f(\boldsymbol{\Theta}|H, b) f(H|b) f(b) \\ &= f(\mathbf{n}|\boldsymbol{\Theta}, H) f(\boldsymbol{\Theta}|H) f(H|b) f(b) \\ &= f(\mathbf{n}|R, A, M) f(R|H) f(A|H) f(M|H) f(H|b) f(b) \end{aligned} \tag{2.25}$$

Due to the complexity of joint distribution and prior distribution in Bayesian model, it is difficult to sample from the posterior distribution directly. In our situation, the conditional distributions of the parameters b and H have a nice closed-form distribution so that they can be directly sampled using Gibbs sampler. At the same time, the conditional distributions of parameters (R, A, M) do not have simple conditional distributions. Then the Metropolis-Hasting method is applied to (R, A, M) to draw samples within each Gibbs sampling step.

The Gibbs sampler is a MCMC algorithm that allows to generate a series of Markov chain observations from the conditional distribution of a parameter while other variables are fixed at their current values (Casella and George, 1992). Since it is usually infeasible to directly sample from a multivariate distribution, the Gibbs sampler makes such process easier by sampling a single variable at a time. It is very useful when the conditional distribution, which based on other fixed variables, has a closed-form expression. Variables are sampled iteratively conditioned on the most recent states of the remaining variables.

Sometimes, the closed-form expression may not be attainable for some parameters in Bayesian modeling. The conditional distribution is proportionally known, thus direct sampling is unavailable. The Metropolis-Hasting method is a MCMC algorithm that can draw samples from any probability distributions, given that a function proportional to the target

density is provided. It uses a Markov process that generates a series of states and asymptotically reach to the target distribution. The advantage of Metropolis-Hastings algorithm is that it does not require a known probability distribution of the target variable, but a function that is proportional to the density function. This requirement is easy to satisfy and useful when the normalization factor is hard to measure. We will use both Gibbs sampler and Metropolis-Hastings method to estimate the posterior distribution of the Bayesian model.

2.4.4 Posterior sampling

The posterior distribution will be estimated using Markov Chain Monte Carlo (MCMC) simulation. Parameters are updated iteratively within Gibbs framework and RR, AF, MR are updated with Metropolis-Hastings algorithm. The Markov Chain Monte Carlo simulation is given as follows for a sample $\mathbf{n}$:

1. Initialization of parameters.

   Assign random initial values to $b^0, H^0, \mathbf{\Theta^0}$ in the beginning of the simulation. For each sample $\mathbf{n}$, three chains with different initial status are generated in the Markov Chain Monte Carlo simulation.

2. Update b.

   Given other parameters are fixed, the conditioned posterior distribution at step t can be derived from 2.25 as follows:

$$\begin{aligned} f(b^t|\mathbf{n}, \mathbf{\Theta^t}, H^t) &= \frac{f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)}{\int f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)\ db^t} \\ &= \frac{f(H^t|b^t)f(b^t)}{\int f(H^t|b^t)f(b^t)\ db^t} \\ &\propto (b^t)^{H^t+\alpha_b-1}(1-b^t)^{\beta_b-H^t} \end{aligned} \tag{2.26}$$

   The conditioned posterior distribution is proportional to a Beta distribution with $\alpha_{new} = H^t + \alpha_b, \beta_{new} = \beta_b - H^t + 1$. The next state b^{t+1} can be directly sampled

from the closed-form conditioned posterior distribution based on current state:

$$f(b^{t+1}|H^{(t)}) = Beta(\alpha_{new} = H^{(t)} + \alpha_b,\ \beta_{new} = \beta_b - H^{(t)} + 1) \tag{2.27}$$

Note that means of conditioned posterior distribution and prior distribution satisfy the inequality: $E(f(b^{t+1}|H^{(t)} = 0)) < E(f(b^t)) < E(f(b^{t+1}|H^{(t)} = 1))$. The difference between conditioned posterior means can be represented as:

$$E(f(b^{t+1}|H^{(t)} = 1)) - E(f(b^{t+1}|H^{(t)} = 0)) = \frac{1}{\alpha_b + \beta_b + 1} \tag{2.28}$$

The denominator should be relatively small to make SNP status more distinguishable. Thus, $\alpha_b = 1$ and $\beta_b = 2$ are chosen due to the fact that most SNPs are non-associated.

3. Update H.

Given other parameters are fixed, the conditioned posterior distribution at step t can be derived as follows:

$$\begin{aligned} f(H^t|\ b^t, \mathbf{n}, \mathbf{\Theta^t}) &= \frac{f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)}{\sum_{h=0,1} f(\mathbf{n}, \mathbf{\Theta^t}, H^t = h, b^t)} \\ &= \frac{f(\mathbf{n}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t) f(H^t|b^t) f(b^t)}{\sum_{h=0,1} f(\mathbf{n}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t = h) f(H^t = h|b^t) f(b^t)} \\ &= \frac{f(\mathbf{\Theta^t}|H^t) f(H^t|b^t)}{\sum_{h=0,1} f(\mathbf{\Theta^t}|H^t = h) f(H^t = h|b^t)} \end{aligned} \tag{2.29}$$

Substitute the prior distributions, there are

$$\begin{aligned} f(H^t|\ b^t, \mathbf{\Theta^t}, \mathbf{n}) \propto &\big(Gamma(R^t; k_1, \theta_1) Beta(A^t; \alpha_{A1}, \beta_{A1}) Beta(M^t; \alpha_{M1}, \beta_{M1}) b\big)^{H^t} \cdot \\ &\big(Gamma(R^t; k_0, \theta_0) Beta(A^t; \alpha_{A0}, \beta_{A0}) Beta(M^t; \alpha_{M0}, \beta_{M0}) (1-b)\big)^{1-H^t} \end{aligned} \tag{2.30}$$

Let d_1 be the probability corresponding to $H^t = 1$, and d_0 be the probability corresponding to $H^t = 0$ in the above formula. Thus the conditioned posterior probability density function of H^t follows a Bernoulli distribution. Based on current state, the next state H^{t+1} can be directly sampled with probability $p = \frac{d_1}{d_1 + d_0}$.

4. Update relative risk R.

Since the conditioned probability density function has no closed form, the posterior distribution of relative risk R will be simulated by Metropolis-Hastings algorithm within Gibbs sample framework.

4.1. Conditional posterior distribution.

Given other parameters are fixed, the proportional posterior distribution of relative risk R^t at current state can be derived as follows:

$$\begin{aligned} f(R^t|\mathbf{n}, H^t, A^t, M^t, b^t) &= \frac{f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)}{\int f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)\ dR^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t)\ dR^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t}) f(R^t|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t}) f(R^t|H^t)\ dR^t} \\ &\propto f(\mathbf{n^t}|\mathbf{\Theta^t}) f(R^t|H^t) \end{aligned} \tag{2.31}$$

where $f(R^t|H^t)$ is defined in 2.12.

4.2. Proposal distribution.

The proposal function will generate a value R^* in the neighborhood based on current state R^t through Gamma distribution.

$$q(R^*|R^t) = Gamma(shape = 1 + 5R^t, rate = 5) \tag{2.32}$$

Correspondingly, there is

$$q(R^t|R^*) = Gamma(shape = 1 + 5R^*, rate = 5) \tag{2.33}$$

4.3. Acceptance probability.

Compute the acceptance ratio based on the proposal distribution and joint density

function.

$$\begin{aligned} accept\_ratio &= \frac{f(\mathbf{n^t}|\mathbf{\Theta}^*)f(R^*|H^t)\cdot q(R^t|R^*)}{f(\mathbf{n^t}|\mathbf{\Theta^t})f(R^t|H^t)\cdot q(R^*|R^t)} \\ &= \frac{f(\mathbf{n}|R^*,A^t,M^t)f(R^*|H^t)\cdot q(R^t|R^*)}{f(\mathbf{n}|R^t,A^t,M^t)f(R^t|H^t)\cdot q(R^*|R^t)} \end{aligned} \tag{2.34}$$

Then compute the acceptance ratio

$$r = min\left(1,\ \frac{f(\mathbf{n}|R^*,A^t,M^t)f(R^*|H^t)\cdot q(R^t|R^*)}{f(\mathbf{n}|R^t,A^t,M^t)f(R^t|H^t)\cdot q(R^*|R^t)}\right)$$

The acceptance indicator is from Bernoulli distribution with probability r. If acceptance ratio is greater than 1, let acceptance probability be 1.

4.4. Iterative updates.

Take R^* as the current state, repeat above steps to generate the next state R^{*1}. For Metropolis-Hastings iteration $h = 1, 2, ..., 99$, let $R^{*(h)}$ as current state, repeat above steps to generate next state $R^{*(h+1)}$. Finally, the next state of Gibbs iteration R^{t+1} is updated with the last state of Metropolis-Hastings iteration:

$$R^{t+1} = R^{*(H)}\ where\ H\ is\ the\ last\ state\ in\ MH\ iterations$$

5. Update A^t.

5.1. Conditional posterior distribution.

Given other parameters are fixed, the proportional posterior distribution of relative risk A^t at current state can be derived as follows:

$$\begin{aligned} f(A^t|\mathbf{n},R^t,M^t,H^t,b^t) &= \frac{f(\mathbf{n},\mathbf{\Theta^t},H^t,b^t)}{\int f(\mathbf{n},\mathbf{\Theta^t},H^t,b^t)\ dA^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t})f(\mathbf{\Theta^t}|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t})f(\mathbf{\Theta^t}|H^t)\ dA^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t})f(A^t|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t})f(A^t|H^t)\ dA^t} \\ &\propto f(\mathbf{n^t}|\mathbf{\Theta^t})f(A^t|H^t) \end{aligned} \tag{2.35}$$

5.2. Proposal distribution.

The proposal function generates a candidate value A^* in the neighborhood based on current state A^t through Beta distribution.

$$q(A^*|A^t) = Beta(\alpha = \frac{1+100A^t}{1-A^t}, \beta = 102) \tag{2.36}$$

Correspondingly, there is

$$q(A^t|A^*) = Beta(\alpha = \frac{1+100A^*}{1-A^*}, \beta = 102) \tag{2.37}$$

5.3. Acceptance probability.

Compute the acceptance ratio based on the proposal distribution and joint density function.

$$\begin{aligned} accept\_ratio &= \frac{f(\mathbf{n^t}|\mathbf{\Theta}^*)f(A^*|H^t)\cdot q(A^t|A^*)}{f(\mathbf{n^t}|\mathbf{\Theta^t})f(A^t|H^t)\cdot q(A^*|A^t)} \\ &= \frac{f(\mathbf{n}|R^t,A^*,M^t)f(A^*|H^t)\cdot q(A^t|A^*)}{f(\mathbf{n}|R^t,A^t,M^t)f(A^t|H^t)\cdot q(A^*|A^t)} \end{aligned} \tag{2.38}$$

Accept the proposed candidate with probability of acceptance ratio. If acceptance ratio is greater than 1, let acceptance probability be 1.

$$r = min\left(1,\ \frac{f(\mathbf{n}|R^t,A^*,M^t)f(A^*|H^t)\cdot q(A^t|A^*)}{f(\mathbf{n}|R^t,A^t,M^t)f(A^t|H^t)\cdot q(A^*|A^t)}\right)$$

5.4. Iterative updates.

Take A^* as the current state, repeat above steps to generate the next state A^{*1}. For Metropolis-Hastings iteration $h = 1, 2, ..., 99$, use $A^{*(h)}$ as current state, repeat above steps to generate next state $A^{*(h+1)}$. Finally, the next state of Gibbs iteration A^{t+1} is updated with the last state of Metropolis-Hastings iteration:

$$A^{t+1} = A^{*(H)}\ where\ H\ is\ the\ last\ state\ in\ MH\ iterations$$

6. Update M^t.

6.1. Conditional posterior distribution.

The proportional posterior distribution of current mutation rate M^t conditioned on other parameters and data is as follows:

$$\begin{aligned} f(M^t|\mathbf{n}, R^t, A^t, H^t, b^t) &= \frac{f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)}{\int f(\mathbf{n}, \mathbf{\Theta^t}, H^t, b^t)\ dM^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|H^t)\ dM^t} \\ &= \frac{f(\mathbf{n^t}|\mathbf{\Theta^t}) f(M^t|H^t)}{\int f(\mathbf{n^t}|\mathbf{\Theta^t}) f(M^t|H^t)\ dM^t} \\ &\propto f(\mathbf{n^t}|\mathbf{\Theta^t}) f(M^t|H^t) \end{aligned} \tag{2.39}$$

6.2. Proposal distribution.

The proposal function generates a candidate value M^* in the neighborhood based on current state M^t through Beta distribution.

$$\begin{aligned} q(M^*|M^t) &= Beta(\alpha = \frac{1 + 1000M^t}{1 - M^t},\ \beta = 1002) \\ q(M^t|M^*) &= Beta(\alpha = \frac{1 + 1000M^*}{1 - M^*},\ \beta = 1002) \end{aligned} \tag{2.40}$$

6.3. Acceptance probability.

Compute the acceptance ratio based on the proposal distribution and joint density function.

$$\begin{aligned} accept\_ratio &= \frac{f(\mathbf{n^t}|\mathbf{\Theta}^*)\ f(M^*|H^t)\ q(M^t|M^*)}{f(\mathbf{n^t}|\mathbf{\Theta^t})\ f(M^t|H^t)\ q(M^*|M^t)} \\ &= \frac{f(\mathbf{n}|R^t, A^t, M^*)\ f(M^*|H^t)\ q(M^t|M^*)}{f(\mathbf{n}|R^t, A^t, M^t)\ f(M^t|H^t)\ q(M^*|M^t)} \end{aligned} \tag{2.41}$$

Accept the proposed candidate with probability of acceptance ratio. If acceptance ratio is greater than 1, let acceptance probability be 1.

$$r = min\left(1,\ \frac{f(\mathbf{n}|R^t, A^t, M^*)\ f(M^*)\ q(M^t|M^*)}{f(\mathbf{n}|R^t, A^t, M^t)\ f(M^t)\ q(M^*|M^t)}\right) \tag{2.42}$$

6.4. Iterative updates.

Take M^* as the current state, repeat above steps to generate the next state

M^{*1}. For Metropolis-Hastings iteration $h = 1, 2, ..., 99$, use $M^{*(h)}$ as current state, repeat above steps to generate next state $M^{*(h+1)}$. Finally, the next state of Gibbs iteration M^{t+1} is updated with the last state of Metropolis-Hastings iteration:

$$M^{t+1} = M^{*(H)} \textit{ where } H \textit{ is the last state in } MH \textit{ iterations}$$

7. Burn-in and thinning.

   The burn-in and thinning process is performed to correct the potential bias introduced by the initial status. Some starting points may over sample the regions that are rare events. The beginning samples may not be stabilized to the stationary distribution, thus it should not contribute to the inferences. In this simulation, the initial 20% period of Gibbs sampler iterations is discarded and every 5th value is kept. The remaining sequences are used for inferences after burn-in and thinning process.

8. Convergence and diagnosis.

   A major consideration in MCMC simulations is the convergence of Markov chains. The simulated chains are expected to fully explore the target distribution. Multiple chains initiated with different starting values are adopted. A common method to assess the MCMC convergence is to analyze and compare the differences between multiple chains.

   The Gelman-Rubin statistic is applied to diagnose the MCMC convergence by analyzing the differences between multiple Markov chains. The convergence is evaluated by comparing the estimated between-chains and within-chains variances. Large differences between these variances indicate non-convergence.

   Suppose there are M simulated chains with length T. Given a model parameter θ, let θ_{mt} be the value at the t^{th} updates on the m^{th} simulated chains, where $t = 1, 2, ..., T$ and $m = 1, 2, ..., M$. Let $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the m^{th} chain. The between-chains variance B and within-chains variance W are given by

$$B = \frac{T}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2$$
$$W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2 \tag{2.43}$$

The pooled variance $\widehat{V} = \frac{T-1}{T}W + \frac{M+1}{MT}B$ is an unbiased estimator of the marginal posterior variance of θ under certain stationarity conditions (Gelman and Rubin 1992). The potential scale reduction factor (PSRF) is defined to be the ratio of $\widehat{V}$ and W. If the simulated M chains converge to the target distribution, then PSRF should be close to 1. The PSRF is given as below, where $\hat{d}$ is the degrees of freedom estimate of a t distribution.

$$R_c = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{V}}{W}}$$

According to (Brooks and Gelman, 1998), if for all parameters the conditions $R_c < 1.2$ are satisfied, then it can be concluded that multiple Markov chains are converged. If PSRF reaches beyond this threshold, then simulated chains may not explore the full posterior distribution or longer simulations are needed.

9. Bayesian inferences.

   The SNP status H is estimated by posterior mean. It assesses the probability that the SNP is associated with disease. We will use estimation of SNP status to evaluate the power of model to identify associated SNPs. The hyper-parameter b and other parameters R, A, M are estimated by posterior median. The estimation of these parameters are used to evaluated the prediction accuracy of Bayesian model.

2.5 Simulation study

The simulation study evaluates the performance of single marker analysis using penalized maximum likelihood method and the Bayesian methods. We will consider performance of the

estimators and power of the hypothesis testing. The quality of the estimator is assessed by mean square error (MSE) of the RR estimators. The power of hypothesis testing measures the probability one method identifies a positive variant given that variant is associated. The simulation data are generated in combinations of different settings: $R = 1, 2, 3, A = 0.05, 0.1, 0.2, M = 0.001, 0.005$ and $n = 1000, 3000$. In each setting we generate 400 data sets to assess the performance. A simulated data set will contain a two-way summary counts of normal and tumor genotypes, generated under the settings of the parameters RR, AF, MR and sample size.

The simulated data sets are regarded as independent variants for single marker analysis using penalized MLE and Bayesian methods. We propose to estimate the allelic effect size under additive model, which is one of the most commonly used risk models (Thakkinstian et al., 2005; Attia, 2003). In additive model, the impact of heterozygous genotype is the additive mean of impacts of the homozygous genotypes.

2.5.1 Performance of estimators

The quality of RR estimator is assessed by the MSE. We calculate the MSE of RR estimator with all 400 data under different settings for penalized MLE and single Bayesian methods. Due to sampling variation, the parameter estimation in MLE can be infeasible. The ridge coefficient is added to avoid the RR estimations from being too large, and to keep AF and MR estimations from being on the boundaries. The ridge coefficient is set as 0.05. In Bayesian modeling, the prior distribution $RR|H = 0 \sim Gamma(3, 3)$ centers at 1. The prior $AF|H = 0$ follows Beta distribution that centers at observed allele frequency in sample. The prior $MR|H = 0$ follows Beta distribution that centers at observed mutation rate in sample. Under null hypothesis, the non-associated variant has effect size equaled to 1 and allele frequency remains unchanged from population to control samples. Thus the SNP status tends to be close to 0. Under alternative hypothesis, the prior distribution

$RR|H = 1 \sim Gamma(3, 8)$ has center around 2.6. The prior $AF|H = 1$ follows Beta distribution that has prior mean calculated by the observed allele frequency and effect size in 2.16. The prior $MR|H = 1$ follows Beta distribution that has prior mean calculated by observed mutation rate, observed allele frequency and effect size in 2.21. In alternative hypothesis, the variant has effect size larger than 1 and allele frequency is enriched in control samples. Under this assumption, the prior distribution of AF and MR in population should be adjusted by the formula 2.16-2.21.

The posterior distribution of Bayesian model is drawn from MCMC simulations, which uses Metropolis-Hastings algorithm embedded in the Gibbs sampler structure. Three independent Markov chains are generated with different starting values. The parameters are estimated by the mean or median of posterior distribution after burn-in and thinning. The Figure 2.4 shows the MSE comparison of penalized MLE and Bayesian model under different settings in single marker analysis.

The above comparison shows that Bayesian model systematically has lower MSE than the regularized maximum likelihood estimation. For settings with lower mutation rate, the regularized MLE has much higher estimation errors. Due to few observations in the off-diagonal entries of data matrix, the sampling error can bring high impact on MLE, even with parameter regularization. On the other hand, Bayesian model with predetermined prior distribution has higher accuracy in prediction. In settings with higher mutation rate, regularized MLE and Bayesian model both performs better. In settings with higher mutation rate and higher allele frequency, performance of regularized MLE and Bayesian method is close.

2.5.2 Power analysis

We consider the null hypothesis $H_0 : R = R_0 = 1$ and alternative hypothesis $H_1 : R \neq R_0 = 1$. In MLE method, the Wald test is calculated as 2.8 for RR estimation. The posterior

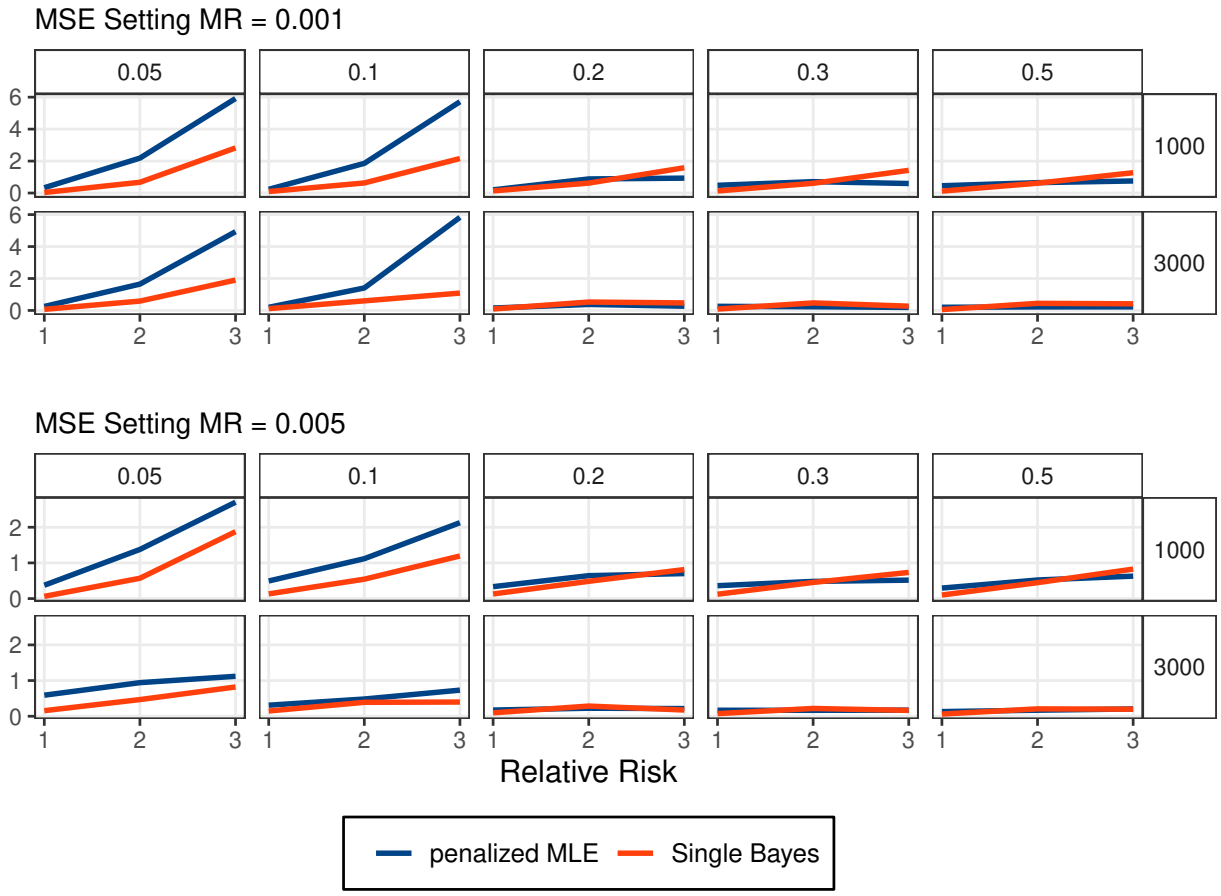

Figure 2.4: Simulation studies of single marker analysis. The MSE is calculated for MLE and Bayesian method.

distribution is sampled from MCMC simulations with three sequences. The probability of SNP status $H = 1$ can be interpreted as the probability of the SNP being associated. The SNP association status is estimated by the posterior median. The Figure 2.5 shows the power comparison of both penalized MLE and Bayesian model under different settings in single marker analysis.

The above comparison shows that the Bayesian model has higher power under settings with higher RR and MR. Under large samples and relative large allele frequency, both performances are similar. The power to identify variants with moderate to large effect size

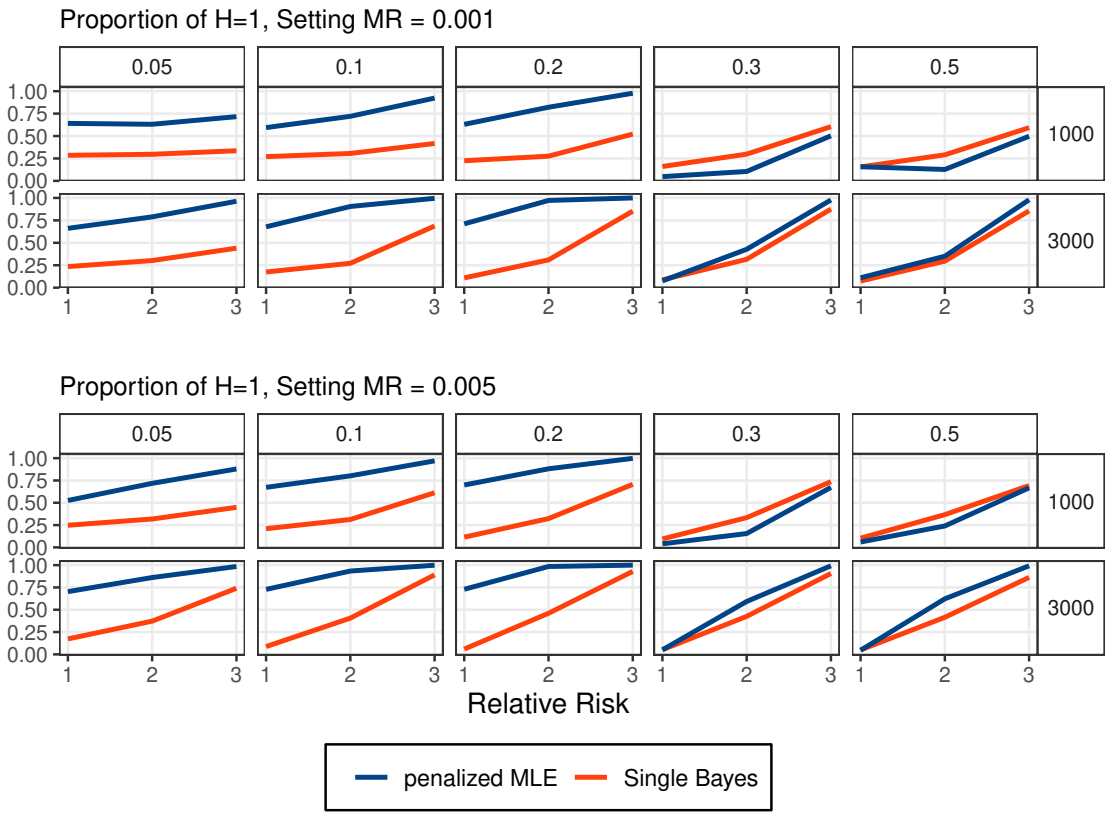

Figure 2.5: Simulation studies of single marker analysis. The MSE is calculated for MLE and Bayesian method.

is high. In the meanwhile, the type I error, where a neutral variant is classified as a risky variant under null hypothesis, is low.

In summary, the simulations reveal that Bayesian model overall performs better in parameter estimation and power analysis. By selecting proper prior distribution associated with SNP status for RR, AF and MR, it can reduce the variance of the estimator and mitigate the impact of sampling variation. Thus MSE of Bayesian estimator is relatively low. By assigning proper prior distribution for SNP status, the posterior distribution can be a good indicator of SNP status under different association scenarios. The power to identify risk variant is relative high while the type I error remains relative low. The penalized MLE method has similar performance in settings with sufficient large sample sizes and high allele frequency. There are several settings where the penalized MLE method has a slightly higher

power and a much higher MSE. The reason is that the penalized MLE is not stable when few data are available and sampling variation presents. From these studies, the Bayesian model has better performance under most scenarios. We will explore multiple marker analysis by extending the Bayesian framework to multiple SNPs in the next chapter.

CHAPTER 3

MULTIPLE MARKERS ANALYSIS

3.1 Introduction

The typical genome-wide association studies have effectively detected large amounts of SNPs that are associated with various diseases, however the identified variants explain a limited portion of the disease heritability (Manolio et al., 2009). There are many associated variants remained undiscovered partly because the single locus GWAS has limited power to reveal the associations. The single locus GWAS typically examines each genetic marker for association and conducts multiple hypothesis testings to get statistically significant variants at the genome level. The high-throughput genotyping technology nowadays has made rapid progress in producing up to millions of candidate SNPs, which are further evaluated and tested individually. In multiple hypothesis testings, the type I error rates should be adjusted using Bonferroni or other correction methods. For hundreds of thousands or even millions of candidate markers, the threshold for each individual marker will be very small, causing very few markers with large effect sizes can be declared significant among all tests. Candidate markers with moderate effect sizes often fail to be identified in the single locus GWAS analysis. In addition, the SNP interactions are ignored in single marker analysis. The consideration of joint effects of a set of related SNPs can play an important role in understanding the complex diseases susceptibility (Li et al., 2014). Thus, limitations in single locus GWAS necessitates a multiple markers analysis, which can jointly examine a SNP-set to improve the detection powers.

We propose a multiple marker model by extending the single-marker Bayesian model. The multiple marker model considers SNP-sets aggregated in a biological meaningful region, such as genes, pathways, or topological 3D structures. For simplicity, we use genes as an example of grouping SNPs into sets in the following sections, with a note that the model applies to any other biologically meaningful ways of grouping SNPs.

3.2 Bayesian Hierarchical Models

The Bayesian hierarchical framework is discussed in Chapter 2 for single marker analysis. In Chapter 3, we extend the Bayesian hierarchical framework to multiple marker analysis. The purpose of developing a Bayesian model is to fully take advantage of all genetic markers by jointly analyze the effect sizes on the gene level. The multiple marker model is applied to determine the association status of a gene, given that the samples of genetic markers on the gene are provided. The multi-locus Bayesian structure is similar to the single-locus Bayesian structure. In the rest of this section, we will discuss the prior distribution, joint distribution and posterior sampling of the multi-locus Bayesian hierarchical model.

3.2.1 Prior Distribution

Suppose there are J SNPs on a gene. Let G denote the gene association status: 1 for associated and 0 for not associated. Gene status G is assigned a Bernoulli prior with probability b, where b is assigned a Beta prior. As is well known that only a small proportion of genes are associated with cancer, the prior mean of b is predetermined at 0.2. The distributions of gene status G and probability b are as follows:

$$\begin{aligned} G &\sim Bernoulli(b) \\ b &\sim Beta(\alpha_b, \beta_b) \end{aligned} \tag{3.1}$$

Let H_j represent the SNP association status for the j^{th} SNP on the gene, where $H_j = 1$ indicates a risky marker, $H_j = 0$ indicates an unrelated or neutral marker, $j = 1, ..., J$. The probability of $H_j = 1$ depends on the gene status G. If the gene is associated with the disease, there is a higher probability that the SNPs on the gene associated with the disease, let $H_j|(G = 1) \sim Bernoulli(p_1)$. On the other hand, if the gene is not associated with the disease, the chance that each SNP is associated is low. Let $H_j|(G = 0) \sim Bernoulli(p_0)$,

where $p_0 = 0.1$ and $p_1 = 0.9$. Thus for each SNP the conditional prior distribution is $H_j|G \sim G \cdot Bernoulli(p_1) + (1-G) \cdot Bernoulli(p_0)$, which is equivalent to:

$$H_j|G \sim Bernoulli(p_1)^G \cdot Bernoulli(p_0)^{(1-G)} \tag{3.2}$$

Let $\mathbf{H} = \{H_1, \ldots, H_J\}$ be the set of parameters for SNP association status. Assume that each H_j is conditionally independent, the prior distribution is as follows:

$$f(\mathbf{H}|G) = f(H_1, \ldots, H_J|G) = \prod_{j=1}^{J} f(H_j|G) \tag{3.3}$$

Let R_j denote the relative risk (RR) of the j^{th} SNP on the gene. The prior distribution of RR depends on the association status of the j^{th} SNP. When the j^{th} SNP is risky, the effect size should be greater than 1. Then R_j is assigned a prior $Gamma(k_1, \theta_1)$ with mean predefined at 2.5. On the other hand, when the j^{th} SNP is neutral, the effect size should be 1. Then R_j is assigned a prior $Gamma(k_0, \theta_0)$ with mean at 1. Let $\mathbf{R} = \{R_1, \ldots, R_J\}$ be a set of RR variables on the same gene. Assume $R_j|H_j$ is independent of each other, and R_j is independent of $H_{j'}$ when $j \neq j'$. Then the prior distribution of RR is as follows:

$$\begin{aligned} R_j|(H_j = 1) &\sim Gamma(k_1, \theta_1) \\ R_j|(H_j = 0) &\sim Gamma(k_0, \theta_0) \\ R_j|H_j &\sim Gamma(k_1, \theta_1)^{H_j}\ Gamma(k_0, \theta_0)^{(1-H_j)} \\ f(\mathbf{R}|\mathbf{H}) &= \prod_{j=1}^{J} f(R_j|H_j) \end{aligned} \tag{3.4}$$

Let A_j denote the allele frequency (AF) of the j^{th} SNP on the gene. The prior distribution of AF depends on the association status of the j^{th} SNP. When a SNP is neutral, the allele frequency is neither enriched nor diminished in the disease population. That is, we expect to observe the same the true allele frequency in the disease population. The observed AF at j^{th} SNP is $A_{obs,j} = \frac{n_{10}+n_{11}+n_{12}+2n_{20}+2n_{21}+2n_{22}}{2n}$, which is an estimator of true AF when a SNP

is assumed to be non-associated. Thus $A_j|(H_j = 0)$ is assigned a prior $Beta(\alpha_{A0,j}, \beta_{A0,j})$ with mean at $\mu_{A_{obs,j}}$.

On the other hand, for a risky SNP, the allele frequency is enriched in the disease population. The level of enrichment depends on effect size and mutation rate from equation 2.14. The effect size is estimated to be 2.5 for risky SNP. Let MR parameter be estimated by the observed MR. Then for a risky association status, $A_j|(H_j = 1)$ is assigned a prior $Beta(\alpha_{A1,j}, \beta_{A1,j})$, with prior mean evaluated by formula 2.17. Let $\mathbf{A} = \{A_1, \ldots, A_J\}$ be a set of AF variables on the same gene. Assume that $A_j|H_j$ is independent of each other, and A_j is independent of $H_{j'}$ when $j \neq j'$. Therefore, the distribution of AF is as follows:

$$\begin{aligned} A_j|_{H_j=1} &\sim Beta(\alpha_{A1,j}, \beta_{A1,j}) \\ A_j|_{H_j=0} &\sim Beta(\alpha_{A0,j}, \beta_{A0,j}) \\ f(A_j|H_j) &= Beta(\alpha_{A1,j}, \beta_{A1,j})^{H_j}\ Beta(\alpha_{A0,j}, \beta_{A0,j})^{(1-H_j)} \\ f(\mathbf{A}|\mathbf{H}) &= \prod_{j=1}^{J} f(A_j|H_j) \end{aligned} \tag{3.5}$$

Let M_j represent the mutation rate (MR) of the j^{th} SNP on the gene. The prior distribution of MR depends on the association status of the j^{th} SNP. When the SNP is non-associated, the MR variable $M_j|(H_j = 0)$ is assigned a prior $Beta(\alpha_{M0,j}, \beta_{M0,j})$. From 2.19 , the observed MR $M_{obs,j}$ of the j^{th} SNP is an estimator of true MR since allele frequency does not change in disease population. For a risky SNP, the observed MR depends on the true RR, AF and MR, thus the MR variable $M_j|(H_j = 1)$ is assigned a prior $Beta(\alpha_{M1,j}, \beta_{M1,j})$ with mean calculated from 2.21. Let $\mathbf{M} = \{M_1, \ldots, M_J\}$ be a set of MR variables on the same gene. Assume that $M_j|H_j$ is independent of each other, and M_j is independent of $H_{j'}$ when $j \neq j'$. Therefore, the distribution of MR is as follows:

$$
\begin{aligned}
M_j|_{H_j=1} &\sim Beta(\alpha_{M1,j}, \beta_{M1,j}) \\
M_j|_{H_j=0} &\sim Beta(\alpha_{M0,j}, \beta_{M0,j}) \\
f(M_j|H_j) &= Beta((\alpha_{M1,j}, \beta_{M1,j}))^{H_j}\ Beta(\alpha_{M0,j}, \beta_{M0,j})^{(1-H_j)} \\
f(\mathbf{M}|\mathbf{H}) &= \prod_{j=1}^{J} f(M_j|H_j)
\end{aligned}
\tag{3.6}
$$

3.2.2 Joint Posterior Distribution

Let $\mathbf{\Theta_j} = \{R_j, A_j, M_j\}$ be the set of RR, AF and MR variables on the j^{th} SNP, where $\mathbf{\Theta_j}$ and $\mathbf{\Theta'_j}$ are independent when $j \neq j'$. Let $\mathbf{\Theta} = \{R_1, ..., R_J, A_1, ..., A_J, M_1, ..., M_J\}$ be the set of RR, AF and MR variables on all SNPs of a given gene. Let $\mathbf{H} = \{H_1, ..., H_J\}$ be a set of all SNP association status on a given gene. Under assumption that $R_j|H_j,\ A_j|H_j,\ M_j|H_j$ are independent of each other, the joint distribution of $\mathbf{\Theta}, \mathbf{H}, G, b$ can be derived as follows:

$$
\begin{aligned}
f(\mathbf{\Theta}, \mathbf{H}, G, b) &= f(\mathbf{\Theta}|\mathbf{H}, G, b)\ f(\mathbf{H}, G, b) \\
&= f(\mathbf{\Theta}|\mathbf{H}, G, b)\ f(\mathbf{H}|G, b)\ f(G, b) \\
&= f(\mathbf{\Theta}|\mathbf{H}, G, b)\ f(\mathbf{H}|G, b)\ f(G|b)\ f(b) \\
&= f(\mathbf{\Theta}|\mathbf{H})\ f(\mathbf{H}|G)\ f(G|b)\ f(b) \\
&= f(\mathbf{R}, \mathbf{A}, \mathbf{M}|\mathbf{H})\ f(\mathbf{H}|G)\ f(G|b)\ f(b) \\
&= f(\mathbf{R}|\mathbf{H}) f(\mathbf{A}|\mathbf{H}) f(\mathbf{M}|\mathbf{H})\ f(\mathbf{H}|G)\ f(G|b)\ f(b) \\
&= \left(\prod_{j=1}^{J} f(R_j|H_j) f(A_j|H_j) f(M_j|H_j)\ f(H_j|G) \right)\ f(G|b)\ f(b)
\end{aligned}
\tag{3.7}
$$

Based on the sampling framework in Chapter 2, the j^{th} SNP samples $\mathbf{S_j} = \{s_{00(j)}, s_{01(j)}, s_{02(j)}, s_{10(j)}, s_{11(j)}, s_{12(j)}, s_{20(j)}, s_{21(j)}, s_{22(j)}\}$ follow a Multinoulli distribution with 9 categories, where the expected probability of $P_{k,k'}$ in each category is a function of RR, AF and MR. For each marker, $f(\mathbf{S_j}|\mathbf{\Theta_j}) = \prod_{k,k'}^{0,1,2} P_{k,k'(j)}^{s_{k,k'(j)}}$.

Let $\mathbf{S} = \{\mathbf{n_1}, ..., \mathbf{n_J}\}$ be the set of samples on a given gene, where each column are samples on the j^{th} SNP. The samples $\mathbf{S_j}$ is independent of $\mathbf{\Theta'_j}$ when $j \neq j'$. The likelihood function is as follows:

$$\begin{aligned} f(\mathbf{S}|\mathbf{\Theta}) &= \prod_{j=1}^{J} f(\mathbf{S_j}|\mathbf{\Theta_j}) \\ &= \prod_{j=1}^{J} \prod_{k,k'}^{0,1,2} P_{k,k'(j)}^{s_{k,k'(j)}} \end{aligned} \tag{3.8}$$

Therefore the joint posterior distribution can be derived as follows:

$$\begin{aligned} f(\mathbf{S}, \mathbf{\Theta}, \mathbf{H}, G, b) &= f(\mathbf{S}|\mathbf{\Theta}, \mathbf{H}, G, b)\ f(\mathbf{\Theta}, \mathbf{H}, G, b) \\ &= f(\mathbf{S}|\mathbf{\Theta})\ f(\mathbf{\Theta}, \mathbf{H}, G, b) \\ &= f(\mathbf{S}|\mathbf{\Theta})\ f(\mathbf{R}, \mathbf{A}, \mathbf{M}|\mathbf{H})\ f(\mathbf{H}|G)\ f(G|b)\ f(b) \\ &= f(\mathbf{S}|\mathbf{\Theta}) \left(\prod_{j=1}^{J} f(R_j|H_j) f(A_j|H_j) f(M_j|H_j)\ f(H_j|G) \right)\ f(G|b)\ f(b) \end{aligned} \tag{3.9}$$

where each part is given from equations 3.1, 3.3, 3.4, 3.5, 3.6, 3.8.

The posterior distributions do not have closed-form expressions, thus it is difficult to sample from the posterior distributions directly. Alternatively, conditional posterior distribution can be easily derived. Markov Chain Monte Carlo method is used to draw samples from conditional posterior distribution. The Gibbs sampler and Metropolis-Hasting algorithm are combined to approximate target distribution. For variables having a closed-form conditional posterior distribution, Gibbs sampler is used to approximate target distribution.

3.2.3 Posterior sampling

The posterior distribution is estimated using Markov Chain Monte Carlo (MCMC) simulation. Parameters are updated iteratively by Gibbs sampling framework. Variables $b, G, \mathbf{H}$

are sampled directly from closed-form conditional posterior distribution, while other variables are sampled from Metropolis-Hastings algorithms within each update step in the Gibbs framework. For a gene with J SNPs, the Markov Chain Monte Carlo simulation is given as follows:

1. Initialization of parameters.

   Assign random initial values to $b^0, G^0, \mathbf{H^0}, \mathbf{R^0}, \mathbf{A^0}, \mathbf{M^0}$ parameters. For each gene, three chains with different initial status are generated in Markov Chain Monte Carlo simulation.

2. Update b^t.

   Given other parameters are fixed, the conditioned posterior distribution is derived as follows:

$$\begin{aligned} f(b^t|\mathbf{S}, G^t, \mathbf{H^t}, \mathbf{\Theta^t},) &= \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{\int f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)\ db^t} \\ &= \frac{f(G^t|b^t)f(b^t)}{\int f(G^t|b^t)f(b^t)\ db^t} \\ &\propto f(G^t|b^t)f(b^t) \\ &\propto (b^t)^{G^t+\alpha_b-1}(1-b^t)^{\beta_b-G^t} \end{aligned} \tag{3.10}$$

   The right hand side is proportional to a Beta distribution. Thus conditioned posterior distribution of b is a Beta distribution with parameters $\alpha_{new} = G^{(t)} + \alpha_b,\ \beta_{new} = \beta_b - G^{(t)} + 1$. Update b^t by sampling from the Beta distribution based on current state:

$$b^{t+1} \sim Beta(\alpha_{new} = G^{(t)} + \alpha_b,\ \beta_{new} = \beta_b - G^{(t)} + 1)$$

3. Update G.

$$
\begin{aligned}
f(G^t|\mathbf{S}, b^t, \mathbf{H^t}, \mathbf{\Theta^t}) &= \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, b^t)} \\
&= \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{\sum_{g=0,1} f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t = g, b^t)} \\
&= \frac{f(\mathbf{S}|\mathbf{\Theta^t})\ f(\mathbf{\Theta^t}|\mathbf{H^t})\ f(\mathbf{H^t}|G^t)\ f(G^t|b^t)\ f(b^t)}{\sum_{g=0,1} f(\mathbf{S}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|\mathbf{H^t}) f(\mathbf{H^t}|G^t = g) f(G^t = g|b^t) f(b^t)} \\
&= \frac{f(\mathbf{H^t}|G^t) f(G^t|b^t)}{\sum_{g=0,1} f(\mathbf{H^t}|G^t = g) f(G^t = g|b^t)} \\
&\propto f(\mathbf{H^t}|G^t) f(G^t|b^t)
\end{aligned}
\tag{3.11}
$$

Substitute the previous equations, there are

$$
f(G^t|\ b^t, \mathbf{H^t}, \mathbf{\Theta^t}, \mathbf{S}) \propto \left[b^t p_1^{\sum H_j^t} (1-p_1)^{J - \sum H_j^t} \right]^{G^t} \cdot \left[(1-b^t) \cdot p_0^{\sum H_j^t} (1-p_0)^{J - \sum H_j^t} \right]^{(1-G^t)} \tag{3.12}
$$

The conditioned posterior probability density function of G^t is proportional to a Bernoulli distribution. The next state G^{t+1} can be sampled from Gibbs sampler based on current state:

$$
G^{t+1} \sim Bernoulli \left(\frac{b^t p_1^{\sum H_j^t} (1-p_1)^{J - \sum H_j^t}}{b^t p_1^{\sum H_j^t} (1-p_1)^{J - \sum H_j^t} + (1-b^t) \cdot p_0^{\sum H_j^t} (1-p_0)^{J - \sum H_j^t}} \right) \tag{3.13}
$$

4. Update $\mathbf{H} = H_1, H_2, ..., H_J$.

Let $H_j = 0$ *or* 1 denote the association status of each j^{th} SNP on the gene. When $H_j = 1$, the j^{th} SNP is associated. Otherwise the j^{th} SNP is not associated. Given other parameters are fixed, the conditioned posterior distribution at step t can be derived as follows:

$$
\begin{aligned}
&f(H_j^t|\mathbf{S}, b^t, G^t, H_1^t, ..., H_{j-1}^t, H_{j+1}^t, ..., H_J^t, \mathbf{\Theta^t}) \\
&= \frac{f(\mathbf{S}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|\mathbf{H^t}) f(\mathbf{H^t}|G^t) f(G^t|b^t)\ f(b^t)}{\sum_{H_j^t=0,1} f(\mathbf{S}|\mathbf{\Theta^t}) f(\mathbf{\Theta^t}|\mathbf{H^t}) f(\mathbf{H^t}|G^t) f(G^t|b^t)\ f(b^t)} \\
&= \frac{f(\mathbf{S}|\mathbf{\Theta^t}) \left(\prod_{l=1}^{J} f(R_l^t|H_l^t) f(A_l^t|H_l^t) f(M_l^t|H_l^t) f(H_l^t|G^t)\right) f(G^t|b^t)\ f(b^t)}{\sum_{H_j^t=0,1} f(\mathbf{S}|\mathbf{\Theta^t}) \left(\prod_{l=1}^{J} f(R_l^t|H_l^t) f(A_l^t|H_l^t) f(M_l^t|H_l^t) f(H_l^t|G^t)\right) f(G^t|b^t)\ f(b^t)} \\
&= \frac{f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t) f(H_j^t|G^t)}{\sum_{H_j^t=0,1} f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t) f(H_j^t|G^t)} \\
&\propto f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t) f(H_j^t|G^t)
\end{aligned}
\tag{3.14}
$$

The conditional posterior distribution of H_j^t is proportional to Bernoulli distribution. The next state can be directly sampled:

$$
\begin{aligned}
&f(H_j^t|\mathbf{S}, b^t, G^t, H_1^t, ..., H_{j-1}^t, H_{j+1}^t, ..., H_J^t, \mathbf{\Theta^t}) \\
\propto &\Big(Gamma(R_j^t; k_1, \theta_1) Beta(A_j^t; \alpha_{A1,j}, \beta_{A1,j}) Beta(M_j^t; \alpha_{M1,j}, \beta_{M1,j}) {p_1}^{G^t} {p_0}^{1-G^t}\Big)^{H_j^t} \cdot \\
&\Big(Gamma(R_j^t; k_0, \theta_0) Beta(A_j^t; \alpha_{A0,j}, \beta_{A0,j}) Beta(M_j^t; \alpha_{M0,j}, \beta_{M0,j}) (1-p_1)^{G^t} (1-p_0)^{1-G^t}\Big)^{1-H_j^t}
\end{aligned}
\tag{3.15}
$$

5. Update relative risk $\mathbf{R^t} = R_1^t, R_2^t, ..., R_J^t$.

   Parameters are updated one by one in the order $R_1^0, R_2^0, ..., R_J^0, R_1^1, R_2^1, ..., R_J^1, ..., R_1^t, R_2^t, ..., R_J^t, R_1^{t+1}, R_2^{t+1}, ..., R_J^{t+1}$, conditioning on all other parameters being fixed at their current values. Since there is no closed-form for conditional posterior distribution, the desired conditional posterior distribution will be sampled by Metropolis-Hastings algorithm within each Gibbs step.

   5.1. Conditional posterior distribution.

   The proportional posterior distribution of R_j^t conditioned on other parameters is derived as follows:

$$
\begin{aligned}
&f(R_j^t|\mathbf{S}, b^t, G^t, \mathbf{H^t}, R_1^t, ..., R_{j-1}^t, R_{j+1}^t, ..., R_J^t, \mathbf{A^t}, \mathbf{M^t}) \\
&\quad = \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{\int f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)\ dR_j^t} \\
&\quad = \frac{\left(\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right) f(G^t|b^t) f(b^t)}{\int \left(\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right) f(G^t|b^t) f(b^t)\ dR_j^t} \\
&\quad = \frac{f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)}{\int f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)\ dR_j^t} \\
&\quad = \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)\ dR_j^t} \\
&\quad = \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t)\ dR_j^t} \\
&\quad \propto f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t)
\end{aligned}
\tag{3.16}
$$

The marginal distribution can hardly be integrated. Thus the proportional posterior distribution is used and Metropolis-Hastings algorithm is applied.

5.2. Proposal distribution.

Gamma distribution is used as the proposal distribution to generate a candidate value R_j^* in the neighborhood based on the value of current state R_j^t. The transition probability and reverse transition probability are as follows:

$$q(R_j^*|R_j^t) = Gamma(shape = 1 + 5R_j^t, rate = 5)$$

$$q(R_j^t|R_j^*) = Gamma(shape = 1 + 5R_j^*, rate = 5)$$

5.3. Acceptance probability.

Compute the acceptance ratio and determine the acceptance probability:

$$acceptance\ ratio = \frac{f(\mathbf{n_j^t}|R_j^*, A_j^t, M_j^t) f(R_j^*|H_j^t) \cdot q(R_j^t|R_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) \cdot q(R_j^*|R_j^t)} \tag{3.17}$$

Then compute the acceptance probability:

$$r = min\left\{1, \frac{f(\mathbf{n_j^t}|R_j^*, A_j^t, M_j^t) f(R_j^*|H_j^t) \cdot q(R_j^t|R_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) \cdot q(R_j^*|R_j^t)}\right\}$$

The acceptance indicator is generated from $Bernoulli(r)$ distribution with probability equals r. If acceptance indicator equals to 1, the next state $R_j^{*1} = R_j^*$, otherwise $R_j^{*1} = R_j^t$.

5.4. Iterative updates.

For Metropolis-Hastings iterations $h = 1, 2, ..., 99$, take R_j^{*h} as the current state, repeat the above steps to update the next state in Metropolis-Hastings sequences $R_j^{*1}, R_j^{*2}, ..., R_j^{*100}$. Finally, the next state of Gibbs iterations R_j^{t+1} is updated with the last state in Metropolis-Hastings iterations:

$$R_j^{t+1} = R_j^{*H} \textit{ where } H \textit{ is the last state in } MH \textit{ iterations}$$

6. Update allele frequency $\mathbf{A^t} = A_1^t, A_2^t, ..., A_J^t$.

Parameters are updated one by one in the order $A_1^0, A_2^0, ..., A_J^0, A_1^1, A_2^1, ..., A_J^1, ..., A_1^t, A_2^t, ..., A_J^t, A_1^{t+1}, A_2^{t+1}, ..., A_J^{t+1}$, given that all other parameters are fixed at their current values. Within each Gibbs iteration, the Metropolis-Hastings algorithm proposes and updates the parameter for 100 times.

6.1. Conditional posterior distribution.

The proportional posterior distribution of A_j^t conditioning on other parameters is derived as follows:

$$
\begin{aligned}
&f(A_j^t|\mathbf{S}, b^t, G^t, \mathbf{H^t}, \mathbf{R^t}, A_1^t, ..., A_{j-1}^t, A_{j+1}^t, ..., A_J^t, \mathbf{M^t}) \\
&= \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{\int f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)\ dA_j^t} \\
&= \frac{\left[\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right] f(G^t|b^t) f(b^t)}{\int \left[\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right] f(G^t|b^t) f(b^t)\ dA_j^t} \\
&= \frac{f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)}{\int f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)\ dA_j^t} \\
&= \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)\ dA_j^t} \\
&= \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(A_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(A_j^t|H_j^t)\ dA_j^t} \\
&\propto f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(A_j^t|H_j^t)
\end{aligned}
\tag{3.18}
$$

The marginal distribution of A_j^t can hardly be integrated. Thus the proportional posterior distribution is used and Metropolis-Hastings algorithm is applied.

6.2. Proposal distribution.

Beta distribution is used as the proposal distribution to generate a candidate value A_j^* in the neighborhood based on the value of current state A_j^t. The transition probability and reverse transition probability are as follows:

$$
\begin{aligned}
q(A_j^*|A_j^t) &= Beta(\alpha = \frac{1 + 10A_j^t}{1 - A_j^t}, \beta = 10) \\
q(A_j^t|A_j^*) &= Beta(\alpha = \frac{1 + 10A_j^*}{1 - A_j^*}, \beta = 10)
\end{aligned}
\tag{3.19}
$$

6.3. Acceptance probability.

Compute the acceptance ratio and determine the acceptance probability:

$$
acceptance\ ratio = \frac{f(\mathbf{n_j^t}|R_j^t, A_j^*, M_j^t) f(A_j^*|H_j^t) \cdot q(A_j^t|A_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(A_j^t|H_j^t) \cdot q(A_j^*|A_j^t)} \tag{3.20}
$$

Then acceptance probability is determined as:

$$r = min\left\{1, \quad \frac{f(\mathbf{n_j^t}|R_j^t, A_j^*, M_j^t)f(A_j^*|H_j^t)\cdot q(A_j^t|A_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t)f(A_j^t|H_j^t)\cdot q(A_j^*|A_j^t)}\right\}$$

The acceptance indicator is generated from $Bernoulli(r)$ distribution with probability equals to r. If acceptance indicator equals to 1, the next state in Metropolis-Hastings iterations is updated $A_j^{*1} = A_j^*$. Otherwise $A_j^{*1} = A_j^t$.

6.4. Iterative updates.

For Metropolis-Hastings iterations $h = 1, 2, ..., 99$, take A_j^{*h} as the current state, repeat the above steps to update the next state in Metropolis-Hastings sequences $A_j^{*1}, A_j^{*2}, ..., A_j^{*100}$. Finally, the next state of Gibbs iteration A_j^{t+1} is updated with the last state in Metropolis-Hastings iterations:

$$A_j^{t+1} = A_j^{*H} \ where \ H \ is \ the \ last \ state \ in \ MH \ iterations$$

7. Update mutation rate $\mathbf{M} = M_1, M_2, ..., M_J$.

Parameters are updated one by one in the order $M_1^0, M_2^0, ..., M_J^0, M_1^1, M_2^1, ..., M_J^1$, $..., M_1^t, M_2^t, ..., M_J^t, M_1^{t+1}, M_2^{t+1}, ..., M_J^{t+1}$ when all other parameters are fixed. Within each Gibbs iteration, the Metropolis-Hastings algorithm proposes and updates the parameter for 100 times. Then the next Gibbs iteration is updated from the Metropolis-Hastings iterations.

7.1. Conditional posterior distribution.

The proportional posterior distribution of M_j^t conditioning on other parameters is derived as follows:

$$
\begin{aligned}
&f(M_j^t|\mathbf{S}, b^t, G^t, \mathbf{H^t}, \mathbf{R^t}, \mathbf{A^t}, M_1^t, ..., M_{j-1}^t, M_{j+1}^t, ..., M_J^t) \\
&= \frac{f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)}{\int f(\mathbf{S}, \mathbf{\Theta^t}, \mathbf{H^t}, G^t, b^t)\ dM_j} \\
&= \frac{\left[\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right] f(G^t|b^t) f(b^t)}{\int \left[\prod_{l=1}^{J} f(\mathbf{n_l^t}|\mathbf{\Theta_l^t}) f(\mathbf{\Theta_l^t}|H_l^t) f(H_l^t|G^t)\right] f(G^t|b^t) f(b^t)\ dM_j^t} \\
&= \frac{f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)}{\int f(\mathbf{n_j^t}|\mathbf{\Theta_j^t}) f(\mathbf{\Theta_j^t}|H_j^t)\ dM_j^t} \\
&= \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(R_j^t|H_j^t) f(A_j^t|H_j^t) f(M_j^t|H_j^t)\ dM_j^t} \\
&= \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(M_j^t|H_j^t)}{\int f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(M_j^t|H_j^t)\ dM_j^t} \\
&\propto f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(M_j^t|H_j^t)
\end{aligned}
\tag{3.21}
$$

The marginal distribution of M_j^t can hardly be integrated. Thus the proportional posterior distribution is used and Metropolis-Hastings algorithm is applied.

7.2. Proposal distribution.

Beta distribution is used as the proposal distribution to generate a candidate value M_j^* in the neighborhood based on the value of current state M_j^t. The transition probability and reverse transition probability are as follows:

$$
\begin{aligned}
q(M_j^*|M_j^t) &= Beta(\alpha = \frac{1 + 1000 M_j^t}{1 - M_j^t}, \beta = 1002) \\
q(M_j^t|M_j^*) &= Beta(\alpha = \frac{1 + 1000 M_j^*}{1 - M_j^*}, \beta = 1002)
\end{aligned}
\tag{3.22}
$$

7.3. Acceptance probability.

Compute the acceptance ratio and determine the acceptance probability:

$$
acceptance\ ratio = \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^*) f(M_j^*|H_j^t) \cdot q(M_j^t|M_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t) f(M_j^t|H_j^t) \cdot q(M_j^*|M_j^t)}
\tag{3.23}
$$

Then acceptance probability is determined as:

$$r = min\left\{1, \quad \frac{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^*)f(M_j^*|H_j^t)\cdot q(M_j^t|M_j^*)}{f(\mathbf{n_j^t}|R_j^t, A_j^t, M_j^t)f(M_j^t|H_j^t)\cdot q(M_j^*|M_j^t)}\right\}$$

The acceptance indicator is generated from $Bernoulli(r)$ distribution with probability equals to r. If acceptance indicator equals to 1, the next state in Metropolis-Hastings iterations is updated $M_j^{*1} = M_j^*$. Otherwise $M_j^{*1} = M_j^t$.

7.4. Iterative updates.

For Metropolis-Hastings iterations $h = 1, 2, ..., 99$, take M_j^{*h} as the current state, repeat above steps to update the next state in Metropolis-Hastings sequences $M_j^{*1}, M_j^{*2}, ..., M_j^{*100}$. Finally, the next state of Gibbs iteration M_j^{t+1} is updated with the last state in Metropolis-Hastings iterations:

$$M_j^{t+1} = M_j^{*H} \ where\ H\ is\ the\ last\ state\ in\ MH\ iterations$$

8. Burn-in and thinning.

The burn-in and thinning process is performed to correct the potential bias introduced by the initial status. Some starting points may over sample the regions that are rare events. The beginning samples may not be stabilized to the stationary distribution, thus it should not contribute to the inferences. In this simulation, the initial 20% period of Gibbs sampler iterations is discarded and every 5th value is kept. The remaining sequences are used for inferences after burn-in and thinning process.

9. Convergence and diagnosis.

A major consideration in MCMC simulations is the convergence of Markov chains. The simulated chains are expected to fully explore the target distribution. Multiple chains initiated with different starting values are adopted. A common method to assess the MCMC convergence is to analyze and compare the differences between multiple chains.

The Gelman-Rubin statistic is applied to diagnose the MCMC convergence by analyzing the differences between multiple Markov chains. The convergence is evaluated by comparing the estimated between-chains and within-chains variances. Large differences between these variances indicate non-convergence.

Suppose there are M simulated chains with length T. Given a model parameter θ, let θ_{mt} be the value at the t^{th} updates on the m^{th} simulated chains, where $t = 1, 2, ..., T$ and $m = 1, 2, ..., M$. Let $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the m^{th} chain. The between-chains variance B and within-chains variance W are given by

$$\begin{aligned} B &= \frac{T}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2 \\ W &= \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}_m^2 \end{aligned} \tag{3.24}$$

The pooled variance $\widehat{V} = \frac{T-1}{T}W + \frac{M+1}{MT}B$ is an unbiased estimator of the marginal posterior variance of θ under certain stationarity conditions (Gelman and Rubin 1992). The potential scale reduction factor (PSRF) is defined to be the ratio of $\widehat{V}$ and W. If the simulated M chains converge to the target distribution, then PSRF should be close to 1. The PSRF is given as below, where $\hat{d}$ is the degrees of freedom estimate of a t distribution.

$$R_c = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{V}}{W}}$$

According to (Brooks and Gelman, 1998), if for all parameters the conditions $R_c < 1.2$ are satisfied, then it can be concluded that multiple Markov chains are converged. If PSRF reaches beyond this threshold, then simulated chains may not explore the full posterior distribution or longer simulations are needed.

3.3 Simulation study

The simulation studies are carried out under different settings to assess the performance of multi-marker Bayesian model, single-marker Bayesian model and penalized MLE methods for paired tumor-normal data. Similar to simulation studies in single marker analysis, we consider the quality of estimators and hypothesis testing. The quality of estimator is assessed by the mean square error (MSE) of RR estimator. The hypothesis testing measures the type I error under null hypothesis, and the power to identify an associated variant under alternative hypothesis. The simulation data is generated in combinations of different setting: $R = 1, 2, 3, A = 0.05, 0.1, 0.2, M = 0.001, 0.005$ and sample size $n = 1000, 3000$. In each setting there are 400 data sets generated to provide adequate data to assess performance. Each simulated data set is a two-way summary counts of normal and tumor genotypes generated under setting parameters RR, AF, MR and sample size. The simulated data is independent variant without linkage disequilibrium (LD). The allelic effect size is evaluated by additive genetic model, where heterozygous genotype impact is the additive mean of homozygous genotype impact (Ziegler and König, 2010).

We consider the situation where 4 similar SNPs are grouped together to evaluate the joint effect of the SNP set. In multi-marker Bayesian model, the 4 SNPs share the same gene statue while in single-marker Bayesian model the individual gene statue is identical to the SNP status. In each settings 400 SNPs data sets are generated. There are 100 artificial genes formed in multi-marker Bayesian model. The Figure 3.1 shows the SNPs and genes in simulation studies.

In multiple marker simulation studies, we extend the hierarchical Bayesian model to combine neighboring biomarkers and inference the status of the SNP set by jointly evaluating the impact of multiple SNPs. Compared to single locus association analysis in a typical GWAS, combining SNPs that share similar biological characteristics has shown advantages in aggregating individual moderate effects.

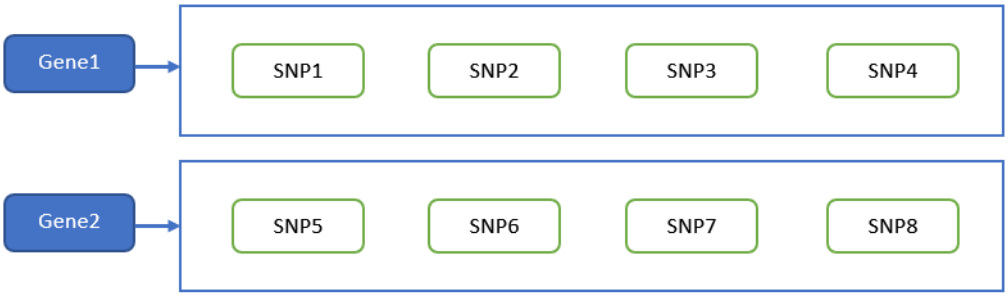

Figure 3.1: Simulation for multiple marker analyzes on a SNP set of 4 markers and estimate aggregated gene status of each group

3.3.1 Quality of estimator

The quality of RR estimator is assessed by MSE. We calculate the MSE with the simulated 400 data sets under different settings. Comparison of MSE is conducted for penalized MLE, the single Bayesian model and the multiple Bayesian model. In the penalized MLE method, we let ridge coefficient = 0.05 to regularize the RR estimation.

In Bayesian modeling, the prior distribution under given neutral SNP status is $RR|H = 0 \sim Gamma(3, 3)$. $AF|H = 0$ follows Beta distribution with mean at the observed allele frequency. $MR|H = 0$ follows Beta distribution with mean at observed mutation rate. Under null hypothesis, the neutral variant has allelic effect size equal to 1. The allele frequency is not enriched or diminished through sampling for a certain disease , and thus remains unchanged from population to case samples. The mutation rate remains unchanged from population to samples under null hypothesis as well. Therefore, the observed AF and observed MR is assumed to be the prior mean given neutral SNP status. Under alternative hypothesis, the prior $RR|H = 1 \sim Gamma(3, 8)$ has center around 2.6. $AF|H = 1$ follows Beta distribution with mean calculated by the observed AF and a moderate effect size in 2.16. $MR|H = 1$ follows Beta distribution with mean calculated by the observed MR, observed AF and a moderate effect size 2.21. When variant is associated with the disease, the allele frequency is generally enriched and mutation rate is impacted in the case samples. The calculations 2.16, 2.21 are to adjust for the prior mean under alternative hypothesis.

MCMC simulations are used to drawn posterior distribution for Bayesian model parameters. To handle more variables, the Metropolis-Hastings algorithm is embedded in Gibbs sampler framework to sample from complicated distribution. Three chains are generated for each variable. The Bayesian estimation is derived by the mean or median of conditional posterior distribution after burn-in and thinning. The Figure 3.2, 3.3 shows the quality of estimators in penalized MLE, single Bayesian model and multiple Bayesian model.

The results show that multiple-marker Bayesian model has the lowest MSE in most settings. This indicates that multiple Bayesian models can have better RR estimation than single SNP by considering the joint effect of variants on neighboring locations. The improvement of estimation performance is much better in settings without sufficient sample size, allele frequency and mutation rate. This advantage in multiple Bayesian model will be helpful in GWAS with limited sample size and allele frequency.

3.3.2 Power analysis

Let the null hypothesis be $H_0 : R = R_0 = 1$ and alternative hypothesis $H_1 : R \neq R_0 = 1$. The Wald test is used in penalized MLE method. In single and multiple Bayesian models, let the SNP status H represent the probability of variant has association. The prior distribution of SNP status $H|G = 1 \sim Bernoulli(p1)$ and $H|G = 1 \sim Bernoulli(p0)$ is dependent on gene status. In single Bayesian model, $p1 = 0.99$ and $p0 = 0.01$ are assigned such that G is equivalent to H. In multiple Bayesian model, the value of p1 and p0 is determined by an assumption of probability of positive variants on the gene. The gene status has prior distribution $G \sim Bernoulli(b)$. The posterior distribution of G and H is drawn by Gibbs sampler with burn-in and thinning. The Figure 3.4, 3.5 show the power comparison of the three methods.

The type I error of multiple Bayesian model is lowest among all three methods in most settings. In moderate risk $R = 2$ settings, the power of multiple Bayesian exceeds single

Bayesian model in some settings. From $R = 2$ to $R = 3$ settings, the power improvement of multiple Bayesian is much higher than the other two methods. Their performance are similar in settings with large sample size and risk allele frequency.

Overall, from both MSE and power analysis results, the multiple Bayesian model outperforms the other two methods in most settings. The Bayesian model provides a more stable estimation than penalized MLE. In scenarios where data are limited by insufficient sample sizes, small allele frequencies or low mutation rates, the multiple Bayesian model remains a relative low type-I-error rate, and has a higher power improvement under the alternative hypothesis.

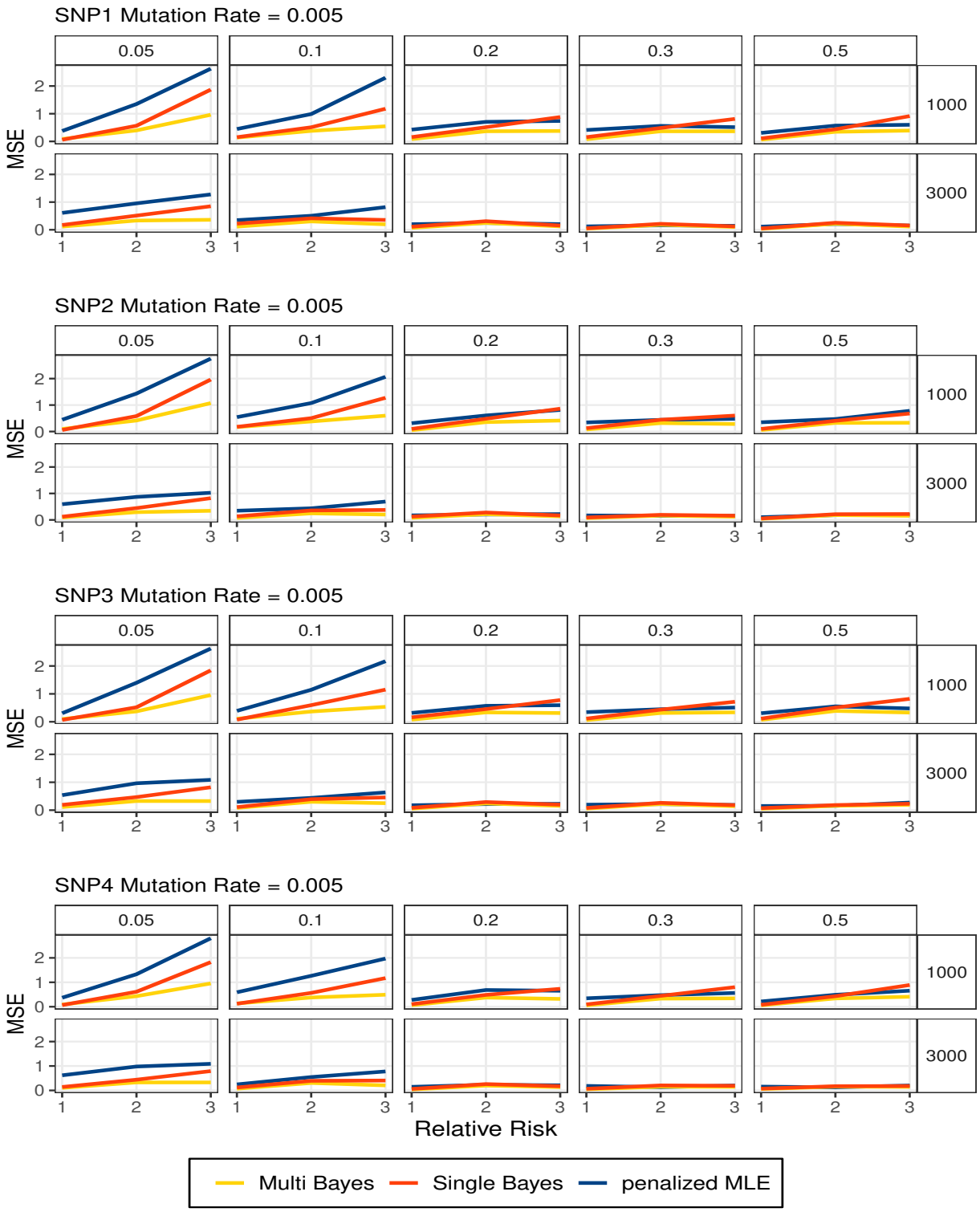

Figure 3.2: Mean square error comparison of penalized MLE, Single Bayesian and Multiple Bayesian model in setting MR = 0.005. Comparison shows that multiple Bayesian model has lowest MSE in most settings.

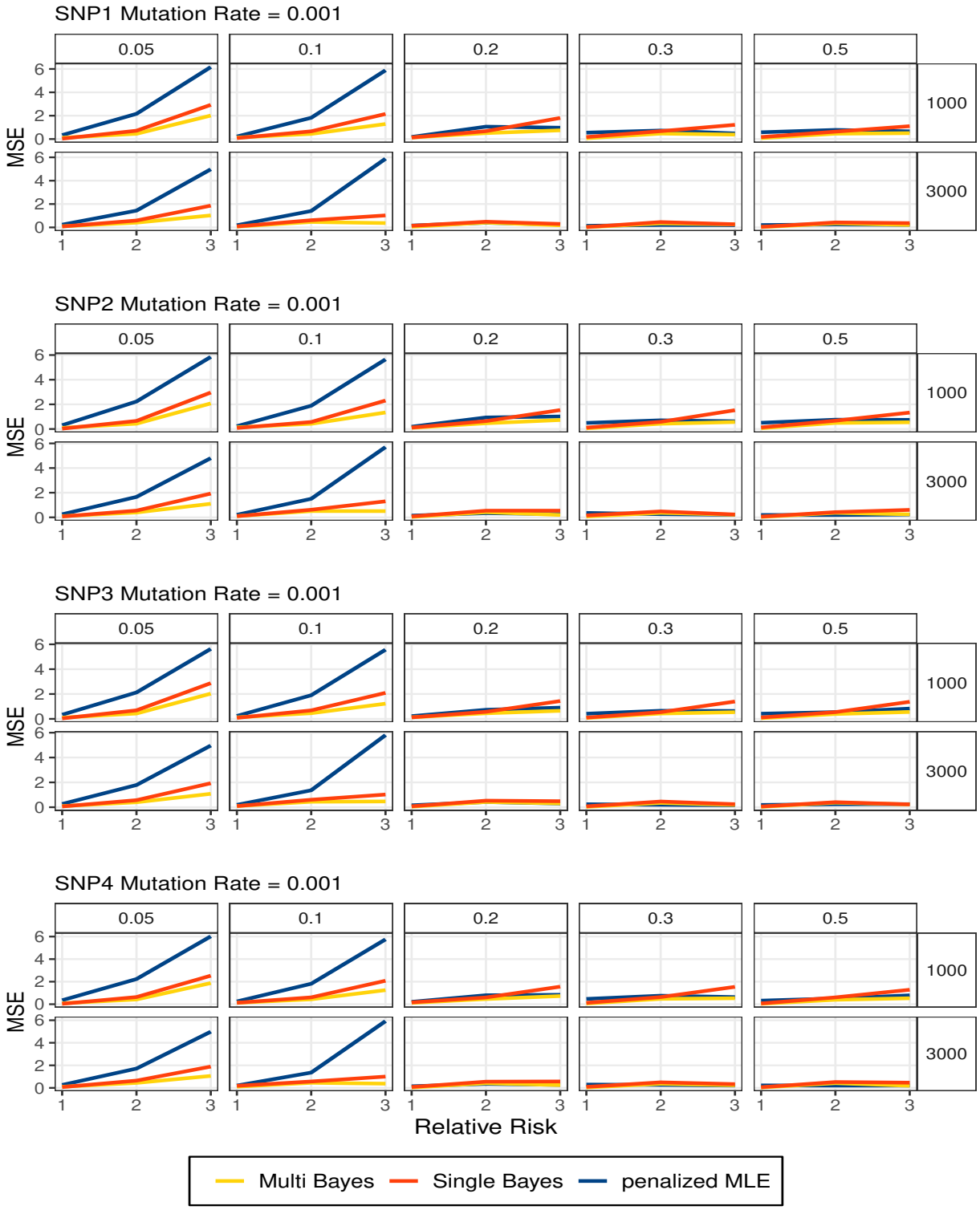

Figure 3.3: Mean square error comparison of penalized MLE, Single Bayesian and Multiple Bayesian model in setting MR = 0.001. Comparison shows that multiple Bayesian model has lowest MSE in most settings.

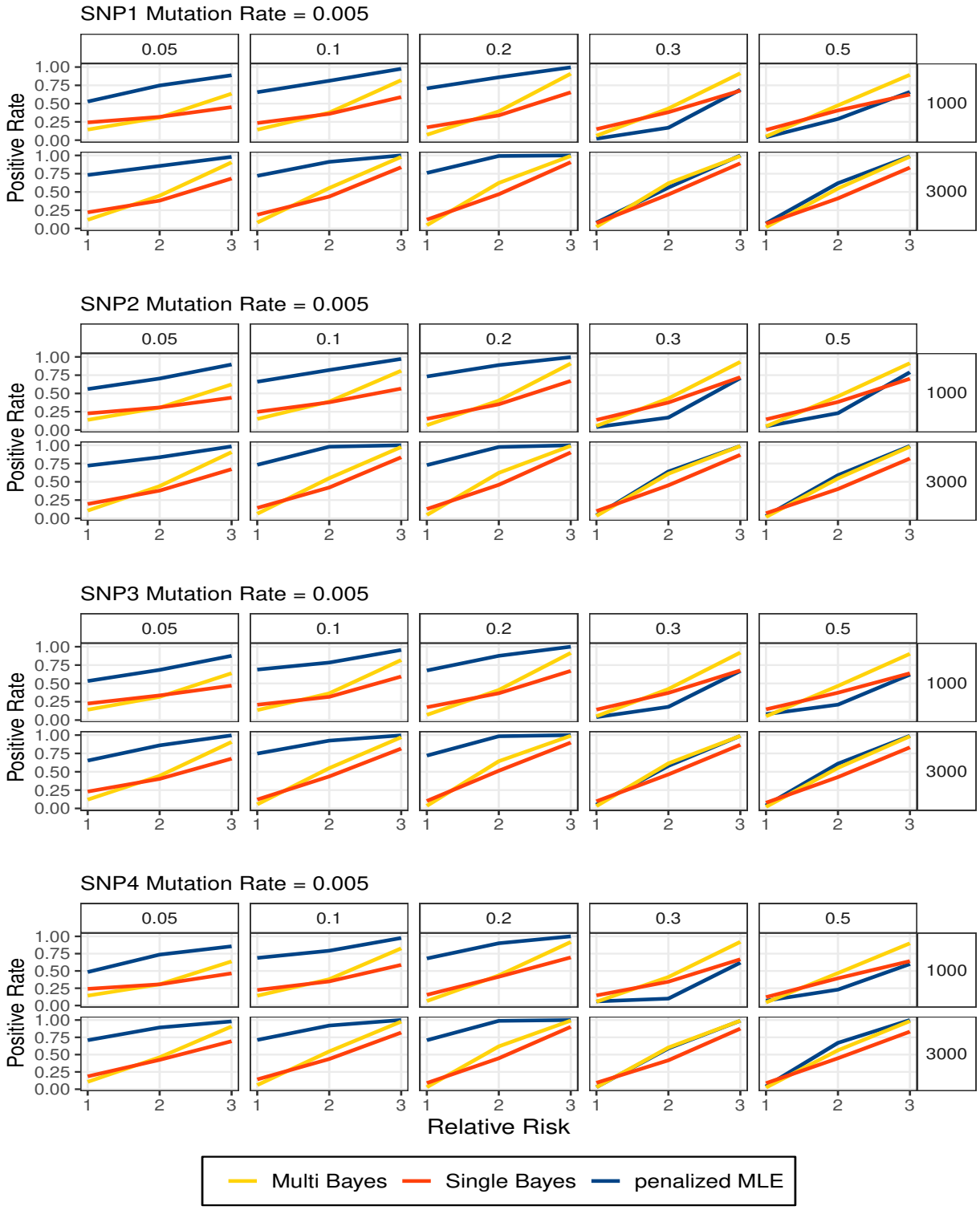

Figure 3.4: Power analysis of penalized MLE, Single Bayesian and Multiple Bayesian model in setting MR = 0.005. Comparison shows that multiple Bayesian model has highest power in most settings.

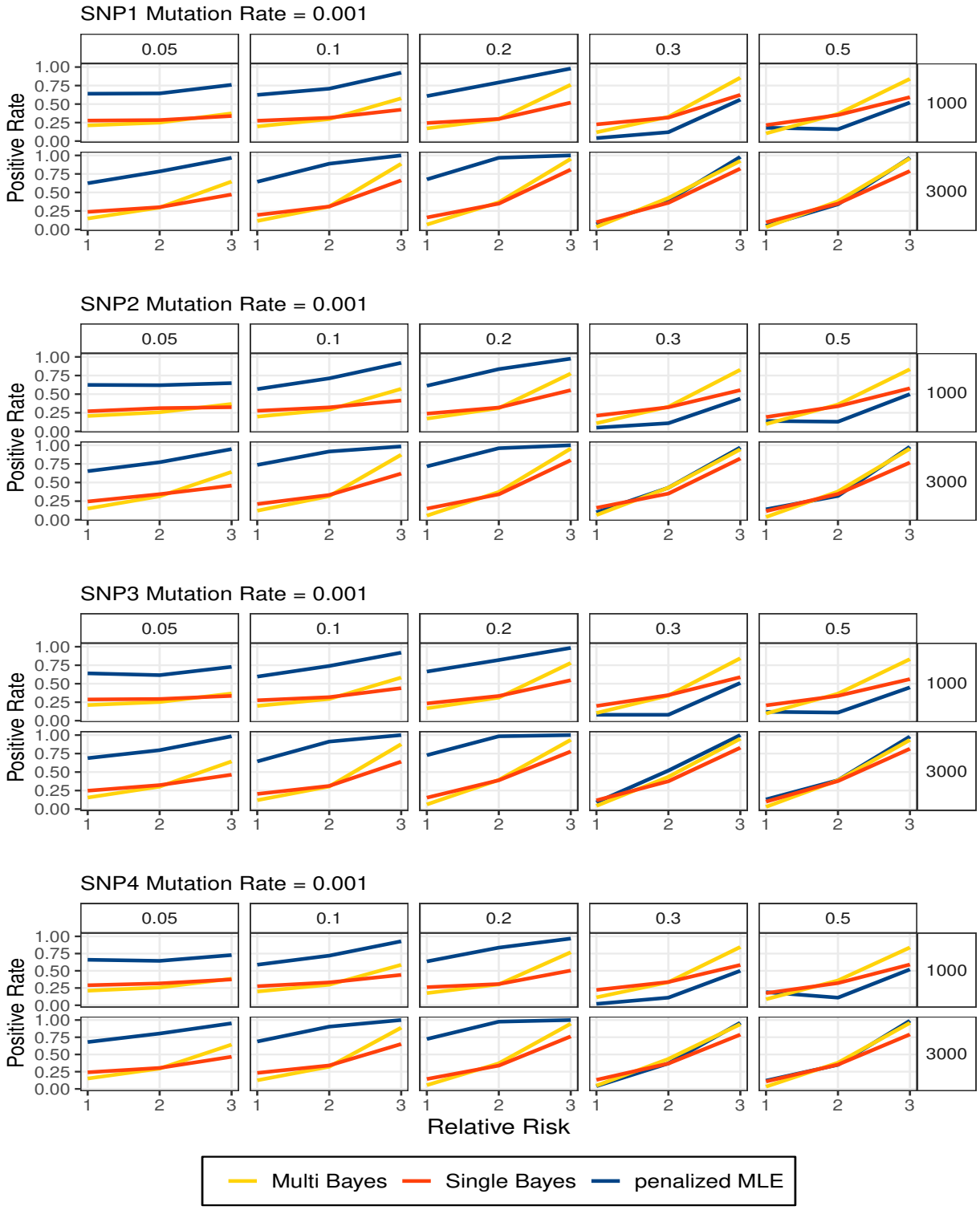

Figure 3.5: Power analysis of penalized MLE, Single Bayesian and Multiple Bayesian model in setting MR = 0.001. Comparison shows that multiple Bayesian model has highest power in most settings.

CHAPTER 4

REAL DATA APPLICATION

4.1 Application to matched-pair Breast Cancer Data

We consider the tumor-normal matched-pair data from The Cancer Genome Atlas (TCGA) on breast cancer. The total number of SNPs available is 905461 and number of matched samples is 1070. We apply common quality control methods to remove invalid SNP genotypes. First, Hardy-Weinberg equilibrium test (Emigh, 1980) is applied and p-value = 0.05 is a cut-off threshold. SNPs with missing genotypes and allele frequency less than 0.05 are removed. After quality control step, there are 614883 SNPs.

Second, the principal Component Analysis (PCA) is applied to genotype data to determine patient population structure. Among the 1070 samples, there are 725 White, 176 Black, 60 Asians, 95 Others and 109 Unknown. We observe that people with different ethnicity groups are highly distinguishable using the top two principal components as shown in Figure 4.1. Emerging studies in GWAS have shown that population heterogeneity can produce spurious associations if sub-population structure is not properly adjusted (Price et al., 2010). We use 95% confidence interval of each group to impute missing racial ancestry for the unknown patients. To adjust for population stratification, we extract the major ethnicity group that has the largest population for subsequent association analysis. 807 samples classified as white people are used in the analysis. The normal and tumor matched data are converted to Multinoulli data Table 2.10 by counting the frequencies in each combined genotype.

We consider the refFlat gene annotation (UCSC hg19) for human genome references. The gene region is determined using transcription start and end positions. We consider SNPs on a gene when they locate within 1000 base pairs upstream and downstream of a gene region. A total of 58545 genes are available. 20353 unique genes contains biomarkers and 220268

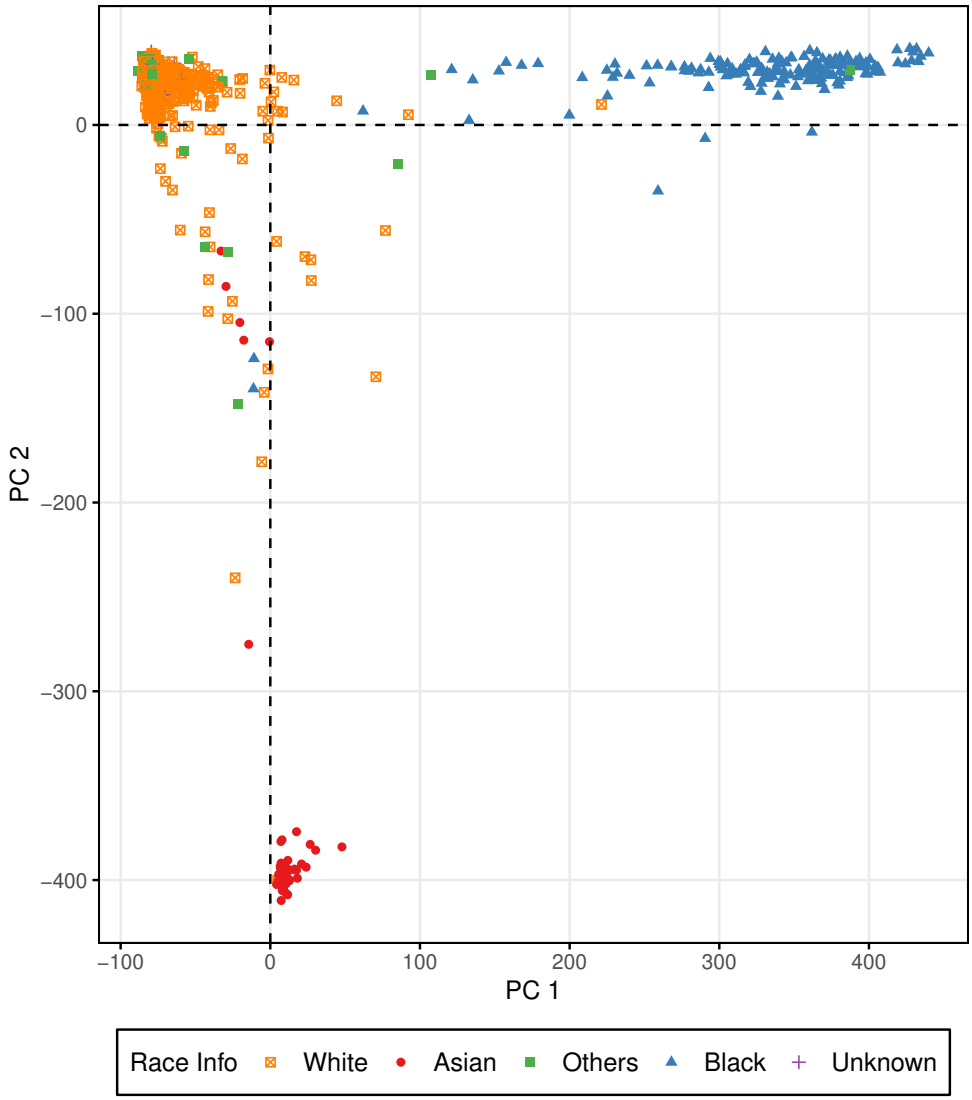

Figure 4.1: principal Component Analysis on breast cancer data

SNPs are mapped to these genes. The summary statistics about SNP counts and gene length in base pairs across all genes are provided in Table 4.1.

Table 4.1: Summary of SNP counts, gene length on all genes

Summary	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
SNP Counts	1	3	8	42	39	1242
Gene Length	21	16215	47282	159292	166206	2.32×10^6

The SNPs tend to be widely separated across a long gene region, where linkage disequilibrium can hardly be observed. To measure the joint effects of adjacent biomarkers, we consider to split large genes evenly into small gene segments that contains 32000 base pairs or less. After segmentation, there are 58161 gene segments. Summary statistics about SNP counts on each gene segment is given in Table 4.2.

Table 4.2: Summary of SNP counts on gene segments

Summary	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
SNP counts	2	3	5	6	8	40

We consider to apply the maximum likelihood estimation and Bayesian single marker analysis to all 220268 SNPs to investigate individual SNP effect. Then multi-marker Bayesian analysis method is applied to all 58161 gene segments to jointly measure adjacent SNP effects.

4.1.1 Single Marker Analysis

After preliminary analysis, the tumor-normal matched breast cancer data from TCGA contains 807 samples from major ethnicity group and 220268 SNPs after common quality control process. The counts of genotypes from tumor and normal tissues can summarized as in Table 2.10, where row counts and column counts correspond to normal genotypes and tumor genotypes, respectively. According to the sampling framework described in Chapter 2, the observed genotype counts have a Multinoulli distribution with parameters R, A, M. The expected probabilities can be expressed as a function of parameters R, A, M as in 2.2.1. In maximum likelihood method, the maximum likelihood estimation with ridge penalty described in 2.7 is applied to evaluate parameters R, A, M. We conduct a hypothesis testing for each SNP with null hypothesis $H_0 : R = R_0 = 1$ and alternative hypothesis $H_1 : R \neq R_0 = 1$. The Wald test is applied to determine significance of SNP effects.

The single marker analysis is applied to the breast cancer data using the Bayesian model. The individual SNP status is estimated by the mean of the posterior distribution. The

SNP relative risk, allele frequency and mutation rate is estimated by the mean of posterior distribution. To obtain aggregated scores for genes, we considered a SNP on a gene segment if its locus resides within the gene segment region. The aggregated gene status of a gene segment is derived as the mean of SNP status of all SNPs on the segment. Let the gene status be represented by the maximum gene segment status. We rank the all the genes based on aggregated posterior gene status.

Table 4.3: Genes with highest gene status estimated by posterior median in single marker analysis

Gene Name	**Chr**	**Gene Status**
IL7	chr8	0.975
TIAM1	chr21	0.972
CKAP2L	chr2	0.961
TTC28	chr22	0.956
NDST4	chr4	0.952
EIF2AK2	chr2	0.952
CACNB4	chr2	0.95
PARD3B	chr2	0.946
TMEM117	chr12	0.944
ATP6V0D1	chr16	0.943

Table 4.3 shows the top genes that are identified by single marker Bayesian model. The Appendix B lists the top 100 genes identified by single-marker Bayesian analysis. Among them, many genes have been reported to be cancer related. Multiple recent studies have indicated a positive correlation between boosted TIAM1 expression level and higher grade of human breast cancer (Minard et al., 2004; Adam et al., 2001). The TIAM1 gene and the encoded protein has been implicated cell proliferation, migration, invasion and tumor progression in a variety of human cancers (Walch et al., 2008; Engers et al., 2006; Minard et al., 2006; Ding et al., 2009). NDST4 genetic loss is significantly associated with tumor progression and NDST4 gene is identified as a novel candidate tumor suppressor in human colorectal cancer (Tzeng et al., 2013). A number of studies have suggested that the activation of EIF2AK2 can suppress tumor growth(Meurs et al., 1993; Shir and Levitzki, 2002; Kim

and Cho, 2017), while elevated expression of EIF2AK2 increases carcinoma progression in a variety of human cancers, including breast cancer (Kim et al., 2000; Lee et al., 2019; Garcia et al., 2006). The TMEM117 gene belongs to the TMEM family. Evidences have shown that down- or up-regulated TMEM expression has been identified in tumor tissues compared to adjacent healthy tissues, and suggest some TMEMs as prognostic biomarkers(Schmit and Michiels, 2018).

4.1.2 Multi-Marker Analysis

The multiple marker Bayesian model is also applied to the TCGA breast cancer data. In multiple marker analysis results, the gene segment status is estimated by the median of posterior distribution, and the SNP relative risk, allele frequency and mutation rate is estimated by the mean of posterior distribution. Let the gene status be represented by the maximum value of gene segment status. All the genes are ranked based on the gene status.

Table 4.4: Genes with highest gene status estimated by posterior median in multiple marker analysis

Gene Name	**Chr**	**Gene Status**
LINC00383	chr13	0.999
KIRREL3	chr11	0.999
STX3	chr11	0.999
AGPAT4	chr6	0.999
SYCE1	chr10	0.997
RCBTB1	chr13	0.997
PKNOX2	chr11	0.997
RGS3	chr9	0.997
GCSH	chr16	0.997
CSMD1	chr8	0.996

Table 4.4 shows the top genes in multiple marker Bayesian model ranked by gene status. The Appendix B lists the top 100 genes identified by multi-marker Bayesian analysis. Recent studies have identified highly significant association on KIRREL3 region with breast cancers (Wang et al., 2010). Novel research has shown that STX3 gene plays a potential role in

carcinogenesis via up- or down- regulation in different cancers and promoting breast cancer cell growth (Giovannone et al., 2018; Nan et al., 2018). Higher Agpat4 expression in cancer tissues correlates positively with worse survival rates among colorectal cancer patients (Zhang et al., 2020). Previous studies have indicated that deletion in PKNOX2 region is prone to breast cancer and ovarian cancer malignancies (Launonen et al., 1998; Gentile et al., 2001). Some studies also suggested the possible function of RGS3 protein as a cancer suppressor (Chen et al., 2015). Research in CSMD1 revealed that CSMD1 is a tumor suppressor gene and its low expression is significantly associated with high breast tumor grade (Escudero-Esparza et al., 2016; Kamal et al., 2009).

4.2 Comparison with existing breast cancer research

We use the external oncogenic database, the Catalogue Of Somatic Mutations In Cancer (COSMIC), which provides comprehensive somatic mutations and genes that are associated with all types of breast cancer tissues. The gene list contains gene symbol, mutated samples and total samples. The COSMIC breast cancer genes are sorted by mutated rates. Genes with higher mutation rates tend to have greater risk in breast cancers. The ranked gene list from COSMIC is compared with ranked gene list of multi- and single- marker analysis. From the ranked multi- and single- marker analysis results, we consider the gene status to be the prediction probabilities, and indicators whether the gene belongs to the set of top 500 COSMIC genes to be the true binary labels. We calculate the true positive rate and false positive rates to plot the receiver operating characteristic (ROC) curve. Figure 4.2 shows the ROC curves of multi- and single- marker analysis and the area under curve (AUC) of both methods. The AUC of multi-marker analysis is 0.86 and the AUC of single-marker analysis is 0.83.

We consider the somatic copy-number alternation (SCNA) regions that are significantly associated with the breast invasive carcinoma. The highly significant SCNA regions are

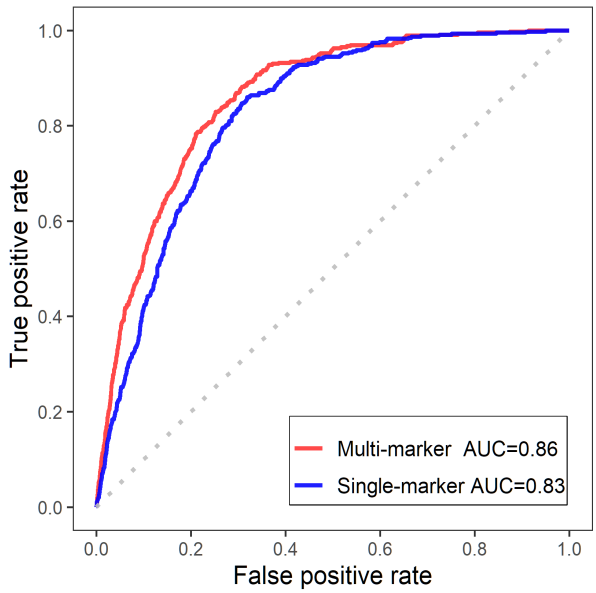

Figure 4.2: ROC curves of multiple and single marker models

identified using TCGA breast cancer data by GISTIC (Genomic Identification of Significant Targets in Cancer), a tool to identify genomic regions that are significantly gained or lost across a set of tumors (Mermel et al., 2011). There are 28 significant focal amplifications and 42 significant focal deletions found in the TCGA breast genome analysis (Broad Institute TCGA Genome Data Analysis Center, 2016). We perform a comparison analysis between significant SCNA regions and non-significant SCNA regions on gene segment level and SNP level.

On gene segment level, we consider a gene segment to be located in significant SCNA regions if they overlap. Otherwise the gene segments are located in non-significant SCNA regions. The non-parametric Kolmogorov–Smirnov (KS) test and Mann–Whitney U test are used to compare the distribution of gene segment association status predicted by multiple marker Bayesian model among these two groups. The KS test states the null hypothesis H_0 that the distributions of gene segment scores for the two groups are equal and alternative hypothesis H_1 that distributions are not equal. The Mann–Whitney U test states the null hypothesis H_0 that two groups have same distribution and alternative hypothesis H_1 that

one group has larger or smaller values than the other. The Table 4.5 shows the comparison tests between two groups of gene segments. The test results conclude that distributions of segment association status are different for significant SCNA regions and non-significant SCNA regions. The Figure 4.3 compares the gene segment score distributions for significant SCNA regions and non-significant SCNA regions. Each interval corresponds to 10th percentile of all gene segment scores. It shows that the high risk gene segments are enriched on deletion and amplification cytobands, which genomic associated predicted by multiple marker Bayesian model is aligned with GISTIC analysis.

Table 4.5: Test on gene segment level

Significant Regions	**Test**	**Statistics**	**P Value**
Deletions	KS test	0.07589	$< 2.2e-16$
Deletions	Mann–Whitney U test	141020239	$< 2.2e-16$
Amplifications	KS test	0.04444	0.004762
Amplifications	Mann–Whitney U test	39894790	0.1943

On SNP level, we compare the relative risk of SNPs on significant deletion and amplification regions and SNPs outside these regions. The KS test and Mann–Whitney U test are used to compare the distribution of relative risk for the two groups. The Table 4.6 shows that the distribution of relative risk of SNPs on deletion and amplification cytobands are different than the distribution of SNPs outside these cytobands. The Figure 4.4 shows that high risk SNPs are enriched in deletion cytobands while the enrichment can not be observed in amplification cytobands. It indicates that the relative risk predicted by single marker Bayesian model is more aligned with significant deletion regions identified by GISTIC.

We also explore another external resources from the Genomic Data Commons (GDC) Data Portal to compare with the gene lists provided by our Bayesian models. The data contains mutations that have been reported to associated with breast cancers. The impact of breast cancer mutations are classified on the basis of the severity of the variant consequences by three tools: Ensembl Variant Effect Predictor (VEP), Polymorphism Phenotyping

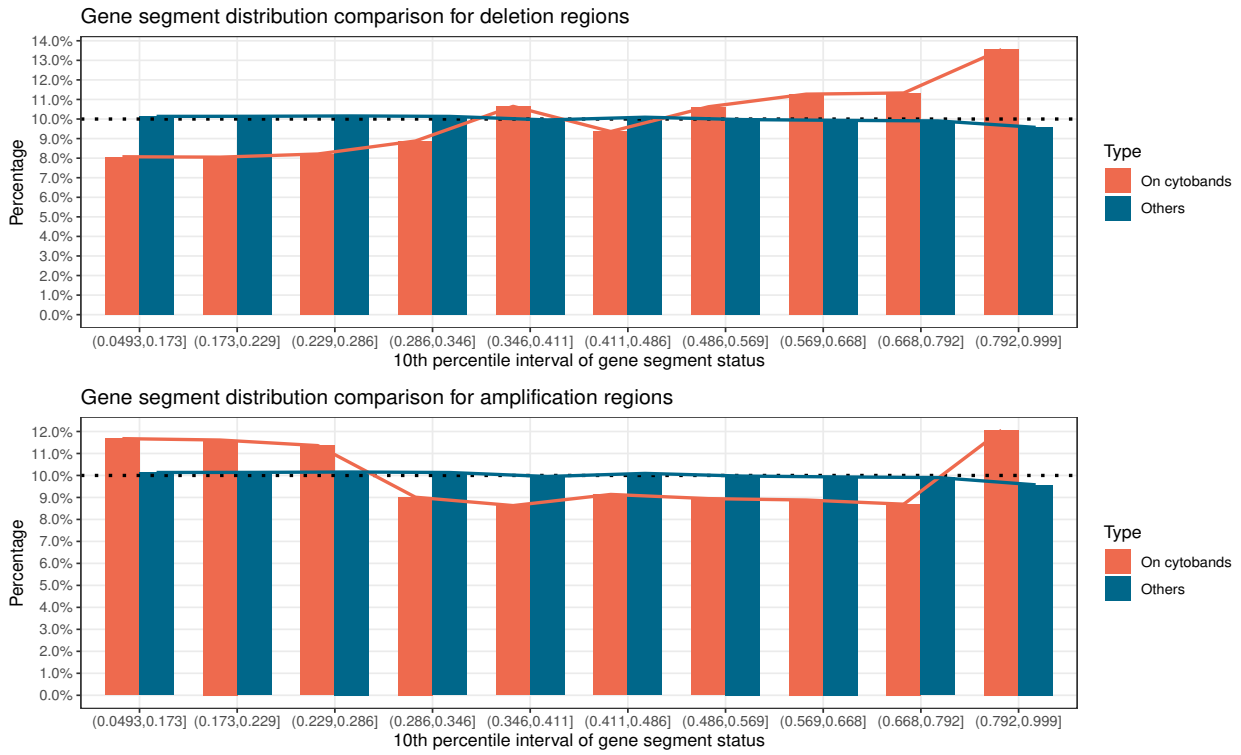

Figure 4.3: Comparison of gene segment scores for significant SCNA regions and non-significant SCNA regions.

Table 4.6: Test on SNP level

Significant Regions	**Test**	**Statistics**	**P Value**
Deletions	KS test	0.06212	$< 2.2e-16$
Deletions	Mann–Whitney U test	1954578458	$< 2.2e-16$
Amplifications	KS test	0.02395	0.002515
Amplifications	Mann–Whitney U test	573484616	0.002193

(PolyPhen) and Sorting Intolerant From Tolerant (SIFT). The Ensembl VEP tool provides the effect of a genomic variant in coding and non-coding regions. The effect levels include high, moderate, low and modifier, which ranges from high impact in protein to no evidence of impact. The SIFT tool predict whether an amino acid substitution will affect protein function and phenotype based on sequence homology. The impact levels are"deleterious", "deleterious low confidence", "tolerated low confidence" and "tolerated", which ranges from very likely to have a phenotypic effect to not likely to have a phenotypic effect. The PolyPhen

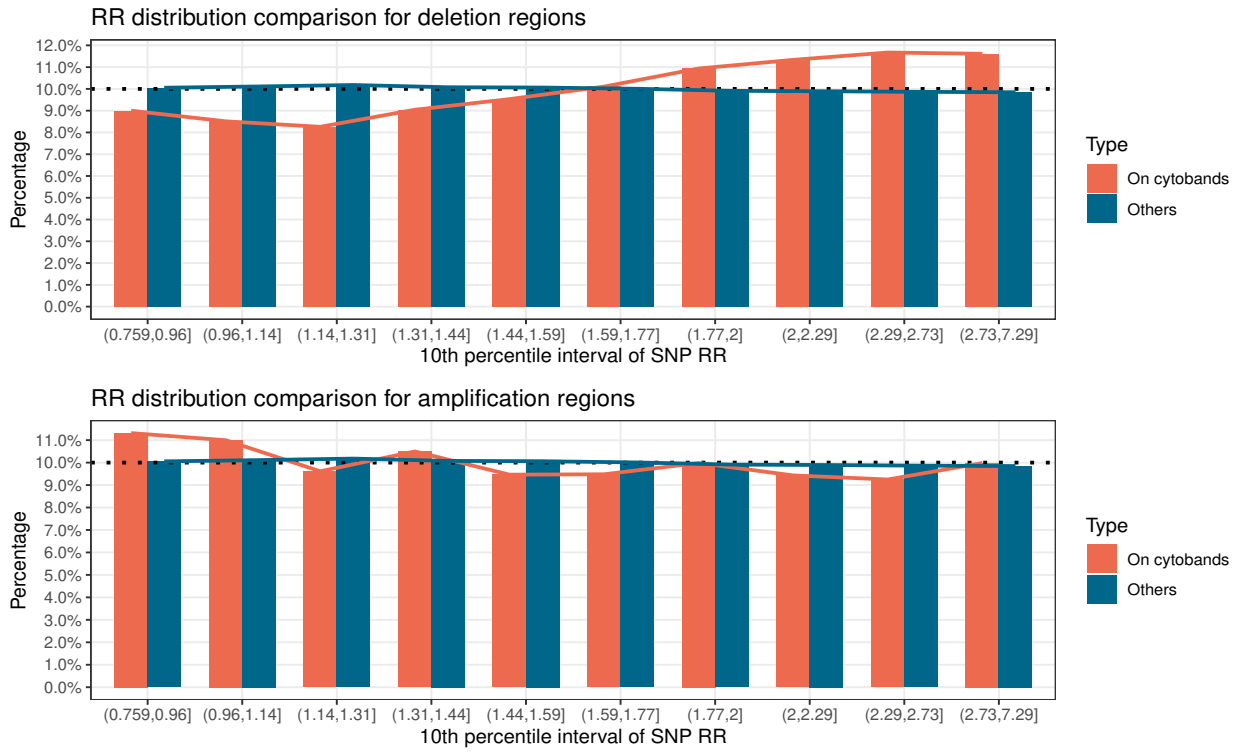

Figure 4.4: Comparison of relative risk for significant SCNA regions and non-significant SCNA regions.

tool predicts the potential impact of an amino acid substitution on human proteins. The impact levels are "probably damaging", "possibly damaging", "benign" and "unknown", which ranges from high confidence to affect protein function or structure, to unknown prediction. In Figure 4.5 we summarise the counts of each impact levels using the variants that located within the top 100 associated genes identified by single-marker Bayesian model and multi-marker Bayesian model. The results show that the variants from the top genes identified by multi-marker model have more classifications as high impact variants, compared to the sets from top genes identified by single-marker model.

In addition, we also explore the differences between association status and variant effect size estimated by multiple Bayesian model and single Bayesian model. In multiple-marker analysis, the segment association status is estimated by the posterior mean of MCMC simulation. In single-marker analysis, the segment association status is derived as the mean of

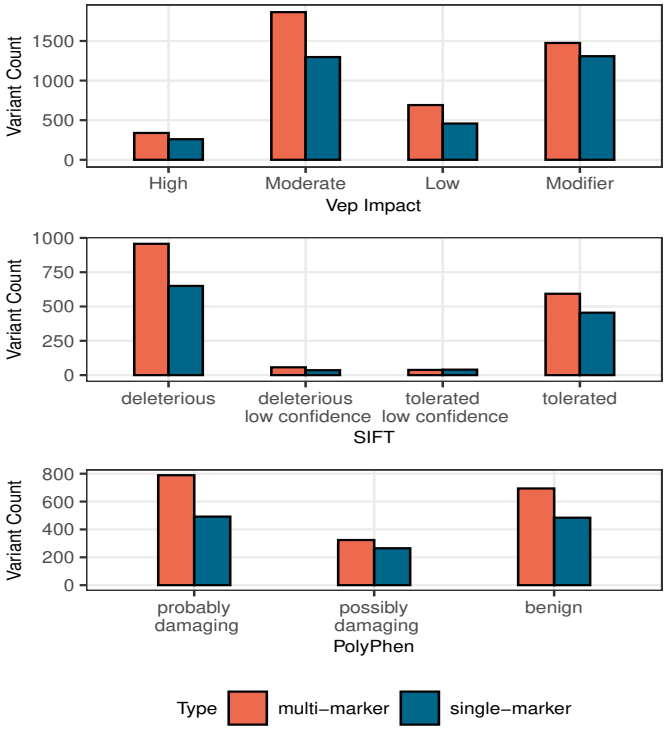

Figure 4.5: The counts of impact levels predicted by three tools: Ensembl VEP, SIFT and PolyPhen using the variants from the top 100 associated genes identified by multiple-marker Bayesian model and single-marker Bayesian model.

estimated association status of individual variants. We compare the estimations of genes TACC2, CSMD1 and CDH13, which have been recognized to associate with breast cancer in multiple literature (Conte et al., 2003; Ma et al., 2009; Toyooka et al., 2001). Figure 4.6-4.8 plot the relative risk distribution, segment association status and variant counts on each gene segment of the target gene. The plots show that the distributions of the estimated allelic relative risk are similar in both models. However, multi-marker Bayesian model is more sensitive to moderate variants by considering the joint effects of neighboring SNPs on a segment. Take the segment 6 and 7 in CSMD1 gene for example, given that the SNP counts are the same, the relative risk distribution has a small increase from segment 6 to segment 7. Correspondingly, the multi-marker model has a sharp jump on association status estimations from around 0.2 to 0.8, while single-marker model has a mild increase on

association status estimations from 0.2 to 0.4. On the other hand, the number of variants also affects the aggregation impact. For example, the segments 44 and 46 of CSMD1 gene have 32 and 9 variants respectively. Both segments have moderate relative risk distributions, multi-marker model has association status 0.8 and 0.6 while single-marker model has both association status around 0.4.

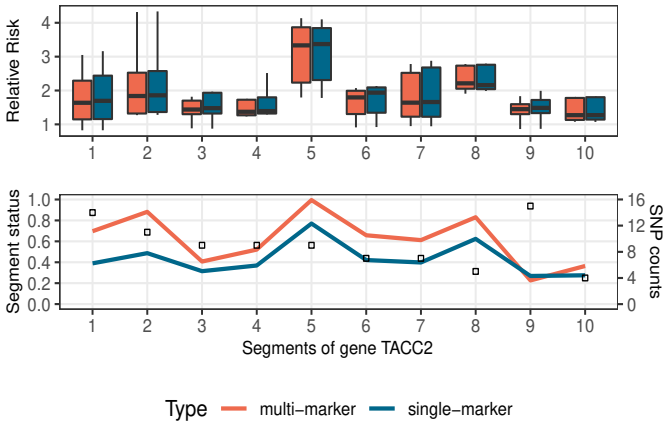

Figure 4.6: TACC2 gene is divided to 10 segments. The relative risk estimations, segment status and segment SNP counts are plotted.

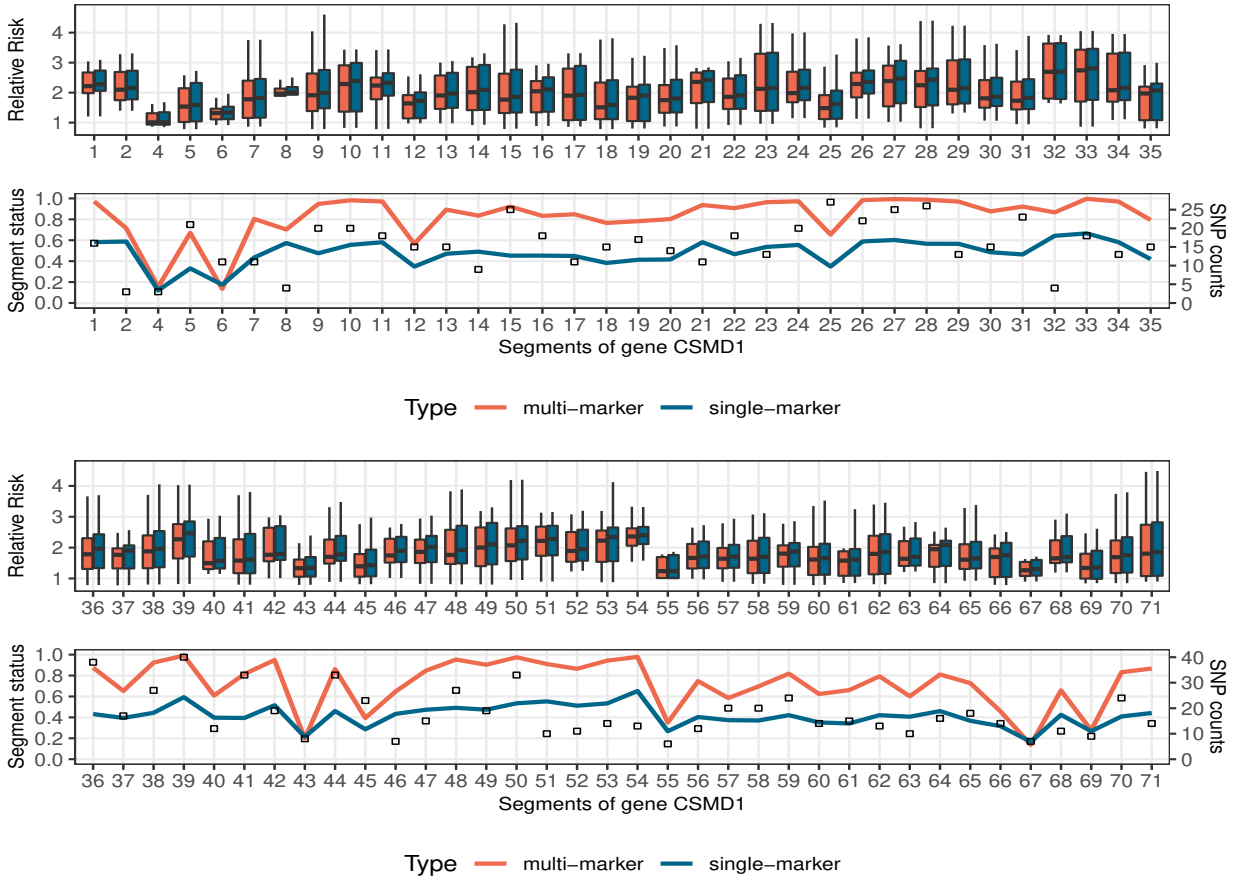

Figure 4.7: CSMD1 gene is divided to 71 segments. The relative risk estimations, segment status and segment SNP counts are plotted.

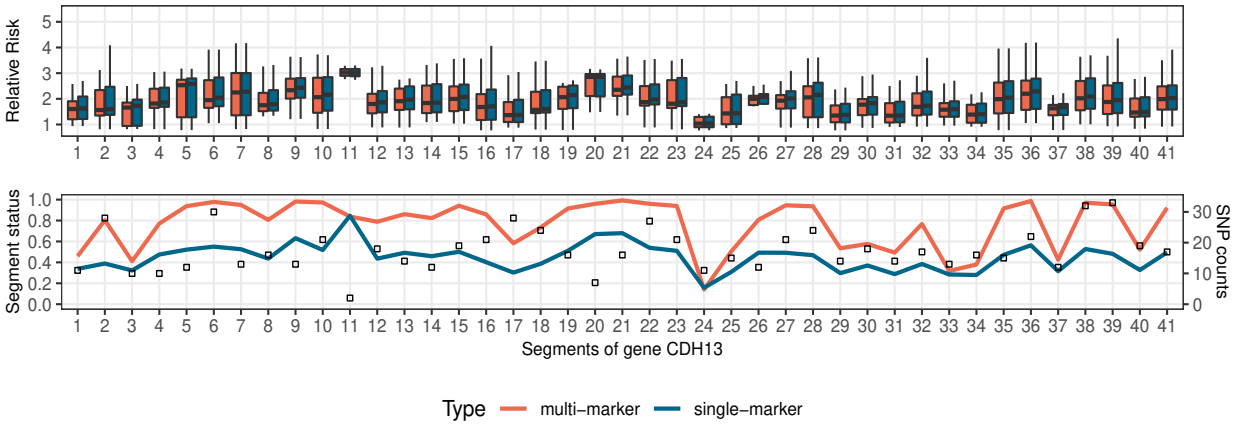

Figure 4.8: CDH13 gene is divided to 41 segments. The relative risk estimations, segment status and segment SNP counts are plotted.

CHAPTER 5

DISCUSSION AND FUTURE WORK

In this dissertation, we propose a novel model framework to analyze variants in tumor and matched normal tissue data in GWAS. This framework establishes a connection between paired genotype data and allelic association, and allows to apply frequentist and Bayesian methods to evaluate the association status. There are limited studies for matched tumor-normal data in traditional GWAS, since most statistical models, such as regression and chi-square test, require the independence of control and case samples(Bush and Moore, 2012). Due to the specific structure of matched data, the traditional GWAS is not applicable and may cause spurious results. Studies have indicated that the progression of cancer is associated with the accumulation of somatic mutations (Alexandrov and Stratton, 2014; Alexandrov et al., 2013; Greenman et al., 2007). Exploring the impact of somatic mutations in carcinoma can be significant in risk prediction, continuous monitoring and early detection of cancer, and can contribute to individualized prevention and therapeutic strategies. By considering the sampling process from precancerous tissue to tumor tissue, the proposed framework assigns a Multinoulli distribution to matched data and connects the distribution with allelic effect size, frequency and mutation rate. Another benefit of the framework is that the association result is more reliable due to the matched data intrinsically control many confounding factors. With this cornerstone, frequentist and Bayesian statistical models can be applied to assess the association of variants in matched data.

The penalized maximum likelihood estimation (MLE) is proposed to provide the individual effect size, allele frequency and mutation rate estimation that maximize the likelihood function with penalty terms. The single-marker hierarchical Bayesian model is proposed to provide more flexibility to assess the individual SNP status in addition to the above allelic estimation. From simulation studies, we find that the single Bayesian model, compared to

penalized MLE method, has relative low mean square error (MSE) in RR estimation and relative high power to identify associated SNPs in different settings. However, in settings with sufficient sample size, high allele frequency and mutation rate, the performance of penalized MLE and single Bayesian method are similar. While the analytical or numerical solution of maximum likelihood function is efficient, the MCMC simulation of Bayesian model requires much more computation. The performance under frequentist method is as good as Bayesian model in settings with large samples, relatively high allele frequencies, high mutation rates and high relative risks. On the other hand, the single Bayesian is better in scenarios with limited sample sizes and low values of the model parameters. In addition, the advantages of single Bayesian model includes flexibility of hierarchical model, posterior inference of variables, as well as the prior knowledge that is taken into account for uncertainty.

The multiple-marker hierarchical Bayesian model is extended from the single-marker Bayesian model by combining SNPs into groups in a biologically meaningful way and performing SNP-set analysis. It allows to collect the joint effects of multiple SNPs and enhance the power to detect SNPs with moderate risk. Simulations have shown that the multiple Bayesian model improves the power to identify a contributory SNP-set while remains a relative low type-I-error rate. Another benefit of multiple Bayesian model is that by reducing the number of hypothesis testings, the threshold to declare significance is less stringent and the power to detect moderate SNPs is further improved. The comparisons of the three proposed methods show that the multiple Bayesian model outperforms in most scenarios. It is worth mention that cost of computation is similar for multiple and single Bayesian model. While penalized MLE is more computational efficient, it requires a better data quality especially a large sample size to achieve a comparable performance. The process to obtain a large population and sequence genotypes can be costly(Spencer et al., 2009).

With the application to breast cancer data from The Cancer Genome Atlas (TCGA), we find that the top genes identified by multiple-marker Bayesian model are mostly cancer

related genes. From literature, many of the identified top genes have been reported to have positive correlation with breast tumor progression, or act as a cancer suppressor. It is worth mentioning that multiple Bayesian highlights the genes that are normally downgraded in single marker analysis due to the vast majority of SNPs having medium effects. The multiple Bayesian model is more advantageous for large genes, whose gene impact can be mitigated by most individual moderate SNPs, when the joint effects are not taken into account. Finally we consider another somatic mutation resources for human breast cancer from the Catalogue of Somatic Mutations in Cancer (COSMIC). By comparing the top associated gene list identified by Bayesian models, we find that the gene lists from multiple-marker analysis is more consistent with the COSMIC data. We also consider the breast cancer variants, classified into different impact level by three predictive tools, from the Genomic Data Commons (GDC) Data Portal for comparison. In addition, the segment status and relative risk distribution are plotted to measure the aggregation of joint effects in multiple-marker analysis. It shows that the multiple-marker analysis are more sensitive to SNP sets with moderate effect size, and the aggregation effect can be improved as the number of SNPs increases.

The data framework and hierarchical Bayesian model proposed in this dissertation has provided a statistical method to analyze tumor and normal matched-paired data in GWAS. The simulation studies and real data application show that multiple-marker analysis has improved power to identify related variants while remaining a relative low type I error. However, this advantage may not be obvious in some situations. When variants are located sparsely or far away from other variants on the genes, the aggregation of joint effect is decreased.

There are possible ways to further improving our proposed framework and models. Our current methods focus on case-control studies in which the phenotypes are dichotomous traits. One extension is to develop the framework and models for quantitative traits in GWAS

with matched pair data. This can be further combined with gene expression profiling, such as expression quantitative trait loci (eQTLs) to identify contributory genes. Extensive research and much effort have been devoted to eQTL-based analysis to reveal the association between gene expressions and cancers (Li et al., 2013; Loo et al., 2012). However, few studies have been conducted on eQTL analysis with matched pair data in GWAS. Having appropriate tools to address somatic alternations using matched data in GWAS would contribute to power improvement and better genetic insights.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

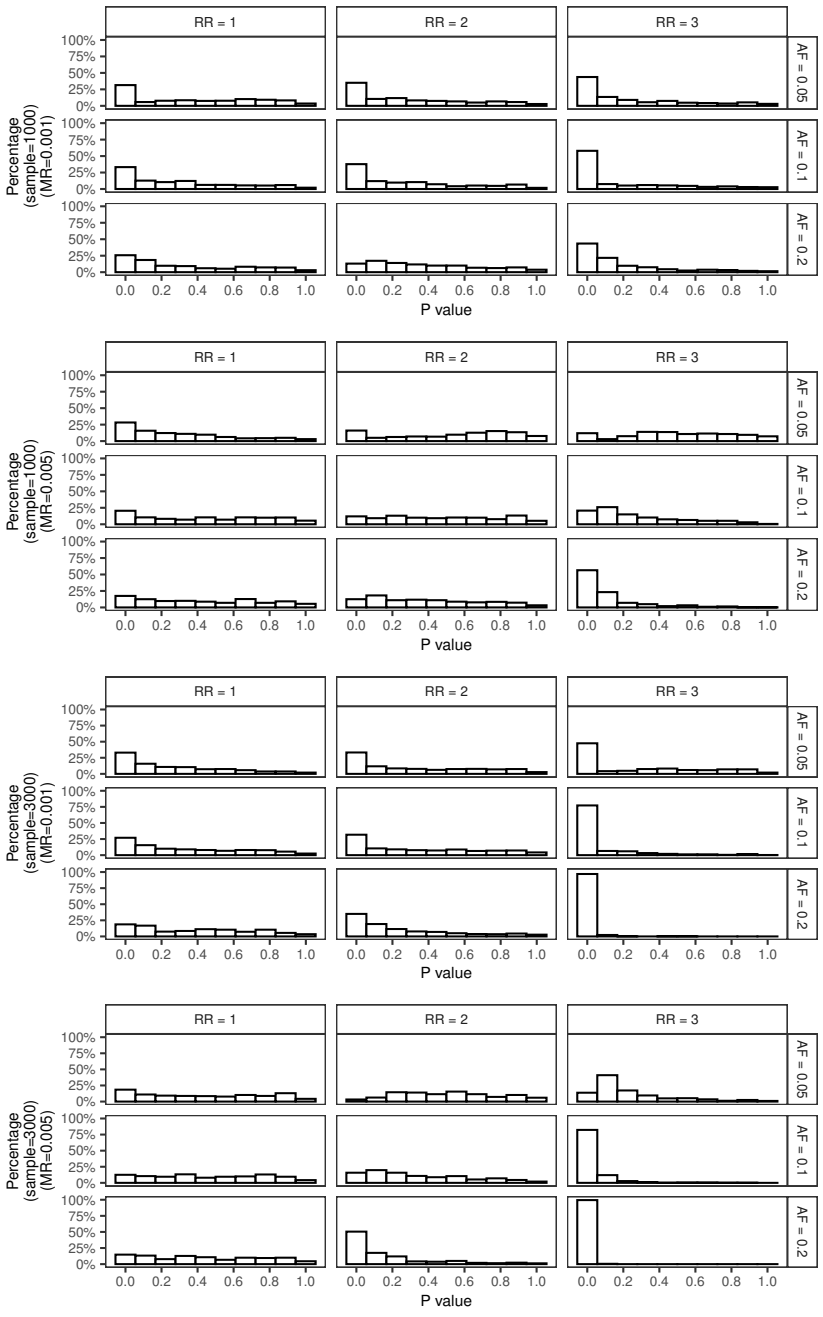

Figure A.1: The distribution of p-values in Wald test using penalized MLE method under different settings of RR, AF, MR and sample size.

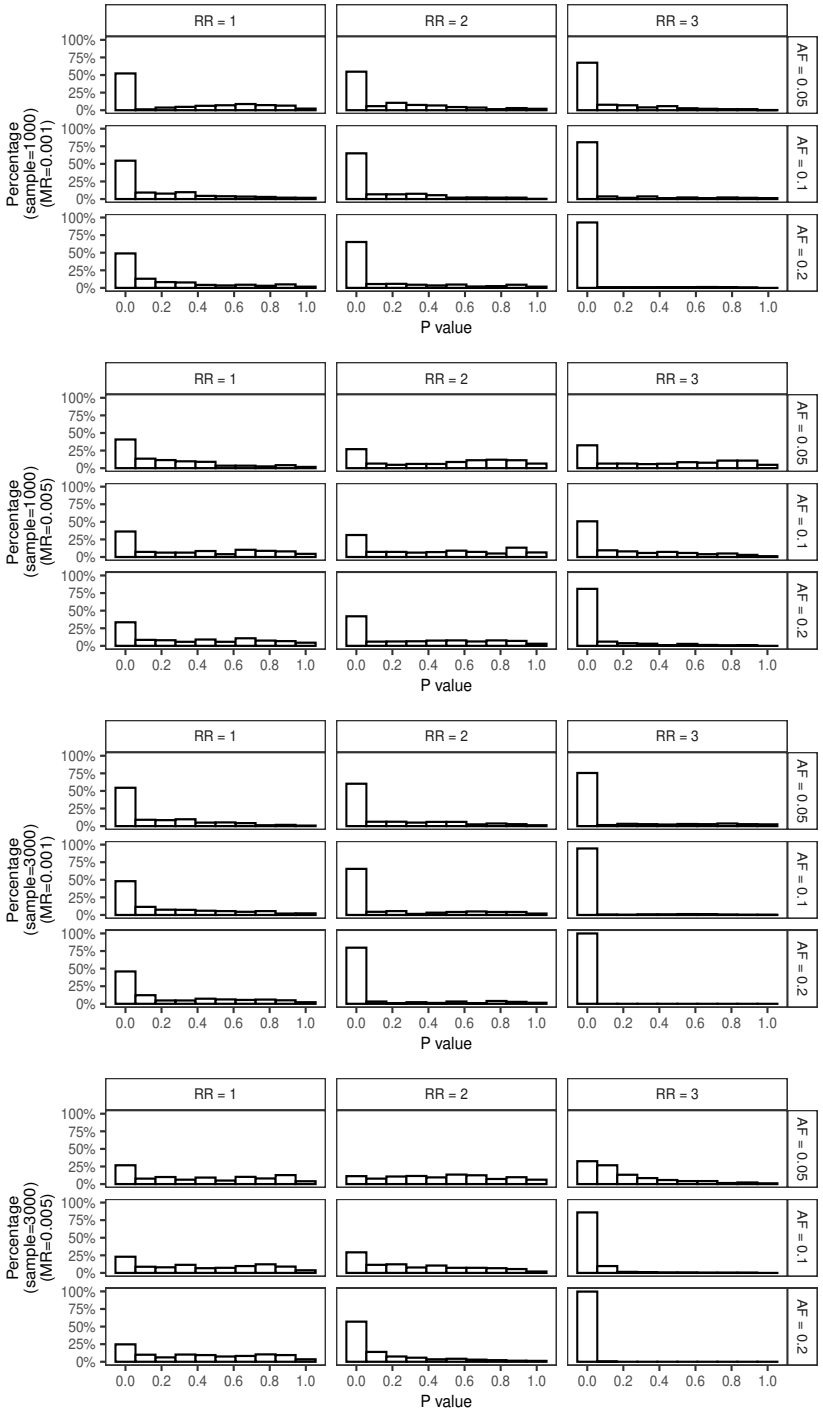

Figure A.2: The distribution of p-values in Score test using penalized MLE method under different settings of RR, AF, MR and sample size.

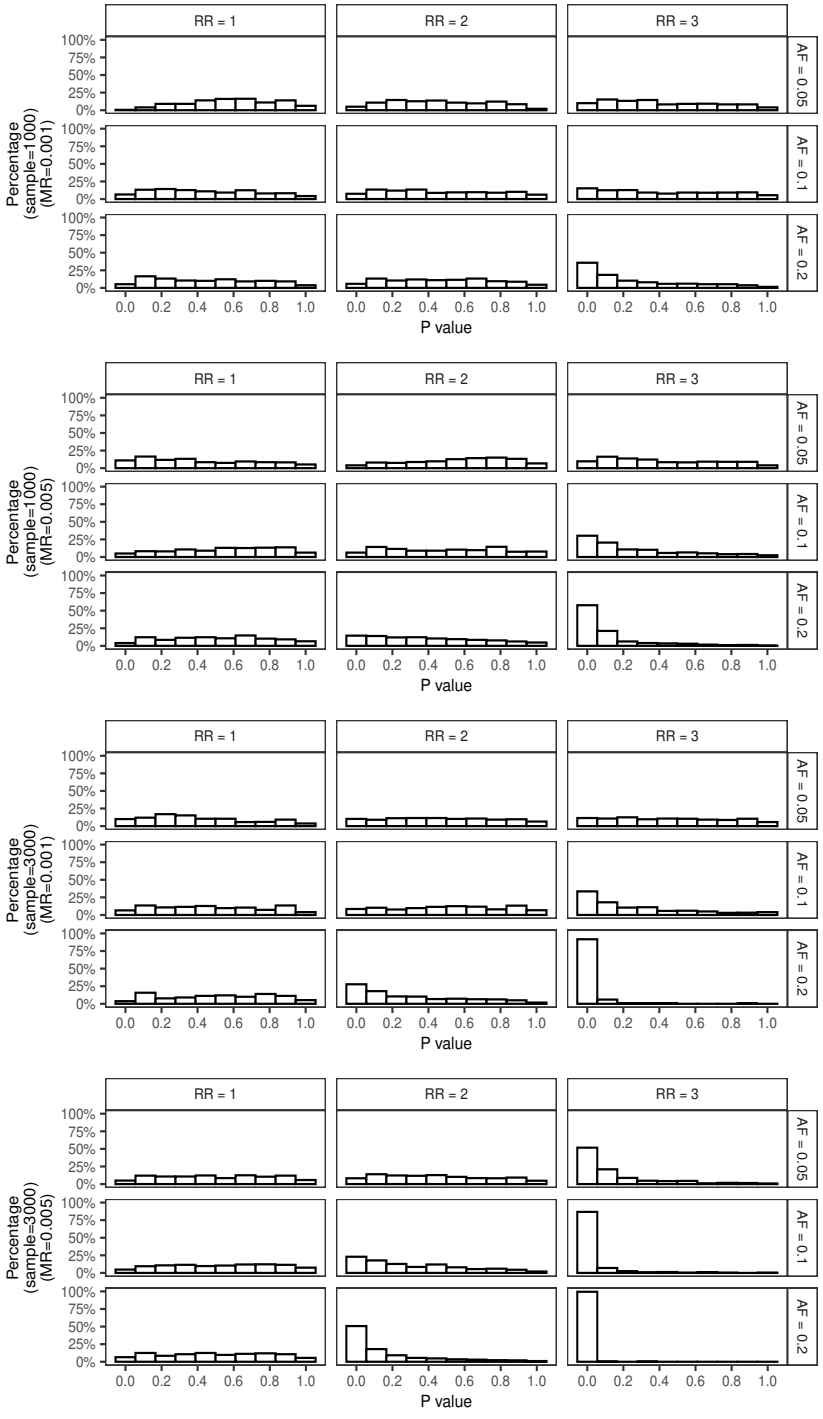

Figure A.3: The distribution of p-values in Likelihood Ratio test using penalized MLE method under different settings of RR, AF, MR and sample size.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Table B.1: Top associated genes identified by multiple-marker analysis. Each gene has been divided into segments based on the length of the gene. The gene status is represented by the highest association status of its segments. In multi-marker analysis, the gene segment status is estimated by the posterior mean of the parameter.

Gene Name	Chromosome	Gene Status
LINC00383	chr13	0.999
KIRREL3	chr11	0.999
STX3	chr11	0.999
AGPAT4	chr6	0.999
SYCE1	chr10	0.997
RCBTB1	chr13	0.997
PKNOX2	chr11	0.997
RGS3	chr9	0.997
GCSH	chr16	0.997
CSMD1	chr8	0.996
WWOX	chr16	0.996
ADAMTS19	chr5	0.996
MSR1	chr8	0.995
ALK	chr2	0.995
TACC2	chr10	0.995
PTPRK	chr6	0.995
PHF21B	chr22	0.994
ADAM28	chr8	0.994
LOC101929294	chr8	0.994
ARSJ	chr4	0.994
DOCK2	chr5	0.994
SGCZ	chr8	0.993
NEBL	chr10	0.993
CDH13	chr16	0.993
PPL	chr16	0.993

COTL1	chr16	0.991
EIF3D	chr22	0.991
UNC5D	chr8	0.99
PBLD	chr10	0.99
PTPRD	chr9	0.99
SYK	chr9	0.99
SMARCA2	chr9	0.989
PALLD	chr4	0.989
MYO5B	chr18	0.989
PIGG	chr4	0.988
NAXD	chr13	0.988
ADAMTS18	chr16	0.988
TANGO6	chr16	0.988
DCLK1	chr13	0.988
LOC101928516	chr6	0.988
LOC102724084	chr16	0.987
NAALADL2	chr3	0.987
MGMT	chr10	0.986
FAM19A5	chr22	0.986
EDNRB-AS1	chr13	0.986
BNC2	chr9	0.986
CCDC88C	chr14	0.985
NTM	chr11	0.985
DDR2	chr1	0.985
KIFC3	chr16	0.985
ZNF664-FAM101A	chr12	0.985
SAXO1	chr9	0.985
DLGAP1	chr18	0.984
FGF14	chr13	0.984
CLYBL	chr13	0.984
ENDOD1	chr11	0.984
KCNC2	chr12	0.984
MDGA2	chr14	0.984
IL17RD	chr3	0.984
ADRA1A	chr8	0.984

SYT1	chr12	0.984
AFAP1	chr4	0.983
DLG2	chr11	0.983
MTMR9	chr8	0.983
MICU3	chr8	0.983
DEC1	chr9	0.983
LOC101927815	chr8	0.983
GPC6	chr13	0.983
CENPN	chr16	0.983
MARCH1	chr4	0.983
SHFM1	chr7	0.983
FAM172A	chr5	0.983
PLCG2	chr16	0.983
NPAS3	chr14	0.982
CTU2	chr16	0.982
OPRM1	chr6	0.982
CAMK4	chr5	0.982
SPATA5	chr4	0.982
GABBR2	chr9	0.982
PSD3	chr8	0.981
MSRA	chr8	0.981
ASIC2	chr17	0.981
TECTA	chr11	0.981
EFCAB6-AS1	chr22	0.981
INPP4B	chr4	0.981
LOC400655	chr18	0.981
LRRC7	chr1	0.981
XIRP2	chr2	0.981
PKP1	chr1	0.981
DPYSL2	chr8	0.981
RGS6	chr14	0.981
EPHX2	chr8	0.98
ZBTB20	chr3	0.98
PTPN13	chr4	0.98
ITSN1	chr21	0.98

PKD1L2	chr16	0.979
CMIP	chr16	0.979
SCTR	chr2	0.979
NCAPG2	chr7	0.979
DPF3	chr14	0.979

Table B.2: Top associated genes identified by single-marker analysis. Each gene has been divided into segments based on the length of the gene. The gene status is represented by the highest association status of its segments. In single-marker analysis, the individual variant association status is estimated by the posterior mean. The gene segment status is derived as the mean of individual status and the gene association status is represented by the highest segment status.

Gene Name	**Chromosome**	**Average Gene Status**
IL7	chr8	0.975
TIAM1	chr21	0.972
CKAP2L	chr2	0.961
TTC28	chr22	0.956
NDST4	chr4	0.952
EIF2AK2	chr2	0.952
CACNB4	chr2	0.95
PARD3B	chr2	0.946
TMEM117	chr12	0.944
ATP6V0D1	chr16	0.943
LOC101928058	chr8	0.943
KIFC3	chr16	0.941
LRRC4C	chr11	0.937
SPATA5	chr4	0.937
C1orf101	chr1	0.937
MIR6126	chr16	0.934
DLGAP1	chr18	0.932
KIAA0556	chr16	0.93
RNF166	chr16	0.928

UPP2	chr2	0.927
CMTM4	chr16	0.927
GCSH	chr16	0.924
GNA14	chr9	0.922
SLC18A2	chr10	0.921
HBB	chr11	0.92
LOC100128317	chr7	0.919
IMPAD1	chr8	0.919
CDKL3	chr5	0.918
INTS6	chr13	0.917
LOC101928775	chr9	0.917
BEAN1	chr16	0.916
DCC	chr18	0.916
APLP2	chr11	0.915
EVL	chr14	0.914
NPNT	chr4	0.913
CTB-12O2.1	chr5	0.912
TMEM232	chr5	0.912
MYLK3	chr16	0.911
LINGO2	chr9	0.91
LINC00456	chr13	0.91
GPALPP1	chr13	0.909
BMP1	chr8	0.909
MCF2L2	chr3	0.908
LEPROTL1	chr8	0.907
SAMD12	chr8	0.907
LOC102724874	chr8	0.907
PSD3	chr8	0.906
PPL	chr16	0.906
MAPK14	chr6	0.905
CLCC1	chr1	0.905
LOC100129307	chr13	0.904
CAMTA2	chr17	0.903
BTNL3	chr5	0.903
GOLM1	chr9	0.902

MYRIP	chr3	0.902
YWHAE	chr17	0.901
SPOPL	chr2	0.899
STX3	chr11	0.899
UNC5D	chr8	0.898
ARID1B	chr6	0.895
TAF4	chr20	0.894
CHODL	chr21	0.894
CNST	chr1	0.893
LINC00606	chr3	0.893
PHF21A	chr11	0.891
NUDC	chr1	0.891
CX3CL1	chr16	0.891
KIRREL3-AS2	chr11	0.891
METTL16	chr17	0.89
HEATR6	chr17	0.889
SLC4A10	chr2	0.887
KIAA0355	chr19	0.886
VCL	chr10	0.886
ATP2C1	chr3	0.885
LOC339874	chr3	0.884
ZGRF1	chr4	0.883
LOC101927849	chr4	0.883
SERINC3	chr20	0.883
APOA1	chr11	0.882
FAM110B	chr8	0.882
KIF6	chr6	0.881
C16orf46	chr16	0.88
ADAMTS20	chr12	0.878
HBS1L	chr6	0.878
DPH6	chr15	0.877
SYNJ2	chr6	0.877
APLF	chr2	0.877
YLPM1	chr14	0.876
DLG2	chr11	0.874

NRXN3	chr14	0.874
MICU3	chr8	0.872
ATG5	chr6	0.871
CMA1	chr14	0.871
PARK2	chr6	0.871
PRKG2	chr4	0.87
LINC00383	chr13	0.869
LOC101928203	chr16	0.868
SLC5A1	chr22	0.867
NRG1	chr8	0.866
IRS1	chr2	0.866

REFERENCES

Adam, L., R. K. Vadlamudi, P. McCrea, and R. Kumar (2001, apr). Tiam1 overexpression potentiates heregulin-induced lymphoid enhancer factor-1/β-catenin nuclear signaling in breast cancer cells by modulating the intercellular stability. *Journal of Biological Chemistry 276*(30), 28443–28450.

Alexandrov, L. B., , S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, , and and (2013, aug). Signatures of mutational processes in human cancer. *Nature 500*(7463), 415–421.

Alexandrov, L. B. and M. R. Stratton (2014, feb). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development 24*, 52–60.

Anastasiadi, Z., G. D. Lianos, E. Ignatiadou, H. V. Harissis, and M. Mitsis (2017, mar). Breast cancer in young women: an overview. *Updates in Surgery 69*(3), 313–317.

Attia, J. (2003, apr). Meta-analyses of molecular association studies: Methodologic lessons for genetic epidemiology. *Journal of Clinical Epidemiology 56*(4), 297–303.

Borecki, I. B. and M. A. Province (2008, sep). Genetic and genomic discovery using family studies. *Circulation 118*(10), 1057–1063.

Broad Institute TCGA Genome Data Analysis Center (2016). Snp6 copy number analysis (gistic2).

Brooks, S. P. and A. Gelman (1998, dec). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Bush, W. S., S. M. Dudek, and M. D. Ritchie (2008, nov). Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Biocomputing 2009*. World Scientific.

Bush, W. S., T. L. Edwards, S. M. Dudek, B. A. McKinney, and M. D. Ritchie (2008). Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics 9*(1), 238.

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

Cantor, R. M., K. Lange, and J. S. Sinsheimer (2010, jan). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics 86*(1), 6–22.

Casella, G. and E. I. George (1992, aug). Explaining the gibbs sampler. *The American Statistician 46*(3), 167–174.

Chen, Z., Y. Wu, Q. Meng, and Z. Xia (2015, sep). Elevated microRNA-25 inhibits cell apoptosis in lung cancer by targeting RGS3. *In Vitro Cellular & Developmental Biology - Animal 52*(1), 62–67.

Conte, N., B. Delaval, C. Ginestier, A. Ferrand, D. Isnardon, C. Larroque, C. Prigent, B. Séraphin, J. Jacquemier, and D. Birnbaum (2003, nov). TACC1–chTOG–aurora a protein complex in breast cancer. *Oncogene 22*(50), 8102–8116.

Ding, Y., B. Chen, S. Wang, L. Zhao, J. Chen, Y. Ding, L. Chen, and R. Luo (2009, feb). Overexpression of tiam1 in hepatocellular carcinomas predicts poor prognosis of HCC patients. *International Journal of Cancer 124*(3), 653–658.

Emigh, T. H. (1980, dec). A comparison of tests for hardy-weinberg equilibrium. *Biometrics 36*(4), 627.

Engers, R., M. Mueller, A. Walter, J. G. Collard, R. Willers, and H. E. Gabbert (2006, sep). Prognostic relevance of tiam1 protein expression in prostate carcinomas. *British Journal of Cancer 95*(8), 1081–1086.

Escudero-Esparza, A., M. Bartoschek, C. Gialeli, M. Okroj, S. Owen, K. Jirström, A. Orimo, W. G. Jiang, K. Pietras, and A. M. Blom (2016, oct). Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer. *Oncotarget 7*(47), 76920–76933.

François, O., H. Martins, K. Caye, and S. D. Schoville (2016, jan). Controlling false discoveries in genome scans for selection. *Molecular Ecology 25*(2), 454–469.

Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton (2004, mar). A census of human cancer genes. *Nature Reviews Cancer 4*(3), 177–183.

Garcia, M. A., J. Gil, I. Ventoso, S. Guerra, E. Domingo, C. Rivas, and M. Esteban (2006, dec). Impact of protein kinase PKR in cell biology: from antiviral to antiproliferative action. *Microbiology and Molecular Biology Reviews 70*(4), 1032–1060.

Gayán, J., A. González-Pérez, F. Bermudo, M. Sáez, J. Royo, A. Quintas, J. Galan, F. Morón, R. Ramirez-Lorca, L. Real, and A. Ruiz (2008). A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics 9*(1), 360.

Gentile, M., A. Wiman, S. Thorstenson, N. Loman, A. Borg, and S. Wingren (2001). Deletion mapping of chromosome segment 11q24-q25, exhibiting extensive allelic loss in early onset breast cancer. *International Journal of Cancer 92*(2), 208–213.

Giovannone, A. J., C. Winterstein, P. Bhattaram, E. Reales, S. H. Low, J. E. Baggs, M. Xu, M. A. Lalli, J. B. Hogenesch, and T. Weimbs (2018, feb). Soluble syntaxin 3 functions as a transcriptional regulator. *Journal of Biological Chemistry 293*(15), 5478–5491.

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. deFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton (2007, mar). Patterns of somatic mutation in human cancer genomes. *Nature 446*(7132), 153–158.

Kamal, M., A. M. Shaaban, L. Zhang, C. Walker, S. Gray, N. Thakker, C. Toomes, V. Speirs, and S. M. Bell (2009, aug). Loss of CSMD1 expression is associated with high tumour grade and poor survival in invasive ductal breast carcinoma. *Breast Cancer Research and Treatment 121*(3), 555–563.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. yee Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010, mar). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics 42*(4), 348–354.

Kim, H. S., J. D. Minna, and M. A. White (2013, jan). GWAS meets TCGA to illuminate mechanisms of cancer predisposition. *Cell 152*(3), 387–389.

Kim, S. H., A. P. Forman, M. B. Mathews, and S. Gunnery (2000, jun). Human breast cancer cells contain elevated levels and activity of the protein kinase, PKR. *Oncogene 19*(27), 3086–3094.

Kim, T.-H. and S.-G. Cho (2017, may). Kisspeptin inhibits cancer growth and metastasis via activation of EIF2ak2. *Molecular Medicine Reports 16*(5), 7585–7590.

Kooperberg, C. and M. LeBlanc (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genetic Epidemiology 32*(3), 255–263.

Larson, N. B. and D. J. Schaid (2013, jul). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology 37*(7), 695–703.

Launonen, V., F. Stenbäck, U. Puistola, R. Bloigu, P. Huusko, S. Kytölä, A. Kauppila, and R. Winqvist (1998, nov). Chromosome 11q22.3-q25 LOH in ovarian cancer: Association with a more aggressive disease course and involved subregions. *Gynecologic Oncology 71*(2), 299–304.

Lee, Y. S., N. Kunkeaw, and Y.-S. Lee (2019, jun). Protein kinase r and its cellular regulators in cancer: An active player or a surveillant? *WIREs RNA 11*(2).

Li, P., M. Guo, C. Wang, X. Liu, and Q. Zou (2014, 09). An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics 14*(2), 143–155.

Li, Q., J.-H. Seo, B. Stranger, A. McKenna, I. Pe'er, T. LaFramboise, M. Brown, S. Tyekucheva, and M. L. Freedman (2013, jan). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell 152*(3), 633–641.

Loo, L. W. M., I. Cheng, M. Tiirikainen, A. Lum-Jones, A. Seifried, L. M. Dunklee, J. M. Church, R. Gryfe, D. J. Weisenberger, R. W. Haile, S. Gallinger, D. J. Duggan, S. N. Thibodeau, G. Casey, and L. L. Marchand (2012, feb). cis-expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PLoS ONE 7*(2), e30477.

Ma, C., K. M. Quesnelle, A. Sparano, S. Rao, M. S. Park, M. A. Cohen, Y. Wang, M. Samanta, M. S. Kumar, M. U. Aziz, T. L. Naylor, B. Weber, S. S. Fakharzadeh, G. S. Weinstein, A. Vachani, M. D. Feldman, and M. S. Brose (2009, may). CharacterizationCSMD1in a large set of primary lung, head and neck, breast and skin cancer tissues. *Cancer Biology & Therapy 8*(10), 907–916.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher (2009, Oct). Finding the missing heritability of complex diseases. *Nature 461*(7265), 747–753.

Marees, A. T., H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks (2018, feb). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research 27*(2), e1608.

Martincorena, I. and P. J. Campbell (2015). Somatic mutation in cancer and normal cells. *Science 349*(6255), 1483–1489.

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz (2011, apr). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology 12*(4).

Meurs, E. F., J. Galabru, G. N. Barber, M. G. Katze, and A. G. Hovanessian (1993, jan). Tumor suppressor function of the interferon-induced double-stranded RNA-activated protein kinase. *Proceedings of the National Academy of Sciences 90*(1), 232–236.

Millstein, J., D. V. Conti, F. D. Gilliland, and W. J. Gauderman (2006, jan). A testing framework for identifying susceptibility genes in the presence of epistasis. *The American Journal of Human Genetics 78*(1), 15–27.

Minard, M. E., L. M. Ellis, and G. E. Gallick (2006, nov). Tiam1 regulates cell adhesion, migration and apoptosis in colon tumor cells. *Clinical & Experimental Metastasis 23*(5-6), 301–313.

Minard, M. E., L.-S. Kim, J. E. Price, and G. E. Gallick (2004, mar). The role of the guanine nucleotide exchange factor tiam1 in cellular migration, invasion, adhesion and tumor progression. *Breast Cancer Research and Treatment 84*(1), 21–32.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity 56*(1-3), 73–82.

Mukhopadhyay, I., E. Feingold, D. E. Weeks, and A. Thalamuthu (2009). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, n/a–n/a.

Nan, H., L. Han, J. Ma, C. Yang, R. Su, and J. He (2018, may). STX3 represses the stability of the tumor suppressor PTEN to activate the PI3k-akt-mTOR signaling and promotes the growth of breast cancer cells. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1864*(5), 1684–1692.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006, jul). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*(8), 904–909.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson (2010, jun). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics 11*(7), 459–463.

Pritchard, J. K. (2002, oct). The allelic architecture of human disease genes: common disease-common variant... or not? *Human Molecular Genetics 11*(20), 2417–2423.

Reich, D. E. and E. S. Lander (2001, sep). On the allelic spectrum of human disease. *Trends in Genetics 17*(9), 502–510.

Schmit, K. and C. Michiels (2018, dec). TMEM proteins in cancer: A review. *Frontiers in Pharmacology 9*.

Shir, A. and A. Levitzki (2002, aug). Inhibition of glioma growth by tumor-specific activation of double-stranded RNA–dependent protein kinase PKR. *Nature Biotechnology 20*(9), 895–900.

Smith, J. E., A. R. Clark, and A. T. Staggemeier (2009). A genetic approach to statistical disclosure control. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, GECCO '09, New York, NY, USA, pp. 1625–1632. ACM.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini (2009, may). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics 5*(5), e1000477.

Stadler, Z. K., P. Thom, M. E. Robson, J. N. Weitzel, N. D. Kauff, K. E. Hurley, V. Devlin, B. Gold, R. J. Klein, and K. Offit (2010). Genome-wide association studies of cancer. *Journal of Clinical Oncology 28*(27), 4255.

Thakkinstian, A., P. McElduff, C. D'Este, D. Duffy, and J. Attia (2005). A method for meta-analysis of molecular association studies. *Statistics in Medicine 24*(9), 1291–1306.

The Wellcome Trust Case Control Consortium (2007, jun). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661–678.

Tibshirani, R. (1996, jan). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Timpson, N. J., J. H. Tobias, J. B. Richards, N. Soranzo, E. L. Duncan, A.-M. Sims, P. Whittaker, V. Kumanduri, G. Zhai, B. Glaser, J. Eisman, G. Jones, G. Nicholson, R. Prince, E. Seeman, T. D. Spector, M. A. Brown, L. Peltonen, G. D. Smith, P. Deloukas, and D. M. Evans (2009, jan). Common variants in the region around osterix are associated with bone mineral density and growth in childhood. *Human Molecular Genetics 18*(8), 1510–1517.

Toyooka, K. O., S. Toyooka, A. K. Virmani, U. G. Sathyanarayana, D. M. Euhus, M. Gilcrease, J. D. Minna, and A. F. Gazdar (2001, June). Loss of expression and aberrant methylation of the cdh13 (h-cadherin) gene in breast and lung carcinomas. *Cancer research 61*, 4556–4560.

Tzeng, S.-T., M.-H. Tsai, C.-L. Chen, J.-X. Lee, T.-M. Jao, S.-L. Yu, S.-J. Yen, and Y.-C. Yang (2013, jun). NDST4 is a novel candidate tumor suppressor gene at chromosome 4q26 and its genetic loss predicts adverse prognosis in colorectal cancer. *PLoS ONE 8*(6), e67040.

Voight, B. F., , L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. A. McCarroll, C. Langenberg, O. M. Hofmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E. Boerwinkle, L. L. Bonnycastle, K. B. Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. F. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U. Jackson, P. R. V. Johnson, T. Jørgensen, W. H. L. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. B. Perry, A.-K. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. W. Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. M. van Dam, T. W. van Haeften, T. van Herpt, J. V. van Vliet-Ostaptchouk, G. B. Walters, M. N. Weedon, C. Wijmenga, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. J. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. A. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, B. O. Boehm, H. Campbell, M. J. Daly, A. T. Hattersley, F. B. Hu, J. B. Meigs, J. S. Pankow, O. Pedersen, H.-E. Wichmann, I. Barroso, J. C. Florez, T. M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, and M. I. M. and (2010, Jun). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics 42*(7), 579–589.

Walch, A., S. Seidl, C. Hermannstädter, S. Rauser, J. Deplazes, R. Langer, C. H. von Weyhern, M. Sarbia, R. Busch, M. Feith, S. Gillen, H. Höfler, and B. Luber (2008, feb). Combined analysis of rac1, IQGAP1, tiam1 and e-cadherin expression in gastric cancer. *Modern Pathology 21*(5), 544–552.

Wang, X., V. S. Pankratz, Z. Fredericksen, R. Tarrell, M. Karaus, L. McGuffog, P. D. Pharaoh, B. A. Ponder, A. M. Dunning, S. Peock, M. Cook, C. Oliver, D. Frost, O. M. Sinilnikova, D. Stoppa-Lyonnet, S. Mazoyer, C. Houdayer, F. B. Hogervorst, M. J. Hooning, M. J. Ligtenberg, A. Spurdle, G. Chenevix-Trench, R. K. Schmutzler, B. Wappenschmidt, C. Engel, A. Meindl, S. M. Domchek, K. L. Nathanson, T. R. Rebbeck, C. F. Singer, D. Gschwantler-Kaulich, C. Dressler, A. Fink, C. I. Szabo, M. Zikan, L. Foretova, K. Claes, G. Thomas, R. N. Hoover, D. J. Hunter, S. J. Chanock, D. F. Easton, A. C. Antoniou, and F. J. Couch (2010, apr). Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Human Molecular Genetics 19*(14), 2886–2897.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin (2010, jun). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics 86*(6), 929–942.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler (2005, dec). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics 38*(2), 203–208.

Zeng, P., Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu, and F. Chen (2015, jul). Statistical analysis for genome-wide association study. *Journal of Biomedical Research*.

Zhang, D., R. Shi, W. Xiang, X. Kang, B. Tang, C. Li, L. Gao, X. Zhang, L. Zhang, R. Dai, and H. Miao (2020, mar). The agpat4/LPA axis in colorectal cancer cells regulates antitumor responses via p38/p65 signaling in macrophages. *Signal Transduction and Targeted Therapy 5*(1).

Zhang, Y. and J. S. Liu (2007, aug). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics 39*(9), 1167–1173.

Ziegler, A. and I. R. König (2010, mar). *A Statistical Approach to Genetic Epidemiology*. Wiley-VCH Verlag GmbH & Co. KGaA.

BIOGRAPHICAL SKETCH

Yashi Bu was born and grew up in Guangdong, China. From 2010 to 2014, she attended the South China University of Technology and completed the Bachelor of Science degree in Applied Mathematics. In August 2014, she joined the PhD program in Statistics in The University of Texas at Dallas. In spring 2016, she passed the PhD Qualifying Exam, and received the Master of Science degree in Statistics. Under the guidance of Professor Min Chen, she started her research on statistical approaches for tumor and normal matched-paired data in genome-wide association studies (GWAS).

CURRICULUM VITAE

Yashi Bu

November 1, 2020

Contact Information:

Department of Mathematical Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080, U.S.A.

Email: Yashi.Bu@utdallas.edu

Educational History:

Ph.D., Statistics, The University of Texas at Dallas, 2020
M.S., Statistics, The University of Texas at Dallas, 2016
B.S., Applied Mathematics, South China University of Technology, 2014

Employment History:

Teaching Assistant, The University of Texas at Dallas, 2016 – 2019

Presentations:

Poster presentation, Southern Regional Council on Statistics, June 2018
Poster presentation, American Society of Human Genetics, 2018
Small Grants travel award, The University of Texas at Dallas, 2018

Professional Memberships:

American Statistical Association
American Mathematical Society