A Bayesian Hierarchical Model with SNP-level Functional Priors Applied to a PWAS.

by

Weizi Huang

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Advisor

_____
Elizabeth R. Hauser

_____
Jeanette Mccarthy

_____
Merlise Clyde

_____
Kenneth Kreuzer

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2010

Abstract

(Computational Biology and Bioinformatics)

A Bayesian Hierarchical Model with SNP-level Functional Priors Applied to a PWAS.

by

Weizi Huang

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Advisor

_____
Elizabeth R. Hauser

_____
Jeanette Mccarthy

_____
Merlise Clyde

_____
Kenneth Kreuzer

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2010

Abstract

Tremendous effort has been put into study of the etiology of complex diseases including the breast cancer, type 2 diabetes, cardiovascular diseases, and prostate cancers. Despite large numbers of reported disease-associated loci, few associated loci have been replicated, and some true associations does not belong to the group of the most significant loci reported to be associated. We built a Bayesian hierarchical model incorporated with SNP-level functional data that can help identify associated SNPs in pathway-wide association studies. We applied the model to an association study for the serous invasive ovarian cancer based on the DNA repair and apoptosis pathways. We found that using our model, blocks of SNPs located in regions enriched for missense SNPs or gene inversions were more likely to be identified as candidates of the association.

Contents

List of Tables

List of Figures

Acknowledgements

To Professor Edwin S. Iversen, thank you for being my thesis supervisor in the past few years. This thesis would not have been completed if without your constant encouragements, supports, and guidance. Thank you for introducing me to such an exciting field of research; thank you for your endless patience in leading me through my graduate studies; and thank you for giving me great inspirations and help in doing research.

To my thesis committee members, Professor Jeanette Mccarthy, Professor Elizabeth R. Hauser, Professor Merlise Clyde, and Professor Kenneth Kreuzer, thank you for your kind challenges and useful advices on my thesis.

To my parents and my husband, thank you for your continuous supports during this special period of my life. Even though we are far apart, you are one of my sources of motivations to complete my work here. To all of my friends at Duke and in China, you have brought so many happiness and enjoyable experience to my life.

Introduction

In the past few decades, mapping genes to phenotypes has received tremendous attention because it helps people understand the function of genes and uncover the mechanisms of life, as well as provide clues for the development of new treatments and therapies when phenotypes of interest are diseases. Great success has been achieved in mapping thousands of genetic loci to simple monogenic Mendelian disorders [1], including Huntington's disease [2]. Until recently, the Online Mendelian Inheritance in Man (OMIM), an updated online catalog of human genes and diseases [3], had recorded over 2200 disease related genes [4].

Complex traits are phenotypes that appear not to follow the simple dominant-recessive models of Mendelian inheritance, where inheritance can be explained by a single gene [5]. They have the property that a common genotype leads to different phenotypes or different genotypes but result in the same phenotype [5]. The susceptibility to complex diseases is one type of complex traits. Different from studies for Mendelian diseases, the primary obstacle in studying such disease susceptibility comes from the difficulty in linking a genetic marker to the disease trait, due to the incomplete penetrance of the disease-causing allele (the proportion of affected individuals among people carrying the genotype), genetic heterogeneity (mutations at different genes can cause the disease independently), or polygenic inheritance (multiple alleles at different genes must be present together to cause the disease) [5]. To address these problems, linkage analysis has been conducted on larger samples [6], large

family pedigrees [7], and more densely placed genetic markers [8] [9] [10] [11] [12]. However, linkage analysis has generated inconsistent results for some diseases [13], and it is less powerful than an association study for detecting loci with modest effects [14]. Instead of tracing disease transmissions in family pedigrees, association studies look for over-represented patterns of genetic markers in patients compared to controls. In 1996, the Genome Wide Association Study (GWAS) was suggested by Risch [14], Lander [5] and Collins [15]. The goal of the study is to conduct a whole-genome association scan to look for genetic variants that are associated to the disease susceptibility. Broadly speaking, there are two inferential paradigms used GWAS: the Frequentist approach and the Bayesian approach.

The most common Frequentist approach is to test the null hypothesis of no association (H_0) against the alternative hypothesis of association (H_1) one at a time for each SNP. If a test statistic (a value calculated based on the data and hypotheses) is significant (the test statistics is an extreme value under H_0), one may conclude that H_0 is not supported by the data and it is rejected. Frequentist approaches include Pearson's χ^2 test, Fisher's exact test, the Armitage trend test, and the score test. These methods have been successful in identifying associations between susceptibility loci and complex disease, such as type 2 diabetes [16] [17], inflammatory bowel disease [18], Crohn disease [19], breast cancer [20] [21], prostate cancer [22]. In Frequentist method, "p-value" is the commonly reported test statistic. A p-value smaller than 5×10^{-8} will be considered as significantly associated. However, the p-value does not provide any information on how likely H_1 is correct. It is possible that given a significant p-value, the H_1 is not supported by the data either.

The Bayesian approach is an alternative method that compares H_1 and H_0 by reporting a "Bayes Factor", a ratio of the probability of observing the data under H_1 and H_0. A landmark GWAS study conducted by the Wellcome Trust Case Control Consortium reported both p-values and the log Bayes Factor in summarizing the

evidence of association for individual SNPs [23]. When provided with priors on H_0 and H_1, the posterior odds of H_1 and H_0 conditional on the data can be calculated. When a SNP is more likely to be associated given the data, it will have a larger posterior odds. Several Bayesian methods have been developed for association studies including SNPTEST [24] and BIMBAM [25].

It has been the case that results from most association studies, from both Frequentist and Bayesian approaches, are not replicable, and sometimes true associations are not those with highest significance [4]. Therefore, we decided to develop a Bayesian hierarchical model that can incorporate biology-relevant knowledge into the process of determining SNP associations. We hope that by adding the biology information as "priors" into association studies, we will be able to identify association candidates which may be missed if without the annotation data. Our effort will be an exploration in the possibility of constructing such model and extracting useful information from the SNP functional annotation data for association studies. The model will also be a complementary to the current widely used methods in GWASs.

In Chapter 1, we describes how we conducted a case-control analysis on the SNP data downloaded from Johnson's database [26] to identify which functional variables are more critical in suggesting SNP associations. The procedure for the data collection, the study design, and analysis and results will be presented. In Chapter 2, we develop and implement a Bayesian hierarchical model incorporating SNP-level functional annotation data for applications in pathway-wide association studies. In Chapter 3, we apply the model to a pathway based association study on the data set from a Phase I GWAS of ovarian cancer with a total of 3994 subjects. The conclusion is given in Chapter 4.

1

Analysis of a Public Database

1.1 Introduction

Johnson et al. (2009) created an open access database of SNPs showing significant associations in 118 GWAS studies published through March 1, 2008 [26]. A total of 56411 SNPs have been recorded in Johnson's database, and 68% of them are within 60 kb of RefSeq genes. The database also documents the study information, such as the genotyping platforms, size of studies, and so on. SNPs are updated with their reference ID, chromosomal locations, associated phenotypes, p-values from the study, etc. It will be interesting to know whether the SNPs in Johnson's database share any common attributes. Hindorff et al. (2009) addressed the question in his own collection of GWAS results. He conducted a functional and evolutionary study on 465 unique SNPs with p-value less than 5×10^{-8}. These SNPs were extracted from 151 GWAS studies that assayed at least 100,000 SNPs in initial scans. Hindorff et al. (2009) found that LD blocks containing these SNPs are enriched in nonsynonymous sites and promoters, and less frequently in intergenic regions and microRNA target sites [27].

Since Hindorff's conclusion was not readily incorporable into our hierarchical model, we conducted our own analysis on the SNPs in Johnson's database. In this chapter, we selected a subset of SNPs from Johnson's database, and studied what functional variables may have the ability to suggest the association of these SNPs to phenotypes. We were interested in variables with SNP functional implications. We constructed a matched case-control data set, and conducted comparative analysis using both unconditional and conditional logistic regressions. We also ran a Bayesian model selection to identify the single model with the best predictive ability. Johnson's database was downloaded in May, 2009.

1.2 Data Collection

1.2.1 Construction of Case Blocks and Control Blocks

We have constructed a case-control study in which SNPs in Johnson's database meeting our definition of genome-wide significance define the "cases". Among 56411 SNPs in the database, 3061 of them were genotyped using the Illumina Hap 550 chip, and 44180 were genotyped using the Affymetrix 500K chip; the remaining ones were genotyped by custom or other platforms. To avoid the possibility of confounding due to chip design, we decided to focus on the studies using the Affymetrix 500K chip. We defined the "cases" as those SNPs with a p-value less than or equal to 5×10^{-8} in a first stage scan, or 0.05 in a replication study. 1557 out of 44180 SNPs were classified as cases. We chose our control SNPs from the SNPs on the Affymetrix 500K chip that did not appear in Johnson's database. Because of LD, an associated SNP can either be a disease-causing SNP or in strong LD with a causal variant. Therefore, we also included SNPs that were in strong LD ($r^2 \geq 0.8$) with the "case" and "control" SNPs. To summarize these SNPs, we introduced terms "block" and "LD partners". A block refers to a case or control SNP and its LD partners. If two case blocks overlapped, we merged them into a single case block. Same procedure

has been applied to control blocks. Control blocks were further chosen so as not to overlap with the case blocks.

Blocks vary in chromosomes, the number of array SNPs, and size. These three factors may become confounders in block associations. Thus, in the subsequent analysis, we matched each case block with a control block based on number of case/control SNPs and LD partners. We had constructed a total of 466 case blocks and 466 control blocks with 12365 SNPs and 12143 SNPs, respectively.

1.2.2 Functional Annotation Data for SNPs

Empirical data and context based predictions of SNP function may indicate how likely a SNP is disease-causing. We refer to SNP-level data on such properties as functional annotation data, and each type of functional data as a functional variable. In general, we focus on three types of functional variables, including: 1) variables applied to SNPs in protein-coding regions; 2) variables applied to SNPs in cis-acting regulatory regions; and 3) variables that are generally applicable to all types of SNPs, especially to SNPs of unknown functions.

Functional Variables Used in the Analysis

Some functional variables apply only to SNPs in particular contexts. In order to match SNPs with functional variables, we classified SNPs based on eight "func" categories adopted from dbSNP130 [28]. Table 1.1 lists eight types of SNPs and their dbSNP130 definitions. Functional variables are grouped by the type of SNPs they can be applied to in Table 1.2. As shown in the last two columns in Table 1.2, the number of case SNPs and control SNPs is umbalanced among different func types. The majority of the SNPs are located in introns or have no known function, while a few SNPs are positioned in the 5 prime untranslated region of known genes. To balance the data, we kept the most interesting func categories unchanged (missense

and coding-synonymous SNPs), aggregated small categories of SNPs in potential regulatory regions into the "nonCoding" group, and divided the unknown SNPs into sub-groups based on their relative proximity to nearby genes. Here the nonCoding SNPs include SNPs in untranslated regions (untranslated-5 and untranslated-3), gene flanking regions (near-gene-3 and near-gene-5), and introns. Table 1.3 shows seven new func categories and their definitions, and Table 1.4 lists theses new categories with the functional variables that can be applied to them. In following analysis, we used the new func categories, and the functional variables as shown in Table 1.3 and 1.4.

func Categories	**Definition**
missense	Change of amino acid with respect to reference assembly.
coding-synonymous	No change of amino acid with respect to reference assembly.
untranslated-3	In 3 prime end of the transcript, but not coding .
untranslated-5	In 5 prime end of the transcript, but not coding.
intron	In intron, excluding splicing sites.
near-gene-3	2kb downstream of a gene.
near-gene-5	2kb upstream of a gene.
unknown	No known dbSNP functional classification.

Table 1.1: The *func* categories of SNPs and their definitions from the NCBI dbSNP database, build 130.

Descriptions on Functional Variables

For coding SNPs, the "Polyphen" variable is uniquely applicable to missense SNPs. This variable is an indicator variable summarizing the PolyPhen score [29]. If a missense SNP is predicted to result in a deleterious change in the protein structure or function (eg. an amino acid change at a protein catalytic site may lead to protein malfunctioning), then the Polyphen value of the SNP will be "damaging".

For nonCoding SNPs, applicable variables include "oRegAnno", "RegPotential", and "TFBS". We constructed the variable oRegAnno based on experimentally iden-

func Type SNPs	**Applicable Functional Variables**	**Case SNPs**	**Control SNPs**
missense	Polyphen	152	63
coding-synonymous	–	135	67
untranslated-3	oRegAnno, RegPotential, TFBS, RNA.Sanger	235	120
untranslated-5	oRegAnno, RegPotential, TFBS	46	10
intron	oRegAnno, RegPotential, TFBS	4260	4901
near-gene-3	oRegAnno, RegPotential, TFBS	189	81
near-gene-5	oRegAnno, RegPotential, TFBS	201	91
unknown	dis_cat	6878	6735
Applicable to all SNPs: laminB1, dgv_cnv, dgv_indel, dgv_inv_invBP, RP.na.			

Table 1.2: Functional variables that can be applied to each *func* category of SNPs are listed in column 2. The number of case SNPs and control SNPs in each category are listed in the last two columns.

tified regulatory sequences recorded in the ORegAnno database [30] [31]. If a SNP is located in a regulatory sequence, the value of its oRegAnno score will be equal to 1, otherwise it is 0. The RegPotential score measures how likely it is that a sequence will have a regulatory function based on alignments of seven species including human, chimpanzee, macaque, mouse, rat, dog and cow [32] [33]. A score less than

New func Categories	**Definition**
more50kb	unknown SNP, dis_cat=more50kb
10-50kb	unknown SNP, dis_cat=10-50kb
less10kb	unknown SNP, dis_cat=less10kb
InRefGene	unknown SNP, dis_cat=InRefGene
coding-syn	no change of amino acid with respect to reference assembly
missense	change of amino acid with respect to reference assembly
nonCoding	SNPs located at intron, UTRs, near gene regions

Table 1.3: New *func* categories are generated by collapsing or redefining dbSNP *func* categories. The definition of each new category is listed in the right column.

New func Type SNPs	**Applicable Functional Variables**
more50kb	–
10-50kb	–
less10kb	–
InRefGene	–
coding-syn	–
missense	Polyphen
nonCoding	oRegAnno, RegPotential, TFBS, RNA.Sanger
Applicable to all SNPs: laminB1, dgv_cnv, dgv_indel, dgv_inv_invBP,RP.na.	

Table 1.4: Functional variables applied to SNPs in each new *func* category are listed in the right column.

0 indicates that the alignment patterns of the target sequence resembles a pattern typical for neutral DNAs, and a score larger than 0.1 indicates a very remarkable resemblance to the pattern of regulatory elements. Because RegPotential scores are generally much larger in coding regions of the genome, we applied this score only to nonCoding SNPs. TFBS indicates whether a SNP is located in a transcription factor binding site. These binding sites are found by a method developed by Xu and Taylor (submitted). For SNPs in 3' untranslated regions, we also checked whether they were located in a microRNA target site. The "RNA.Sanger" variable is generated by SNPinfo based on the miRBase database [34].

in the case of SNPs of unknown func type, we characterized them with the variable "disCat" to measure the relative distance of these SNPs from their closet reference genes based on SNPinfo data. The relative distance has four levels: in the reference gene, within 10kb flanking regions of a gene, in 50kb flanking regions but further away than 10kb, and outside the 50kb flanking regions. Based on these four levels, unknown SNPs are divided into four groups, as shown in the first four rows in Table 1.3.

The functional variables "laminB1", "dgv_cnv", "dgv_indel", and "dgv_inv_invBP" are not gene-centric, and they may cover relatively large chromosomal segments. Therefore, these variables can be applied to all eight types of SNPs. laminB1 indicates whether a SNP is located in a laminB1-associated domain (LAD) in lung cell fibroblasts. LADs tend to have lower gene expression levels and represent repressed segments in the genome [35]. dgv_cnv, dgv_indel, and dgv_inv_invBP indicate whether a SNP is located in a copy number variant, a region of insertion or deletion, or in a region with gene inversion or inversion break points, respectively. Data on these genomic structure variants is retrieved from the Database of Genomic Variants [36].

Download of Functional Annotation Data

Many SNP annotation databases are available nowadays [37] [38] [39]. We have chosen the databases which are updated and allow bulk downloads of data. The UCSC Genome Browser [40] and SNPinfo database [41] are updated regularly and have aggregated the most widely used functional annotations in the field from multiple external resources. All functional variables retrieved were under "NCBI/hg18" coordinates in the human reference genome assembly [42] [43]. Annotation data was last downloaded from UCSC Genome Browser and SNPinfo in July, 2009.

1.3 Study Design and the Model

In what follows, we describe a model for the relationship between functional variables and association status using a data set housing the results from past association studies. Since the association status will be described as a binary variable, the analysis will adopt the framework of a case-control study. In this analysis, cases and controls are blocks of SNPs, including genotyped SNPs and their LD partners. The association status of individual blocks are termed $Y = \{Y_b\}$, where $b = 1, ..., n$ and

n is the total number of blocks. $Y_b = 1$ indicates that block b is associated, while $Y_b = 0$ indicates that block b is unassociated. W is the design matrix with the first column equal to 1 and the remaining being the functional variables; W is an $n \times p$ matrix, where p-1 is the number of functional variables. Row b in W, $W_{b.}$, refers to the intercept and functional data for block b. ω, a $p \times 1$ vector, is the vector of the coefficient for the design matrix W with its first element ω_1 being the intercept.

We made the following assumptions: 1) the Y_b are conditionally independent of each other, given W and the parameter ω; 2)the probability of block b being associated to its case-control status depends on the biological function of the SNPs within the block; that is, $P(Y_b = 1)$ should be a function of $W_{b.}$. Because Y_b is a binary variable, we model the relationship between Y_b and $W_{b.}$ using logistic regression.

We first fit the data using an unconditional logistic regression. The likelihood function is

$$\begin{aligned} \mathcal{L}(Y|W,\omega) &= \prod_{b=1}^{n} P(Y_b|W_{b.},\omega) \\ &= \prod_{b=1}^{n} \Big[\frac{\exp(W_{b.} \times \omega)}{1+\exp(W_{b.} \times \omega)}\Big]^{Y_b=1} \Big[\frac{1}{1+\exp(W_{b.} \times \omega)}\Big]^{Y_b=0}, \end{aligned} \tag{1.1}$$

where the prior distribution for the intercept is

$$\omega_1 \sim Normal(0,1), \tag{1.2}$$

the priors for the seven new func variables (listed in Table 1.3) are

$$\omega_i \sim Normal(0,2), i = 2,...,8, \tag{1.3}$$

and the priors for the remaining functional variables are specified as

$$\omega_i \sim Normal(0,1), i = 9,...,17. \tag{1.4}$$

Here we assumed that the majority of the variables have little or no effects in predicting block associations, so we assigned them Normal priors centered at zero. However,

we think the seven new func variables may have a greater ability to determine the block associations, so we specified relatively defused priors for these variables. When the block level functional variables are standardized, the meaning of ω can be explained as: 1) $\exp(\omega_1)$ is the expected odds of a SNP being associated when the function variables are not taken into account; 2) $\exp(\omega_i)$ is the fold of change in the odds of association when the function variable i increases 1 unit on the standard deviation scale.

We also applied the conditional logistic regression on the data. Since we only have one-to-one matched case and control pairs, the likelihood function for the conditional logistic regression is

$$\prod_{b \in S} \frac{1}{1 + \exp \sum_i \omega_i (W_{b^*}[i] - W_b[i])}, \tag{1.5}$$

where S is the set of indexes for the case blocks, and $W_{b^*.}$ is the predictor variables of the matched control block for case block b. The conditional logistic regression is available as a function called *clogit* in the R package named *survival* [44].

1.4 Analysis and Results

1.4.1 Unconditional Versus Conditional Logistic Regression

If unmatched analysis is applied to a matched data set, we can get more conservative estimates of the relative risk [45] [46]. Table 1.5 presents the estimates and standard errors from Frequentist unconditional and conditional logistic regression analyses. Comparing the results from the two methods, we observe that the majority of the estimates from the unconditional analysis are biased towards zero, which indicates smaller odds ratio effects. Coefficients in bold are exceptions: the less10kb and dgv variables. These variables are overestimated in the unconditional analysis. Agreements are found between the results: variables dgv_inv_invBP, nonCoding, and more50kb are associated with an increased odds of association; variables RP.na

and laminB1 tend to decrease such odds; variables InRefGene and oRegAnno show ambiguities in their contributions to the odds.

	Unconditional glm		**Conditional glm**	
Variables	Mean	S.E.	Mean	S.E.
Intercept	0.064	0.072		
more50kb	0.47	0.186	0.483	0.195
0-50kb	0.15	0.2	0.245	0.208
less10kb	-0.216	0.145	-0.184	0.151
InRefGene	-0.021	0.1	-0.058	0.11
nonCoding	0.514	0.277	0.677	0.294
coding_syn	0.157	0.091	0.13	0.097
missense:benign	0.119	0.093	0.155	0.1
missense:damaging	0.056	0.083	0.036	0.086
oRegAnno	-0.108	0.082	-0.097	0.087
TFBS	-0.24	0.119	-0.257	0.123
RNA.Sanger	-0.107	0.079	-0.108	0.083
RegPotential	-0.188	0.156	-0.188	0.158
RP.na	-0.69	0.284	-0.729	0.291
laminB1	-0.285	0.273	-0.381	0.291
dgv_indel	0.31	0.104	0.29	0.101
dgv_inv_invBP	0.893	0.285	0.886	0.302
dgv_cnv	0.197	0.232	0.097	0.237

Table 1.5: Mean and standard errors (S.E.) from the analysis on the matched data set using unconditional and conditional logistic regressions. Variables with larger mean from the unconditional logistic regression are emphasized in bold.

1.4.2 Bayesian Model Selection

In this section, we investigate the predictive utility of 17 functional variables within the framework of Bayesian model averaging. We estimate the posterior probability for each model defined by a unique subset of all functional variables, and then select the model whose predictions are closest to the Bayesian Model Averaging (BMA) [47] [48] predictions.

Posterior Probability of the Model Given Data

The posterior probability of a model is given by

$$P(M|Y,W) = \frac{P(Y|M,W)P(M)}{\sum_M P(Y|M,W)P(M)}. \tag{1.6}$$

In order to calculate the posterior probability, $P(M|Y,W)$, we need to know the priors on the model, $P(M)$, and the marginal likelihoods of the data given the model, $P(Y|M,W)$. There was no prior preference on the models, so a uniform prior was used on all the models and $P(M) = \frac{1}{2^{17}-1}$. The marginal likelihood is calculated as an integral

$$P(Y|M,W) = \int P(Y|M,W,\omega)P(\omega)d\omega. \tag{1.7}$$

Because the product of the likelihood function, $P(Y|M,W,\omega)$, and the prior on parameter, $P(\omega)$, cannot be integrated with respect to ω in the closed form, the value of $P(Y|M,W)$ cannot be directly calculated. A Laplace approximation is often used to approximate such integrals [49]. It involves two steps: 1) find the maximum of the integrand; 2) use a second-order Taylor expansion to approximate the logarithm of the integrand [49].

Models are compared based on their posterior probabilities. Using the function *image.bma* in package *BAS* [50], the top 1024 models are plotted in Figure 1.1 according to their posterior probabilities and ranks. Given the posterior probabilities, several models surpass the rest, and the majority of the models have posterior probabilities around zero. The top model, with a posterior probability of 0.125, is more than 125 times more likely than any models ranked below the top 117, whose individual posterior probabilities drop below 0.001. In addition, the data does not favor models consisting of large numbers of functional variables. The top ranked model only includes three variables: less10kb, dgv_inv_invBP, and dgv_indel. The top 10

models generally have three to six functional variables. Variables dgv_inv_invBP, dgv_indel, missense, Polyphen, and RP.na have the largest posterior inclusion probabilities. This finding indicates that these variables are more important than the rest in predicting the association status of blocks.

Select the Best Model

We take into account both the model uncertainty and the predictive ability to look for a model that is closet to the true model. BMA has better performance in its predictions comparing to the single models [47]. BMA generates a sum of predictions which are weighted by the posterior probability of all averaged models that make the prediction. With this in mind, we decided to find out which model makes predictions closest to the BMA predictions, and take this model as our best model. The procedure used to select the best model is divided into four parts in Figure 1.2. First, samples of the coefficients are drawn from their posterior distributions for each model, and the predictions of a model are calculated by taking an average over its samples. The posterior distributions over the parameters, ω, are approximated by normal distributions with means and the variance-covariance matrix calculated in the model selection. By conducting multiple sampling, we are hoping to reduce the effect of sampling variability in the analysis. Second, BMA predictions are calculated with the models' posterior probabilities and their averaged predictions. Third, we use the Mean Square Error (MSE) between the BMA predictions and the model predictions to evaluate how closely each model approximates the BMA predictions. The best model is identified as the one with the minimum MSE to BMA predictions. To establish convergence, we ran seven analyses with different numbers of sampled coefficients. The sample sizes we tried included 100, 200, 400, 1000, 2000, 4000, and 6000. The best model we identified contains four variables: less10kb, missense, dgv_inv_invBP, and dgv_indel.

1.4.3 WinBUGS Predictions for the Best Model

The best model was implemented with WinBUGS to estimate the posterior distribution for the coefficients. The analysis used four variables: *less10kb*, *missense*, *dgv_inv_invBP*, and *dgv_indel*. As above, the model was the one used in the conditional logistic regression, where the likelihood function is given in (1.5). We assign a hierarchial normal prior on *less10kb* and *missense* variables, and allow for flexibility without constraining the prior mean at a given value. Variables *dgv_inv_invBP* and *dgv_indel* are assigned normal priors centering at 0. In particular, the priors are:

$$\mu_1, \mu_2 \sim Normal(0, 1), \tag{1.8}$$

$$\omega_1 \sim Normal(\mu_1, 1), \tag{1.9}$$

$$\omega_2 \sim Normal(\mu_2, 1), \tag{1.10}$$

$$\omega_3, \omega_4 \sim Normal(0, 1). \tag{1.11}$$

Here ω_1 and ω_2 are the coefficients for less10kb and missense, and ω_3 and ω_4 are the coefficients for dgv_inv_invBP and dgv_indel.

The WinBUGS chain consists of 100,000 burnin and 300,000 after burinin iterations with a thinning equal to 10. The chain passed the convergence tests, and the posterior summaries are listed as the last three columns in Table 1.6. Except for *less10kb*, the other three variables have positive posterior means. This may imply that when a block has missense SNPs or has SNPs located in regions of insertions, deletions, and gene inversions, the block will have a higher chance of being associated. The negative coefficient for the variable *less10kb* is counter-intuitive and may be due to how it is defined. *less10kb* SNPs are located within 10kb of genes but outside 2kb. Therefore, when a block has less10kb SNPs, the block tends to be located outside the 2kb flanking region of genes, and the block may have less chance to have function implications and less likely to be associated.

	From Laplace Approximation			From MCMC Updates		
Variables	Mean	HPD 2.5%	HPD 97.5%	Posterior Mean	HPD 2.5%	HPD 97.5%
less10kb	-0.046	-0.071	-0.02	-0.047	-0.072	-0.021
missense	0.2316	0.012	0.451	0.24	0.018	0.464
dgv_indel	0.241	0.058	0.424	0.256	0.078	0.448
dgv_inv_invBP	0.217	0.101	0.333	0.229	0.116	0.351

Table 1.6: Results of fitting the best model to the data using the Laplace approximation and the Markov Chain Monte Carlo method.

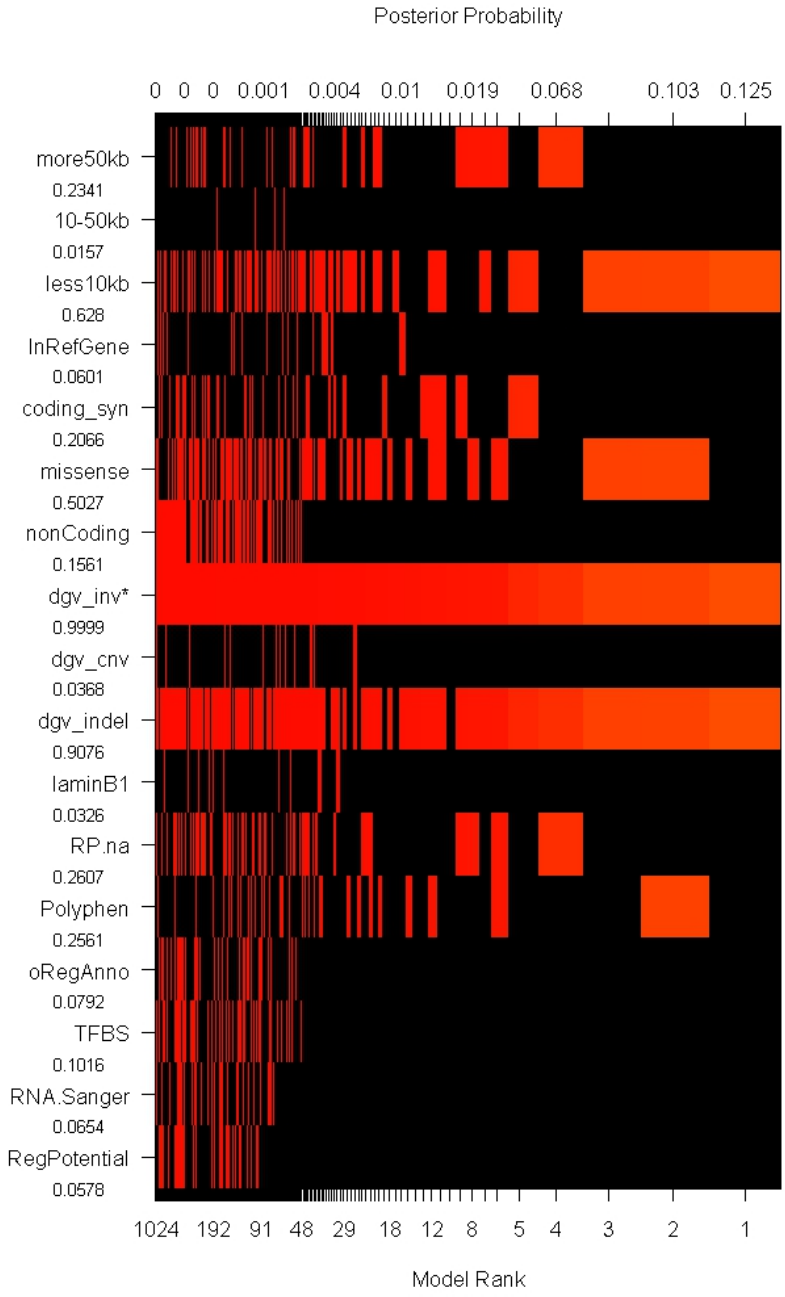

FIGURE 1.1: The plot of the top 1024 models ranked by their posterior probabilities. Each row specifies a variable, and each column represents a model. From the right to the left, models decrease in their posterior probabilities and ranks. The width of each column is proportional to its posterior probabilities. An intersection colored in red indicates that the variable is included in the model; otherwise, it is colored in black. Variable inclusion probabilities are listed below variable names. Note that variable *dgv_inv_invBP* is abbreviated as "dgv_inv*" in the plot.

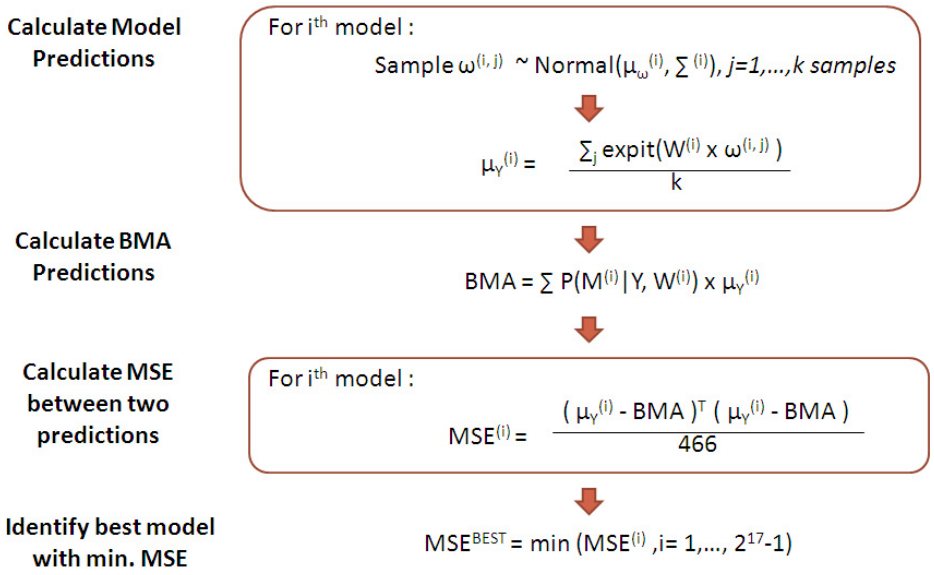

FIGURE 1.2: The procedure of selecting the best model by choosing a model that predicted closest to BMA predictions. $\omega^{(i,j)}$ refers to the j^{th} sample for the coefficient vector $\omega^{(i)}$ of model i. $\mu_\omega^{(i)}$ and $\Sigma^{(i)}$ are the mean and variance matrix for $\omega^{(i)}$. $\mu_Y^{(i)}$ is the expected probability of association for each SNP under model i.

2

Design and Implementation of a Bayesian Hierarchical Model for Pathway Wide Association Studies

2.1 Introduction

Hierarchical modeling has been applied in case-control epidemiology studies to discover the relationships between cancers and exposure variables or genetic variants [51] [52] [53]. The method is built on the framework of logistic regression with disease status as the response variable and genetic variants as predictor variables. Based on this framework, hierarchical models are able to incorporate prior assumptions on the coefficients of predictor variables. In this chapter, we develop a Bayesian hierarchical model for applications in pathway-wide association studies. Our response variable is disease status, but our predictor variables are function variables derived from the SNP-level functional annotation data. Priors will be assigned to these functional variables. Probability of association for each SNP given the data and functional variables will be estimated. The implementation of the model also will be described.

2.2 Design of a Bayesian Hierarchical Model

2.2.1 Description of Model Terminology

We assume that there are n individuals whose disease indicators are denoted by D_k, where $k = 1, ..., n$. The allele frequency data for the SNPs is stored in X, an $n \times m$ matrix, where m is the total number of SNPs. $X_{.i}$ refer to the i^{th} column of X. The functional annotation data is in W, an $m \times n_q$ design matrix, where column 1 is the intercept column and the remaining n_q-1 columns are functional variables. The i^{th} row of W, $W_{i.}$ contains the intercept and functional data for SNP i. ω, an $n_q \times 1$ vector, is the vector of coefficients for the design matrix W with its first element ω_1 being the intercept. M, an $m \times 4$ matrix, records the indicator variable that identifies which genetic model each SNP follows. This i^{th} row of M, $M_{i.}$ is a vector with four elements: the null model of no association ($M_{i0} = 1$), the dominant model ($M_{i1} = 1$), the recessive model ($M_{i2} = 1$), and the log-additive model ($M_{i3} = 1$). Z refer to confounder variables, and Z_k is the value of such variables for individual k.

2.2.2 The Likelihood Function

We begin with logistic regression models for disease given each SNP (taken individually) under each of three genetic models. We denote these by $P(D|X_{.i}, Z, \vec{\beta}_i, M_{ij})$, where $j = 0, 1, 2, 3$ and $\vec{\beta}_i$ is the coefficient vector for the regression on SNP i under model M_{ij} given the confounder Z. After incorporating the prior distribution, we integrate over $\vec{\beta}_i$ to calculate the marginal likelihood of the data under model M_{ij}, $P(D|X_{.i}, Z, M_{ij})$. Our interest is in inferring M_{ij} for each i. To this end, we specify the conditional prior $P(M_{ij}|W_{i.}, \omega)$ and write:

$$P(D|X_{.i}, Z, W_{i.}, \omega) = \sum_{j=0,1,2,3} P(D|X_{.i}, Z, M_{ij}) P(M_{ij}|W_{i.}, \omega). \tag{2.1}$$

The linear predictors in the logistic models are:

$$\text{logit}\Big(P(D_k = 1|X_{k,i}, Z_k, M_{i.}, \vec{\beta_i})\Big) = \begin{cases} \beta_{00} + \beta_{20}Z_k, & M_{i0} = 1; \\ \beta_{01} + \beta_{11}I\{X_{k,i} = 1 \vee 2\} + \beta_{21}Z_k, & M_{i1} = 1; \\ \beta_{02} + \beta_{12}I\{X_{k,i} = 2\} + \beta_{22}Z_k, & M_{i2} = 1; \\ \beta_{03} + \beta_{13}X_{k,i} + \beta_{23}Z_k, & M_{i3} = 1. \end{cases} \tag{2.2}$$

2.2.3 Assumptions in Distributions and Priors

Let μ_i specifies the probability of SNP i being associated given its functional variables $W_{i.}$ and ω. We assume that given that SNP i is associated, it has an equal probability of following the dominant model, the recessive model, or the log-additive model. We also assume that ω_1 is normally distributed with mean μ_ω and standard deviation τ, and that ω_q, $q = 2, ..., n_q$, are independent and normally distributed with mean 0 and standard deviation σ. Here μ_ω, τ, and σ are hyperparameters. The prior specification is flexible, and hierarchical Normal priors can also be assigned to the coefficients. The marginal likelihoods are specified in a G matrix with dimension equal to $m \times 4$. $G_{i.}$ are the marginal likelihoods of the four models calculated for SNP i based on the data. In particular, let

$$logit(\mu_i) = W_{i.} \times \omega \tag{2.3}$$

$$\pi_{i.} = [1 - \mu_i, \frac{\mu_i}{3}, \frac{\mu_i}{3}, \frac{\mu_i}{3}] \cdot G_{i.} \tag{2.4}$$

$$M_{i.} \sim Multinomial(1, \pi_{i.}) \tag{2.5}$$

$$\omega_1 \sim \mathcal{N}(\mu_\omega, \tau^2) \tag{2.6}$$

$$\omega_q \sim \mathcal{N}(0, \sigma^2), q = 2, ..., n_q. \tag{2.7}$$

2.3 Implementation of the Model

2.3.1 Markov Chain Monte Carlo Method

The model was implemented in R (version 2.8, [54]) using a hybrid Markov Chain Monte Carlo (MCMC) [55] algorithm employing both Random Walk (RW) and Gibbs updates. We are interested in estimating the posterior summaries for ω and M; however, we cannot achieve this by directly calculating their posterior distributions, because doing so requires evaluating extremely high dimensional integrals. MCMC is a commonly employed family of algorithms used to estimate summaries of the desired posterior distributions. The method constructs a Markov Chain whose stationary distribution is the distribution of interest. When the Markov Chain reaches equilibrium, the chain can then be used to sample from the desired distribution. In our model, we can only sample ω from its full conditionals, not M. Thus, we use RW updates to sample the components of ω, and Gibbs updates to sample M.

2.3.2 Details in Model Implementation

The MCMC algorithm is outlined in Figure 2.1. The algorithm requires starting values for the coefficients ($\omega^{(initial)}$) and the model indicators ($M^{(initial)}$). The initial values are generated by sampling from their prior distributions. The analysis outputs the following summaries: 1) M.avg, the averaged observed frequency of each genetic model for all SNPs calculated after burnin; 2) G.avg, the averaged probability of each genetic model for all SNPs after burnin; 3) values of the ω samples in the MCMC chain; 4) acceptance rate, the frequency of accepting a proposed value in the MCMC updating of ω. Since we update one coefficient at a time, the acceptance rate for each coefficient is recorded separately.

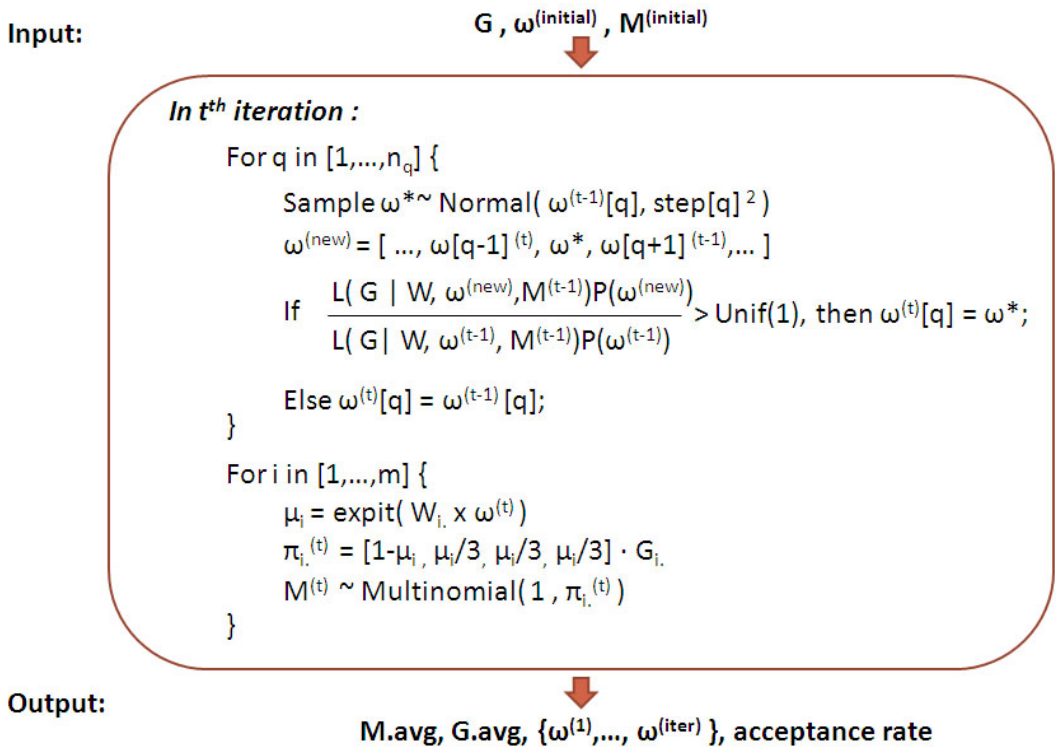

FIGURE 2.1: The diagram description on the model implementation in R.

2.3.3 Special Features of the Implementation

In designing the implementation, we also incorporate some features to facilitate mixing. First, we allow a different step size for each coefficient, shown as $step[q]$ for the q^{th} coefficient in Figure 2.1. A properly set step size is necessary for an efficient RW MCMC chain, because it controls how far from the existing value it is allowed to propose a new value. Before running a long chain for the final analysis, we run several preliminary chains and adjust the step size until the acceptance rates fall between 0.3 and 0.4. Second, we permit the flexibility to scale up or down the variance of our priors. We may choose to allow more prior uncertainty by increasing the prior variance. In addition, by conducting comparative analysis using differently scaled priors, we can also examine how the analysis is affected by the priors. We have evaluated our MCMC code using simulated data sets, and these analyses indicated

that the code was accurate.

3

Application of the Model in a Pathway-wide Association Study

3.1 Introduction

In this chapter, we conducted a pathway-wide association study using the Bayesian Hierarchical model described in Chapter 2. The data used in the study were collected from a Phase I GWAS of ovarian cancer, comprising genotype data on a total of 3994 subjects divided into 1206 serous only cases and 2042 controls. In the full GWAS, 559,179 SNPs passed the quality control, and 2,543,887 SNPs were imputed by MACH version 1.0.16 [56], using phased HapMap haplotypes for the CEU population. In the present study, we focused on 2017 blocks of SNPs in 181 genes belonging to the DNA repair and apoptosis pathways that are critical in tumor prohibitions [57] [58]. The annotation data for the SNPs was downloaded from the SNPinfo FuncPred database [41] and the UCSC Genome Browser [40] on April 12^{th}, 2010. Four functional variables were used in the study: *less10kb*, *missense*, *dgv_indel*, and *dgv_inv_invBP*. Priors on the coefficients of the four functional variables were learned from the results of the analysis in Chapter 1. A sensitivity analysis

and a comparative analysis were conducted to assess the influence from prior specifications and from the annotation data on identifying association candidates. We also monitored whether the annotation data on SNPs in the TP53 gene would change their probabilities of association.

3.2 Construction of Functionally Annotated Block Data

Figure 3.1 summarizes how the LD block data set was constructed. There are 181 genes of interest in the DNA repair and apoptosis pathways. The block data was generated by first identifying array SNPs located in or near these pathway genes, and then constructing LD blocks for these array SNPs using their LD partners. We downloaded the transcript start and end positions of the 181 genes from the UCSC Tables. We specified which table to use by choosing the assembly "Mar.2006(NCBI36/hg18)", the group "Gene and Gene Prediction Tracks", the track "RefSeq Genes", and the table "refGene". Only chromosomal position information in the reference assembly was used. We expanded the gene regions by including 10,000 base pairs upstream and downstream of each gene. The expended gene regions were then used to retrieve SNPs of interest from the UCSC Tables. We selected the table by choosing the assembly "Mar.2006(NCBI36/hg18)", the group "Variation and Repeats", the track "SNPs(126)", and the table "snp126". Then we selected "define regions" to upload our gene regions. This identified 62,402 SNPs of interest and provided us with their chromosomal locations. Among these, 2078 SNPs are array SNPs on the Illumina 610k chip. We then identified LD partners of these 2078 array SNPs by looking for SNPs with r^2 larger or equal to 0.8 using the HapMap release 27 [59] LD map of CEU population. 465 of the 2078 array SNPs had no LD partners. A total of 1613 blocks were generated for the remaining array SNPs.

We annotated the individual SNPs first, and then aggregated for each block to get block-level annotations. First, SNPs were submitted to the SNPinfo database

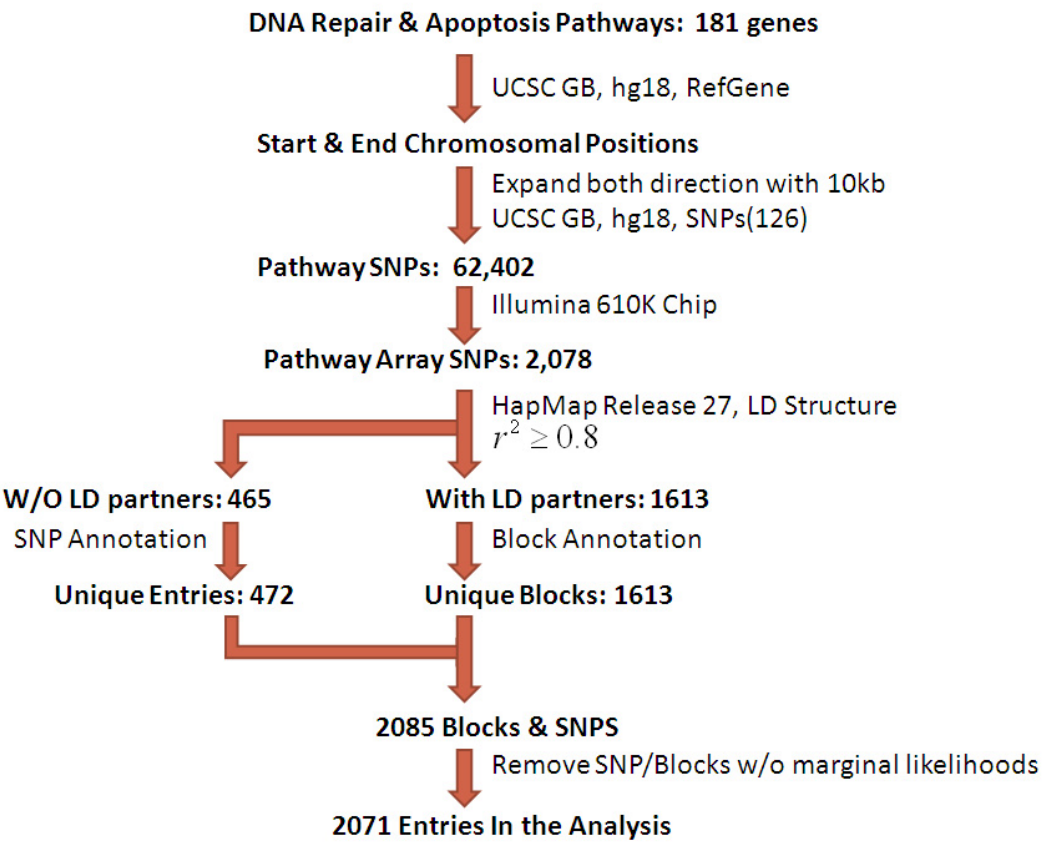

FIGURE 3.1: The diagram description on how the data was collected for a pathway-wide association study.

to identify each SNP's nearby genes. Since LD partners can extend beyond gene regions, we needed to define the relative position of SNPs to their nearby genes. In addition, SNPs with multiple genomic locations were deleted. Second, SNPs' updated chromosomal locations and func annotations were retrieved from the UCSC Tables. The table was chosen as 'Mar.2006(NCBI36/hg18)", the group "Variation and Repeats", the track "SNPs(130)", and the table "snp130". Third, with updated chromosomal locations, we look for SNPs that are located in regions known to contain structural variations, including insertions/deletions and gene inversions. These variables were further processed and aggregated in Table 1.3 and 1.4 to generate four function variables: *less10kb*, *missense*, *dgv_indel*, *dgv_inv_invBP*. Missense SNPs

without Polyphen data were deleted. A total of 2085 unique entries of SNPs and blocks were obtained.

Marginal likelihoods are calculated for each SNP on the three types of genetic models using serous only cases and controls. For entries of single SNPs, their marginal likelihoods are directed used; block marginal likelihoods are specified as those of their array SNPs on which the blocks were constructed. 14 SNPs were shown as monomorphic in the genotyping data, and they are deleted in the analysis. Finally, 2071 entries of SNPs and blocks remained in our data set.

3.3 Running the Model with Pathway Data

In this section, we describe our application of the model outlined in Section 2.2 to the pathway dataset. Here we used a multivariate normal prior for the coefficient vector (excluding the intercept), instead of independent normal priors for each coefficient. The prior assumptions for the intercept and other coefficients are:

$$\omega_1 \sim Normal(\log(\frac{1}{999}), [\log(\frac{1}{499}) - \log(\frac{1}{999})]^2), \tag{3.1}$$

$$\omega_{-1} \sim MultivariateNormal(\mu.vec, \Sigma), \tag{3.2}$$

$$\mu.vec = [-0.047, 0.240, 0.256, 0.230], \tag{3.3}$$

and

$$\Sigma = \begin{bmatrix} 1.71e-4 & -4.88e-4 & -3.14e-4 & -1.65e-4 \\ -4.88e-4 & 1.30e-2 & 4.79e-4 & 7.46e-6 \\ -3.14e-4 & 4.79e-4 & 9.01e-3 & 4.50e-4 \\ -1.65e-4 & 7.46e-6 & 4.51e-4 & 3.65e-3 \end{bmatrix}. \tag{3.4}$$

These priors are learnt from the posterior summaries on MCMC updates for our best model in the analysis in Chapter 1. Scale parameters and step sizes are listed in Table 3.1. Each MCMC chain was run for 400,000 burnin iterations and 600,000 post-burnin updates. The approximate runtime was 500 iterations per minute. The

MCMC chains passed the convergence diagnostics, and the acceptance rates fell between 0.3 and 0.4.

	Step Size s parameter				
	ω_0	ω_1	ω_2	ω_3	ω_4
scale = 1	1.3	0.18	0.55	0.5	0.4
scale = 25	2.3	0.45	1.25	1.2	0.85

Table 3.1: Scale parameters and step sizes are listed in the table. There are two level of scales: 1 and 25. Each $\omega_{\cdot}$ was assigned a different step size under the two scales.

We assessed the influence of the prior distribution on ω and of the functional data by comparative analysis, in which $Model_{intercept}$ and $Model_{best}$ were compared when Σ was scaled by a factor of 1 and 25. $Model_{intercept}$ refers to the model using only an intercept in the annotation data; $Model_{best}$ is our best model with the four additional function variables. The prior distribution used for $Model_{intercept}$ is:

$$\omega_1 \sim \mathcal{N}(\log(\frac{1}{999}), [\log(\frac{1}{499}) - \log(\frac{1}{999})]^2). \tag{3.5}$$

The scale parameters and step sizes are listed in the first column in Table 3.1. As when fitting $Model_{best}$, each MCMC chain for $Model_{intercept}$ has 400,000 burnins and 600,000 after burnin updates. The MCMC chains passed convergence diagnostics, and the acceptance rates fell between 0.3 to 0.4.

3.4 Analysis and Results

3.4.1 Sensitivity Analysis

For the sensitivity analysis, we compared the results obtained using a prior scaling of 1 (the estimates derived from the analysis in Chapter 1), and 25, (the Σ scaled by a factor of 25). Posterior and prior summaries are listed in Tables 3.2 and 3.3. Under both scales, the posterior mean of the coefficients are similar to their prior

means; however, in both cases the posterior mean of the intercept is smaller than the prior mean. This suggests that the data is not informative for the coefficients, and there are fewer associated SNPs than our prior expectation, 1 out of 1000. In addition, we observed a decrease in the posterior mean of the intercept from -7.39 to -10.235 when the scale was increased from 1 to 25. This indicates the intercept is sensitive to the prior distribution.

	Posterior Summaries, scale=1				**Prior Summaries, scale=1**			
Variables	**Mean**	**SD**	**2.50%**	**97.50%**	**Mean**	**SD**	**2.50%**	**97.50%**
Intercept	-7.39	0.573	-8.513	-6.266	-6.907	0.694	-8.267	-5.546
less10kb	-0.047	0.013	-0.072	-0.021	-0.047	0.013	-0.073	-0.021
missense	0.238	0.113	0.016	0.46	0.24	0.114	0.017	0.464
dgv_indel	0.256	0.095	0.07	0.442	0.256	0.095	0.07	0.442
dgv_inv_invBP	0.226	0.06	0.11	0.343	0.229	0.06	0.111	0.348

Table 3.2: Results of fitting the best model to the pathway data when the scale of the prior variance is equal to 1. Posterior summaries are compared to the prior summaries of each coefficient.

	Posterior Summaries, scale=25				**Prior Summaries, scale=25**			
	Mean	SD	2.50%	97.50%	Mean	SD	2.50%	97.50%
Intercept	-10.235	2.145	-14.439	-6.031	-6.907	3.47	-13.709	-0.104
less10kb	-0.045	0.065	-0.172	0.082	-0.047	0.065	-0.1.75	0.081
missense	0.204	0.544	-0.862	1.27	0.24	0.569	-0.875	1.356
dgv_indel	0.25	0.474	-0.679	1.179	0.256	0.475	-0.674	1.186
dgv_inv_invBP	0.174	0.268	-0.351	0.699	0.229	0.302	-0.363	0.821

Table 3.3: Results of fitting the best model to the pathway data when the scale of the prior variance is equal to 25. Posterior summaries are compared to the prior summaries of each coefficient.

3.4.2 Effects of Annotation Data on Association

While the pathway data set does not help us update the relationship between the functional data and association status, comparative analysis suggests that the annotation data can help us identify potentially interesting candidates that may have

otherwise been ignored. The logarithm of the observed frequency of associations, $\log(1 - M.avg|Model)$, obtained using $Model_{intercept}$, are plotted against those obtained using $Model_{best}$, under the two scales in Figure 3.2 and Figure 3.3. The two figures show a similar pattern: the majority of the points are distributed around the line $y = x$; a funnel shape is observed when the values are close to -13; and some points tend to form small local clusters that are deviated from the line $y = x$. A straight line $y = x - 1$ is also plotted in the figures to demonstrate the degree of deviation. The pattern implies that the evidence of association for the majority of the SNPs/blocks is not significantly influenced by the annotation data; for those with small association probabilities, sampling variations are observed in estimating their probabilities of association, explaining the funnel shape in the plot. The points that are significantly below line $y = x$ are association candidates whose posterior probabilities of association increased given the annotation data. Clusters of points indicate that association candidates may be closely located or overlap with one another. For example, the three blocks located on the right-most corner below $y = x - 1$ in Figure 3.2 are constructed based on three different array SNPs rs17153785, rs4840584, and rs8191606; however, since these array SNPs are closely located, the three blocks are identical, including the same group of LD SNPs.

We take a further look at the blocks colored in red in Figure 3.2 and Figure 3.3 to find out why the rankings of these blocks are greatly elevated by the functional annotation data. The genomic coordinates and the approximate block size in "kb", (kilobase pairs), of the 12 blocks are listed in Table 3.4. Note that some blocks are highly overlapped. For example, blocks 3 to 5 and block 8 on chromosome 8 are nested together. Another observation is that some of the large blocks are ten times as large as the smaller ones. Since the analysis was not conditioned on the block size, this observation may indicate the block size is not a confounder in the block association. Table 3.5 lists the array SNPs that tag each of the 12 blocks. 9 out of

12 blocks are tagged by more than one array SNP, and the largest block, block 11, is tagged by 9 array SNPs. This implies that the regions where large blocks are located may be subject to less recombination; array SNPs located in these regions may be redundantly genotyped based on the LD structure.

	Block	**Chrom.**	**Block Start**	**Block End**	**Block Length (kb)**
Both Scales	1	7	5894482	6037641	143.2
	2	7	6002033	6032531	30.5
	3	8	11638919	11739531	100.6
	4	8	11648909	11692205	43.3
	5	8	11648909	11674271	25.4
	6	15	38628978	38818978	190
scale=1	7	8	11648909	11660865	12
	8	8	11654245	11681324	27.1
scale=25	9	1	15686773	15785708	99
	10	4	2115518	2235898	120.4
	11	17	38430214	38595510	165.3
	12	19	50600890	50614163	13.3

Table 3.4: List of the unique 12 most deviated blocks (blocks colored in red in Figure 3.2 and 3.3). Blocks 1 to 6 are found deviated under both scales; blocks 7 and 8 are found deviated only under the scale of 1, and blocks 9 to 12 are found deviated only under the scale of 25. The chromosomal locations and block size of these blocks are shown in columns 2 to 5.

Known oncogenes are found in the 12 blocks. Table 3.6 lists all the reference genes found in the block regions. Genes with records in the OMIM database are highlighted in bold font. Table 3.7 lists the reference genes with MorbidMap descriptions that document to which diseases they are related. Cancer susceptibility genes include: BRCA1 (Breast Cancer 1 Gene), RAD51 (Recombination Protein A), and PMS2 (Postmeiotic Segregation Increased 2). Mutations in BRCA1 identified in patients with familial breast-ovarian cancer syndrome include deletions, insertions,

Block	Array SNPs Within Each Blocks
1	rs1860459
2	rs2286680,rs3779107
3	rs17153785, rs4840584,rs8191606
4	rs2686187,rs1466785
5	rs809204,rs804256
6	rs2619681,rs2304580
7	rs3203358
8	rs804292
9	rs4646018,rs1052571,rs2042370,rs1862710
10	rs1745335,rs529966,rs9328764,rs6830513, rs7659386,rs10011549,rs10018786
11	rs9911630,rs11657053,rs8176273,rs8176265, rs1799966,rs1060915,rs16942,rs799917, rs16940
12	rs735482,rs2336219,rs3212964

Table 3.5: Array SNPs that tag each of the 12 block are listed.

missense substitutions, regulatory mutations, and frameshifts [60] [61] [62]. Mutations in PMS2 found in HNPCC4 patients include deletions, missense mutations, and genomic rearrangements [63] [64] [65]. Other types of mutations in PMS2, including a truncating mutation, were discovered in studies of the Mismatch Repair Cancer Syndrome, such as the early-onset brain tumor [66] and colon cancer [67]. In RAD51, a missense mutation was found in patients with familial breast cancer [68], and the 135G-C SNP is associated with increased breast cancer risk in the carriers of mutations in BRCA1 and BRCA2 genes [69] [70].

The 12 blocks have an increased number of SNPs located in regions of gene inversions and/or a number of missense SNPs. Table 3.8 lists the functional variables and the number of SNPs in each block. Almost all SNPs in blocks 1 to 5 and 7 to 8 are located in a region with gene inversions, and blocks 6 and 9 to 12 contain three to four missense SNPs. Therefore, blocks located in a region of gene inversion or containing more missense SNPs are more likely to be proposed as candidates of

Block	Genes Located in the Block Region
1	**PMS2**,**JTV1**,EIF2AK1,C7orf28A,RSPH10B
2	**PMS2**,**JTV1**,EIF2AK1
3	**GATA4**,**FDFT1**,**CTSB**,NEIL2
4	**GATA4**,NEIL2
5	**GATA4**,NEIL2
6	**RAD51**,**FAM82A2**,C15orf57,RPUSD2,CASC5
7	**GATA4**
8	**GATA4**,NEIL2
9	**ELA2B**,**CASP9**,DNAJC16, AGMAT
10	**POLN**,HAUS3,MXD4
11	**BRCA1**,**NBR1**,NBR2,RND2
12	**ERCC1**,PPP1R13L,CD3EAP

Table 3.6: Reference genes that are located in or overlapped with each block are listed. These genes are found in the "refGene" table under the "NCBI/hg18" reference assemble. Gene names in bold are documented in the OMIM database.

association based on our best model.

3.4.3 Analysis of Blocks Covering the TP53 Gene

The product of the TP53 gene is a transcription factor that plays an anti-cancer role by inducing cell cycle arrest, DNA repair, and apoptosis. Deletions or muta-

Gene Name	OMIM MorbidMap Description
BRCA1	Familial breast-ovarian cancer
ERCC1	Cerebrooculofacioskeletal syndrome 4
GATA4	Atrial septal defect-2
PMS2	Hereditary Nonpolyposis Colorectal Cancer Type 4 (HNPCC4) Mismatch repair cancer syndrome
RAD51A	Susceptibility to Breast cancer

Table 3.7: Some OMIM reference genes are found related to certain diseases. This information was found in OMIM Morbid map description.

Block	less10kb	missense	dgv_indel	dgv_inv_invBP	SNP Number
1	3	0	0	**9**	9
2	0	0	0	**6**	7
3	1	0	0	**13**	13
4	5	0	0	**9**	10
5	1	0	0	**6**	6
6	9	**4**	0	**8**	33
7	2	0	0	**5**	5
8	1	0	0	**5**	5
9	0	**4**	0	0	45
10	0	**3**	0	0	49
11	7	**4**	0	0	82
12	0	**3**	0	0	7

Table 3.8: Functional annotation data for the 12 most deviated blocks.

tions in the TP53 gene are commonly found in many human cancers [71] [72]. The meta-analysis including 13 case-control studies in the Ovarian Cancer Association Consortium has identified two SNPs that are associated with serous invasive ovarian cancer [73]. In our analysis, there are 4 blocks covering the TP53 gene, and one array SNP located in the gene. Our annotation data has little influence on the probability of association for the blocks and the SNP. We observe that the five points (colored in red) are located on or near the line $y = x$ in Figure 3.4 and Figure 3.5. Table 3.9 lists the functional annotation data for the array SNP (the first row) and the four blocks (the last four rows). Since their values of functional variables are mostly zero, their probabilities of association are similar with or without the annotation data.

Array SNP	less10kb	missense	dgv_indel	dgv_inv_invBP	SNP Num.
rs12951053	0	0	0	0	1
rs2909430	0	0	0	0	2
rs8079544	0	0	0	0	2
rs2078486	0	0	0	0	5
rs2287497	0	1	0	0	5

Table 3.9: Functional annotation data of blocks that cover or overlap with the TP53 gene.

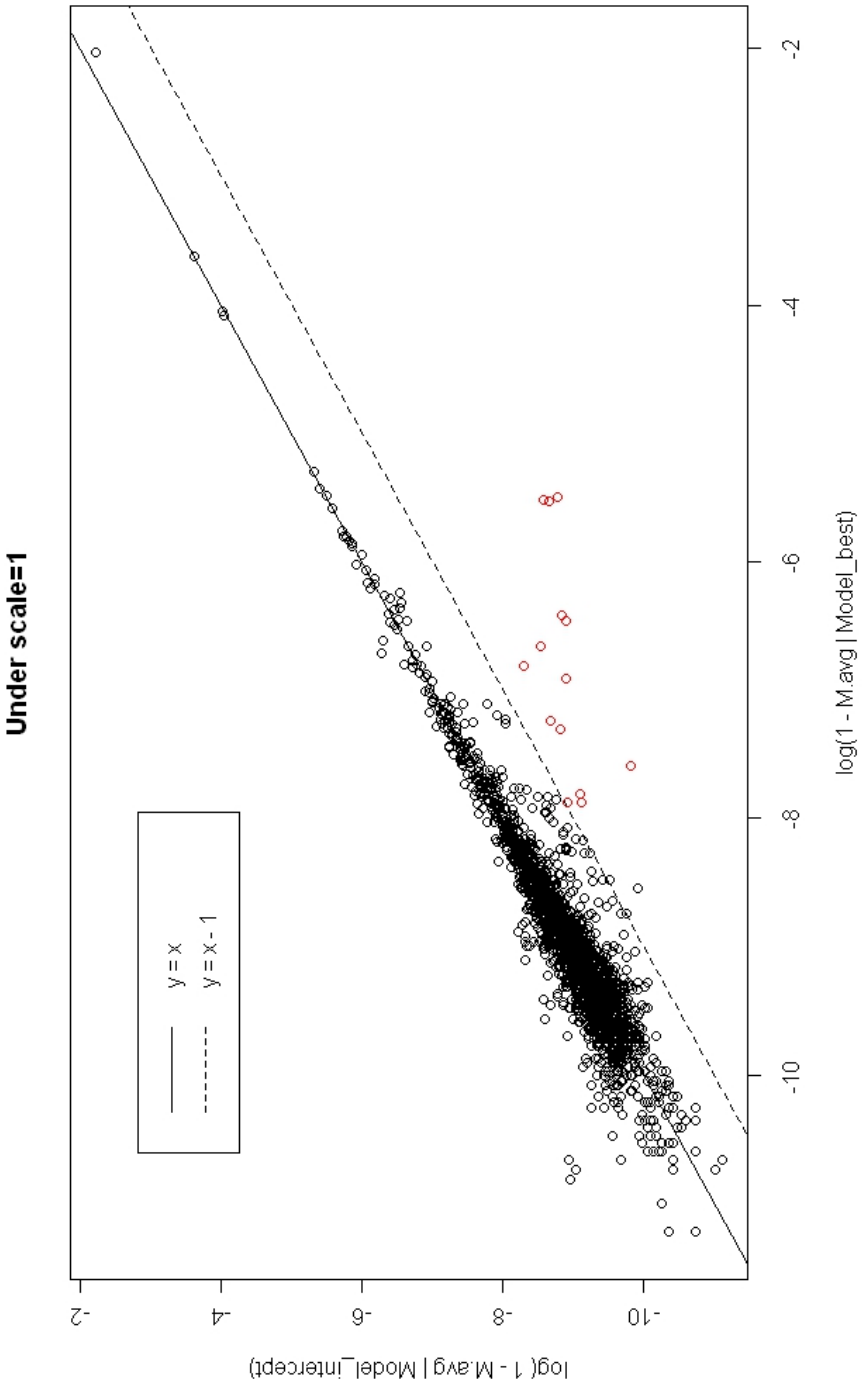

FIGURE 3.2: The logarithm of the averaged frequency of observed association obtained by fitting the best model (M_{best}) is plotted against that calculated by fitting the model with the intercept only ($M_{intercept}$). A scale of 1 is applied to the prior variance for the intercept and coefficients. Blocks colored in red are analyzed further in detail.

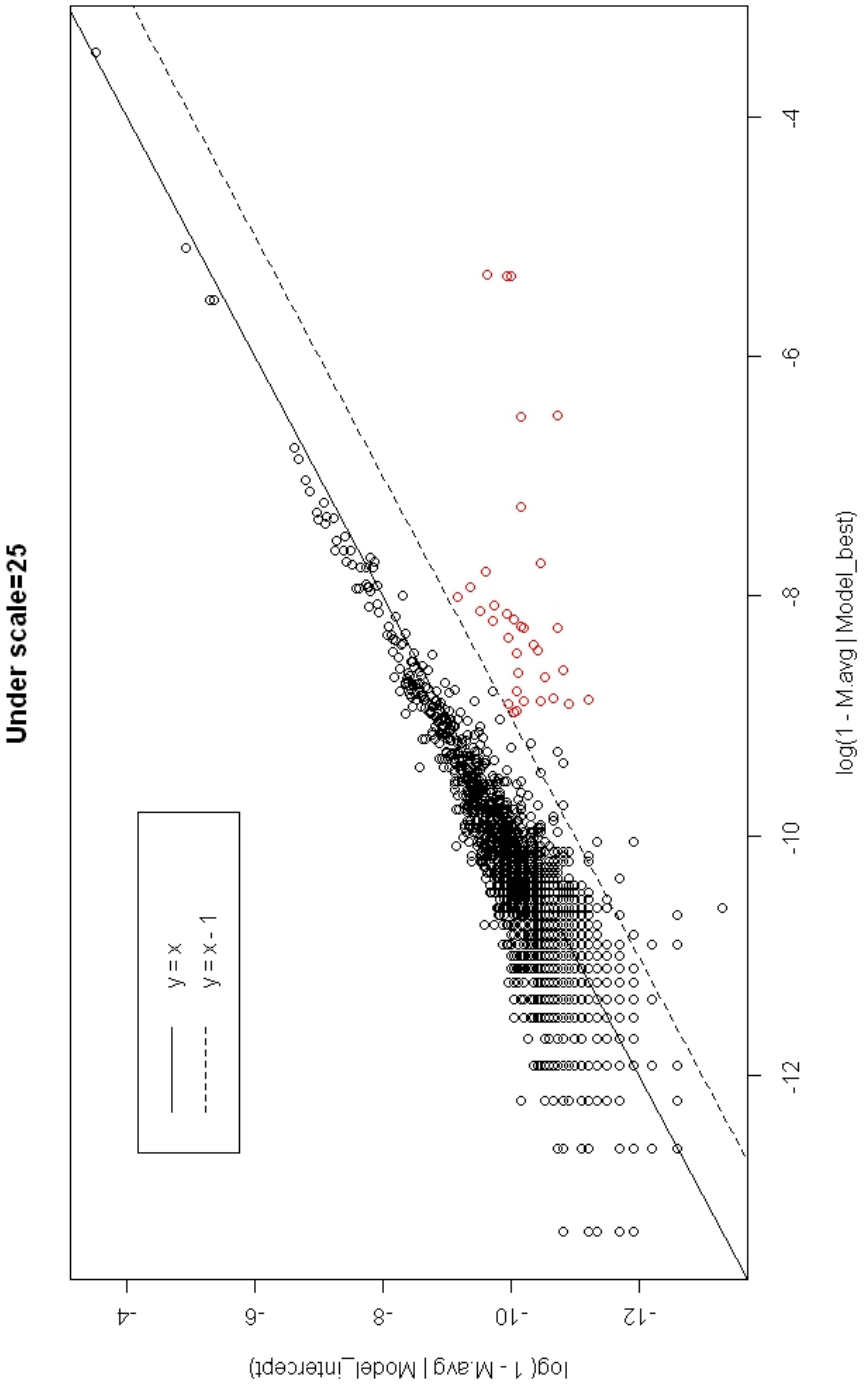

FIGURE 3.3: The logarithm of the averaged frequency of observed association obtained by fitting the best model (M_{best}) is plotted against that calculated by fitting the model with the intercept only ($M_{intercept}$). A scale of 25 is applied to the prior variance for the intercept and coefficients. Blocks colored in red are analyzed further in detail.

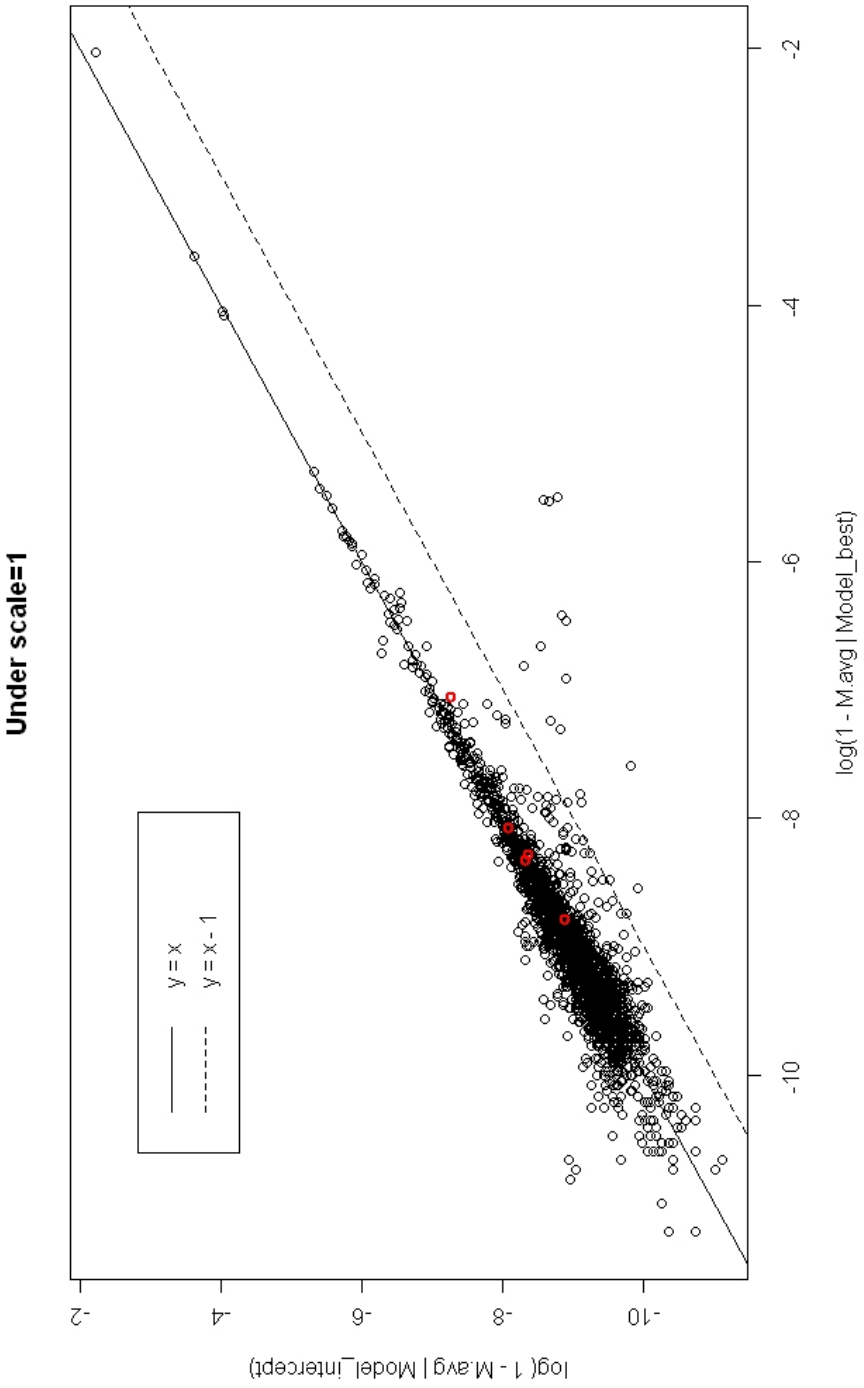

FIGURE 3.4: The logarithm of the averaged frequency of observed association for blocks covering the TP53 gene (colored in red) are compared with and without annotation data. A scale of 1 is applied to the prior variance for intercept and coefficients.

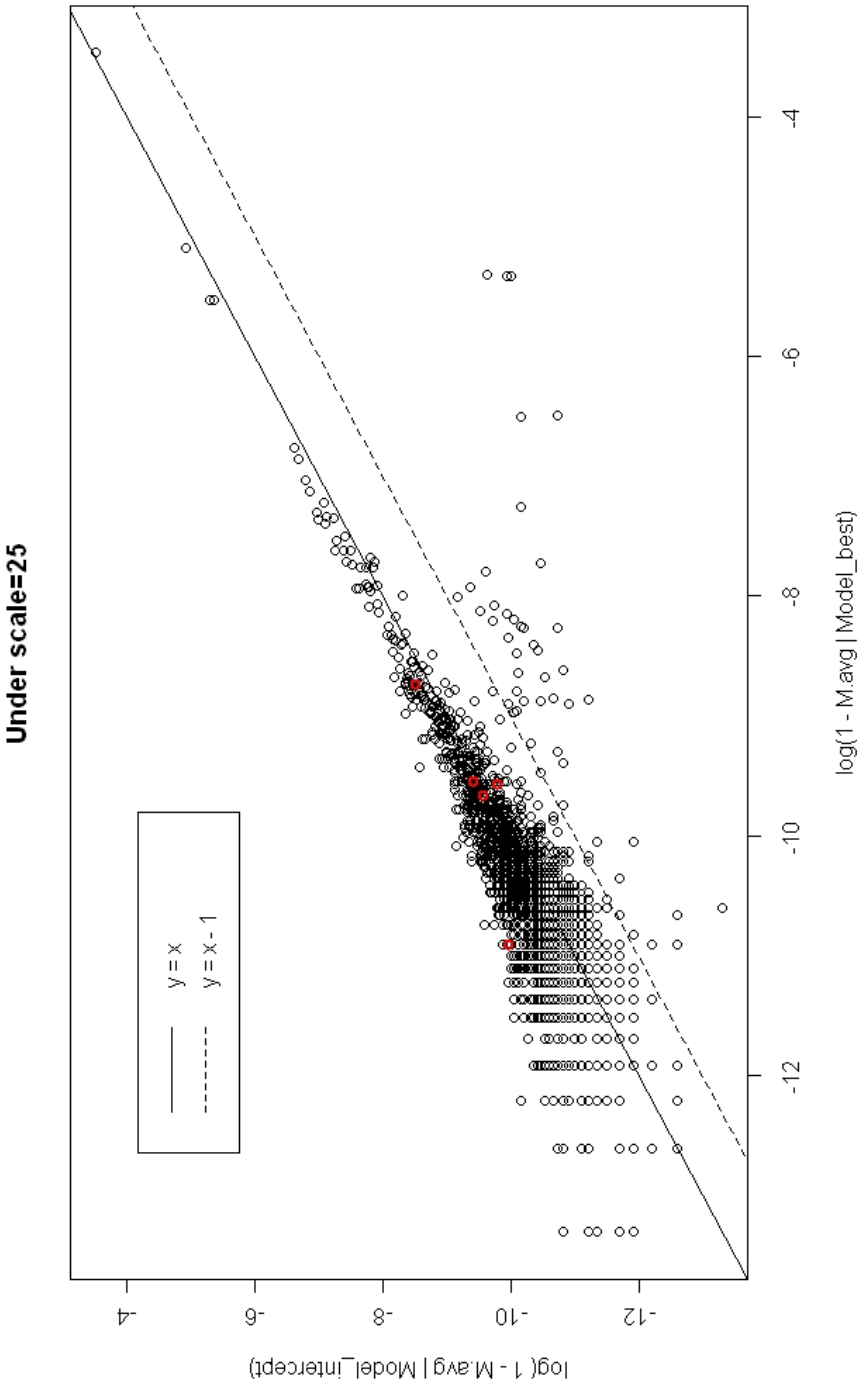

FIGURE 3.5: The change of the logarithm of averaged frequency of observed association for blocks covering the TP53 gene (colored in red) are compared with and without annotation data. A scale of 25 is applied to the prior variance for intercept and coefficients.

4

Conclusion

In this thesis, we have conducted a case-control analysis on SNPs in a database comprised of significant SNP-phenotype associations from collections of GWAS results [26]. From the result of the analysis, we identified four SNP functional variables that are more critical in predicting SNP associations. We also developed a Bayesian hierarchical model that can incorporate SNP-level functional annotation data into pathway-wide association studies. We implemented the model in R (version 2.8) and confirmed that the implementation was accurate using simulated data sets. We applied our model to an ovarian cancer data set and identified certain blocks of SNPs that may be candidates for association with the serous invasive ovarian cancer. These identified candidates may be missed without the functional annotation data.

Bibliography

[1] Jimenez-Sanchez, G., Childs, B., Valle, D. Human disease genes. *Nature*, 409(6822):853–5, 2001.

[2] Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonilla, E., Martin, J. B. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 306:234–238, Nov 1983.

[3] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Online mendelian inheritance in man, omim (tm). *World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/*, 2009.

[4] Altshuler, D., Daly, M.J., Lander, E.S. Genetic mapping in human disease. *Science*, 322(5903):881–8, 2008.

[5] Lander, E. S., Schork, N. J. Genetic dissection of complex traits. *Science*, 265(5181):2037–48, 1994.

[6] Cox, N.J.,Wapelhorst, B., Morrison, V.A., Johnson, L., Pinchuk, L., Spielman, R.S., Todd, J.A., Concannon, P. Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am J Hum Genet*, 69(4):820–30, 2001.

[7] Blangero, J. Localization, identification of human quantitative trait loci: king harvest has surely come. *Curr Opin Genet Dev*, 14(3):233–40, 2004.

[8] Levy, D., DeStefano, A. L., Larson, M. G., O'Donnell, C. J., Lifton, R. P., Gavras, H., Cupples, L. A., Myers, R. H. Evidence for a gene influencing blood pressure on chromosome 17. genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension*, 36(4):477–83, 2000.

[9] Sawcer, S. J. and Maranian, M. and Singlehurst, S. and Yeo, T. and Compston, A. and Daly, M. J. and De Jager, P. L. and Gabriel, S. and Hafler, D. A. and Ivinson, A. J. and Lander, E. S. et al. Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum Mol Genet*, 13(17):1943–9, 2004.

[10] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., Marks, D. S. Human microrna targets. *PLoS Biol*, 2(11):e363, 2004.

[11] Middleton, F. A., Pato, M. T., Gentile, K. L., Morley, C. P., Zhao, X., Eisener, A. F., Brown, A., Petryshen, T. L., Kirby, A. N., Medeiros, H., Carvalho, C., Macedo, A., Dourado, A. et al. Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (snp) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet*, 74(5):886–97, 2004.

[12] Evans, D.M., Cardon, L.R. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am. J. Hum. Genet.*, 75(4):687–92, 2004.

[13] Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H., Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet*, 69(5):936–50, 2001.

[14] Risch, N., Merikangas, K. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 1996.

[15] Collins, F.S., Guyer, M.S., Charkravarti, A. Variations on a theme: cataloging human dna sequence variation. *Science*, 278(5343):1580–1, 1997.

[16] Scott, L. J. and Mohlke, K. L. and Bonnycastle, L. L. and Willer, C. J. and Li, Y. and Duren, W. L. and Erdos, M. R. and Stringham, H. M. and Chines, P. S. and Jackson, A. U. and Prokunina-Olsson, L. et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–5, 2007.

[17] Sladek, R. and Rocheleau, G. and Rung, J. and Dina, C. and Shen, L. and Serre, D. and Boutin, P. and Vincent, D. and Belisle, A. and Hadjadj, S. and Balkau, B. and Heude, B. and Charpentier, G. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–5, 2007.

[18] Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S. Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos,

T., Bitton, A., Yang, H., Targan, S., Datta, L.W., Kistner, E.O., Schumm, L.P., Lee, A.T., Gregersen, P.K., Barmada, M.M., Rotter, J.I., Nicolae, D.L., Cho, J.H. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, 314(5804):1461–3, 2006.

[19] Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W., Shugart, Y. Y. et al. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet*, 39(5):596–604, 2007.

[20] Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D, Thompson, D., Ballinger, D.G. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–93, 2007.

[21] Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C. et al. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, 39(7):870–4, 2007.

[22] Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., Helgason, A., Rafnar, T., Bergthorsson, J. T., Agnarsson, B. A. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet*, 39(5):631–7, 2007.

[23] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[24] Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 2007.

[25] Servin, B. and Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114, 2007.

[26] Johnson, A. D., O'Donnell, C. J. An open access database of genome-wide association results. *BMC Med Genet*, 10:6, 2009.

[27] Hindorff, L. A. ,Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., Manolio, T. A. Potential etiologic and functional implications of

genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 2009.

[28] Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. Database of single nucleotide polymorphisms (dbsnp). *Available from: http://www.ncbi.nlm.nih.gov/SNP/*, (dbSNP Build ID: 130).

[29] Ramensky, V., Bork, P., Sunyaev, S. Human non-synonymous snps: server and survey. *Nucleic Acids Res*, 30(17):3894–900, 2002.

[30] Montgomery, S. B., Griffith, O. L., Sleumer, M. C., Bergman, C. M., Bilenky, M., Pleasance, E. D., Prychyna, Y., Zhang, X., Jones, S. J. Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22(5):637–40, 2006.

[31] Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E. , Donaldson, I. J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M. S., Wasserman, W., Hardison, R., Bergman, C. M., Jones, S. J. Oreganno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, 36(Database issue):D107–13, 2008.

[32] Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., Chiaromonte, F. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*, 14(4):700–7, 2004.

[33] King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., Hardison, R. C. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*, 15(8):1051–60, 2005.

[34] Griffiths-Jones, S., Saini, H. K., van Dongen, S., Enright, A. J. mirbase: tools for microrna genomics. *Nucleic Acids Res*, 36(Database issue):D154–8, 2008.

[35] Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B.,Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., van Steensel, B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–51, 2008.

[36] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., Lee, C. Detection of large-scale variation in the human genome. *Nat Genet*, 36(9):949–51, 2004.

[37] Mooney, S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform*, 6(1):44–56, 2005.

[38] Karchin, R. Next generation tools for the annotation of human snps. *Brief Bioinform*, 10(1):35–52, 2009.

[39] C. Fong, D. C. Ko, M. Wasnick, M. Radey, S. I. Miller, and M. Brittnacher. Gwas analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics*, 26(4):560–4, 2010.

[40] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D. The human genome browser at ucsc. *Genome Res*, 12(6):996–1006, 2002.

[41] Xu, Z., Taylor, J. A. Snpinfo: integrating gwas and candidate gene information into functional snp selection for genetic association studies. *Nucleic Acids Res*, 37(Web Server issue):W600–5, 2009.

[42] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K. et al.

[43] Kent, W. J., Haussler, D. Assembly of the working draft of the human genome with gigassembler. *Genome Res*, 11(9):1541–8, 2001.

[44] Terry Therneau and original R port by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.*, 2009. R package version 2.35-7.

[45] Fienberg S.E. Holland P.W. Bishop, Y.M.M. Discrete multivariate analysis: Theory and practice. 1975.

[46] P. Armitage. The use of the cross-ratio in aetiological survey. *Perspectives in Probability and Statistics*, pages 349–355, 1975.

[47] Madigan D. Raftery A.E. Volinsky C. Hoeting, J.A. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[48] George E.I. Clyde, M. Model uncertainty. *Statistical Science*, 19(1):81–94, 2001.

[49] Kass, R.E., Raftery, A.E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[50] Merlise Clyde. *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*, 2010. R package version 0.85.

[51] J. S. Witte, S. Greenland, R. W. Haile, and C. L. Bird. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, 5(6):612–21, 1994.

[52] R. J. Hung, P. Brennan, C. Malaveille, S. Porru, F. Donato, P. Boffetta, and J. S. Witte. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev*, 13(6):1013–21, 2004.

[53] R. J. Hung, M. Baragatti, D. Thomas, J. McKay, N. Szeszenia-Dabrowska, D. Zaridze, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, L. Foretova, V. Janout, V. Bencko, A. Chabrier, N. Moullan, F. Canzian, J. Hall, P. Boffetta, and P. Brennan. Inherited predisposition of lung cancer: a hierarchical modeling approach to dna repair and cell cycle control pathways. *Cancer Epidemiol Biomarkers Prev*, 16(12):2736–44, 2007.

[54] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[55] Richardson S. Spiegelhalter D.J. Gilks, W.R. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[56] Abecasis G.R. Li, Y. Mach 1.0: rapid haplotype reconstruction and missing genotype inference [abstract 2290/c]. *Am. J. Hum. Genet.*, 2006.

[57] J. H. J. Hoeijmakers. Genome maintenance mechanisms for preventing cancer. *Nature*, 411(6835):366–374, 2001.

[58] M. B. Kastan and J. Bartek. Cell-cycle checkpoints and cancer. *Nature*, 432(7015):316–323, 2004.

[59] The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–96, 2003.

[60] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, and et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

[61] L. S. Friedman, E. A. Ostermeyer, C. I. Szabo, P. Dowd, E. D. Lynch, S. E. Rowell, and M. C. King. Confirmation of brca1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat Genet*, 8(4):399–404, 1994.

[62] J. Simard, P. Tonin, F. Durocher, K. Morgan, J. Rommens, S. Gingras, C. Samson, J. F. Leblanc, C. Belanger, F. Dion, and et al. Common origins of brca1 mutations in canadian breast and ovarian cancer families. *Nat Genet*, 8(4):392–8, 1994.

[63] N. C. Nicolaides, N. Papadopoulos, B. Liu, Y. F. Wei, K. C. Carter, S. M. Ruben, C. A. Rosen, W. A. Haseltine, R. D. Fleischmann, C. M. Fraser, and et al. Mutations of two pms homologues in hereditary nonpolyposis colon cancer. *Nature*, 371(6492):75–80, 1994.

[64] Z. Q. Yuan, B. Gottlieb, L. K. Beitel, N. Wong, P. H. Gordon, Q. Wang, A. Puisieux, W. D. Foulkes, and M. Trifiro. Polymorphisms and hnpcc: Pms2-mlh1 protein interactions diminished by single nucleotide polymorphisms. *Hum Mutat*, 19(2):108–13, 2002.

[65] Y. M. Hendriks, S. Jagmohan-Changur, H. M. van der Klift, H. Morreau, M. van Puijenbroek, C. Tops, T. van Os, A. Wagner, M. G. Ausems, E. Gomez, M. H. Breuning, A. H. Brocker-Vriends, H. F. Vasen, and J. T. Wijnen. Heterozygous mutations in pms2 cause hereditary nonpolyposis colorectal carcinoma (lynch syndrome). *Gastroenterology*, 130(2):312–22, 2006.

[66] M. De Vos, B. E. Hayward, S. Picton, E. Sheridan, and D. T. Bonthron. Novel pms2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome. *Am J Hum Genet*, 74(5):954–64, 2004.

[67] J. Auclair, D. Leroux, F. Desseigne, C. Lasset, J. C. Saurin, M. O. Joly, S. Pinson, X. L. Xu, G. Montmain, E. Ruano, C. Navarro, A. Puisieux, and Q. Wang. Novel biallelic mutations in msh6 and pms2 genes: gene conversion as a likely cause of pms2 gene inactivation. *Hum Mutat*, 28(11):1084–90, 2007.

[68] M. Kato, K. Yano, F. Matsuo, H. Saito, T. Katagiri, H. Kurumizaka, M. Yoshimoto, F. Kasumi, F. Akiyama, G. Sakamoto, H. Nagawa, Y. Nakamura, and Y. Miki. Identification of rad51 alteration in patients with bilateral breast cancer. *J Hum Genet*, 45(3):133–7, 2000.

[69] E. Levy-Lahad, A. Lahad, S. Eisenberg, E. Dagan, T. Paperna, L. Kasinetz, R. Catane, B. Kaufman, U. Beller, P. Renbaum, and R. Gershoni-Baruch. A

single nucleotide polymorphism in the rad51 gene modifies cancer risk in brca2 but not brca1 carriers. *Proc Natl Acad Sci U S A*, 98(6):3232–6, 2001.

[70] A. C. Antoniou, O. M. Sinilnikova, J. Simard, M. Leone, M. Dumont, S. L. Neuhausen, J. P. Struewing, D. Stoppa-Lyonnet, L. Barjhoux, D. J. Hughes, I. Coupier, M. Belotti, C. Lasset, V. Bonadona, Y. J. Bignon, T. R. Rebbeck, T. Wagner, H. T. Lynch, S. M. Domchek, K. L. Nathanson, J. E. Garber, J. Weitzel, S. A. Narod, G. Tomlinson, O. I. Olopade, A. Godwin, C. Isaacs, A. Jakubowska, J. Lubinski, J. Gronwald, B. Gorski, T. Byrski, T. Huzarski, S. Peock, M. Cook, C. Baynes, A. Murray, M. Rogers, P. A. Daly, H. Dorkins, R. K. Schmutzler, B. Versmold, C. Engel, A. Meindl, N. Arnold, D. Niederacher, H. Deissler, A. B. Spurdle, X. Chen, N. Waddell, N. Cloonan, T. Kirchhoff, K. Offit, E. Friedman, B. Kaufmann, Y. Laitman, G. Galore, G. Rennert, F. Lejbkowicz, L. Raskin, I. L. Andrulis, E. Ilyushik, H. Ozcelik, P. Devilee, M. P. Vreeswijk, M. H. Greene, S. A. Prindiville, A. Osorio, J. Benitez, M. Zikan, C. I. Szabo, O. Kilpivaara, H. Nevanlinna, U. Hamann, F. Durocher, A. Arason, F. J. Couch, D. F. Easton, and G. Chenevix-Trench. Rad51 135g–¿c modifies breast cancer risk among brca2 mutation carriers: results from a combined analysis of 19 studies. *Am J Hum Genet*, 81(6):1186–200, 2007.

[71] F. Toledo and G. M. Wahl. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer*, 6(12):909–23, 2006.

[72] K. H. Vousden and D. P. Lane. p53 in health and disease. *Nat Rev Mol Cell Biol*, 8(4):275–83, 2007.

[73] Goode E.L. Clyde M.A. Iversen Jr E.S. Moorman P. Berchuck A. Marks J. et al. Schildkraut, J.M. Single nucleotide polymorphisms in tp53 and susceptibility to invasive epithelial ovarian cancer. *Cancer Research*, 69(6):2349–2357, 2009.