A Simulation Study to Compare Gene Set Analysis Methods

Andrew James Pfeiffer

Thesis submitted for the degree of

Master of Philosophy

in

Statistics

at

The University of Adelaide

(School of Mathematical Sciences)

May 19, 2016

Contents

List of Tables

List of Figures

List of Acronyms

ALIGATOR Association LIst Go AnnoTatOR.

AUC area under the ROC curve.

BH Benjamini-Hochberg.

bp base pairs.

BY Benjamini-Yekutieli.

CATT Cochran-Armitage Trend Test.

DAVID the Database for Annotation, Visualization and Integrated Discovery.

dbSNP the Database of Short Genetic Variations.

DNA deoxyribonucleic acid.

DSL disease susceptibility locus.

EASE Expression Analysis Systematic Explorer.

ES enrichment score.

EVA Exploratory Visual Analysis.

FDR false discovery rate.

FET Fisher's exact test.

FPR false positive rate.

FWER family-wise error rate.

GLM general linear model.

GO gene ontology.

GSA gene set analysis.

GSE gene set enrichment.

GSEA Gene Set Enrichment Analysis.

GSEAPR GSEAPreranked.

GWA study genome-wide association study.

HapMap CEU Data genomic data collected from CEPH (Utah residents with ancestry from northern and western Europe) (abbreviation: CEU) as part of the International HapMap Project.

HWE Hardy-Weinberg Equilibrium.

IPA QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, `www.quiagen.com/ingenuity`).

KEGG Kyoto Encyclopaedia of Genes and Genomes.

LD linkage disequilibrium.

LE linkage equilibrium.

LS Li and Stephens.

MGSEA Modified GSEA.

MPEVA MinP-EVA.

NCBI the National Center for Biotechnology Information.

NES normalised enrichment score.

OR overrepresentation.

PANTHER Protein ANalysis THrough Evolutionary Relationships.

PARIS Pathway Analysis by Randomization Incorporating Structure.

RNA ribonucleic acid.

ROC receiver operating characteristic.

rs reference SNP.

SNP single nucleotide polymorphism.

the SRT the SNP Ratio Test.

TPR true positive rate.

Abstract

Genome-wide association studies (GWA studies) identify alleles that are associated with a disease. These allele variations are called single nucleotide polymorphisms (SNPs). However, GWA studies do not account for interaction between SNPs. Gene set analysis (GSA) is used in GWA studies to account for interaction. GSA methods map SNPs to gene sets and identify gene sets that are associated with a disease. Comprehensive reviews of GSA exist in the literature. However, these reviews do not compare specific methods or implement them on data.

In this thesis, we compare six GSA methods. We use seven factors highlighted by the reviews as important in GSA to compare these methods. For example, we analyse how each method accounts for parameters that could affect the analysis. These parameters include gene size and SNP interaction. We consider the null hypothesis tested by each method. We also analyse the sensitivity of methods to individual SNPs with small p-values. In contrast, the marginal effect of many SNPs that cause diseases is often small. The p-values of such SNPs need not be small.

We conduct a simulation study to compare four GSA methods. We investigate the sensitivity of these methods to SNPs with very small p-values. We use Manhattan plots to display gene sets that were assigned disparate p-values by different methods. We also use receiver operating characteristic curves to compare the performance of each method. Finally, we recommend a method that gave excellent performance.

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....

To my Father God, my Lord Jesus Christ, and the Holy Spirit. Thank you for foreknowing me, predestining me to be conformed to the image of Jesus, calling me, justifying me and glorifying me (Romans 8:29,30).

Acknowledgements

I begin by thanking my supervisors, Associate Professor Gary Glonek and Dr Jonathan Tuke. Gary and Jonathan have patiently taught me so much over the last two years. Of course, they have taught me so much about statistics. Yet even more importantly, they have also taught me many skills that I will continue to use, even when I am no longer using statistics in my career. Soft skills such as writing well, managing time well, being organised and communicating information clearly to others are not skills that students are taught explicitly. Yet they are skills that need to be learned to make the most of the short time that each of us has on this earth. I thank Gary and Jonathan for helping me to develop in these skills. Whether it be through taking notes and emailing them out after every meeting with a list of action points, learning how to use tools such as Evernote, OmniFocus and Sublime Text or growing in my understanding of programming languages such as LaTeX, Python, R and shell scripting, I have learned so much from both of you.

I especially thank Jonathan for rekindling my love for statistics. After completing Year 12 Mathematical Studies in 2006, I told myself that I would never study statistics again. Nonetheless, I remember going to lectures for Probability and Statistics II in 2011, feeling Jonathan's excitement for statistics, and being encouraged by the genuine care he had for students. He even gave students free lollipops when they came to his office to ask him a question. If it weren't for Jonathan, I would not be finishing a Master of Philosophy in statistics.

To my friends in the postgraduate room, thank you. The life of a postgraduate student is not always easy, but it has been both comforting and encouraging to share parts of this crazy ride with you.

To everyone at Adelaide University Evangelical Students (ES), thank you so much. ES is a Christian group on campus, passionate about sharing the good news of Jesus Christ with everyone, and God has used it to completely change my life while I have been at university. I recently attended the Leavers' Day and Dinner; we were given the opportunity to share something that we wanted to praise God for about ES, and some advice we'd like to offer continuing members of ES. And like many over-analytic perfectionists, I kept wondering if I said the right things. I posted the following on Facebook the morning after the dinner, and I repeat it here.

> First, I praise God for the staffworkers: Geoff, Reuben, Laura, Emily, Matthew, Mark, Oliver and Dave. As I am beginning to understand, vocational ministry involves many, many sacrifices, including a two-year apprenticeship (for many), three or more years at Bible college, and foregoing a secular career that would have probably paid considerably better. And yet, you've chosen the path of vocational ministry: you've chosen to share your lives with us, mentor us, teach us the Bible, help us to read it for ourselves, help us to teach it to other people, help us read it with our not-yet-Christian friends, train us in evangelism, and train us in leadership in other ways, whether it be by leading teams or serving on committees. I am so immensely thankful.
>
> For indeed, God created us to share the good news of Jesus Christ with other people wherever we go. As a uni student, one of my primary tasks has been to declare the praises of God and what he has done on the university campus - it's been my mission field. And the staffworkers have helped to train us up

to be missionaries on the university campus. But the awesome thing is, the way that they have trained us in Bible reading, evangelism, and serving the kingdom in other ways is not just limited to the university campus. We'll be able to use everything that they have taught us wherever we go in life: whether the workplace, vocational ministry, cross-cultural mission, and so on. In this sense, among many others, God has used the staffworkers, through ES, to change my life completely.

So how could I not want to get involved with such a brilliant group of people, whose vision is "reaching every uni student on North Terrace with the good news of Jesus Christ to present everyone mature in Him." This vision is the vision all of God's people need to have, just applied to the context of university. Hallelujah!! I'd also like to praise God for the staffworkers because they see people who are keen to be involved with ES, and give them opportunities to serve! I'm so grateful for the opportunity to lead Bible study groups for the last three years, lead the Bible talks team last year, and be the treasurer of ES this year. Bible study leading has stretched me, from the time when I nervously led my first BSG in March 2013 (just ask Tom Worley how nervous I was). It has given me an awesome opportunity to grow in my understanding of the Bible, a passion to teach it to other people, and a desire to mentor younger brothers in Christ in the faith. Leading the Bible Talks team and being Treasurer have been awesome opportunities to develop my administration and behind-the-scenes ministry skills, and to give back to ES in an awesome way. To those of you who, like me, are rather freaked out by things like walk-up evangelism, please don't feel like an inferior Christian. Of course, give walk-up a shot, because God can use even the most nervous Christian to bring people into the

kingdom. But know that serving in ministry behind the scenes is such a necessary part of ministry, to facilitate reaching the university campus with the gospel message.

So to the staffworkers, thank you for sharing your lives with us, for helping us to grow in our relationship with God in such a life-changing way, and for giving us opportunities to serve in ES.

Next, I praise God for my fellow Evangelical Students. (I don't understand why we are called ESers, because it doesn't make sense when the acronym is expanded, or ES students, because it's like "ATM machine" or "PIN number"). That's beside the point, of course. To my fellow ES members: thank you for being such welcoming and encouraging brothers and sisters in Christ. As someone who was bullied in primary school, high school, and the early parts of university, I cannot thank my fellow ES members enough for welcoming me and accepting me for who I am. It has been an honour, privilege and joy to be fellow soldiers for Christ with you all on campus. Whether it be sharing the highs and lows of evangelism, having deep theological or personal conversations (often on camps, but also on the university campus), or just chilling out and doing nothing with you, thank you. Thank you for rebuking me and pulling me up where necessary, for speaking the truth to me in love, and for being brilliant brothers and sisters in Christ. I have to say that I met most of my best friends at ES, and even though I am going into the workforce next year, God willing, I truly hope to stay in touch with you.

In particular, I'd like to thank Tom, Chloe and Hendré for serving on Exec with me this year. It's been a privilege to serve ES with you in this way. Tom and I were chatting last night

about how we worked so well as a team, and how everyone played their part and did what they needed to do. To see you all work so hard in ES on a volunteer basis is a huge testament to your passion for seeing God work on campus. Also, thanks to Chloe, Claire, Hayden and Lachlan, for being willing to serve on Exec next year. I know that under God, you will serve ES faithfully in your respective roles. To Irene, as last year's Treasurer, thank you for teaching me the ropes of treasury this year.

To Hayden, Daniel and George, thank you for serving with me on the Bible talks team last year. It was the first time ES ran the Bible talks team, and you all did what needed to be done exceptionally well. And thanks to Hayden for running the team so well this year - the fact that your work went unnoticed on the whole is a testament to how smoothly you ran everything. To my fellow Bible Study group leaders: Tom, Stephen, Sowmya and Renee (I apologise if I forgot anyone) - thank you for imparting your leadership wisdom as I was learning how to lead BSGs, and for allowing me to do the same. To the people that led O'Weeks, camps and Jesus Weeks over the last few years: Jane, Samantha, David, Jack, Shane, Ellen, Tom, Lachlan, Warwick and Timothy - thank you for all the hard work you put in to make these brilliant events happen for all of us. To the Exec committees over the last few years: Nicholas, Michael, Sasha, Kathryn, Ada, Sarah, Brendan, Liam, Irene, Madeleine and Ryan, thank you for your hard work, too. To the Faculty Leaders this year: Ryan, Naomi, Jack, Lydia, Liam and Bonnie, thank you for the work you put in this year to get the new Faculty system up and running. I know that it has been worth it, and it will continue to reap fruit. And to next year's faculty leaders: Hendre, Sowmya, Oli, Corinne and Warwick, I know that you will do everything to make new

people feel welcome at ES in your respective faculties. To Claire, Maggie, Declan and Phil: thank you for running the finances for the various events this year, and making my life as treasurer so much easier. To John, Lisa, Daniel, Abe and everyone else who came to the maths prayer group last year - thank you for joining me. It was a privilege to spend time with you on Thursday mornings in prayer to our awesome God. I still remember our awesome breakfast catchup at Aroma café at the end of last year.

Finally, to all my friends at ES. Thank you. Thank you for being some of the best friends I have ever had. ES has completely changed my life, and you have been a huge part of that, so thank you.

If you've read this far, then thank you. For those of you who are commencing university studies next year, or continuing members of ES, I encourage you to *get involved*. It's true, everyone says it. But the fact is, the period of time you have at university is brief, compared to the rest of your life. (Well, maybe not so brief for me). I doubt you will ever get another opportunity to be stretched and grown in your faith like you will at ES, or opportunities to learn how to evangelise. So try everything. If you're asked to be on a team, or lead a team, or do something for ES, say yes. Unless God tells you to say no, of course. Try doing walk-up evangelism, even if it freaks you out. Try helping to run a camp, or an event, or anything. And if you'd like to be involved more, just chat to a staffworker. They'd love to catch up with you for coffee and chat to you about how you can be more involved with ES. And even if you don't think you have time, *make time*. Like I said, it's a brief period of your life. Keep serving at your local church, of course, because that should be

the number one place you serve. But beyond that, serve as much as you possibly can with ES. You will *not* regret it.

One of the reasons that I love ES so much is that it is such a strategic ministry. Thousands and thousands of (mostly) young people are going onto campus to learn about various subjects. And it is during young adulthood when people are thinking for themselves about what it is they are going to believe, including strong Christians, nominal Christians who just go to church each Sunday because their parents do, and not-yet-Christians. The ministry that ES do can make the most of this melting pot of worldviews at university by presenting the gospel message. And if you don't know what the gospel message (the good news) is, then keep reading. Soon after God created mankind, we rebelled against him. We deserve God's judgment. But in his infinite mercy, God sent his son, Jesus Christ, into the world, to take the punishment for people's sins, so that anyone who believes in him can be in right relationship with God for all eternity. If you don't believe this good news yet, but you'd like to find out more, please just ask me.

Also, to those of you who think that ES is a bunch of weirdos, or are put off because you think we're pushy about the gospel message, please reconsider your scepticism. We just want everyone to know the good news about Jesus Christ, and we don't want anyone to fall under God's judgment. Forgive us for the times when we have been unloving in our portrayal of the gospel message, but please understand that we only share the gospel message because we love you.

Finally, to those of you who resonate with what I've said, and who are passionate about the ministry of ES, one more piece

> of advice. Seriously consider serving ES in full-time vocational ministry. You know how God has used the staffworkers in ES to change our lives, and it would be an amazing privilege to be used by God to change the lives of the next generation of university students. If you'd like to consider this challenge more, chat to one of the staffworkers about it, chat to me about it, or come along to CV Conference in 2016. Ultimately, nothing is more important than sharing the gospel message with everyone that does not know Jesus, and discipling those that do know Jesus, so that they can disciple others.
>
> In one sense, I'm sad that my time at ES has drawn to a close. Nonetheless, I know that now is the right time for this season, amazing as it has been, to come to an end. For the next few years, I will be working with the ATO in Adelaide. But soon, God willing, I will come back to serve ES as a ministry apprentice. And then, after getting some formal theological education and training at Bible College, I will come back to serve ES full-time, for as long as I am physically able. And I can't wait. Bring. It. On. Praise God for everything.

To my parents, thank you for so much. Thank you, first and foremost, for raising me to know my Lord, Jesus Christ. Thank you for supporting me throughout my university studies. Thank you for providing so much for me. Thank you for loving me so much. Thank you also to my mother for proofreading my thesis and finding approximately 15 typographical errors.

Finally, to my Father God, my Lord Jesus Christ, and the Holy Spirit. Thank you for foreknowing me, predestining me to be conformed to the image of Jesus, calling me, justifying me and glorifying me (Romans 8:29,30). Only by your grace do I exist; only by your grace do I have the skills that I needed to complete this thesis, and only by your grace did Jesus die for me, so that I can be in right

relationship with you. Please help me to keep giving my life to you, for your glory. Amen.

Chapter 1

Introduction

1.1 Background

A genome-wide association study (GWA study) is a commonly used study in the field of statistical genetics. The aim of a GWA study is to identify allele variations known as single nucleotide polymorphisms (SNPs) that are associated with a given disease. This information can be used to elucidate how genetic variation causes the disease, which can help improve treatments for the disease. GWA studies have identified SNPs that are associated with many diseases, such as type 1 diabetes (Polychronakos and Li, 2011), Crohn's disease (Franke *et al.*, 2010) and multiple sclerosis (International Multiple Sclerosis Genetics Consortium and Wellcome Trust Case Control Consortium 2, 2011). However, GWA studies have yielded many results that have not been replicated in independent studies (Laird and Lange, 2011). Furthermore, they have only identified a small proportion of the genetic variation that is associated with most diseases (Maher, 2008). Two primary factors that contribute to these problems are *small effect sizes* and *epistasis* (Hong *et al.*, 2009; Mooney *et al.*, 2014; O'Dushlaine *et al.*, 2009; Wang *et al.*, 2010; Wang *et al.*, 2011; Yaspan *et al.*, 2011).

SNPs have a *small effect size* if they have a real but small effect on the probability of disease in an individual. However, GWA studies often fail to identify SNPs with small effect sizes due to the *multiple testing problem*. In a typical GWA study, many SNPs are simultaneously tested for association with a given disease. Consequently, using a common significance level such as $\alpha = 0.05$ for each SNP can result in many false positive findings. Reducing the significance level is a common solution to this problem, but this can reduce the power of GWA studies to detect such SNPs.

The other factor that impacts GWA studies is epistasis. Epistasis occurs when the combined effect of two genetic factors (such as SNPs) is not additive. Historically, GWA studies have only considered SNPs individually, so such studies could not account for epistasis. Epistasis occurs because genetic factors can interact with each other in a complex way. Consequently, a number of novel methods have been developed that analyse gene sets for association with a given disease. Fridley and Biernacka (2011), Mooney *et al.* (2014), and Wang *et al.* (2011) refer to these methods collectively as gene set analysis (GSA) methods.

There are many GSA methods in the literature, and they are diverse. For example, GSA methods may differ in null hypothesis that they test, the way that they map SNPs to genes, the way that they calculate test statistics of gene sets and so on. Furthermore, GSA methods should account for factors that can affect the analysis, such as gene size and *linkage disequilibrium (LD)*, which describes the dependency structure between SNPs on the same chromosome. Consequently, a number of reviews of GSA methods exist in the literature, such as Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011). However, none of these reviews compare specific GSA methods in detail or test their conclusions by implementing GSA methods on simulated data.

In this thesis, we review six GSA methods by Askland *et al.* (2009), Holmans *et al.* (2009), Hong *et al.* (2009), O'Dushlaine *et al.* (2009), Wang *et al.* (2007),

and Yaspan *et al.* (2011). We compare the properties of each method, including whether or not each method accounts for LD and gene size. We also implement four of these methods on simulated genetic data to test our comparisons. We compare the performance of each method at identifying gene sets that are associated with a disease.

1.2 Outline

We detail the necessary background for GWA studies in Chapter 2. In Section 2.1, we introduce the necessary biological background, including the definition of a GWA study. In Section 2.2, we detail the statistical background that is necessary to understand GWA studies. We review categorical data analysis in Subsection 2.2.1, we apply the principles of categorical data analysis to GWA study in Subsection 2.2.2, and we review multiple hypothesis testing in Subsection 2.2.3. Finally, in Section 2.3, we discuss the dependency structure between SNPs on the same chromosome.

In Chapter 3, we review the GSA methods by Askland *et al.* (2009), Holmans *et al.* (2009), Hong *et al.* (2009), O'Dushlaine *et al.* (2009), Wang *et al.* (2007), and Yaspan *et al.* (2011). We provide a more detailed critique of traditional GWA studies in Section 3.1. In Section 3.2, we discuss the ways that the GSA methods map SNPs to gene sets. We then describe the six GSA methods in detail in Section 3.3. In Section 3.4, we discuss seven important properties of GSA methods as detailed by Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011), and we use each property to compare the six GSA methods.

In Chapter 4, we detail the methods that we used to compare the performance of GSA methods on genetic data. In section 4.1, we discuss the advantages of using simulated genetic data to compare GSA methods. We also discuss three methods of simulating genetic data: the LS model (Li and Stephens, 2003), HAPGEN

(Spencer *et al.*, 2009) and HAPGEN2 (Su *et al.*, 2011). In Section 4.2 we discuss our implementation of HAPGEN2 – the method that we used to simulate genetic data – and four GSA methods. In particular, we justify the choices that we made for the values of the parameters that we could vary in our simulation study. Finally, we detail the procedures that we used to compare the performance of the GSA methods in Section 4.3.

We display and discuss our results in Chapter 5. In Section 5.1, we analyse the effect of each method parameter on the performance of the method. In Section 5.2, we compare the performance of the four GSA methods at identifying gene sets associated with a disease. Finally, in Section 5.3, we investigate the sensitivity of GSA methods to SNPs with very small p-values.

Chapter 2

Genome-Wide Association Studies

In this chapter, we summarise the well-established theory of *genome-wide association studies (GWA studies)*. We give a brief overview of GWA studies in Section 2.1. We then discuss the statistical theory necessary to perform GWA studies in Section 2.2, including a discussion of categorical data analysis and multiple hypothesis testing. Finally, we give a mathematical formulation of the dependency structure in genomes in Section 2.3.

2.1 Biological Background

A detailed biological background can be found in literature such as Foulkes (2009), Gonick (1991), and Laird and Lange (2011). We summarise the important points here.

Three classes of macromolecules are *deoxyribonucleic acid (DNA)*, *ribonucleic acid (RNA)* and *protein*. DNA encodes the information to form proteins in its sequence of nucleotides (A, C, G and T). According to the *central dogma of molecular biology*, cells create proteins by *transcribing* the appropriate sequence of DNA to messenger RNA, which is then *translated* to the amino acid sequence of the protein. A *gene* is a sequence of nucleotides that codes for a particular protein.

In many living organisms, DNA is partitioned into long strands, called *chromosomes*. In *diploid* organisms, most cells contain pairs of chromosomes. For example, most cells in humans contain 23 pairs of chromosomes. In contrast, *haploid* organisms only have one copy of each chromosome in all of their cells. In this thesis, we assume that all organisms are diploid. When DNA is copied or *replicated*, mistakes can occur, which are often referred to as *mutations*. A *single nucleotide polymorphism (SNP)* is a single nucleotide position with more than one possible nucleotide. These different nucleotides are called *alleles*. A *genetic locus* is a particular location in the chromosome that has more than one possible allele. In this thesis, we assume that there are two possible alleles at a genetic locus. The *genotype* of an individual at a genetic locus is the pair of alleles present at that genetic locus. We say that an individual is *homozygous* (*heterozygous*) at a genetic locus if the two alleles at that locus are the same (different).

A variation in an organism's DNA might change the structure of an important protein produced by the DNA, and this change could increase the probability of a given disease. For example, in humans, sickle-cell anaemia is caused by a mutation on Chromosome 11 in the gene that codes for haemoglobin, a protein that transports oxygen around the body through the bloodstream. A genetic locus where a mutation affects the function of a protein such that the probability of disease is increased is referred to as a *disease susceptibility locus (DSL)*. Consequently, ongoing research exists that is seeking to find DSLs for various diseases. One of the aims of such research is to better understand how mutations cause various diseases, so that we can develop better treatments for them.

A GWA study is a study designed to locate SNPs on the genome associated with a given disease. However, not all SNPs on the genome are tested in a GWA study, because there is a complex dependency structure that exists between nucleotides on the genome. For example, there are 3×10^9 base pairs on the human genome. Foulkes (2009) notes that genotyping platforms used in GWA studies such as the Affymetrix and Illumina chips, which can genotype 5×10^5 to 10^6

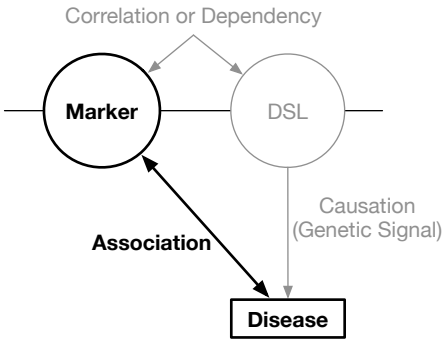

Figure 2.1: Relationship between DSL, marker and disease

SNPs simultaneously, can sufficiently characterise genetic variation in humans. Consequently, the SNPs that are found to be associated with a given disease are not necessarily the DSLs themselves. Such SNPs are correlated with a DSL, and they are referred to as *marker SNPs*. When we test SNPs for association with a given disease, we say that we are trying to detect the *genetic signal* from the DSL. However, we cannot observe the DSL, the genetic signal, or the correlation between the DSL and the marker directly. We can only infer the genetic signal from the association between the marker and the disease. We summarise this information in Figure 2.1.

To perform a GWA study, a fixed number of controls (subjects without the disease) and cases (subjects with a given disease) are genotyped at each genetic locus using a genotyping platform such as the Affymetrix or Illumina chips. Each locus is then tested for association with the disease. In Section 2.2 we discuss methods of testing the association between a single SNP and a given disease. We also discuss the multiple testing problem in the context of GWA studies, because many SNPs are tested in a GWA study simultaneously. Then, in Section 2.3, we give a mathematical formulation of the dependency structure between nucleotides in the genome.

2.2 Statistical Background

In this thesis, we assume that the reader has a background in statistical inference, such as that found in Casella and Berger (2002). In particular, we assume that the reader has a good grasp of the common discrete and continuous probability distributions (Chapters 1 to 4), hypothesis testing (Chapter 8) and linear regression (Chapter 11).

2.2.1 Categorical Data Analysis

Notation

In this subsection, we give an overview of the techniques from categorical data analysis that are used in GWA studies to test for association between a single genetic locus and a disease. The material is based on Agresti (2013) and Zheng *et al.* (2012).

Suppose X and Y are two categorical random variables, such that X can take values in I categories and Y can take values in J categories, where I and J are positive integers. We write $X = x$ and $Y = y$ for $x \in \mathbb{Z}_I = \{0, 1, \ldots, I-1\}$ and $y \in \mathbb{Z}_J = \{0, 1, \ldots, J-1\}$ respectively.

Now, suppose that a data set is the result of many realisations of the joint distribution of X and Y. We can then display this data set in a table with I rows and J columns, where each cell (x, y) of the table contains the number of realisations $m_{x,y}$ of the joint event that $X = x$ and $Y = y$. Then the marginal frequencies are

$$m_{x,+} = \sum_{y \in \mathbb{Z}_J} m_{x,y}$$

and

$$m_{+,y} = \sum_{x \in \mathbb{Z}_I} m_{x,y},$$

	Y			
X	0	$\dots$	$J-1$	Total
0	$\pi_{0,0}$	$\dots$	$\pi_{0,J-1}$	$\pi_{0,+}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$I-1$	$\pi_{I-1,0}$	$\dots$	$\pi_{I-1,J-1}$	$\pi_{I-1,+}$
Total	$\pi_{+,0}$	$\dots$	$\pi_{+,J-1}$	1

Table 2.1: Notation for the joint and marginal distributions of categorical variables

and the total sample size is

$$m_{+,+} = \sum_{x \in \mathbb{Z}_I} \sum_{y \in \mathbb{Z}_J} m_{x,y}.$$

We call such a table an $I \times J$ *contingency table*.

We introduce notation for the various probability distributions relating X and Y. Let

$$\pi_{x,y} = \Pr(X = x, Y = y),$$
$$\pi_{x,+} = \Pr(X = x) \text{ and}$$
$$\pi_{+,y} = \Pr(Y = y).$$

Thus

$$\sum_{x \in \mathbb{Z}_I} \pi_{x,y} = \pi_{+,y},$$
$$\sum_{y \in \mathbb{Z}_J} \pi_{x,y} = \pi_{x,+} \text{ and}$$
$$\sum_{x \in \mathbb{Z}_I} \sum_{y \in \mathbb{Z}_J} \pi_{x,y} = 1.$$

We present this notation in Table 2.1, which is based on Table 2.2 of Agresti (2013).

We also consider conditional distributions of X and Y. Write

$$\pi_{y|x} = \Pr(Y = y | X = x); \text{ and}$$
$$\pi_{x|y} = \Pr(X = x | Y = y).$$

Thus, from the definition of conditional probability,

$$\pi_{y|x} = \frac{\pi_{x,y}}{\pi_{x,+}} \text{ and}$$
$$\pi_{x|y} = \frac{\pi_{x,y}}{\pi_{+,y}}.$$

It should also be clear that for all x and y,

$$\sum_{y \in \mathbb{Z}_J} \pi_{y|x} = 1 \text{ and}$$
$$\sum_{x \in \mathbb{Z}_I} \pi_{x|y} = 1.$$

The variables X and Y are said to be *independent* if

$$\pi_{x,y} = \pi_{x,+}\pi_{+,y}$$

for all x and y. However, the true distributions $\pi_{x,y}, \pi_{x,+}$ and $\pi_{+,y}$ are typically unknown. Consequently, some methods of testing the hypothesis that X and Y are independent use the sample proportions, which we denote by replacing π with $\hat{\pi}$. We use the frequencies in the contingency table to define the sample proportions. For example,

$$\hat{\pi}_{x,y} = \frac{m_{x,y}}{m_{+,+}},$$

and the marginal proportions of X and Y respectively are

$$\hat{\pi}_{x,+} = \frac{m_{x,+}}{m_{+,+}}$$

and

$$\hat{\pi}_{+,y} = \frac{m_{+,y}}{m_{+,+}}.$$

Also, the conditional distribution of Y given X can be estimated by

$$\hat{\pi}_{y|x} = \frac{\hat{\pi}_{x,y}}{\hat{\pi}_{x,+}} = \frac{m_{x,y}}{m_{x,+}},$$

and the conditional distribution of X given Y can be estimated by

$$\hat{\pi}_{x|y} = \frac{\hat{\pi}_{x,y}}{\hat{\pi}_{+,y}} = \frac{m_{x,y}}{m_{+,y}}.$$

In what follows, we assume that $m_{x,y}$ is a realisation of the random variable $M_{x,y}$. Let

$$\begin{aligned}\boldsymbol{m}_x &= (m_{x,0}, \ldots, m_{x,J-1}),\\ \boldsymbol{m}_y &= (m_{0,y}, \ldots, m_{I-1,y}) \text{ and}\\ \boldsymbol{m} &= (m_{0,0}, \ldots, m_{I-1,J-1}).\end{aligned}$$

with analogous notation for the corresponding random variables and probabilities. We now discuss various distributions of the entries of contingency tables.

In contingency tables where only the grand total $m_{+,+}$ is fixed, $\boldsymbol{M}$ can be modelled with a multinomial distribution with parameters $m_{+,+}$ and $\boldsymbol{\pi}$, and probability mass function

$$\Pr(\boldsymbol{M} = \boldsymbol{m}) = m_{+,+}! \prod_{x \in \mathbb{Z}_I} \prod_{y \in \mathbb{Z}_J} \frac{\pi_{x,y}^{m_{x,y}}}{m_{x,y}!},$$

where $m_{x,y} \geq 0$ for all $x \in \mathbb{Z}_I, y \in \mathbb{Z}_J$. A contingency table under these conditions is called a *multinomial sample*.

In contingency tables where the row totals $m_{x,+}$ are fixed, $\boldsymbol{M}$ can be modelled with a product multinomial distribution with parameters $m_{x,+}$ and $\pi_{y|x}$, where $x \in \mathbb{Z}_I$, and probability mass function

$$\Pr(\boldsymbol{M} = \boldsymbol{m}) = \prod_{x \in \mathbb{Z}_I} m_{x,+}! \prod_{y \in \mathbb{Z}_J} \frac{\pi_{y|x}^{m_{x,y}}}{m_{x,y}!},$$

where $m_{x,y} \geq 0$ for all $x \in \mathbb{Z}_I, y \in \mathbb{Z}_J$. We can consider the case where the column totals are fixed analogously. A contingency table under these conditions is called a *product multinomial sample*.

Comparing Two Proportions in Categorical Data

We now discuss various ways that we can quantify the dependency between X and Y in a 2×2 contingency table. In what follows, let $\pi_x = \pi_{1|x}$, so that $\pi_{0|x} = 1 - \pi_x$. Then the null hypothesis that X and Y are independent is

$$H_0 : \pi_0 = \pi_1.$$

Thus a measure of the dependency between X and Y should capture the extent to which π_0 and π_1 are different. One such quantity is the *difference of proportions* $\pi_0 - \pi_1$. Since $\pi_x \in [0, 1]$, the difference of proportions can lie between -1 and 1.

However, if the probabilities are very small, then the *relative risk*

$$RR = \frac{\pi_0}{\pi_1}$$

may be a more appropriate measure of comparison. For example, suppose that $\pi_0 = 0.30$ and $\pi_1 = 0.21$. Then the difference of proportions is 0.09 and the relative risk is approximately 1.4. However, if $\pi_0 = 0.10$ and $\pi_1 = 0.01$, then the difference of proportions is still 0.09, but the relative risk is 10.

A third measure commonly used to compare probabilities is the *odds ratio*. First, we define the *odds* Ω of an event as

$$\Omega = \frac{\pi}{1 - \pi},$$

where π is the probability of the event. Then, if Ω_x is the odds corresponding to probability π_x, the *odds ratio* is

$$\theta = \frac{\Omega_0}{\Omega_1} = \frac{\pi_0/(1 - \pi_0)}{\pi_1/(1 - \pi_1)}. \tag{2.1}$$

We can also consider the log of the odds ratio, $\log \theta$, commonly known as the *log odds ratio*.

From these definitions, the following are equivalent:

- X and Y are independent,
- $\pi_0 = \pi_1$,
- The difference of proportions is zero,
- The relative risk $RR = 1$,
- The odds ratio $\theta = 1$ and
- The log odds ratio $\log \theta = 0$.

We can estimate quantities such as the difference of proportions, the relative risk, the odds ratio and the log odds ratio using the sample proportions $\hat{\pi}_{y|x} = \frac{m_{x,y}}{m_{x,+}}$. For example, the sample odds ratio $\hat{\theta}$ is

$$\begin{aligned}
\hat{\theta} &= \frac{\hat{\pi}_0/(1-\hat{\pi}_0)}{\hat{\pi}_1/(1-\hat{\pi}_1)} \\
&= \frac{\hat{\pi}_{1|0}/(1-\hat{\pi}_{1|0})}{\hat{\pi}_{1|1}/(1-\hat{\pi}_{1|1})} \\
&= \frac{\hat{\pi}_{1|0}/\hat{\pi}_{0|0}}{\hat{\pi}_{1|1}/\hat{\pi}_{0|1}} \\
&= \frac{\hat{\pi}_{1|0}\hat{\pi}_{0|1}}{\hat{\pi}_{0|0}\hat{\pi}_{1|1}} \\
&= \frac{m_{1,0}/m_{+,0} \times m_{0,1}/m_{+,1}}{m_{0,0}/m_{+,0} \times m_{1,1}/m_{+,1}} \\
&= \frac{m_{1,0}m_{0,1}}{m_{0,0}m_{1,1}}, \qquad (2.2)
\end{aligned}$$

which has the property that it does not change when any row or column of the contingency table is multiplied by a non-zero constant. Furthermore, the sample odds ratio is symmetric in the first and second indices, and a similar calculation can be used to show that the true odds ratio has the same property (by replacing $\hat{\pi}$ with π). Consequently, the definition of the odds ratio does not depend on the set of conditional probabilities ($\pi_{x|y}$ or $\pi_{y|x}$) that are used to define it.

The difference of proportions, relative risk, odds ratio and log odds ratio can be extended to arbitrary $I \times J$ contingency tables. In such tables, we cannot use one of these summary statistics to estimate the association between X and Y. We can, however, use various sets of these summary statistics. For example, if $J = 2$, then we can estimate the association between X and Y by considering the relative risks

$$RR_x = \frac{\pi_x}{\pi_0},$$

where $x \in \mathbb{Z}_I \backslash \{0\}$.

Large-Sample Inference for Unordered Contingency Tables

We now discuss several methods of testing the null hypothesis that X and Y are independent in a 2×2 contingency table, under the assumption that the frequencies of the cells in the contingency table are large. In particular, we discuss methods that test the hypothesis in terms of the log relative risk and the difference of proportions. We outline Wald statistics for these tests and the score statistic for the difference of proportions. These tests are valid when the entries of the contingency table are realisations of a product multinomial (binomial) sample. These statistics are all asymptotically standard normal, hence confidence intervals and hypothesis tests can be derived from them easily.

Assume that the row totals $m_x = m_{x,+}$ are fixed and let $Y_x = M_{x,1}$. The case where the column totals are fixed can be considered analogously. Then the Wald statistic for the relative risk is

$$\frac{\log r}{\hat{\sigma}(\log r)},$$

where

$$r = \frac{\hat{\pi}_0}{\hat{\pi}_1} = \frac{Y_0/m_0}{Y_1/m_1}$$

is the sample relative risk and

$$\hat{\sigma}(\log r) \approx \sqrt{\frac{1-\hat{\pi}_0}{y_0} + \frac{1-\hat{\pi}_1}{y_1}}$$

is the approximate standard error of $\log r$. Similarly, the Wald statistic for the difference of proportions is

$$\frac{\hat{\pi}_0 - \hat{\pi}_1}{\hat{\sigma}(\hat{\pi}_0 - \hat{\pi}_1)},$$

where

$$\hat{\sigma}(\hat{\pi}_0 - \hat{\pi}_1) = \sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{m_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{m_1}}$$

is the standard error of the sample difference of proportions $\hat{\pi}_0 - \hat{\pi}_1$.

Also, the score test statistic for the difference of proportions is

$$z = \frac{\hat{\pi}_0 - \hat{\pi}_1}{\tilde{\sigma}(\hat{\pi}_0 - \hat{\pi}_1)},$$

where

$$\tilde{\sigma}(\hat{\pi}_0 - \hat{\pi}_1) = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{m_0} + \frac{1}{m_1}\right)},$$

and

$$\hat{\pi} = \frac{y_0 + y_1}{m_0 + m_1}$$

is a pooled estimate of π_0 and π_1. Thus the score statistic uses a pooled estimate of π_0 and π_1, whereas the Wald statistic uses separate estimates.

We also detail two methods of testing the null hypothesis in an $I \times J$ contingency table. These methods can be used when the entries of the contingency table are realisations of a multinomial sample or a product multinomial sample.

The *Pearson chi-squared test* (Pearson, 1900) uses the test statistic

$$X^2 = \sum_{x \in \mathbb{Z}_I} \sum_{y \in \mathbb{Z}_J} \frac{(m_{x,y} - \hat{\mu}_{x,y})^2}{\hat{\mu}_{x,y}},$$

where $\hat{\mu}_{x,y} = m_{+,+}\hat{\pi}_{x,+}\hat{\pi}_{+,y} = \frac{m_{x,+}m_{+,y}}{m_{+,+}}$. When $I = J = 2$, this test statistic is related to the score test statistic for the difference of proportions by $X^2 = z^2$. The *likelihood ratio test* uses the test statistic

$$G^2 = 2 \sum_{x \in \mathbb{Z}_I} \sum_{y \in \mathbb{Z}_J} m_{x,y} \log \frac{m_{x,y}}{\hat{\mu}_{x,y}}.$$

Under the null hypothesis, both test statistics have an asymptotic $\chi^2_{(I-1)(J-1)}$ distribution.

All of these tests rely on asymptotic distributions, and thus they should only be used for sufficiently large values of $m_{x,y}$. There are some rules of thumb in the literature regarding how large these values need to be. The constraints are often given in terms of the values of $\hat{\mu}_{x,y} = \mathrm{E}[M_{xy}] = m_{+,+}\pi_{x,y}$, but since these values are unknown, we apply them to the $m_{x,y}$. For example, for the Pearson chi-squared test and the likelihood ratio test, the Cochran conditions state that for tables where $I > 2$ or $J > 2$, asymptotic tests are appropriate if $\min(\hat{\mu}_{x,y}) \approx 1$, as long as less than 20% of the $\hat{\mu}_{x,y}$ are less than 5 (Cochran, 1954). In other words, if all cells contain positive entries, and more than 80% of the cells are at least 5,

then the Pearson chi-squared test and the likelihood ratio test are appropriate. In general, the Pearson chi-squared test performs better than the likelihood ratio test for tables with small entries, especially when $\frac{m_{+,+}}{IJ} < 5$.

Small-Sample Inference for Unordered Contingency Tables

When the entries of a contingency table are not sufficiently large, methods that use exact distributions often perform better than methods that use asymptotic distributions. The most well-known test that uses the exact distribution of the cell counts is *Fisher's exact test (FET)* (Fisher, 1922). FET is used to test the null hypothesis that X and Y are independent in an $I \times J$ contingency table, and it assumes that both sets of margins are fixed. Under the null hypothesis, the conditional distribution of $\boldsymbol{M}$ is multivariate hypergeometric, with probability mass function

$$\Pr\left(\boldsymbol{M} = \boldsymbol{m}\right) = \frac{\left(\prod_{x \in \mathbb{Z}_I} m_{x,+}!\right)\left(\prod_{y \in \mathbb{Z}_J} m_{y,+}!\right)}{m_{+,+}! \prod_{x \in \mathbb{Z}_I} \prod_{y \in \mathbb{Z}_J} m_{x,y}!},$$

where $m_{x,y} \geq 0$ for all $x \in \mathbb{Z}_I, y \in \mathbb{Z}_J$. It is instructive to consider the case when $I = J = 2$, when the multivariate hypergeometric distribution reduces to the hypergeometric distribution. Since the margins are fixed, the table is characterised by the value of $m_{0,0}$:

$$\Pr(M_{0,0} = m_{0,0}) = \frac{\binom{m_{0,+}}{m_{0,0}}\binom{m_{1,+}}{m_{+,0}-m_{0,0}}}{\binom{m_{+,+}}{m_{+,0}}},$$

where

$$\max(0, m_{0,+} + m_{+,0} - m_{+,+}) \leq m_{0,0} \leq \min(m_{0,+}, m_{+,0}).$$

We can use this distribution to write down the p-values for the one-sided and two-sided versions of FET.

Loosely speaking, the p-value is the probability of obtaining a result at least as extreme as the data under the null hypothesis. To understand what an extreme result is in a one-sided FET, consider the two cases. Suppose that the null

hypothesis is

$$H_0 : \theta \leq 1$$

and the alternative hypothesis is

$$H_A : \theta > 1.$$

Then we have more evidence to reject H_0 if the sample odds ratio (2.2) is large. For fixed margins, this occurs for small $m_{0,0}$, and hence the p-value for this test is

$$p = \Pr(M_{0,0} \leq m_{0,0}).$$

Similarly, if the null hypothesis is

$$H_0 : \theta \geq 1$$

and the alternative hypothesis is

$$H_A : \theta < 1,$$

then the p-value is

$$p = \Pr(M_{0,0} \geq m_{0,0}).$$

For the two-sided case, consider a result at least as extreme as the data if the probability of obtaining the result under the null hypothesis is no greater than the probability of obtaining the data under the null hypothesis. Then the p-value is the sum of the probabilities of all such results that are at least as extreme as the data:

$$p = \sum_{\substack{t: \\ p_t \leq p_{m_{0,0}}}} p_t,$$

where $p_t = \Pr(M_{0,0} = t)$. We can apply the same method to an $I \times J$ contingency table to test the null hypothesis. Let $\boldsymbol{t} = (t_{0,0}, \ldots, t_{I-1,J-1})$. If

$$p_{\boldsymbol{t}} = \Pr\left(\boldsymbol{M} = \boldsymbol{t}\right),$$

then the p-value is

$$p = \sum_{\substack{\boldsymbol{t}: \\ p_{\boldsymbol{t}} \leq p_{\boldsymbol{m}}}} p_{\boldsymbol{t}}. \tag{2.3}$$

The Parametric Bootstrap for Unordered Contingency Tables

The parametric bootstrap can also be used to test the null hypothesis that X and Y are independent in an $I \times J$ contingency table, particularly when it is difficult to calculate the p-value correctly. For example, if I, J and m are large, using FET may be too computationally intensive.

To introduce the parametric bootstrap, assume that the null hypothesis is simple, so that the null distribution is completely specified. Let $\hat{\theta}$ be a statistic calculated from the contingency table. Simulate $B \in \mathbb{N}$ contingency tables from the null distribution and let $\hat{\theta}^*(b)$ be the value of the statistic for simulation $b \in \{1, \ldots, B\}$. If larger values of $\hat{\theta}$ provide more evidence to reject the null hypothesis, then the estimated p-value using the parametric bootstrap is

$$p = \frac{1}{B} \sum_{b=1}^{B} I\{\hat{\theta}^*(b) \geq \hat{\theta}\}. \tag{2.4}$$

Analogously, if smaller values of $\hat{\theta}$ provide more evidence to reject the null hypothesis, then the estimated p-value is

$$p = \frac{1}{B} \sum_{b=1}^{B} I\{\hat{\theta}^*(b) \leq \hat{\theta}\}. \tag{2.5}$$

Note that some authors, such as O'Dushlaine *et al.* (2009), choose to add one to the numerator and denominator in p-values estimated using the parametric bootstrap. The logic is that the data itself is a possible simulation, which should be included in the calculation. And since $\hat{\theta}^*(b) = \hat{\theta}$ for the data, the indicator function in both (2.4) and (2.5) are equal to one for this term. Many authors who do not add one to the numerator and denominator, such as Wang *et al.* (2007), quote p-values of zero as "$< \frac{1}{B}$," where B is the number of bootstrap samples that they used.

However, in many contingency tables, the exact null distribution is unknown. This occurs when the null hypothesis depends on unknown parameters. In such cases, an approximate parametric bootstrap can be used, where contingency tables are sampled from an approximate null distribution. An approximate null

distribution can be generated by replacing the unknown parameters in the exact null distribution with estimators for the parameters, such as the maximum likelihood estimators.

For example, the cell frequencies in a contingency table where only m is fixed may have a multinomial distribution with parameters $m_{+,+}$ and $\pi_{x,y}$, where $x \in \mathbb{Z}_I$ and $y \in \mathbb{Z}_J$. Under the null hypothesis that X and Y are independent, $\pi_{x,y} = \pi_{x,+}\pi_{+,y}$. However, $\pi_{x,+}$ and $\pi_{+,y}$ are unknown. To conduct an approximate parametric bootstrap for this table, $\pi_{x,+}$ and $\pi_{+,y}$ could be replaced with their sample analogues, $\hat{\pi}_{x,+} = \frac{m_{x,+}}{m_{+,+}}$ and $\hat{\pi}_{+,y} = \frac{m_{+,y}}{m_{+,+}}$ respectively. Then approximate p-values could be sampled from this approximate null distribution, in the same way that p-values were calculated from the exact null distribution.

Inference for Ordered Contingency Tables

In some contingency tables, one or both variables could exhibit a natural ordering. For example, consider a study to test for association between a risk factor and a disease. Then the level of exposure to the risk factor exhibits a natural ordering. When this natural ordering is present, tests that take it into account are usually more powerful than the tests that don't (Agresti, 2013). In the following tests, we assume that X is an ordinal variable, and $J = 2$.

General linear models (GLMs) can be used to test the null hypothesis that X and Y are independent, and they take into account the ordering in X. A GLM takes the form

$$g[\pi(x)] = \alpha + \beta x, \tag{2.6}$$

where g is a link function, $\pi(x) = \Pr(Y = 1|x)$, α is the intercept parameter, and β is the slope parameter. Assume that the possible values of X represent the natural ordering present in X. Of course, the model is sensitive to these values.

GLMs are defined by the choice of link function g. The simplest model is the *linear probability model*, which uses the unit link function, $g[\pi(x)] = \pi(x)$.

However, the linear probability model is inappropriate: $\pi(x) \in (0,1)$, but the right-hand side of (2.6) could be any real number. Consequently, a more appropriate GLM is the *logistic regression model*, where the link function is the log odds or *logit*:

$$g[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \text{logit}[\pi(x)].$$

Since $\pi(x) \in (0,1)$, the left-hand size of (2.6) can take any real value, which is consistent with the right-hand side.

We have assumed in GLMs that the response variable Y is random, and the predictor variable x is fixed. This is the case in *prospective studies*, where the number of subjects exposed to each level of the risk factor is fixed. However, in *case-control* or *retrospective studies*, the disease status $y \in \{0,1\}$ of each subject is fixed and X is random.

The hypothesis that X and Y are independent in a logistic regression framework is equivalent to the hypothesis that $\beta = 0$. Consequently, the following result by Prentice and Pyke (1979) demonstrates that it is appropriate to use a prospective logistic regression model in a case-control study to test the hypothesis.

Proposition 2.2.1. *Suppose that the prospective logistic regression model*

$$\log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \tag{2.7}$$

is used to model the effect of a risk factor X on the probability of disease. Furthermore, assume that the event that a subject is sampled and their level of exposure to the risk factor X are independent. Then the slope parameter in the corresponding retrospective logistic regression model is also β.

Proof. Let the random variable Z indicate whether or not a given subject is sampled. Let $Z = 1$ for subjects that are sampled and $Z = 0$ for subjects that are not. Also, let $Y = 1$ denote a case, let $Y = 0$ denote a control, and let $\rho_y = \Pr(Z = 1|Y = y)$ denote the probabilities of sampling a case and a control. Assume that the event that a subject is sampled and their level of exposure to

the risk factor X are independent, so that

$$\rho_y = \Pr(Z = 1|Y = y, X = x) \tag{2.8}$$

for all x and y.

In the prospective logistic regression model, the probability of sampling an individual is not based on the response variable Y. Consequently, we model the probability according to

$$\Pr(Y = 1|X = x) = \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \tag{2.9}$$

where the last equality comes from rearranging (2.7). In contrast, in the retrospective logistic regression model, we condition on the fact that the subjects have already been sampled. Hence we use the probability

$$\begin{aligned}
\Pr(Y = 1|Z = 1, X = x) &= \frac{\Pr(Z = 1|Y = 1, X = x)\Pr(Y = 1|X = x)}{\sum_{y=0}^{1}\Pr(Z = 1|Y = y, X = x)\Pr(Y = y|X = x)} \\
&= \frac{\rho_1 \Pr(Y = 1|X = x)}{\sum_{y=0}^{1}\rho_y \Pr(Y = y|X = x)} \\
&= \frac{\rho_1 \Pr(Y = 1|X = x)}{\rho_0\left[1 - \Pr(Y = 1|X = x)\right] + \rho_1 \Pr(Y = 1|X = x)} \\
&= \frac{\rho_1\left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right]}{\rho_0\left[1 - \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right] + \rho_1\left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right]} \\
&= \frac{\rho_1\left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right]}{\rho_0\left[\frac{1}{1+\exp(\alpha+\beta x)}\right] + \rho_1\left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right]} \\
&= \frac{\rho_1 \exp(\alpha + \beta x)}{\rho_0 + \rho_1 \exp(\alpha + \beta x)} \\
&= \frac{\frac{\rho_1}{\rho_0}\exp(\alpha + \beta x)}{1 + \frac{\rho_1}{\rho_0}\exp(\alpha + \beta x)} \\
&= \frac{\exp\left(\alpha + \log\left[\frac{\rho_1}{\rho_0}\right] + \beta x\right)}{1 + \exp\left(\alpha + \log\left[\frac{\rho_1}{\rho_0}\right] + \beta x\right)},
\end{aligned}$$

which is a logistic regression model with intercept parameter $\alpha + \log\left[\frac{\rho_1}{\rho_0}\right]$ and slope parameter β. Thus the slope parameter in the retrospective model is also β. □

Under the null hypothesis that $\beta = 0$, the maximum-likelihood estimators of logistic regression models are asymptotically normal, which means that we can use Wald, score and likelihood ratio tests to test the null hypothesis.

The last test that we detail is the *Cochran-Armitage Trend Test (CATT)* (Armitage, 1955; Cochran, 1954), which tests the null hypothesis that X and Y are independent in an $I \times 2$ contingency table. Assume product multinomial sampling, such that the column totals $m_{+,y}$ are fixed for $y \in \{0, 1\}$. The CATT statistic is

$$Z_{CATT} = \frac{U}{\sqrt{\hat{\text{var}}(U)}}, \tag{2.10}$$

where

$$U = \sum_{x \in \mathbb{Z}_I} x\{(1-p)m_{x|1} - pm_{x|0}\},$$

and $p = \frac{m_{+,1}}{m_{+,+}}$. The variance of U can be derived as

$$\begin{aligned}\text{var}(U) = &\frac{m_{+,+}}{m_{+,1}^2} p(1-p)^2 \left(\sum_{x \in \mathbb{Z}_I} x^2 \pi_{x|1} - \left[\sum_{x \in \mathbb{Z}_I} x \pi_{x|1} \right]^2 \right) \\ &+ \frac{m_{+,+}}{m_{+,1}^2} p^2(1-p) \left(\sum_{x \in \mathbb{Z}_I} x^2 \pi_{x|0} - \left[\sum_{x \in \mathbb{Z}_I} x \pi_{x|0} \right]^2 \right),\end{aligned}$$

and it can be estimated in two ways. Each $\pi_{x|y}$ in $\text{var}(U)$ can be estimated separately using $\hat{\pi}_{x|y} = \frac{m_{x,y}}{m_{+,y}}$. Alternatively, the null hypothesis $H_0 : \pi_{x|1} = \pi_{x|2}$ can be used to pool the estimates: $\hat{\pi}_{x|1} = \hat{\pi}_{x|2} = \frac{m_{x,+}}{m_{+,+}}$. Both versions of the CATT statistic are asymptotically standard normal under the null hypothesis.

	Y		
X	0	1	Total
0	$m_{0,0}$	$m_{0,1}$	$m_{0,+}$
1	$m_{1,0}$	$m_{1,1}$	$m_{1,+}$
2	$m_{2,0}$	$m_{2,1}$	$m_{2,+}$
Total	$m_{+,0}$	$m_{+,1}$	$m_{+,+}$

Table 2.2: 3×2 contingency table of genotype frequencies at a genetic locus

2.2.2 Categorical Data Analysis in GWA Studies

Genotype-Based Tests

We now discuss categorical data analysis in the context of GWA studies. Recall that in a GWA study, a fixed number of controls and cases are genotyped at each genetic locus. Each locus is then tested for association with the disease. Also, recall our assumption that there are two possible alleles at each genetic locus: the common allele and the rare allele. The genotype of a subject at a genetic locus is then the pair of alleles present at the locus. At a given locus, a subject may have 0, 1 or 2 rare alleles, which means that there are three possible genotypes.

Consequently, for each genetic locus, and for a given subject, denote by X the number of rare alleles that are present at that locus. Also, denote the disease status of the subject by Y: let $Y = 0$ if the subject is a control, and let $Y = 1$ if the subject is a case. Consequently, for each genetic locus, the subjects in the GWA study can be arranged in a 3×2 contingency table according to their disease status and the number of rare alleles present at that locus. We display this data in Table 2.2.

Recall that the aim of a GWA study is to identify SNPs that are associated with a disease. Consequently, at each genetic locus, we test the hypothesis that X and Y are independent. We now discuss the application of the tests that we detailed in Subsection 2.2.1 to GWA studies.

Of the large-sample tests, only the Pearson chi-squared test, the likelihood ratio test and the CATT can be used with the data in Table 2.2, since the table is 3×2. These tests assume that the data in this contingency table are realisations of a product multinomial sample. This assumption is reasonable, because the total number of controls $m_{+,0}$ and cases $m_{+,1}$ are fixed. Recall that the Pearson chi-squared test is more commonly used, because it is superior to the likelihood ratio test. However, these asymptotic tests are only appropriate if all cells contain positive entries, and more than 80% of the cells are at least 5.

FET can be used with the data in Table 2.2, and the p-value for the test is given in (2.3). The permutation test is also often used with the data in Table 2.2. Bootstrap samples are generated from an approximate null distribution by permuting the case and control labels on the subjects in the original data many times, which removes the association between the genotype and the disease (Zheng *et al.*, 2012).

The number of rare alleles X at the genetic locus is an ordinal variable. Consequently, the retrospective logistic regression model and the CATT can be used to test the null hypothesis. Recall that the retrospective logistic regression model assumes that y is fixed and X is random, which is a valid assumption, because the number of cases and controls in the study is fixed. Similarly, the CATT assumes that the column totals of a contingency table are fixed, which is true for the data in Table 2.2.

It is instructive to discuss the assumptions used by many of these tests. For example, consider tests that assume product multinomial sampling and FET, in which the null distribution is multivariate hypergeometric. These tests assume that each observation is independent. Since each observation is a subject, these tests assume that the genotype of a subject is completely uninformative about the genotypes of other subjects. However, the population from which the subjects were sampled may have features that invalidate this assumption. For example, if the sample contains a parent and their child, then the genotype of the parent is

informative about the genotype of the child. In this case, these tests would not be valid. However, we do not consider such departures from independence in this thesis.

Consequently, many tests from Subsection 2.2.1 can be used to test the null hypothesis that X and Y are independent. However, we also detailed tests that can only be used with data in a 2×2 contingency table, such as the one-sided FETs. We now discuss two methods that can be used to display the data in Table 2.2 in a 2×2 contingency table. One method aggregates the data in Table 2.2, and the other method uses the frequency of the *alleles* in the analysis. Of course, the following methods make certain assumptions about the genotype data and/or the underlying probabilities. Consequently, the accuracy of tests used with the data in a 2×2 contingency table depends on whether or not these assumptions are accurate.

Using Genetic Models in Genotype-Based Tests

In statistical genetics, four primary genetic models are often used to describe the way that a genotype can affect the probability of disease. These models are often described in terms of the conditional probabilities

$$\pi_x = \pi_{1|x} = \Pr(Y = 1|X = x)$$

or the relative risks

$$RR_x = \frac{\pi_x}{\pi_0}.$$

We refer to π_x as the *penetrance* of x rare alleles.

We present these models in Table 2.3. In all of these models, we assume that adding rare alleles to the genotype does not decrease the penetrance ($\pi_2 \geq \pi_1 \geq \pi_0$). Also, assume that the penetrance of having two disease alleles is strictly greater than the penetrance of having no disease alleles ($\pi_2 > \pi_0$).

Model	Penetrance Relationship	Relative Risk Relationship
Recessive (REC)	$\pi_1 = \pi_0$	$RR_1 = 1$
Additive (ADD)	$\pi_1 = \frac{\pi_0+\pi_2}{2}$	$RR_1 = \frac{1+RR_2}{2}$
Multiplicative (MUL)	$\pi_1 = \sqrt{\pi_0\pi_2}$	$RR_1 = \sqrt{RR_2}$
Dominant (DOM)	$\pi_1 = \pi_2$	$RR_1 = RR_2$

Table 2.3: Commonly used genetic models in statistical genetics

In the additive and multiplicative models, the penetrances increase with the number of rare alleles additively or multiplicatively. We do not use these models in genotype-based tests, but we use them later in this thesis.

In the recessive model, the rare allele "recedes" in the presence of the common allele, so the penetrance does not increase with one rare allele. The penetrance only increases in the recessive model when there are two rare alleles. In contrast, in the dominant model, the rare allele "dominates" the common allele. The penetrance increases with one rare allele, but it does not increase any more if there are two rare alleles.

If we assume that the dominant or recessive model holds at a genetic locus, then we can aggregate the genotypes into two categories: *low-risk* and *high-risk*. For example, under the dominant model, subjects with no rare alleles are low-risk, whereas subjects with at least one rare allele are high-risk. In contrast, under the recessive model, subjects with two rare alleles are high-risk, and subjects with less than two rare alleles are low-risk. Let $X = 0$ and $X = 1$ denote low-risk and high-risk genotypes respectively. Then to test the null hypothesis under the assumption of the dominant model or the recessive model, aggregate the data in Table 2.2 by combining the frequencies of subjects with high-risk or low-risk genotypes respectively. We display the aggregated frequencies for the dominant model and the recessive model in Tables 2.4 and 2.5 respectively.

Since the genotype data is displayed in a 2×2 contingency table, any of the tests in Subsection 2.2.1 can be used with it to test the null hypothesis.

	Y		
X	0	1	Total
0	$m_{0,0}$	$m_{0,1}$	$m_{0,+}$
1	$m_{1,0} + m_{2,0}$	$m_{1,1} + m_{2,1}$	$m_{1,+} + m_{2,+}$
Total	$m_{+,0}$	$m_{+,1}$	$m_{+,+}$

Table 2.4: 2×2 contingency table of aggregated genotype frequencies to use under the assumption of a dominant model

	Y		
X	0	1	Total
0	$m_{0,0} + m_{1,0}$	$m_{0,1} + m_{1,1}$	$m_{0,+} + m_{1,+}$
1	$m_{2,0}$	$m_{2,1}$	$m_{2,+}$
Total	$m_{+,0}$	$m_{+,1}$	$m_{+,+}$

Table 2.5: 2×2 contingency table of aggregated genotype frequencies to use under the assumption of a recessive model

Allele-Based Tests

To test the null hypothesis at a genetic locus, we can also consider the frequency of the alleles in the analysis, rather than the frequencies of the genotypes in the subjects. For example, let $X = 0$ and $X = 1$ represent the common allele and the rare allele respectively. Then use the data in Table 2.2 to display the allele frequencies in controls and cases in Table 2.6. Note that since each of the $m_{+,+}$ subjects in the GWA study has two alleles, the total number of alleles in the table is $2m_{+,+}$.

Since the allele data is displayed in a 2×2 contingency table, any of the tests in Subsection 2.2.1 can be used to test the null hypothesis. Recall that these tests assume that the observations in a contingency table are independent. For genotype-based tests, this is equivalent to the assumption that the genotypes of the subjects in the analysis are independent. However, since each observation

	Y		
X	0 (Case)	1 (Control)	Total
0	$2m_{0,0} + m_{1,0}$	$2m_{0,1} + m_{1,1}$	$2m_{0,+} + m_{1,+}$
1	$m_{1,0} + 2m_{2,0}$	$m_{1,1} + 2m_{2,1}$	$m_{1,+} + 2m_{2,+}$
Total	$2m_{+,0}$	$2m_{+,1}$	$2m_{+,+}$

Table 2.6: 2×2 contingency table of allele frequencies at a genetic locus

in Table 2.6 is an *allele*, we need to assume that the alleles in the analysis are independent.

Suppose that at a given genetic locus, the two possible alleles are B and b. Furthermore, denote by p the probability that a randomly selected allele at the locus is B. Within an individual, if the two alleles at a genetic locus are independent, then the probability of obtaining each possible genotype is

$$\Pr(G) = \begin{cases} p^2 & \text{if } G = BB \\ 2p(1-p) & \text{if } G = Bb \\ (1-p)^2 & \text{if } G = bb. \end{cases}$$

These equations are known as the *Hardy-Weinberg proportions* (Hardy, 1908; Weinberg, 1908).

During reproduction, an offspring receives one chromosome from each parent. Consequently, to ensure that the alleles at a genetic locus are independent for all subjects in a population, we need to assume that the mating in a population is random. Hardy-Weinberg Equilibrium (HWE) is a commonly used condition in statistical genetics that provides a set of stronger assumptions than the Hardy-Weinberg proportions. Many of the tests in Subsection 2.2.1 can be used to test the null hypothesis that HWE holds in a population. The interested reader is referred to literature such as Foulkes (2009), Laird and Lange (2011), and Zheng *et al.* (2012) for more details.

2.2.3 Multiple Hypothesis Testing

In Subsections 2.2.1 and 2.2.2, we detailed a number of different methods that can be used to test the null hypothesis that a genetic locus and a disease are independent. Typically, a testing procedure fixes the probability of falsely rejecting the null hypothesis (a false positive or type I error) at $\alpha \in (0, 1)$, which is known as the significance level. The procedure then rejects the null hypothesis if the p-value is less than α. Often a significance level of $\alpha = 0.05$ is used, but other values may be more appropriate in certain circumstances.

In a typical GWA study, however, up to 10^6 SNPs are tested simultaneously. Consequently, using this testing procedure could result in many false positive findings. In general, this problem is known as the Multiple Testing Problem, and extensive research exists in the literature that details various ways to overcome this problem. In the remainder of this subsection, we give a brief historical overview of some of the procedures that have been developed to tackle the Multiple Testing Problem. Many of these procedures aim to control a measure of error that is a generalisation of the significance level. We define two such measures of error and detail some commonly used procedures that control these quantities. This material is based on summaries by Foulkes (2009) and Ge *et al.* (2003), as well as literature such as Benjamini and Hochberg (1995) and Storey (2003).

Measures of Error in Multiple Hypothesis Testing

Suppose that we test $m \in \mathbb{N}$ null hypotheses, H_i, where $i \in \{1, \ldots, m\}$. Denote by p_i the p-value obtained from testing hypothesis H_i, and suppose that m_0 of the m null hypotheses are true. Let

- S be the number of correctly rejected null hypotheses (true positives),
- U be the number of correctly accepted null hypotheses (true negatives),
- V be the number of incorrectly rejected null hypotheses (false positives) and

Null Hypotheses	Accepted	Rejected	Total
True	U	V	m_0
False	T	S	$m - m_0$
Total	$m - R$	R	m

Table 2.7: Contingency table in multiple hypothesis testing

- T be the number of incorrectly accepted null hypotheses (false negatives).

Also, let $R = V + S$ be the total number of rejected null hypotheses. We display this notation in Table 2.7, which is Table 1 of Benjamini and Hochberg (1995).

In this framework, accepting or rejecting a null hypothesis is based on the p-value of the test, which is a random variable. Consequently, R, S, T, U and V are random variables. Furthermore, since we do not know which null hypotheses are true, S, T, U and V are unobservable random variables. However, we know how many null hypotheses we reject in total, so R is an observable random variable.

We use this notation to define the measures of error that are commonly controlled in multiple hypothesis testing.

Definition 2.2.2 (Family-Wise Error Rate)**.** *The* family-wise error rate (FWER) *is the probability of at least one false positive. That is,* $\text{FWER} = \Pr(V > 0)$. $\triangle$

In situations where even one false positive could have disastrous consequences, controlling the FWER makes sense. However, Benjamini and Hochberg (1995) note that often controlling the FWER in a multiple testing situation is not needed. For example, in microarray experiments, it is acceptable if a small number of genes are falsely classified as differentially expressed (Ge *et al.*, 2003). Storey (2003) also comments that often, the role of the statistician is to find "as many interesting features in a data set as possible", rather than worrying about the (almost inevitable) chance of making at least one type 1 error.

Consequently, Benjamini and Hochberg (1995) defined the false discovery rate. It is approximately the expected ratio of the number of false positives to the total

Control Type	Condition
Weak	All null hypotheses are true
Exact	The correct set of true and false null hypotheses
Strong	For all 2^m combinations of true and false null hypotheses

Table 2.8: Weak, exact and strong control in multiple hypothesis testing

number of rejected hypotheses, $\mathrm{E}\left[\frac{V}{R}\right]$. However, this expression is undefined if there are no rejected hypotheses ($R = 0$). When this is the case Benjamini and Hochberg (1995) note that there cannot be any false positives, and hence the ratio should be equal to zero. Consequently, we have the following definition:

Definition 2.2.3 (False Discovery Rate)**.** *Let*

$$Q = \begin{cases} \frac{V}{R}, & R > 0 \\ 0, & R = 0. \end{cases}$$

Then the false discovery rate (FDR) *is* $\mathrm{E}[Q]$. △

We need to be precise about what it means to control one of these error rates. A procedure controls the FWER or FDR at level α if it guarantees that the error rate is less than α. Furthermore, a procedure is said to have *weak*, *exact*, or *strong* control of the FWER or FDR if it controls the error rate

- under the condition that all null hypotheses are true,
- under the correct set of true and false null hypotheses and
- for all 2^m combinations of true and false null hypotheses, respectively.

We display these conditions in Table 2.8. Foulkes (2009) comments that since we do not know which hypotheses are indeed true, strong control is usually desirable. In contrast, weak control is usually undesirable, especially in situations like microarray experiments where null hypotheses are often false.

We can also define *adjusted p-values* for procedures that control the FWER or FDR. The adjusted p-value corresponding to hypothesis H_i under some multiple testing procedure is

$$\tilde{p}_i = \inf\{\alpha' : \text{The procedure rejects } H_i \text{ at FWER or FDR } \alpha'\}.$$

The testing procedure would then reject hypothesis H_i if $\tilde{p}_i < \alpha$, for some significance level α.

We now discuss methods of controlling the FWER and FDR, since some of them are used in the literature that we discuss in Chapter 3.

Methods of controlling the FWER

We may divide methods of controlling the FWER into *single-step* and *step-down* procedures. A commonly used single-step procedure is the *Bonferroni adjustment* (Bonferroni, 1936). Under the Bonferroni adjustment, the significance level for each test is set at $\alpha' = \frac{\alpha}{m}$. Then the test that rejects H_i if $p_i < \alpha'$ strongly[1] controls the FWER at level α. Laird and Lange (2011) comment that the Bonferroni adjustment makes *no assumptions about the independence of the events*. However, since the Bonferroni adjustment is based on controlling the FWER, its power to detect associations is limited. The *Bonferroni adjusted p-values* for hypothesis H_i are given by

$$\tilde{p}_i = \min(mp_i, 1).$$

More precisely, the $\tilde{p}_i$ are conservative lower bounds for the adjusted p-values, which cannot be calculated more accurately without further assumptions.

Step-down procedures such as the *Holm procedure* (Holm, 1979) generally provide more power than single step procedures, while still providing strong control of the FWER. Denote by $p_{(1)} \leq \ldots \leq p_{(m)}$ the ordered p-values and denote by $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. Let

$$i^* = \arg\min_i \left\{ p_{(i)} > \frac{\alpha}{m - i + 1} \right\}.$$

[1]Defined in Table 2.8.

The Holm procedure rejects hypotheses $H_{(1)}, \ldots, H_{(i^*-1)}$ if i^* exists, and it rejects all hypotheses otherwise. The *Holm step-down adjusted p-values* are given by

$$\tilde{p}_{(i)} = \max_{k \in \{1,\ldots,i\}} \left\{ \min[(m-k+1)p_{(k)}, 1] \right\}.$$

Since the p-values are multiplied by $m-k+1 \leq m$, the Holm procedure is less conservative than the Bonferroni procedure. However, in a typical GWA study, $m > 10^5$ SNPs are tested for association with a disease, and the number of rejected hypotheses $i^* - 1$ is small. Consequently, there is little benefit in using the Holm procedure over the Bonferroni adjustment in GWA studies.

Methods of controlling the FDR

We now discuss two methods of controlling the FDR. The *Benjamini-Hochberg (BH) procedure* (Benjamini and Hochberg, 1995) assumes that the p-values corresponding to the true null hypotheses are independent. Let

$$i^* = \arg\max_i \left\{ p_{(i)} \leq \frac{i}{m}\alpha \right\}.$$

If i^* exists, the BH procedure rejects hypotheses $H_{(1)}, \ldots, H_{(i^*)}$, and it accepts all hypotheses otherwise. The adjusted p-values for the BH procedure are

$$\tilde{p}_{(i)} = \min_{k \in \{1,\ldots,m\}} \left\{ \min\left(\frac{m}{k} p_{(k)}, 1 \right) \right\}.$$

Benjamini and Yekutieli (2001) note that one of the shortcomings of using the BH procedure is that often, the p-values corresponding to the true null hypotheses are not independent due to experimental considerations. Consequently, Benjamini and Yekutieli (2001) developed a method of controlling the FDR that allows for dependent test statistics. Let

$$i^* = \arg\max_i \left\{ p_{(i)} \leq \frac{i}{m \sum_{l=1}^{m} l^{-1}}\alpha \right\}.$$

The *Benjamini-Yekutieli (BY) procedure* rejects hypotheses $H_{(1)}, \ldots, H_{(i^*)}$ if i^* exists, and it accepts all hypotheses otherwise. The adjusted p-values for the BY procedure are

$$\tilde{p}_{(i)} = \min_{k \in \{1,\ldots,m\}} \left\{ \min\left(\frac{m \sum_{l=1}^{m} l^{-1}}{k} p_{(k)}, 1 \right) \right\}.$$

2.3 Linkage and Linkage Disequilibrium

We now return to a formal discussion and mathematical formulation of the dependency structure between different genetic loci on a chromosome.

Most cells in humans contain pairs of chromosomes. However, sperm and egg cells, which are collectively known as *gametes*, only contain one copy of each chromosome. If parents passed on one of their chromosomes to their offspring without any genetic modification, the potential for genetic variation would be quite limited. However, during *meiosis*, the cell division process through which gametes are formed in organisms, *crossover events* can occur that increase the number of combinations of alleles that a parent may pass on to its offspring. These combinations of alleles are known as *haplotypes*. We demonstrate meiosis and crossover events with an example, illustrated in Figure 2.2.

Suppose that a parent has four genetic loci on a pair of autosomal chromosomes, and that there are two possible alleles at each locus. We label the alleles A and a; B and b; C and c; and D and d. Furthermore, suppose that the parent has haplotypes $ABCD$ and $abcd$ on the two chromosomes. In meiosis, each autosomal pair of chromosomes duplicates, as shown in the top and middle panels of Figure 2.2. The duplicated chromosomes are referred to as *sister chromatids*, and any two chromosomes that are not duplicates are referred to as *non-sister chromatids*. Crossover events may occur between two non-sister chromatids. During each crossover event, the chromosomes overlap, and all of the DNA on one side of the crossover event is exchanged between them. A crossover event is shown in the middle and bottom panels of Figure 2.2 between the B/b and C/c loci on the middle two chromosomes. Many crossover events may occur between pairs of non-sister chromatids in a meiosis. Consequently, a meiosis produces four gametes, each containing a single chromosome, and the number of possible haplotypes on each chromosome is large.

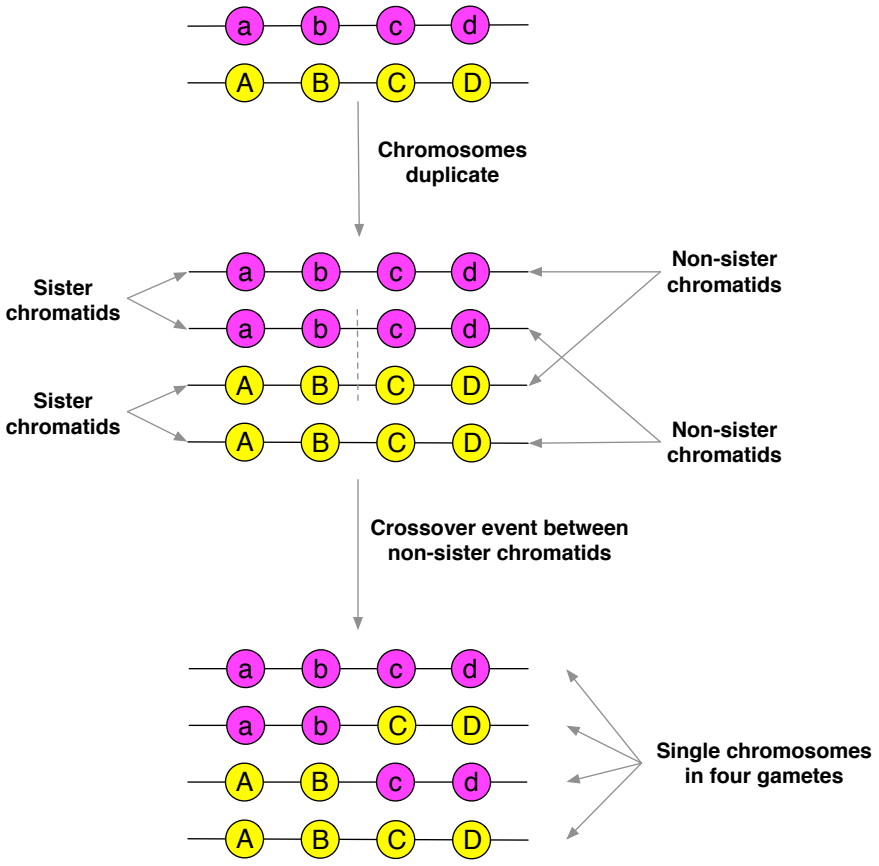

Figure 2.2: Example – crossing-over during the formation of gametes in meiosis

Since we cannot observe crossover events during meiosis directly, we infer them through *recombination events*. A recombination event occurs between two genetic loci if the alleles at the two loci were inherited from different chromosomes. We refer to a chromosome on which a recombination event has occurred as a *recombinant*. We illustrate crossover and recombination events using the parent in Figure 2.2 with haplotypes $ABCD$ and $abcd$.

For simplicity, we assume that crossover events only occur between one pair of non-sister chromatids, and we only consider the two gametes that are produced as a result of these crossover events. Of course, the other two gametes that are produced are $ABCD$ and $abcd$, as in the bottom panel Figure 2.2. If there is a crossover event between the B/b and C/c loci, as in Figure 2.3, then a recombination event has occurred between these loci, and the two gametes are $ABcd$ or $abCD$. In contrast, if two crossover events occur between the B/b and C/c loci, as in Figure 2.4, then no recombination event has occurred between these loci, and the two gametes are $ABCD$ or $abcd$. However, multiple crossover events could occur between *different* pairs of loci. For example, consider Figure 2.5. The first crossover event occurs between the A/a and B/b loci, which exchanges the alleles at the B/b, C/c and D/d loci. The second crossover event occurs between the C/c and D/d loci, which exchanges the alleles at the D/d locus. Thus recombination events have occurred between the A/a and B/b loci, and between the C/c and D/d loci. In this example, the two gametes are $AbcD$ and $aBCd$.

We can see from these examples that a recombination event occurs between two genetic loci if and only if the number of crossover events that occur between them is odd. The *recombination fraction* is the probability that a recombination occurs between two genetic loci, and it is denoted by θ. We say that *linkage* is present between these loci if $\theta < \frac{1}{2}$. If linkage is present between two loci, then the alleles present at these loci are not independent. In contrast, if $\theta = \frac{1}{2}$, then the alleles at these loci are independent.

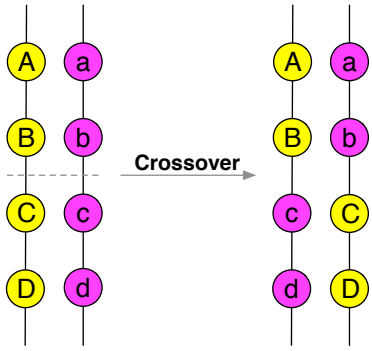

Figure 2.3: Example – two of the four gametes produced when one crossover event occurs between non-sister chromatids

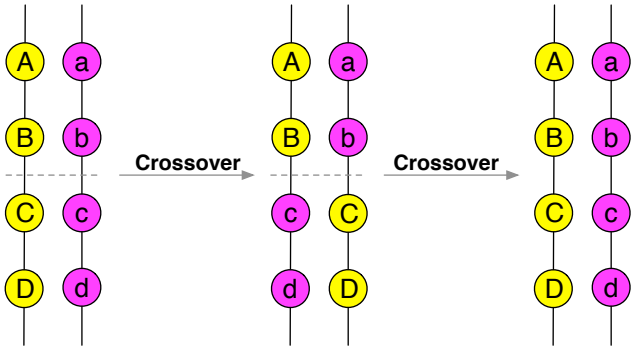

Figure 2.4: Example – two of the four gametes produced when two crossover events occur between the same loci on non-sister chromatids

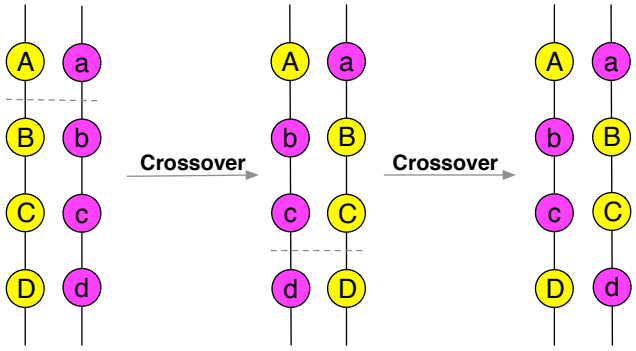

Figure 2.5: Example – two of the four gametes produced when two crossover events occur between different loci on non-sister chromatids

Proposition 2.3.1 (Mather's Law)**.** *The recombination fraction θ can be written in terms of the probability P_0 that no crossover events occur between two loci:*

$$\theta = \frac{1 - P_0}{2}.$$

Proof. Denote by X the total number of crossover events that occur between two genetic loci on non-sister chromatids in a meiosis. Furthermore, suppose that each pair of non-sister chromatids has equal probability of participating in each crossover event. Suppose also that after all of the crossover events, we sample one chromosome from the four chromosomes produced during the meiosis with equal probability.

If $X = 0$, then $P_0 = 1$ and the probability of recombination θ is zero. Consequently, with probability $1 - P_0$, $X \geq 1$. Under this assumption, we show that $\Pr(\text{selected chromosome is a recombinant}) = \frac{1}{2}$.

For a randomly sampled chromosome, the probability that it participates in an arbitrary crossover event is $\frac{1}{2}$. Since there are X independent crossover events between the genetic loci in total, the number of crossover events B that occur between the genetic loci on the chromosome has a binomial distribution with parameters X and $\frac{1}{2}$. It is straightforward to show that $\Pr(B \text{ odd}) = \frac{1}{2}$. A recombination event between two loci occurs if and only if the number of crossover events that occur between the loci is odd. Thus the probability that a recombination event occurs between the two loci is $\frac{1}{2}$, which completes the proof. □

Genetic Distance

In statistical genetics, it is often useful to plot a map of genetic loci on a chromosome which includes a measure of distance. One such measure of distance is *physical distance*, which is simply the number of base pairs (bp) between the two loci.

In general, there is strong linkage between genetic loci that are close together (in terms of physical distance), and hence θ is small. In contrast, genetic loci

that are far apart are nearly independent, and θ is close to $\frac{1}{2}$. However, there are other factors that can influence linkage between genetic loci. For example, there are locations on chromosomes where crossover events are likely to occur, known as *recombination hotspots*. Consequently, it is useful to define a distance measure based on the recombination fraction between pairs of genetic loci. In particular, the *genetic distance* between two loci is the expected number of crossovers that occur between them, per gamete. The unit of genetic distance is the *Morgan* (M).

Recall from Figure 2.2 that one crossover between non-sister chromatids produces two gametes with a crossover, and two gametes without a crossover. Thus the genetic distance between two loci is

$$L = \frac{1}{2}\,\mathrm{E}[X], \tag{2.11}$$

where X is the total number of crossover events that occur between the two loci on non-sister chromatids in a meiosis. Consequently, genetic distance is additive along the chromosome.

As an example, we consider Haldane's measure of genetic distance (Haldane, 1919). Haldane assumed that X has a Poisson distribution, and since its mean is $2L$, X has probability mass function

$$\Pr(X = k) = \frac{e^{-2L}(2L)^k}{k!}$$

for nonnegative integers. Consequently, we have that $P_0 = e^{-2L}$, and hence

$$\theta = \frac{1 - e^{-2L}}{2}$$

using Mather's Law. Rearranging for L, we find that

$$L(\theta) = -\frac{1}{2}\log(1 - 2\theta).$$

Laird and Lange (2011) comment that a Poisson distribution is likely to be inaccurate due to factors such as recombination hotspots. However, Haldane's measure still has some nice properties:

	B	b	Total
A	$p_{AB} = p_A p_B + D$	$p_{Ab} = p_A p_b - D$	p_A
a	$p_{aB} = p_a p_B - D$	$p_{ab} = p_a p_b + D$	p_a
Total	p_B	p_b	1

Table 2.9: Haplotype frequencies under LD

1. Since $\theta \in [0, \frac{1}{2})$, $L(\theta) \geq 0$, which is desirable for a distance measure.

2. If the recombination fraction θ is close to zero, then $L(\theta)$ is close to zero.

3. If θ is close to $\frac{1}{2}$, which indicates that the genetic loci are almost independent, then $L(\theta)$ is large. In fact, $\lim_{\theta \to \frac{1}{2}} L(\theta) = +\infty$.

4. $L(\theta)$ is increasing in θ. That is, as the recombination fraction increases and the loci become more independent, the genetic distance increases.

Linkage Disequilibrium

Linkage disequilibrium (LD) is commonly used to measure the degree of association between alleles on a chromosome. Denote by A and a the two possible alleles at one genetic locus, and denote by B and b the two possible alleles at a second genetic locus. Furthermore, let p_A, p_a, p_B and p_b be the frequency of alleles A, a, B and b at the respective genetic loci. We say that the alleles are in *linkage equilibrium (LE)* if they are independent.

However, when the alleles at the loci are not independent, $D = p_{AB} - p_A p_B \neq 0$. We say that the two alleles are in LD, and D is an *LD Coefficient* (Laird and Lange, 2011). Note that since the margins of Table 2.9 are fixed, the magnitude of D does not change, no matter which difference we use to define it.

Unfortunately, D is not ideal as a measure of the strength of association between allele frequencies. For example, if one allele frequency is low, then the magnitude of D is also low. Consequently, D should be standardised.

From the diagonal entries of Table 2.9, $D \geq -p_A p_B$ and $D \geq -p_a p_b$. Thus

$$D \geq \max\{-p_A p_B, -p_a p_b\} = -\min\{p_A p_B, p_a p_b\}.$$

Similarly, from the off-diagonal entries

$$D \leq \min\{p_a p_B, p_A p_b\}.$$

Consequently, define $D_{\min} = -\min\{p_A p_B, p_a p_b\}$ and $D_{\max} = \min\{p_a p_B, p_A p_b\}$, which are the minimum and maximum values that D can have respectively. Then define

$$D' = \begin{cases} \frac{D}{D_{\max}} & \text{if } D > 0 \\ \frac{D}{D_{\min}} & \text{if } D < 0. \end{cases}$$

Thus $D = 0$ implies $D' = 0$, and $D' = 1$ if any haplotype has zero frequency.

However, Laird and Lange (2011) note that even if D' is large, we cannot necessarily use the allele frequencies at one locus to predict the allele frequencies at the other locus with high accuracy. Nonetheless, if we let

$$r^2 = \frac{D^2}{p_A p_B p_a p_b}, \tag{2.12}$$

then $r^2 = 0$ if and only if $D = 0$, and $r^2 = 1$ only if $p_A = p_B$ and $p_a = p_b$. In other words, the alleles are independent if and only if $r^2 = 0$, and if $r^2 = 1$, then the allele frequencies at one locus predict the allele frequencies at the other locus with perfect accuracy.

2.4 Chapter Summary

In this chapter, we have outlined the necessary background of GWA studies. As we discussed in Section 2.1, the aim of a GWA study is to identify SNPs that are associated with a given disease. In Section 2.2, we discussed the statistical techniques used to perform GWA studies. In a GWA study, controls and cases are genotyped at a large number of genetic loci. Techniques from categorical

data analysis are then used at each genetic locus to test the null hypothesis that disease status and genotype/allele are independent. We gave a general overview of these techniques in Subsection 2.2.1, and we discussed the application of these techniques to GWA studies in Subsection 2.2.2. However, since over 10^5 SNPs are tested simultaneously, techniques from multiple hypothesis testing need to be used to adjust the p-value obtained for each SNP. We discussed these techniques in Subsection 2.2.3. Finally, in Section 2.3, we discussed in detail the concepts of linkage and LD, which describe the dependency structure between the SNPs on a chromosome. As we discuss in Chapter 3, linkage and LD between SNPs need to be accounted for when the results of a GWA study are interpreted.

Chapter 3

Gene Set Analysis Methods

In Section 3.1, we introduce *gene set analysis (GSA)* methods, and we explain their advantages over traditional GWA studies. In Sections 3.2 and 3.3, we review six GSA methods in the literature, which were developed by Askland *et al.* (2009), Holmans *et al.* (2009), Hong *et al.* (2009), O'Dushlaine *et al.* (2009), Wang *et al.* (2007), and Yaspan *et al.* (2011). These methods were compared qualitatively by Yaspan *et al.* (2011). We include some necessary preliminary material in Section 3.2, and then we explain the methods in Section 3.3. In Section 3.4, we review the issues that need to be addressed in GSA methods, as detailed by Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011). We then use these issues to guide a detailed theoretical comparison of the six GSA methods.

3.1 Motivation for GSA

In their review of GWA studies, Visscher *et al.* (2012) note that numerous SNPs have been identified as associated with various diseases, including Crohn's disease, prostate cancer and breast cancer in at least one study. However, it is important to check that an association between a SNP and a given disease is *replicable* (Laird and Lange, 2011). That is, we can only say with confidence that a SNP is associ-

ated with a disease if the result can be replicated in multiple *independent* studies. Many associations have been replicated in this way, such as SNPs associated with type 2 diabetes (Visscher *et al.*, 2012). However, many GWA studies have yielded unreplicated results. Also, GWA studies have only identified a small proportion of the genetic variation that is associated with many diseases (Maher, 2008). Two primary factors that contribute to this problem are *small effect sizes* and *epistasis* (Hong *et al.*, 2009; Mooney *et al.*, 2014; O'Dushlaine *et al.*, 2009; Wang *et al.*, 2010; Wang *et al.*, 2011; Yaspan *et al.*, 2011). We now discuss each of these effects.

Many disease SNPs have a *small effect size*. That is, the occurrence of the disease allele leads only to a small increase in the risk of disease. For such a SNP, the p-value obtained from testing the null hypothesis that it is not associated with the disease is often not very small. Consequently, a GWA study is less likely to classify disease SNPs with a small effect size as significant. Increasing the number of cases and controls in GWA studies can improve their power to find disease SNPs with small effect sizes. However, the issue of small effect size is exacerbated by the necessity of using multiple testing procedures in GWA studies.

In a typical GWA study, between 10^5 and 10^6 SNPs are tested for association with a disease simultaneously. Laird and Lange (2011) comment that "given the many false positive findings in the history of genetic association studies", procedures that conservatively control the FWER such as the Bonferroni correction are preferred to procedures that seek to maximise statistical power. In particular, if the unadjusted significance level in a GWA study is $\alpha = 0.05$ and the GWA study contains 10^6 SNPs, then the Bonferroni-adjusted significance level is $\alpha' = \frac{\alpha}{10^6} = 5 \times 10^{-8}$. Consequently, SNPs that modestly increase the probability of a given disease are unlikely to be identified as significant by a GWA study.[1]

Changing the significance level from $\alpha = 0.05$ to $\alpha = 5 \times 10^{-8}$ increases the number of controls $m_{+,0}$ and cases $m_{+,1}$ that are necessary to achieve the

[1]The consensus now is that a fixed significance level of 5×10^{-8} should be used, regardless of the number of SNPs genotyped (personal communication with Professor David Balding)

same statistical power in SNP association tests. We illustrate this effect with an example, assuming that $m_{+,0} = m_{+,1}$. Suppose that the probability of disease for an individual with no disease alleles is $\pi_0 = 0.001$, and the probability of disease for an individual with at least one disease allele is $\pi_1 = 0.01$ (we are assuming a dominant model here). Suppose also that we use a two-sided test to test the null hypothesis that the SNP is not associated with the disease ($\pi_0 = \pi_1$), and that the entries of the contingency table are realisations of a product binomial sample. Then for a given significance level, we can use the normal approximation to the binomial distribution to calculate the required number of controls and cases to achieve a statistical power of 0.8. For example, if the significance level $\alpha = 0.05$, then we require $m_{+,0} = m_{+,1} = 1059$. However, if the significance level $\alpha = 5 \times 10^{-8}$, then we require $m_{+,0} = m_{+,1} = 5346$.

The other issue that GWA study methods fail to account for is *epistasis*. Epistasis occurs when the combined effect of two genetic factors (such as SNPs) is not additive. Many GWA studies only consider SNPs individually, so they cannot account for epistasis. Epistasis occurs because genetic factors can interact with each other in a complex way. However, considering all possible sets of genes is statistically and combinatorially prohibitive. Consequently, attention is restricted to predefined sets of genes, such as *biological pathways*. A biological pathway can be defined as "a set of interacting genes ... that together perform a specific biological function" (Mooney *et al.*, 2014). A number of databases of biological pathways exist, including the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and Protein ANalysis THrough Evolutionary Relationships (PANTHER) (Mi *et al.*, 2013).

Biological pathways may have certain levels of *redundancy*. We illustrate the concept of redundancy with a simplified example. Consider the two pathways illustrated in Figure 3.1. The aim of each pathway is to produce protein 3. In pathway 1, the cell needs protein 1 to produce protein 2, and it needs protein 2 to produce protein 3. Similarly, in pathway 2, the cell needs proteins 1a and 1b to

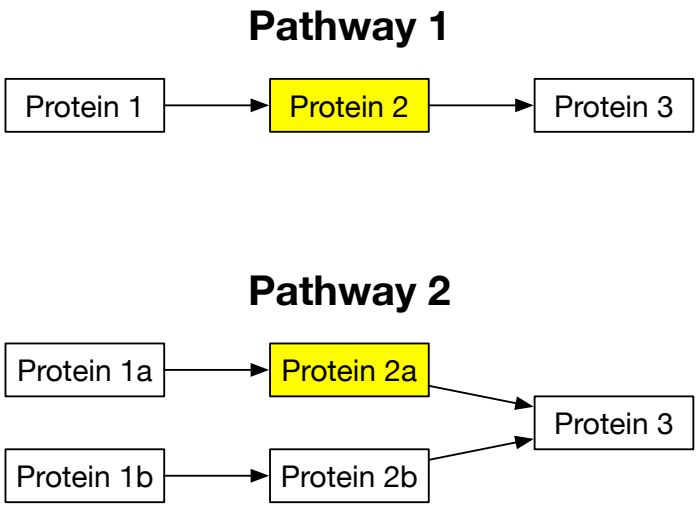

Figure 3.1: Illustration – pathway redundancy and dysfunction. If protein 2 dysfunctions in pathway 1, then it cannot produce protein 3. However, if protein 2a dysfunctions in pathway 2, then the pathway can still use protein 2b to produce protein 3.

produce proteins 2a and 2b respectively. However, if either protein 2a or protein 2b is present, then the cell can produce protein 3.

In the example in Figure 3.1, consider the event that a SNP affects protein 2 in pathway 1, such that it can no longer produce protein 3. When this occurs, we say that the pathway is *dysfunctional*. However, if a SNP affects protein 2a in pathway 2, the pathway can still use protein 2b to produce protein 3, and hence the pathway is not dysfunctional. We say that there is redundancy present in pathway 2, because a SNP can stop a protein in the pathway from functioning correctly without causing the entire pathway to become dysfunctional. However, there is no redundancy present in pathway 1, because the dysfunction of any protein in pathway 1 causes the entire pathway to become dysfunctional.

The aim of GSA is to estimate the degree of association between a *gene set* and a disease. Unlike GWA studies, GSA can account for the interaction between

SNPs and interaction between genes in the gene set. Consequently, GSAs can help to elucidate the underlying biology of the disease in ways that GWA studies are unable to.

A number of other methods exist in the literature that can be used to account for SNP interaction. For example, Zheng *et al.* (2012) detail a logistic regression approach that uses interaction terms to account for SNP interaction. However, even if the genetic models are fully specified at each locus, a model that only includes interactions up to second-order has

$$1 + M + \binom{M}{2}$$

terms, where M is the number of SNPs in the GWA study. The number of subjects in a GWA study is typically in the order of 10^3 to 10^4, but the number of SNPs in a GWA study is typically at least 10^5. Consequently, the number of parameters in the model would be orders of magnitude greater than the number of observations, which renders logistic regression impossible.

Zheng *et al.* (2012) also detail Multifactor Dimensionality Reduction, a method which uses the original genotype data to group each combination of genotypes into "high-risk" and "low-risk". As the name suggests, this reduces the number of parameters in the model. However, since the number of interaction terms is so large, using this method in a GWA study is still infeasible. Consequently, we do not consider such methods any further. Instead, we turn our attention to the GSA methods developed by Askland *et al.* (2009), Holmans *et al.* (2009), Hong *et al.* (2009), O'Dushlaine *et al.* (2009), Wang *et al.* (2007), and Yaspan *et al.* (2011). However, we first need to discuss some preliminary procedures that are used in some of these GSA methods.

3.2 The Mapping Problem

GSA methods can be divided into *one-step* and *two-step* methods (Fridley and Biernacka, 2011; Mooney *et al.*, 2014). In one-step methods, SNPs are directly

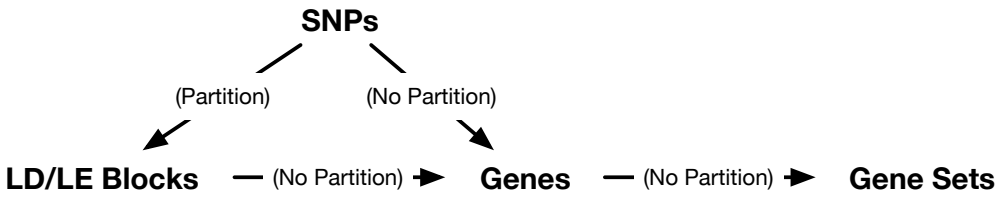

Figure 3.2: Illustration – maps between SNPs, blocks, genes and gene sets. SNPs are partitioned into LD and LE blocks, but SNPs and blocks may be in zero genes, one gene or many genes. Similarly, genes may be in zero gene sets, one gene set or many gene sets.

mapped to gene sets. However, in two-step methods, SNPs are mapped to genes or *blocks*. An *LD block* is defined as a set of SNPs in high LD with each other, and an *LE block* is defined as an individual SNP in LE with other SNPs. Furthermore, the blocks can themselves be mapped to genes. The genes are then mapped to the gene sets that are analysed using each method.

In the methods that map SNPs to blocks, the blocks *partition* the SNPs. Each SNP in the GWA study is mapped to exactly one block. However, when SNPs or blocks are mapped to genes, and genes are mapped to gene sets, the mapping is *not* a partition. For example, a SNP or block can be mapped to no genes, one gene, or many genes. Note that a SNP can be mapped to multiple genes because genes can overlap on a chromosome, and because a SNP can be mapped a gene even if the SNP does not lie within the gene. Similarly, a gene can be mapped to no gene sets, one gene set, or many gene sets. We illustrate these maps in Figure 3.2.

Consequently, many of the six GSA methods use various procedures that calculate the p-value or test statistic of a unit (such as a gene, block or gene set) from the p-value or test statistic of each subunit that is mapped to the unit. In this section, we detail some of these procedures that were developed independently of

the GSA method(s) that use them. Denote by S the gene set that we are calculating or estimating a p-value for, denote by g the total number of genes in the study, and denote by g_S the number of genes in S.

3.2.1 MaxT and MinP

The maxT and minP procedures are simple procedures to calculate the test statistic or p-value of a unit from the test statistic of each subunit that is mapped to the unit. They are equivalent if the distribution of each test statistic under the null hypothesis is the same, and if a larger test statistic provides more evidence to reject the null hypothesis. The maxT procedure calculates the test statistic of the unit as the maximum of the subunit test statistics. Similarly, the minP procedure calculates the p-value of the unit as the minimum of the subunit p-values. Many GSA methods use the minP method to calculate the p-value of each gene from the p-value of each SNP that is mapped to the gene (Fridley and Biernacka, 2011; Holmans, 2009; Ramanan *et al.*, 2012; Wang *et al.*, 2011).

3.2.2 Gene Set Enrichment Analysis (GSEA)

The rest of the procedures that we detail in this section calculate the p-value of a gene set from the p-value or test statistic of each gene mapped to the gene set. Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) calculates a p-value for a gene set using the test statistic of each gene in the gene set. Subramanian *et al.* (2005) used it in the context of gene expression. However, we are interested in its use in GSA.

Denote by r_j a test statistic for gene G_j, and assume that a larger test statistic provides more evidence to reject the null hypothesis. Also, assume that $r_1 >$

$r_2 \ldots > r_g$.[2] Then the *enrichment score (ES)* of S is

$$\mathrm{ES}(S) = \max_{1 \le j \le g} \left\{ \frac{1}{N_R} \sum_{\substack{1 \le k \le j \\ G_k \in S}} |r_k|^p - \frac{1}{g - g_S} \sum_{\substack{1 \le k \le j \\ G_k \notin S}} 1 \right\}, \tag{3.1}$$

where

$$N_R = \sum_{k: G_k \in S} |r_k|^p.$$

Here, p is a parameter that weights the importance of genes with large test statistics. Subramanian *et al.* (2005) recommend using $p = 1$. When $p = 0$, the ES reduces to the Kolmogorov-Smirnov statistic.

According to Subramanian *et al.* (2005), we can estimate the significance of an ES as follows. Simulate P data sets by permuting the case and control labels in the original genotype data, and re-calculate the ES for each simulated data set. Let $\mathrm{ES}(S, \pi_i)$ be the ES obtained using the i^{th} simulated data set. The empirical p-value is then the proportion of simulations where the ES is larger than the ES calculated from the real data.

We demonstrate calculating an ES in GSEA with an example, illustrated in Figure 3.3. In this illustration, the yellow circles represent genes in a gene set, and the white circles represent genes not in the gene set. Thus there are five genes in the GWA study and the test statistic for each gene is

$$r_1 = 10, \qquad r_2 = 7, \qquad r_3 = 6,$$
$$r_4 = 4, \qquad r_5 = 1.$$

Using these results, we calculate the ES of the gene set defined by the yellow circles in Figure 3.3. Note that in the figure, we have ordered the genes by their test statistics (descending order), to better illustrate the calculation of the ES. To perform the calculation, note that there are $g = 5$ genes in the GWA study and the number of genes in the gene set is $g_S = 3$. Also, we assume that $p = 1$, which

[2] We were unable to find documentation regarding the appropriate procedure to use if two test statistics are equal.

j	$f_T(j)$	$f_U(j)$	$f_T(j) - f_U(j)$
1	$\frac{10}{15}$	0	$\frac{2}{3}$
2	$\frac{10}{15}$	$\frac{1}{2}$	$\frac{1}{6}$
3	$\frac{10}{15}$	$\frac{1}{2} + \frac{1}{2}$	$-\frac{1}{3}$
4	$\frac{10}{15} + \frac{4}{15}$	$\frac{1}{2} + \frac{1}{2}$	$-\frac{1}{15}$
5	$\frac{10}{15} + \frac{4}{15} + \frac{1}{15}$	$\frac{1}{2} + \frac{1}{2}$	0

Table 3.1: Example – using GSEA to calculate the ES of a gene set. For each value of j, the value of the two terms inside the braces in (3.1) is shown in the middle two columns, and the difference of these two terms is shown in the right column. The ES is the maximum of the differences over all j.

means that

$$N_R = \sum_{k:G_k \in S} |r_k|^p = 10 + 4 + 1 = 15.$$

To calculate the ES, note that as we increase j, we add the next-highest test statistic to one of the two sums. If the corresponding gene is in S, then we add a term to the first sum, and we add a term to the second sum otherwise. For notational simplicity, let

$$f_T(j) = \frac{1}{N_R} \sum_{\substack{1 \le k \le j \\ G_k \in S}} |r_k|^p = \frac{1}{15} \sum_{\substack{1 \le k \le j \\ G_k \in S}} |r_k|$$

and

$$f_U(j) = \frac{1}{g - g_S} \sum_{\substack{1 \le k \le j \\ G_k \notin S}} 1 = \frac{1}{2} \sum_{\substack{1 \le k \le j \\ G_k \notin S}} 1.$$

We manually perform the calculation of the ES as in Table 3.1. The ES is then the maximum of $f_T(j) - f_U(j)$. From Table 3.1, the first sum in the ES (the second column) is equal to the proportion of genes in the gene set that have been included in the sum, weighted by their test statistics. And the second sum in the ES (the third column) is the unweighted proportion of genes not in the gene set that have been included in the sum. Thus the ES is $\frac{2}{3}$.

GSEA also includes two methods that adjust the empirical p-value of each gene set when multiple gene sets are tested simultaneously. In particular, these

methods control the FDR and the FWER respectively. For a given S, let

$$\overline{\mathrm{ES}}(S) = \frac{1}{P}\sum_{i=1}^{P}\mathrm{ES}(S,\pi_i)$$

be the sample mean of the ESes over all simulations π_i. Then the *normalised enrichment score (NES)* of a gene set S calculated using the real data is

$$\mathrm{NES}(S) = \frac{\mathrm{ES}(S)}{\overline{\mathrm{ES}}(S)}$$

and the NES of a gene set S calculated using the data simulated from permutation π is

$$\mathrm{NES}(S,\pi) = \frac{\mathrm{ES}(S,\pi)}{\overline{\mathrm{ES}}(S)}.$$

For a given S, the FWER-adjusted p-value is

$$\frac{1}{P} \times \left|\left\{\pi : \max_{S'}\mathrm{NES}(S',\pi) > \mathrm{NES}(S)\right\}\right|,$$

and the FDR-adjusted p-value is

$$\begin{aligned}&\frac{\%(S',\pi) \text{ with } \mathrm{NES}(S',\pi) \geq \mathrm{NES}(S)}{\%\text{ of observed } S' \text{ with } \mathrm{NES}(S') \geq \mathrm{NES}(S)}\\ &= \frac{1}{P} \times \frac{|\{(S',\pi) : \mathrm{NES}(S',\pi) \geq \mathrm{NES}(S)\}|}{|\{S' : \mathrm{NES}(S') \geq \mathrm{NES}(S)\}|}.\end{aligned}$$

Subramanian *et al.* (2005) also provide an alternative to GSEA, GSEAPreranked (GSEAPR). In GSEAPR, each simulated data set is obtained by assigning each gene to a random gene test statistic from the real data. The enrichment scores $\mathrm{ES}(S,\pi)$ and normalised enrichment scores $\mathrm{NES}(S,\pi)$ for each simulation are then calculated accordingly.

3.2.3 Exploratory Visual Analysis (EVA)

Exploratory Visual Analysis (EVA) (Reif *et al.*, 2005) uses FET to calculate the p-value of S from a list of significant and nonsignificant genes in the study. Denote by h_S the number of significant genes in S and denote by h the total number of significant genes in the study. We display the gene-level association results in

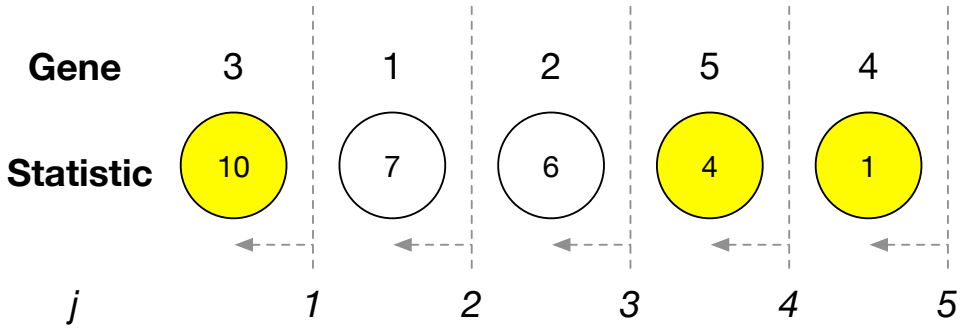

Figure 3.3: Example – using GSEA to calculate the ES of a gene set. Yellow circles represent genes in the gene set, and white circles represent genes not in the gene set. The genes are arranged such that the corresponding test statistics are in decreasing order from left to right.

Number of Genes	In S	Not in S	Total
Significant	h_S	$h - h_S$	h
Insignificant	$g_S - h_S$	$g - g_S - h + h_S$	$g - h$
Total	g_S	$g - g_S$	g

Table 3.2: Gene-wide significance results to use in EVA

Number of Genes	In S	Not in S	Total
Significant	$h_S - 1$	$h - h_S$	$h - 1$
Insignificant	$g_S - h_S$	$g - g_S - h + h_S$	$g - h$
Total	$g_S - 1$	$g - g_S$	$g - 1$

Table 3.3: Gene-wide significance results to use in EASE

Table 3.2. EVA can also calculate a p-value for the gene set using a simulation procedure. EVA does not account for testing multiple gene sets simultaneously.

In EVA, the simulation procedure works as follows. For each simulation, sample g_S genes from the study without replacement. The empirical p-value of S is then the proportion of simulations where the number of significant genes is no less than h_S.

We assume that EVA uses the one-sided FET, because it tests S for an enrichment of significant genes. Using the notation displayed in Table 3.2, the one-sided FET p-value is

$$p_S = \sum_{t \geq h_S} \frac{\binom{g_S}{t}\binom{g-g_S}{h-t}}{\binom{g}{h}}. \tag{3.2}$$

3.2.4 DAVID and EASE

Expression Analysis Systematic Explorer (EASE) (Hosack *et al.*, 2003) is a test provided by the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis Jr *et al.*, 2003) that performs a one-sided FET to calculate the p-value of a gene set. However, EASE reduces the number of significant genes in each gene set by one. We display the modified contingency table that EASE uses in Table 3.3. The p-value for this test is then obtained by deducting 1 from g, h, g_S and h_S in (3.2).

EASE can adjust for multiple testing in a number of different ways, including the Bonferroni correction and bootstrap methods for estimating the FDR. We refer interested readers to Hosack *et al.* (2003) for more details.

3.2.5 Ingenuity Pathway Analysis (IPA)

QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, `www.quiagen.com/ingenuity`) performs a one-sided FET identically to EVA (Reif *et al.*, 2005). It also performs the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to adjust the p-value assigned to each gene set if multiple gene sets are tested simultaneously.

3.3 Review of Six GSA Methods

In this section, we detail the GSA methods developed by Askland *et al.* (2009), Holmans *et al.* (2009), Hong *et al.* (2009), O'Dushlaine *et al.* (2009), Wang *et al.* (2007), and Yaspan *et al.* (2011). Each method calculates a p-value for a gene set using either the original genotype data, or the results of a GWA study in the form of a p-value assigned to each SNP in the GWA study. We defer a comparison of these methods to Section 3.4.

3.3.1 Pathway Analysis by Randomization Incorporating Structure (PARIS)

Pathway Analysis by Randomization Incorporating Structure (PARIS) (Yaspan *et al.*, 2011) partitions the SNPs in the study into LD and LE blocks. As an example of this partitioning procedure, consider Figure 3.4. Each black circle in this figure represents a SNP on a chromosome. The red squares indicate pairs of SNPs in high LD, and the pink squares indicate SNPs in low LD. For example, high LD exists between SNPs 1 and 2, between SNPs 1 and 3, and between SNPs 5 and 6. The black borders around groups of SNPs indicate block boundaries.

In PARIS, a block is mapped to a gene if the block contains a SNP that is mapped to the gene. A gene set S is then a collection of blocks that map to genes in S. The *structure* of S is the number and size of the blocks in S. For example,

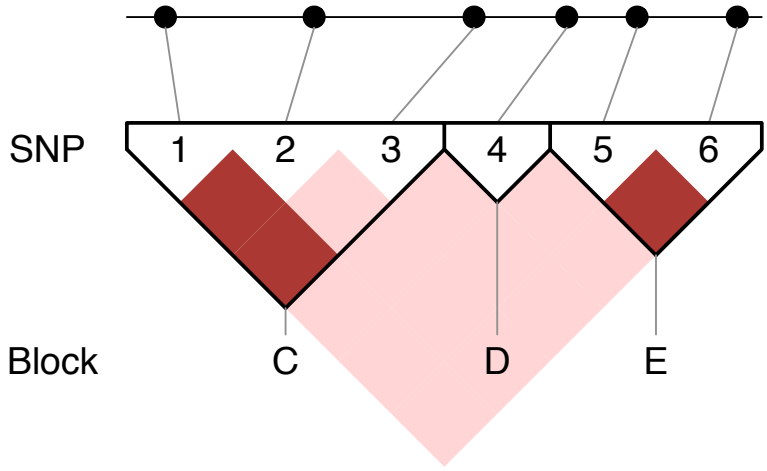

Figure 3.4: Example – Grouping SNPs into blocks in PARIS. Black circles represent SNPs on the chromosome (numbered 1 to 6), red squares indicate high LD between two SNPs, and pink squares indicate low LD between two SNPs. Black borders around groups of SNPs indicate block boundaries. We use the letters C, D and E to refer to the blocks.

suppose that the six SNPs in Figure 3.4 make up S. This gene set contains block C (which is the LD block of size three containing SNPs one, two and three), block D (which is the LE block of size one containing SNP four) and block E (which is the LD block of size two containing SNPs five and six).

To test S for association with a disease, PARIS compares S with random collections of blocks from the rest of the genome, such that each block collection has a similar structure to S, in terms of the number and size of the blocks.

However, there are more small blocks in the genome than large blocks, which complicates sampling random block collections. PARIS accounts for the non-uniform distribution of block sizes in the genome as follows. The list of LD blocks in the genome is sorted by size, and the order of blocks of the same size is random. The first B blocks in the list are assigned to bin 1, the next B blocks in the list are assigned to bin 2, and so on. Yaspan *et al.* (2011) use bins containing

approximately $B = 10000$ blocks. However, LE blocks are all assigned to the same bin.

Denote by n_k the number of blocks in both S and bin k. To compare S with random block collections, PARIS removes the blocks in S from the bins, and then creates P random block collections. Each block collection contains n_k blocks from bin k, for all k. Sampling an individual block collection is performed *without* replacement, but sampling the N block collections is performed *with* replacement. In other words, the same block cannot be present twice within an individual block collection, but the same block may be present in different block collections. Yaspan *et al.* (2011) use $N = 1000$ block collections.

PARIS defines a block to be significant if it contains at least one significant SNP. Yaspan *et al.* (2011) use the significance level $\alpha = 0.05$. The empirical p-value of S is the proportion of simulated block collections that contain more significant blocks than S. PARIS does not account for testing multiple gene sets simultaneously.

We now give an example of using PARIS to calculate the p-value of a gene set S. Suppose that in Figure 3.5, the top rectangle containing red and black sub-rectangles represents the genome. Each sub-rectangle represents a block, and the size of each sub-rectangle indicates the number of SNPs in the block. Sub-rectangles with red borders indicate significant blocks, while sub-rectangles with black borders indicate nonsignificant blocks. Gene set S is highlighted in yellow. We remove S from the genome, and divide the blocks in the rest of the genome into bins of size $B = 3$. We note that S contains $n_k = 1$ block from bin k, for $k \in \{1, 2, 3\}$. Consequently, to create each of the $N = 3$ block collections, we sample one block from each bin. Finally, S contains two significant blocks, but none of the three block collections contain more than two significant blocks. Thus the empirical p-value of S is $\frac{0}{3} = 0$.

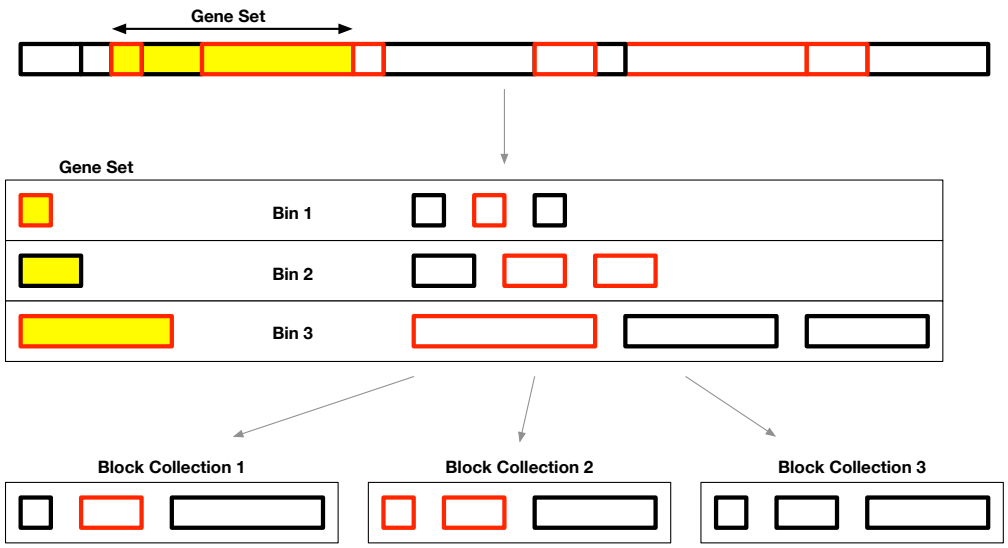

Figure 3.5: Example – using PARIS to estimate the p-value of a gene set. The top rectangle containing red and black sub-rectangles represents the genome. Each sub-rectangle represents a block, and the size of each sub-rectangle indicates the number of SNPs in the block. Red sub-rectangles represent significant blocks, and black sub-rectangles represent non-significant blocks. Highlighted blocks represent blocks that are in the gene set. The blocks not in the gene set have been arranged into bins, as shown in the middle of the figure. We display random block collections in the bottom of the figure, and the number of significant blocks in each collection is compared to the number of significant blocks in the gene set to obtain the p-value of gene set.

3.3.2 The SNP Ratio Test (SRT)

The SNP Ratio Test (the SRT) (O'Dushlaine *et al.*, 2009) performs a GWA study in the usual way by using an association test such as FET at each genetic locus. Then for a given significance level $\alpha \in [0, 1]$, denote by m the total number of significant SNPs in the GWA study. O'Dushlaine *et al.* (2009) recommend using $\alpha \in \{0.001, 0.005, 0.01, 0.05\}$ in various situations. Denote by m_S the number of significant SNPs in S.

The SRT simulates $P \in \mathbb{N}$ data sets by permuting the case and control labels on the original genotype data. O'Dushlaine *et al.* (2009) use $P = 1000$. A GWA study is then performed on each simulated data set. Let $m_S^{(k)}$ be the number of significant SNPs in S in simulation k. However, in the simulations, the SNPs with the smallest m p-values are defined as significant, instead of using the significance level α. Let s be the number of simulations where S contains at least as many significant SNPs as in the real data:

$$s = \left| \left\{ k : m_S^{(k)} \geq m_S \right\} \right|.$$

Then the empirical p-value of S is

$$p = \frac{s + 1}{P + 1}. \tag{3.3}$$

The SRT does not correct for testing multiple gene sets simultaneously.

As an example, illustrated in Figure 3.6, suppose that three cases and three controls are genotyped at six genetic loci in a GWA study. The genotype data are given in the top panel of Figure 3.6. We calculate the p-value of gene set S defined by SNPs 4, 5 and 6. Suppose that we apply an association test to this data, which gives the result that SNPs 1, 4 and 5 are significant. We then create $P = 2$ simulated data sets by permuting the case and control labels in the original genotype data, and perform a GWA study on each simulated data set. The results of these analyses are shown in the bottom panels of Figure 3.6. In the first simulation, SNPs 2, 4 and 6 are significant, and in the second simulation,

SNPs 4, 5 and 6 are significant. Thus the number of significant SNPs in S for the real data is $m_S = 2$, and in the simulations, $m_S^{(1)} = 2$ and $m_S^{(2)} = 3$. Consequently,

$$s = |\{k : m_S^{(k)} \geq m_S\}| = 2,$$

and hence the p-value of S in this example is

$$p_S = \frac{2+1}{2+1} = 1.$$

3.3.3 MinP and Exploratory Visual Analysis (MPEVA)

Askland *et al.* (2009) calculate the p-value of each gene from the SNP p-values using the minP procedure. The genes with the smallest 10% of p-values are then classified as significant. EVA is used to calculate the p-value of each gene set from the p-value of each gene in the gene set. Askland *et al.* (2009) used a one-sided FET and the simulation procedure with $P = 10^5$ simulations. Finally, the p-value of each gene set is adjusted using the Bonferroni correction.

3.3.4 ProxyGeneLD

ProxyGeneLD assumes that the set of SNPs in a GWA study is a subset of the set of all SNPs in a database. Hong *et al.* (2009) refer to these sets as the *study SNPs* and *HapMap SNPs* respectively. The HapMap SNPs are then partitioned into LD and LE blocks. Hong *et al.* (2009) generate each LD block iteratively by adding SNPs to the block if they are in high LD with any other SNP in the block, according to the threshold $r^2 \geq 0.8$. Recall that r^2 is the measure of correlation between alleles at two genetic loci given in (2.12). ProxyGeneLD calculates the p-value of each gene by using the minP procedure on the set of all study SNPs in blocks that contain at least one study SNP that maps to the gene. The p-value is then adjusted by multiplying it by the number of blocks included in the calculation (the *adjustment factor*).[3]

[3]In summarising this part of ProxyGeneLD, we have followed the description by Hong *et al.* (2009) in the main text. However, in the example in Figure 1, Hong *et al.* (2009) are inconsistent.

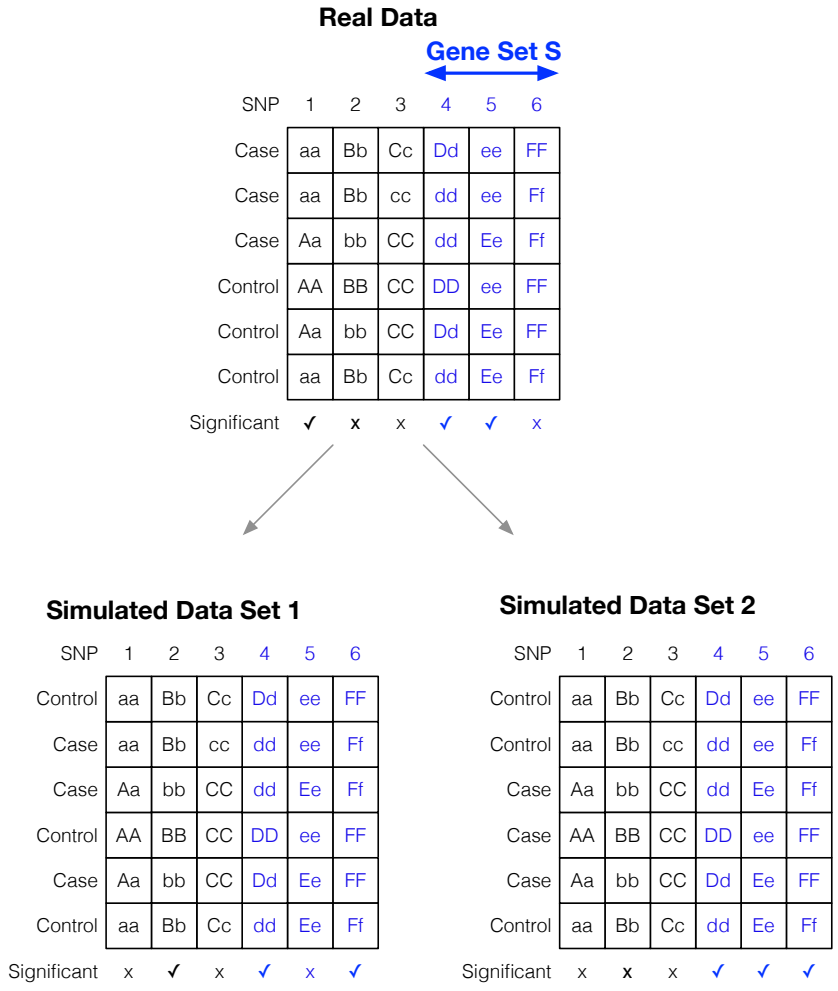

Figure 3.6: Example – using the SRT to estimate the p-value of a gene set. Genotype data for three cases and three controls at six loci (numbered 1 to 6) are shown in the top row of the figure. The gene set contains loci 4 to 6, and the corresponding data are shown in blue. A significance criterion is used which identifies three loci as significant; we use ticks to indicate significant loci, and crosses to indicate nonsignificant loci. We use the same notation in the bottom row of the figure, except the case and control labels in the original genotype data have been permuted to obtain simulated data sets. To obtain the p-value of the gene set, the number of significant SNPs in the gene set for the real data set is compared with the number of significant SNPs in the gene sets for the simulated data sets.

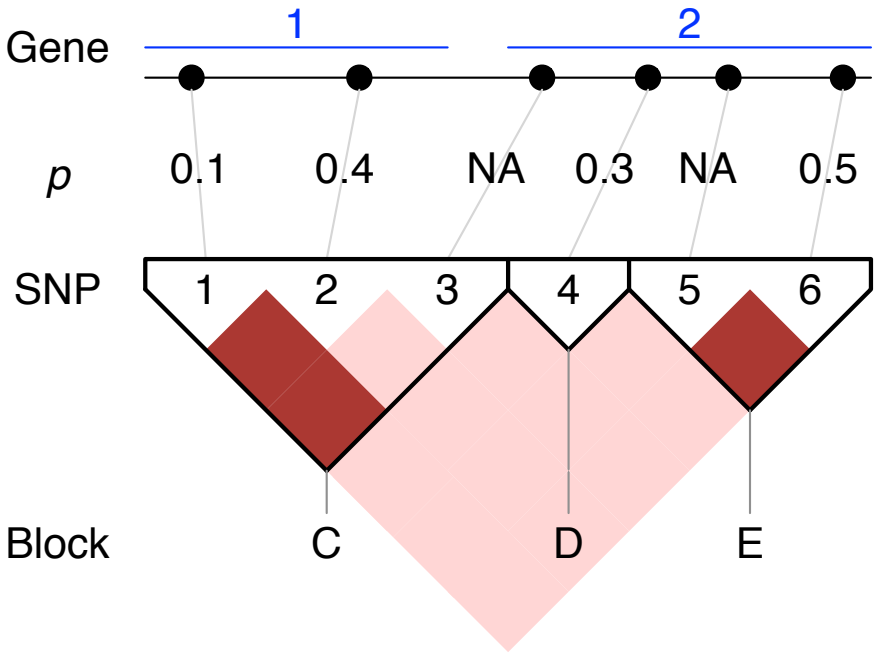

Figure 3.7: Example – calculating the p-values of genes in ProxyGeneLD. Black circles represent SNPs on the chromosome (numbered 1 to 6), red squares indicate high LD between two SNPs, and pink squares indicate low LD between two SNPs. Black borders around groups of SNPs indicate block boundaries. We use the letters C, D and E to refer to the blocks. We represent genes with blue lines at the top of the figure. The variable p is the p-value of each SNP, where NA indicates that the SNP is not a study SNP.

We give an example to demonstrate how ProxyGeneLD calculates gene p-values. We illustrate this example in Figure 3.7. Suppose that there are six HapMap SNPs, where the LD structure is given by the squares in Figure 3.7. As in Figure 3.4, a dark red square indicates high LD between SNPs, and a pink square indicates low LD. Suppose also that the p-value of each SNP is as given in Figure 3.7, where "NA" indicates that the SNP is not a study SNP. Also, suppose that SNPs 1 and 2 have been mapped to gene 1, and SNPs 3 to 6 have been mapped to gene 2.

Gene 1 only contains study SNPs in block A (1 and 2), so the adjustment factor is 1 and the p-value is $\min\{0.1, 0.4\} \times 1 = 0.4$. Gene 2 contains study SNPs in blocks B and C (4 and 6). Note that gene 2 does contain SNP 3, but since it is not a study SNP, we do not include block A in our calculation of the p-value of gene 2. Thus the adjustment factor is 2 and the p-value is $\min\{0.3, 0.5\} \times 2 = 0.6$.

ProxyGeneLD provides three different methods of estimating the p-value of a gene set S from the p-value of each gene in S. One of these methods is GSEA (Subramanian *et al.*, 2005). Yaspan *et al.* (2011) notes that ProxyGeneLD does not require the original genotype data, so we assume that ProxyGeneLD uses GSEAPR, which permutes the gene test statistics rather than the case and control labels on the original genotype data. In GSEA, genes with test statistics larger in magnitude have a greater level of association with the disease. Using a test statistic of $-\log_{10}(p_G)$ for each gene, where p_G is the p-value of the gene G, is inappropriate. This is because ProxyGeneLD multiplies the unadjusted p-value of each gene by the adjustment factor, which may result in genes having p-values greater than one. Consequently, $-\log_{10}(p_G)$ may be negative, and hence the level of association between each gene and the disease does not increase with the magnitude of $-\log_{10}(p_G)$ in general.

The p-value of gene 1 is calculated correctly, but the p-value of gene 2 is not. Hong *et al.* (2009) only consider the study SNPs that are assigned to gene 2. Study SNPs in proxy cluster 1 that are not assigned to gene 2 are not considered, even though proxy cluster 1 contains a study SNP that is assigned to gene 2.

Thus ProxyGeneLD uses a test statistic of

$$r_G = -\log_{10}(p_G) + \log_{10}(p_{\max})$$

for each gene, where $p_{\max}$ is the maximum of the adjusted p-values of all of the genes in the study. Since $r_G \geq 0$ for all genes, this test statistic satisfies the assumptions of GSEA. In the analysis using GSEAPR, Hong *et al.* (2009) used $P = 5000$ permutations and $p = 1$. Hong *et al.* (2009) restricted analysis to gene sets containing 15 to 500 genes, and investigated all gene sets with FDR < 0.25.

ProxyGeneLD also uses DAVID and IPA to calculate the p-value of S from the p-value of each gene in S. Hong *et al.* (2009) classified each gene as significant if its p-value was in the smallest $x\%$ of p-values of all genes. For the analysis using DAVID, Hong *et al.* (2009) used $x \in \{1, 2, 3\}$, and for the analysis using IPA, Hong *et al.* (2009) used $x = 1$.

3.3.5 Association LIst Go AnnoTatOR (ALIGATOR)

Association LIst Go AnnoTatOR (ALIGATOR) (Holmans *et al.*, 2009) classifies each gene as significant if it contains at least one significant SNP. We refer to the set of significant genes as the *significant gene list*. P gene lists are then simulated as follows. Each gene list is simulated by randomly sampling SNPs from the GWA study without replacement and adding to the gene list the genes that the sampled SNP maps to. For each gene list, SNPs are sampled until the number of genes in the gene list is the same as the number of significant genes in the study.[4] Holmans *et al.* (2009) simulate $P_1 = 5000$ gene lists in this way.

Denote by $g_S^{(k)}$ the number of genes that are in both S *and* gene list $k \in \{1, \ldots, P\}$. The p-value of S is then

$$p_S = \frac{1}{P} \times \left| \left\{ k : g_S^{(k)} \geq g_S \right\} \right| \tag{3.4}$$

[4]We assume that if a SNP maps to multiple genes such that the number of genes in the simulated gene list is greater than the number of genes in the significant gene list, then no more SNPs are sampled.

the proportion of simulated gene lists where the number of genes that are in both S and the simulated gene list is no less than the number of genes that are in both S and the significant gene list.

To adjust the p-value of each gene set for testing multiple gene sets simultaneously, ALIGATOR generates B bootstrap sets (of simulated gene lists) as follows. Each set is generated by randomly selecting one of the P simulated gene lists to be the "observed data". P gene lists are then sampled from the rest of the gene lists *with* replacement. Holmans *et al.* (2009) generate $B = 1000$ bootstrap sets in this way. For each bootstrap set, the p-value of S is calculated in the same way as the unadjusted p-value of S. In (3.4), the number of genes that are in both S and the observed data gene list replaces the number of genes that are in both S and the significant gene list, and the bootstrap set of simulated gene lists replaces the original set of simulated gene lists. Denote by $p_S^{(b)}$ the p-value of S calculated using bootstrap set b. The adjusted p-value for gene set S is then

$$\frac{1}{B} \times \left| \left\{ b : \min_{S'} p_{S'}^{(b)} \leq p_S \right\} \right|,$$

the proportion of bootstrap sets where the minimum of the p-values of *all* gene sets is no greater than unadjusted p-value of gene set S.

We demonstrate ALIGATOR with an example, illustrated in Figure 3.8. Suppose that in a GWA study, SNPs in five genes are analysed. The genes are labelled one to five, and significant SNPs are highlighted in yellow. We use ALIGATOR to calculate the p-value of gene sets 1 and 2. We simulate $P = 3$ gene lists by randomly sampling SNPs from the study without replacement, which are shaded in pink. Since there are three genes in the significant gene list, we sample SNPs from the study until each simulated gene list contains three genes. We summarise the number of genes that are in both a given gene set and a given gene list in Table 3.4.

To calculate the p-value of gene set 1, note that there is one gene that is in both gene set 1 and the significant gene list. Furthermore, for two of the three simulated gene lists ($k \in \{2, 3\}$), there is at least one gene in both gene set 1 and

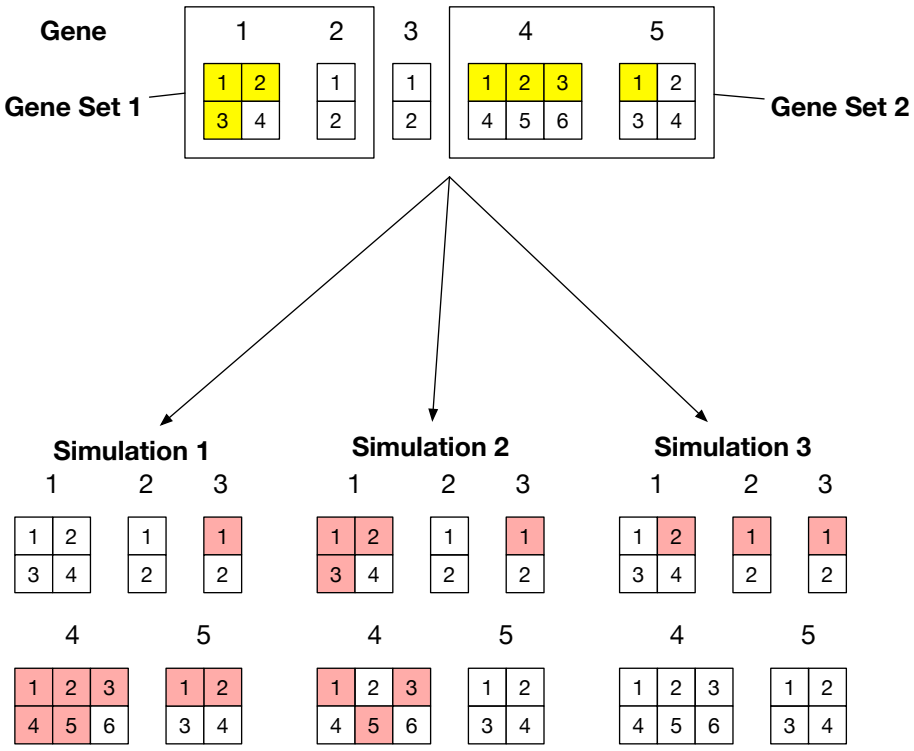

Figure 3.8: Example – calculating the raw p-value of gene sets in ALIGATOR. In the top row of the figure, the SNPs in the genes numbered 1 to 5 are represented by small numbered squares, and significant SNPs are highlighted in yellow. The gene sets are represented by black rectangles around the genes that they contain. In the bottom of the figure, we show three simulated data sets. In each data set, SNPs that were randomly sampled from the study are shaded in pink.

the simulated gene list. Consequently, the p-value of gene set 1 is $\frac{2}{3}$. Similarly, there are two genes that are in both gene set 2 and the significant gene list. Furthermore, for one of the three simulated gene lists ($k = 1$), there is at least two genes in both gene set 2 and the simulated gene list. Consequently, the p-value of gene set 2 is $\frac{1}{3}$.

We also illustrate adjusting these p-values for multiple testing. We use the same example, but we now refer to Figure 3.9, in which we have used slightly different notation. Here, genes in the significant gene list are shaded in yellow,

No. Genes in Gene Set and Gene List	Real Data	$k = 1$	$k = 2$	$k = 3$
Gene set 1	1	0	1	2
Gene set 2	2	2	1	0

Table 3.4: Example – calculating the raw p-value of gene sets in ALIGATOR. Each row corresponds to a gene set, and each column corresponds to a different data set that we use to calculate the raw p-value. The second column corresponds to the real data set, and the three rightmost columns correspond to the three simulated data sets. The entry in each cell of the table is the number of genes in both the gene list (for the given data set) and the given gene set.

and genes in the simulated gene list are shaded in pink. To calculate the adjusted p-value of each gene set, we create $B = 3$ bootstrap sets. For each bootstrap set and each gene set, we calculate the bootstrap p-values $p_S^{(b)}$ in the same way that we calculated the unadjusted p-values. The results of these calculations are shown in Table 3.5. The adjusted p-value for a given gene set is then the proportion of bootstrap sets where the minimum p-value over all gene sets is less than or equal to the unadjusted gene set p-value. For both gene sets 1 and 2, all three bootstrap sets have minimum p-values less than or equal to the unadjusted p-value, so the adjusted p-value of both gene sets is 1.

3.3.6 Modified Gene Set Enrichment Analysis (MGSEA)

Wang *et al.* (2007) use the maxT algorithm to assign a test statistic to each gene in the GWA study from the test statistic assigned to each SNP in the GWA study. GSEA is then used to calculate the p-value of a gene set from the test statistic assigned to each gene in the gene set. GSEA adjusts the p-value of each gene set for testing multiple gene sets simultaneously.

However, Wang *et al.* (2007) comment that the definition of the NES by Subramanian *et al.* (2005) fails to account for the variability of the ESes over all permutations. Consequently, Wang *et al.* (2007) modify the definition of the NES

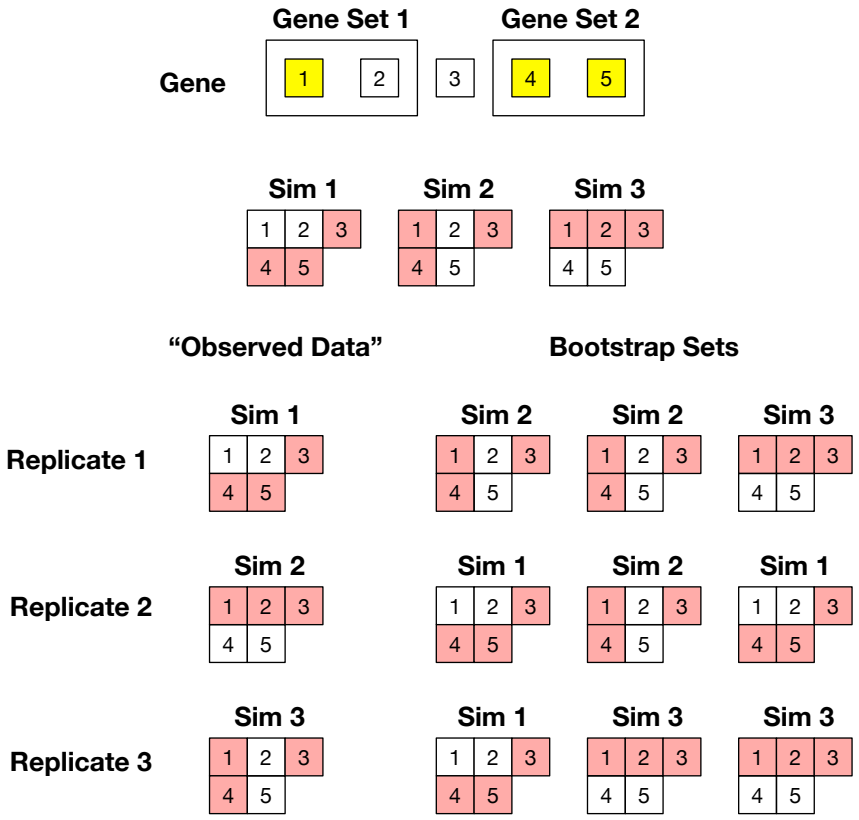

Figure 3.9: Example – correcting for multiple testing in ALIGATOR. We use the same example as in Figure 3.8, but we represent the data in a slightly different way. In the top row of the figure, the numbered squares represent genes, and genes highlighted in yellow are in the significant gene list. Gene sets are represented by boxes around the genes that they contain. The second row from the figure displays the simulated data sets. For each data set, genes highlighted in pink are in the simulated gene list. The remainder of the figure displays three bootstrap replicates in three rows. In each row, the simulated gene list that is selected to be the "observed data" is shown on the left, and the (bootstrap) set of gene lists randomly sampled from the remainder of the simulations is shown on the right.

p-value	Real Data	$b = 1$	$b = 2$	$b = 3$
Gene set 1	$\frac{2}{3}$	1	0	$\frac{2}{3}$
Gene set 2	$\frac{1}{3}$	0	1	$\frac{1}{3}$
Minimum	-	0	0	$\frac{1}{3}$

Table 3.5: Example – correcting for multiple testing in ALIGATOR. Each row, apart from the bottom row, corresponds to a gene set, and each column corresponds to a different data set used in the calculation. The three rightmost columns correspond to the three bootstrap replicates. The entry in each cell of the table, apart from the bottom row, is the raw p-value calculated using the given data set for the given gene set. In the bottom row, we display for each bootstrap replicate the minimum of the p-values over all gene sets.

in Subramanian *et al.* (2005) to account for this variability. In addition to the notation established in Subsection 3.2.2, let

$$s_{\mathrm{ES}(S)} = \left\{ \frac{1}{P-1} \sum_{i=1}^{P} \left[\mathrm{ES}(S, \pi_i) - \overline{\mathrm{ES}}(S) \right]^2 \right\}^{1/2}$$

be the sample standard deviation of the ESes over all permutations π_i. Wang *et al.* (2007) then define the NES of a gene set S calculated using the real data as

$$\mathrm{NES}(S) = \frac{\mathrm{ES}(S) - \overline{\mathrm{ES}}(S)}{s_{\mathrm{ES}(S)}},$$

and the NES of a gene set S calculated using the data simulated from permutation π as

$$\mathrm{NES}(S, \pi) = \frac{\mathrm{ES}(S, \pi) - \overline{\mathrm{ES}}(S)}{s_{\mathrm{ES}(S)}}.$$

3.4 Theoretical Comparison of GSA Methods

We now conduct a theoretical comparison of the six GSA methods that we detailed in Section 3.3. We structure our comparison around seven questions highlighted in reviews of GSA by Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011):

1. How do we define meaningful gene sets to test for association with a disease?
2. What null hypothesis should we test, and how do we calculate a test statistic to test it?
3. Do we need the original genotype data, or only the p-value of each SNP obtained in a GWA study?
4. Do we map SNPs to gene sets directly, or do we map SNPs to intermediaries such as genes or blocks which are then mapped to gene sets?
5. How do we map SNPs to genes?
6. How do we prevent factors such as LD, the number of SNPs in genes and the number of genes in gene sets from impacting the results of GSA in unwanted ways?
7. How do we correct the results of our analysis for testing multiple gene sets simultaneously?

3.4.1 Defining Gene Sets

Recall that the purpose of GSA is to test for association between gene sets and a given disease, in order to elucidate the underlying biology of the disease. Consequently, it is critical that the gene sets analysed in GSA are biologically meaningful.

Two classes of gene set that are commonly analysed using GSA are gene ontologies (GOs) (Ashburner *et al.*, 2000) and biological pathways, such as those found in KEGG (Kanehisa and Goto, 2000) and PANTHER (Mi *et al.*, 2013). Recall that a biological pathway is a set of interacting genes that together perform a specific biological function. In contrast, GOs contain genes that have similar functions, but the relationship between genes in a GO is not known in general. Mooney *et al.* (2014) comments that other choices for gene sets include *biological networks* and *disease biomarkers*.

A biological network, such as a protein-protein interaction network, describes relationships between genes or proteins. However, unlike biological pathways, the genes or proteins in a biological network do not contribute to a biological function. On the other hand, disease biomarkers are sets of genes that have been individually identified as associated with a particular disease. However, genes in a disease biomarker need not interact with each other.

Most GSA methods can theoretically be used to analyse *any* gene set, including PARIS, the SRT, MinP-EVA (MPEVA), ProxyGeneLD, ALIGATOR and Modified GSEA (MGSEA). Consequently, while we recognise the importance of ensuring that gene sets are biologically meaningful, we do not discuss this question any further.

3.4.2 Choosing a Null Hypothesis and Calculating a Gene Set Test Statistic

In any statistical method that calculates a p-value to measure the significance of an association, a precise null hypothesis is critical. In the context of gene expression, Tian *et al.* (2005) formulated two distinct null hypotheses for testing the association between a gene set and a disease:

1. The genes in the gene set show the same level of association with the disease as the genes in the rest of the genome.

2. The genes in the gene set are *not* associated with the disease.

Goeman and Bühlmann (2007) subsequently named these tests *competitive* and *self-contained*, respectively. In other words, competitive tests depend on genes outside of the gene set, whereas self-contained tests do not depend on genes outside of the gene set. We can also consider analagous hypotheses in terms of SNPs or LD blocks.

Many competitive methods can be further divided into *overrepresentation (OR) methods*, and *gene set enrichment (GSE) methods* (Holmans, 2009). These classes describe how a test statistic is calculated for each gene set. In OR methods, units such as SNPs, genes or blocks are classified as significant or nonsignificant based on a criterion such as a significance level α. The number of significant and nonsignificant units in the gene set is then compared with significant and nonsignificant units outside of the gene set. In contrast, GSE methods use the test statistic or p-value of each unit directly to calculate the gene set p-value, without using a significance criterion. The original GSE procedure is GSEA (Subramanian *et al.*, 2005). It is apparent that GSEA is a competitive method, because (3.1) includes both genes in the gene set, and genes not in the gene set.

One advantage of using GSE methods is that they do not require the user to choose an arbitrary significance criterion to classify units such as genes as significant or nonsignificant (Holmans, 2009). In fact, GSEA was designed to overcome this shortcoming (Subramanian *et al.*, 2005). However, GSEA also has a significant weakness compared to OR methods. In GSEA, the ES of a gene set that contains a highly significant SNP or gene is likely to be large (Holmans, 2009; Wang *et al.*, 2007). Consequently, GSEA is more likely to identify gene sets that contain a single, highly significant SNP or gene, and less likely to identify gene sets that contain a number of moderately significant SNPs or genes. This is a weakness of GSEA, because one of the aims of GSA is to identify gene sets that contain a number of moderately significant SNPs or genes, that cannot be identified by GWA studies. In contrast, the results obtained by using OR methods will not be affected by the presence of a single, highly significant SNP or gene. Consequently, OR methods that use liberal significance levels such as $\alpha = 0.05$ may be more suited to GSA (Holmans *et al.*, 2009; O'Dushlaine *et al.*, 2009).

From Section 3.3, PARIS, MPEVA, ProxyGeneLD and ALIGATOR use OR methods; ProxyGeneLD and MGSEA use GSE methods; and the SRT is a self-contained method.

3.4.3 What Input Data are Required for GSA?

Most GSA methods can be classified into two groups: methods that only require the results of a GWA study in the form of a p-value assigned to each SNP, and methods that require the original genotype data of each subject at each genetic locus (Wang *et al.*, 2010; Yaspan *et al.*, 2011). It is important to analyse the type of data that each method requires as input for a number of reasons. For example, the original data is not always available (Holmans *et al.*, 2009; Wang *et al.*, 2010; Yaspan *et al.*, 2011). Furthermore, many methods that require the original data permute the case and control labels on the original data and then perform a GWA study on each simulated data set. Depending on the number of permutations P used, this operation can be computationally expensive (Holmans *et al.*, 2009; Wang *et al.*, 2011; Yaspan *et al.*, 2011).

PARIS, MPEVA, ProxyGeneLD and ALIGATOR only require the results of a GWA study, but the SRT and MGSEA require the original genotype data.

3.4.4 One-Step and Two-Step Methods

Recall that one-step GSA methods calculate a test statistic for each gene set directly from the test statistic or p-value of each SNP, without using intermediaries such as genes or blocks. These methods are also known as *SNP-based* GSA methods (Wang *et al.*, 2011). In contrast, two-step GSA methods calculate a test statistic for each intermediary first. These methods then use these test statistics to calculate a test statistic for each gene set.

One of the benefits of two-step GSA methods that calculate a test statistic for each gene in the analysis is that when the results of the analysis are interpreted, the genes in a gene set can be ranked according to their significance (Fridley and Biernacka, 2011; Mooney *et al.*, 2014). Ranking genes in this way can highlight genes that warrant further investigation. However, such *gene-based* GSA methods require the calculation of a test statistic for each gene. Such calculations can

Method	b
PARIS	50000
The SRT	Not specified
MPEVA	Not specified
ProxyGeneLD	1000
ALIGATOR	$\{0, 20000\}$
MGSEA	500000

Table 3.6: Gene boundary extensions used in different methods

impact the analysis in unwanted ways, due to factors such as the LD structure of the SNPs in the genome and the number of SNPs in genes. We discuss these effects in Subsection 3.4.6.

From Section 3.3, PARIS is a two-step method that uses blocks as intermediaries; the SRT is a one-step method; and MPEVA, ProxyGeneLD, ALIGATOR and MGSEA are two-step methods that use genes as intermediaries.

3.4.5 Mapping SNPs to Genes

The aim of GSA is to estimate the degree of association between a gene set and a disease. Consequently, all GSA methods need to map SNPs to genes. The basic map assigns a SNP to a gene if the position of the SNP is between the position endpoints that define the gene.[5] However, some of the DNA just outside gene boundaries has been shown to play an important role in promoting or inhibiting the production of proteins from genes. Consequently, some methods also map SNPs to genes if the SNP is within b bp of the gene. We display the gene boundary extension b used by the GSA methods in Table 3.6.

For methods where the gene boundary extension $b > 0$, SNPs may map to multiple genes. This can affect the results obtained by the method in an unwanted

[5]Gene boundaries can be found in various databases, such as the Ensembl database (Cunningham *et al.*, 2015).

way. For example, consider calculating the p-value of a gene set containing two genes, where both genes contain a highly significant SNP. If the method calculates a test statistic for each gene, then the significant SNP may be double-counted in the calculation of the gene set p-value. In particular, this problem would affect MGSEA if no amendments were made to it, because it calculates gene test statistics.

However, Wang *et al.* (2007) use the following amendment in MGSEA to circumvent this problem in most circumstances. Assume that a SNP would ordinarily map to two genes when the gene boundary extension is taken into account. If the SNP is not located in either gene (when the gene boundary extension is not taken into account), map the SNP to the closest gene. If the SNP is located in one gene, map the SNP to that gene. However, if the SNP is located in both (overlapping) genes, then map the SNP to both genes. Wang *et al.* (2007) comment that the situation where a SNP is located in both genes occurs very rarely. Consequently, this amendment will almost always be effective.

3.4.6 Accounting for LD and Gene Size

Several factors can impact the results of GSA in unwanted ways, such as the LD structure of the SNPs in the genome and the number of SNPs in genes (Fridley and Biernacka, 2011; Holmans, 2009; Mooney *et al.*, 2014; Ramanan *et al.*, 2012; Wang *et al.*, 2010; Wang *et al.*, 2011). Yaspan *et al.* (2011) confirmed experimentally that these issues can impact GSA if they are not considered carefully. In this subsection, we explain why these issues exist, and how some of the procedures that we detailed in Section 3.2 can impact GSA in unwanted ways. For each GSA method, we then discuss whether or not each factor impacts the results of the method.

To illustrate the impact that LD can have on GSA, suppose that two marker SNPs are associated with a given disease. There are two situations in which this can occur. In the first situation, the LD between the two marker SNPs is low,

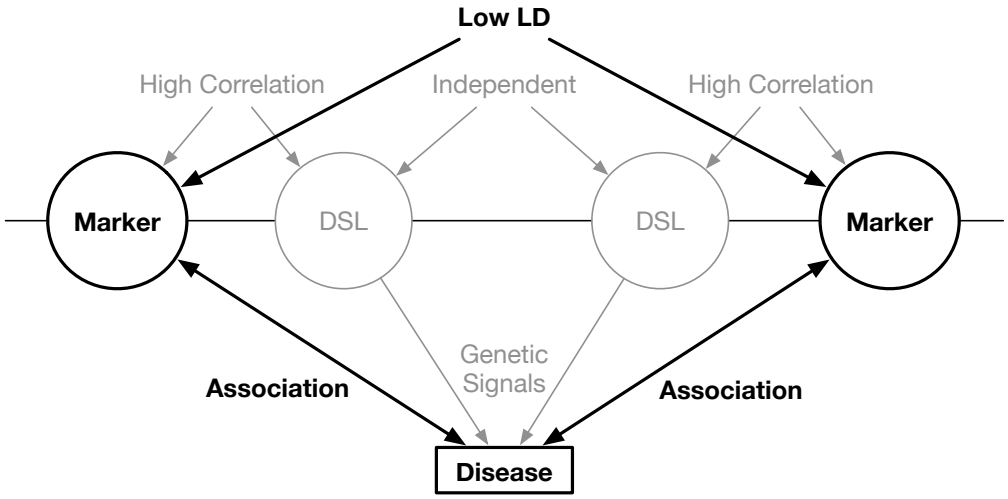

Figure 3.10: Illustration – two SNPs in low LD associated with a disease

as in Figure 3.10. This is usually the case when the two marker SNPs are highly correlated with *different* and functionally independent DSLs. Consequently, the two marker SNPs correspond to separate genetic signals. In the second situation, the LD between the two marker SNPs is high, as in Figure 3.11. This is usually the case when the two marker SNPs are highly correlated with the same DSL. Consequently, the marker SNPs correspond to the same genetic signal.

Furthermore, suppose that the p-value of each marker SNP in the two situations is the same. Then a procedure that combines the p-values of these SNPs without accounting for LD may impact the analysis in unwanted ways. For example, suppose that in each situation, a gene is defined by the two marker SNPs. Then the minP procedure, which is widely used in GSA, assigns the same p-value to both genes (Fridley and Biernacka, 2011; Holmans, 2009; Ramanan *et al.*, 2012; Wang *et al.*, 2011). However, the gene containing the marker SNPs in low LD has a greater degree of association with the disease, because the marker SNPs are highly correlated with different DSLs and correspond to different genetic signals.

To account for LD, some methods, such as PARIS and ProxyGeneLD, partition SNPs into LD and LE blocks. The aim of this procedure is to ensure that each block corresponds to at most one genetic signal. Of course, the effectiveness of

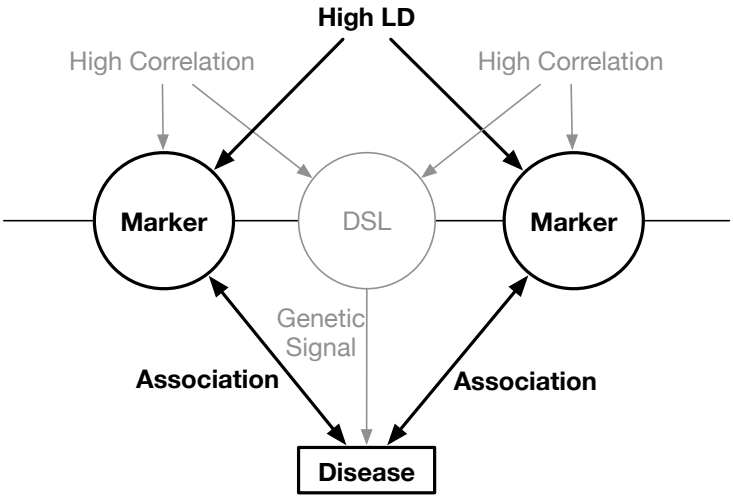

Figure 3.11: Illustration – two SNPs in high LD associated with a disease

this procedure depends on the criteria used to partition the SNPs; this problem is an area of ongoing research.

The number of SNPs in genes and the number of genes in gene sets can also impact the results of GSA in unwanted ways, because larger genes and gene sets are more likely to contain a larger number of significant SNPs and SNPs with smaller p-values. For example, if the minP procedure is used, larger genes are more likely to be assigned smaller p-values by chance (Hong *et al.*, 2009; Wang *et al.*, 2007).

Recall that some GSA methods simulate data sets by permuting the case and control labels on the original data set. These methods then compare a test statistic from the original data to a test statistic calculated using the simulations. This procedure accounts for factors such as LD and gene size, because the calculation of the test statistics uses the same genes and LD structure (Fridley and Biernacka, 2011; Mooney *et al.*, 2014; Ramanan *et al.*, 2012; Wang *et al.*, 2010; Wang *et al.*, 2011).

PARIS

Recall that PARIS partitions SNPs into LD and LE blocks. In particular, in the partitioning method that Yaspan *et al.* (2011) used, a SNP in a block does not need to be in high LD with all other SNPs in the block. Consequently, a single LD block may represent a number of different genetic signals, which may impact the results of PARIS.

A block is then said to be significant if it contains at least one significant SNP. Consequently, blocks with more SNPs are more likely to be significant. However, this does not adversely affect the results of PARIS, because PARIS simulates block collections where the number and size of blocks in each collection are similar to the blocks in the gene set. Also, since PARIS does not calculate gene p-values, it should not be adversely affected by gene size.

The SRT

The SRT simulates data sets by permuting the case and control labels on the original data set. It then compares the number of significant SNPs in each gene set, calculated using the original data, to the number of significant SNPs calculated using the simulated data sets. Consequently, the results of the SRT account for LD and gene size.

MPEVA

MPEVA uses the minP procedure to calculate gene p-values. Consequently, the probability that a randomly chosen gene is significant is not uniform; it depends on LD and gene size (Holmans, 2009). This means that the number of significant genes in each gene set does not follow a hypergeometric distribution. Consequently, the use of FET to calculate gene set p-values is inappropriate; it does not account for the unwanted effects of LD and gene size introduced by the minP procedure.

ProxyGeneLD

Like PARIS, ProxyGeneLD partitions SNPs into blocks. Recall that ProxyGeneLD partitions SNPs into blocks iteratively, where a SNP is added to a block if it is in high LD with any other SNP in the block. Hong *et al.* (2009) use the criterion $r^2 \geq 0.8$ to define high LD. Consequently, the extent to which LD impacts the results of ProxyGeneLD depends on whether or not a single LD block may represent a number of different genetic signals. The results of ProxyGeneLD are also adversely affected by gene size. To see this, recall that the p-value of a gene is calculated using the minP procedure, and adjusted by multiplying by the adjustment factor. This Bonferroni-type adjustment could penalise large genes too much (Yaspan *et al.*, 2011).

ProxyGeneLD uses IPA and EASE to calculate gene set p-values. Recall that these methods use FET, which is inappropriate, since it does not account for the unwanted effect of gene size that has been introduced. ProxyGeneLD also uses GSEAPR to calculate gene set p-values. Since GSEAPR only permutes gene test statistics, it compares the set of test statistics in the gene set to a random set of test statistics. Consequently GSEAPR does not account for the unwanted effect of gene size.

ALIGATOR

ALIGATOR defines a gene to be significant if it contains a significant SNP. Consequently, large genes are more likely to be significant. However, when ALIGATOR simulates gene lists, the probability of adding a gene to the list is proportional to the number of SNPs in the gene. Consequently, ALIGATOR accounts for gene size.

However, Holmans *et al.* (2009) explicitly assume in ALIGATOR that LD between SNPs in a gene set is approximately constant. This assumption is inappropriate, because the LD structure of genomes can be quite complex. Thus the

results of ALIGATOR are adversely affected by LD. Holmans *et al.* (2009) detail an alternative to ALIGATOR which replaces the genes in the simulation with a set of SNPs in low LD with each other. Unlike the partitioning procedures, this procedure omits SNPs from the analysis to create the set of SNPs in low LD. Holmans *et al.* (2009) use a low threshold ($r^2 < 0.2$) to increase the likelihood that no SNPs in the set correspond to the same genetic signal. However, using a low threshold could also reduce the power of the analysis, since fewer SNPs are used.

MGSEA

MGSEA uses the maxT algorithm to calculate gene test statistics. However, the method accounts for LD and gene size, because it compares the ES obtained from the real data with the ESes obtained from data that is simulated by permuting the case and control labels on the real data.

3.4.7 How Does Each Method Correct for Multiple Testing?

Finally, we briefly discuss how some of the GSA methods adjust the p-value assigned to each gene set for testing multiple gene sets simultaneously. Recall that many multiple testing procedures control the FWER, which is the probability of at least one false positive, or the FDR, which is the expected proportion of rejected null hypotheses that are true.

In the context of GSA, controlling the FDR is generally preferable to controlling the FWER (Holmans, 2009). We will not go into further detail here, but the interested reader is encouraged to explore literature such as Holmans (2009), Ramanan *et al.* (2012), and Wang *et al.* (2010) for more details.

PARIS and the SRT do not provide a multiple testing correction. MPEVA uses the Bonferroni correction to test for multiple gene sets, which controls the FWER.

ProxyGeneLD uses GSEA, which provides methods of controlling the FWER and FDR; IPA, which applies the BH procedure (Benjamini and Hochberg, 1995) to control the FDR; and EASE, which provides a variety of different correction procedures, including Bonferroni and bootstrapping procedures. ALIGATOR uses a bootstrapping technique to adjust gene set p-values for multiple testing. However, it is not apparent which measure of error is controlled by the bootstrapping technique used by Holmans *et al.* (2009). Finally, MGSEA uses GSEA, which provides methods of controlling the FWER and FDR.

Of these multiple testing corrections, we favour methods that control the FDR using simulation techniques, such as those used in ALIGATOR and GSEA. The Bonferroni correction conservatively controls the FWER, hence it lacks power to detect gene sets that are associated with the disease. Also, the BH procedure assumes that gene sets are independent, which is inappropriate in GSA.

3.4.8 Overview of GSA Methods

We tabulate the key information about each GSA method that we have discussed in Table 3.7 and make some recommendations. If a self-contained method is preferred, then the SRT is a good choice, because it accounts for LD and gene size. Furthermore, since it uses a SNP significance criterion, it is not sensitive to individual SNPs with very small p-values. However, the SRT requires the original genotype data and it can be computationally intensive. On the other hand, the situation is less clear if a competitive method is preferred. While MGSEA accounts for LD and gene size, it is sensitive to SNPs with very small p-values. Furthermore, like the SRT, it requires the original genotype data and it can be computationally intensive. In contrast, PARIS does not require the original genotype data, and it accounts for gene size. However, PARIS can be affected by LD if a block corresponds to more than one genetic signal. Further research into methods that partition SNPs into blocks, with the aim of ensuring that each block corresponds to at most one genetic signal, is required.

Method	Method Type	Gene Set Test Statistic	Input	Intermediaries
PARIS	OR	Permutation method that accounts for size of blocks	p-values	Blocks
The SRT	Self-Contained	Compares no. of sig. SNPs in gene set from real data with no. from simulations	Genotypes	-
MPEVA	OR	FET	p-values	Genes
ProxyGeneLD	OR and GSE	FET, GSEAPR	p-values	Genes
ALIGATOR	OR	Permutation method where probability of sampling genes proportional to gene size	p-values	Genes
MGSEA	GSE	GSEA	Genotypes	Genes

Method	Gene Boundary Extension	LD[1]	Gene Size[1]	Multiple Testing
PARIS	$b = 50000$	?[2]	✓	-
The SRT	?	✓	✓	-
MPEVA	?	✗	✗	FWER (Bonferroni)
ProxyGeneLD	$b = 1000$	?[2]	✗	FDR (GSEA, BH, Bootstrapping), FWER (GSEA, Bonferroni)
ALIGATOR	$b \in \{0, 20000\}$	✗[3]	✓	Bootstrap method
MGSEA	$b = 500000$	✓	✓	FDR, FWER (GSEA)

[1] Is the method robust to this factor?

[2] Depends on partitioning method.

[3] Original method.

Table 3.7: Theoretical comparison of the six GSA methods

3.5 Chapter Summary

In Section 3.1, we reviewed some of the shortcomings of traditional GWA studies, and we explained why GSA is often used with GWA study data. We explained the mapping problem in Section 3.2, including a discussion of some procedures commonly used in GSA, such as GSEA, FET and the minP procedure. In Section 3.3, we reviewed six GSA methods: PARIS, the SRT, MPEVA, ProxyGeneLD, ALIGATOR and MGSEA. Finally, we conducted a detailed review of these methods in Section 3.4, and we structured our comparison around the reviews by Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011). However, like these reviews, our review has been purely theoretical. Consequently, in Chapter 4, we detail the procedures that we used to implement some of the GSA methods, so that we could compare their performance.

Chapter 4

Procedures to Implement and Compare GSA Methods

The aim of this thesis is to compare PARIS, the SRT, MPEVA, ProxyGeneLD, ALIGATOR and MGSEA. We compared these methods theoretically in Chapter 3. In this chapter, we discuss the procedures that we use to compare the performance of four of these methods at identifying gene sets that are associated with a given disease. In particular, we implement four methods on various sets of simulated gene sets with different properties, and compare the performance of these method for each data set. By simulating gene sets, we are able to choose values for parameters such as the location of genetic loci that are associated with the disease and the relative risks at each locus. In contrast, this information is often not known in real data.

Simulation studies have been used previously in GSA. The study by Jia *et al.* (2011) is the most similar to ours; it compares the SRT, GSEA and a hypergeometric test. However, our study is different for a number of reasons. We compared six methods, including two methods analysed by Jia *et al.*, 2011 (the SRT and GSEA). We also use a different simulation routine to Jia *et al.* (2011). In particular, we use HAPGEN2 (Su *et al.*, 2011) to simulate genotype data, because it can include multiple disease loci on the same chromosome. In contrast,

Jia *et al.* (2011) performed a simulation study by implementing GSA methods on random sets of genes in a real data set. We discuss HAPGEN2 and two methods from which HAPGEN2 was derived in Section 4.1. In Section 4.2, we discuss the values that we choose for parameters in our simulation study, to ensure that the simulated data accurately reflects real data.

To compare GSA methods, we also need to precisely define what is meant by the performance of a method. To do this, we assume that a gene set is associated with a disease if and only if it satisfies particular criteria. For example, if we were analysing a single locus for association with a disease, we could assume that subjects have the disease if and only if they have two disease SNPs (a recessive model) or at least one disease SNP (a dominant model). Similarly, in our simulation study, we assume that a subject with at least x disease SNPs in the gene set has the disease, for some positive integer x. We used empirical receiver operating characteristic (ROC) curves to compare the performance of GSA methods. Empirical ROC curves can be used to evaluate the performance of classification method that classifies a hypothesis as true or false using the p-value of the hypothesis. Other simulation studies have also used empirical ROC curves to compare the performance of GSA methods, including Lu *et al.* (2014) and Zhang *et al.* (2014). We provide more details about the procedures that we used to compare the GSA methods, including empirical ROC curves and our use of them, in Section 4.3.

4.1 Simulating Genotype Data

In Subsection 4.1.1, we detail a method developed by Li and Stephens (2003), the *Li and Stephens (LS) model*, to model the distribution of haplotypes in genotype data. The LS model can be used to simulate haplotypes (sets of alleles on a chromosome) from a set of real haplotypes, such that the LD structure of the real data and the simulated data is similar. This ensures that the LD structure of the simulated data is realistic. In Subsection 4.1.2, we detail HAPGEN (Spencer

et al., 2009), an extension of the LS model that includes one disease SNP in the simulation. HAPGEN2 (Su *et al.*, 2011) is a further extension of HAPGEN that allows for multiple disease SNPs in the simulation. We use HAPGEN2 to simulate genotype data for our simulation study, and we provide details about HAPGEN2 in Subsection 4.1.3.

4.1.1 Simulating Genotype Data with No Disease SNPs

The LS model (Li and Stephens, 2003) can be used to simulate genotype data with a similar LD structure to real genotype data.

In the LS model, suppose that for S biallelic loci on a chromosome, there are n haplotypes $h_1, \ldots, h_n$. We assume that these haplotypes come from $\frac{n}{2}$ diploid individuals, that is, individuals that possess pairs of chromosomes in their cells. The aim of the LS model is to generate the conditional distribution

$$\pi_A(h_{k+1}|h_1, \ldots, h_k)$$

of haplotype h_{k+1} given haplotypes $h_1, \ldots, h_k$, for some $k \in \{1, \ldots, n-1\}$. Consequently, if only haplotypes $h_1, \ldots, h_k$ were known, the LS model could be used to simulate haplotype h_{k+1}.

The LS model uses the following methodology to model the conditional distribution $\pi_A(h_{k+1}|h_1, \ldots, h_k)$. Assume that at each allele, haplotype h_{k+1} copies one of the previous k haplotypes, $h_1, \ldots, h_k$. Let $X_j \in \{1, \ldots, k\}$ be the haplotype that h_{k+1} copies at allele $j \in \{1, \ldots, S\}$. The X_j are referred to as *copying states*. However, the LS model also simulates mutations; haplotype h_{k+1} may copy an allele incorrectly. Consequently, the haplotypes $h_1, \ldots, h_k$ and copying states $X_1, \ldots, X_S$ do not completely determine the new haplotype h_{k+1}.

To motivate the LS model, consider the *Ewens sampling formula* (Ewens, 1972), a well-known model for simulating new haplotypes from a conditional distribution. The Ewens sampling formula assumes that a population satisfies the following:

- The population consists of individuals that randomly select their mate from the population;
- the population has constant size N;
- the population has rate of mutation μ per generation; and
- any mutation results in a completely new haplotype, such that h_{k+1} is not equal to any of $h_1, \ldots, h_k$.

Under these conditions, Ewens (1972) showed that the probability of mutation (and hence a new haplotype) is

$$\frac{\theta}{k+\theta}, \tag{4.1}$$

where $\theta = 4N\mu$. Consequently, the probability that the simulated haplotype is a copy of an existing haplotype is $\frac{k}{k+\theta}$. Ewens (1972) also showed that if an existing haplotype is copied, then the probability of any specific haplotype being chosen follows a discrete uniform distribution, with probability $\frac{1}{k}$.

Li and Stephens (2003) designed the LS model to satisfy the following five properties:

1. Haplotypes that are more common in $h_1, \ldots, h_k$ are more likely to be copied by h_{k+1} if a mutation does not occur.
2. As k, the number of haplotypes, increases, the probability of generating a new haplotype decreases.
3. As θ (which is directly proportional to the mutation parameter μ) increases, the probability of mutation increases.
4. If a mutation occurs that produces a new haplotype, then the new haplotype is far more likely to be similar to the haplotypes that already exist than completely different.

5. These similarities should occur in blocks of alleles, where the size of the block is inversely proportional to the recombination rate in that area of the genome.

Li and Stephens (2003) comment that the Ewens sampling formula captures properties one to three, but it does not satisfy properties four and five. To capture property four, Stephens and Donnelly (2000) designed a model (the *SD model*) such that the number of differences M between the new haplotype and a random existing haplotype followed a geometric distribution, such that $\Pr(M = 0) = \frac{k}{k+\theta}$. Consequently, according to the SD model, one of the haplotypes to copy is chosen randomly (with probability $\frac{1}{k}$), and then the number of mutations is simulated using the geometric distribution. Thus the new haplotype is a (possibly imperfect) copy of one of the existing haplotypes. The model by Fearnhead and Donnelly (2002) (the *FD model*) then extends the SD model to account for property five.

Li and Stephens (2003) use a Markov model to mimic the effects of recombination. In this model, the copying state X_1 has a uniform distribution over all haplotypes, so

$$\Pr(X_1 = x) = k^{-1}$$

for all $x \in \{1, \ldots, k\}$. The Markov transition probabilities are then given by

$$\Pr(X_{j+1} = x' | X_j = x) = k^{-1}(1 - \alpha_{k,j}) + \alpha_{k,j} I\{x' = x\} \tag{4.2}$$

for all $x, x' \in \{1, \ldots, k\}$, where

$$\alpha_{k,j} = \exp\left(-\frac{4Nc_jd_j}{k}\right). \tag{4.3}$$

In other words, with probability $1 - \alpha_{k,j}$, a crossover event occurs between alleles j and $j + 1$. The crossover events split up the haplotype into segments, and the copying state of each segment is sampled uniformly from $\{1, \ldots, k\}$. Note that the copying state of two adjacent segments may be the same.

In (4.3), d_j is the physical distance between loci j and $j+1$ on the chromosome in base pairs and c_j is the average rate of crossover per base pair per meiosis

between loci j and $j+1$. Thus the quantity $c_j d_j$ is the genetic distance between loci j and $j+1$. Also, N is the *effective population size*, which can be thought of as the number of breeding individuals in a population. However, there are a number of definitions of effective population size, and a detailed discussion of them is beyond the scope of this thesis. The interested reader is referred to literature such as Wright (1931) and Wright (1938) for more information.

To simulate mutation events at a given allele on haplotype h_{k+1}, let $h_{i,j}$ denote the j^{th} allele on the i^{th} haplotype, where $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,S\}$. Assume that the alleles $\{h_{k+1,j} : j \in \{1,\ldots,S\}\}$ are conditionally independent, given the haplotypes $h_1,\ldots,h_k$ and the copying states $X_1,\ldots,X_S$. Then the allele $h_{k+1,j}$ is simulated using the conditional probability

$$\Pr(h_{k+1,j} = a | X_j = x, h_1,\ldots,h_k) = \frac{\tilde{\theta}}{2(k+\tilde{\theta})} + \frac{k}{k+\tilde{\theta}} I\{h_{x,j} = a\}, \tag{4.4}$$

where $a \in \{0,1\}$ denotes the allele, $j \in \{1,\ldots,S\}$, and

$$\tilde{\theta} = \left(\sum_{m=1}^{n-1} \frac{1}{m}\right)^{-1} \tag{4.5}$$

is Watterson's point estimate (Watterson, 1975) for $\theta = 4N\mu$. We briefly detail the justification of Li and Stephens (2003) for using (4.4). From (4.1), an estimate for the probability of mutation can be obtained by using Watterson's point estimate: $\frac{\tilde{\theta}}{k+\tilde{\theta}}$. If a mutation occurs, then the allele is sampled uniformly from the wild-type allele and the rare allele. This ensures that as $\tilde{\theta} \to \infty$ and the estimated mutation probability tends to one, the likelihood of both alleles tends to $\frac{1}{2}$ (Li and Stephens, 2003). Thus with probability

$$\nu_k = \frac{1}{2} \times \frac{\tilde{\theta}}{k+\tilde{\theta}} \tag{4.6}$$

the allele is copied incorrectly; the factor of $\frac{1}{2}$ exists because the alleles are sampled uniformly in the event of a mutation. Similarly, with probability

$$\frac{1}{2} \times \frac{\tilde{\theta}}{k+\tilde{\theta}} + \frac{k}{k+\tilde{\theta}} = 1 - \nu_k \tag{4.7}$$

the allele is copied correctly; the first summand corresponds to the event that a mutation occurs, and the second summand corresponds to the event that a mutation does not occur.

As an example, consider the haplotypes in the top panel of Figure 4.1, which is based on Figure 2 of Li and Stephens (2003). The background of each haplotype has been given a different colour, and the black and white circles represent one of the two possible alleles at each genetic locus. In this example, Li and Stephens (2003) simulate haplotypes h_{4A} and h_{4B} from haplotypes h_1, h_2 and h_3. We illustrate two perspectives of the LS model.

In the first perspective, the copying state X_1 at locus 1 has been simulated from a uniform distribution. In the centre-left panel of Figure 4.1, $X_1 = 3$ for both h_{4A} and h_{4B}, as illustrated by the blue rectangles. Then, in the centre-right panel, the rest of the copying states have been simulated from (4.2), so that the colour of the rectangles indicates the copying state at each allele. Thus

$$(X_1, X_2, X_3, X_4, X_5) = \begin{cases} (3,3,2,2,2) \text{ for haplotype } h_{4A} \\ (3,1,2,3,2) \text{ for haplotype } h_{4B}. \end{cases}$$

Alternatively, in the second perspective, crossover events occur between alleles j and $j+1$ with probability $1-\alpha_{k,j}$, which break up the haplotypes into segments. The crossover events are indicated by the black vertical lines in h_{4A} and h_{4B} in the top-right panel. Then, in the centre-right panel, the copying state of each segment is then sampled uniformly from $\{1,2,3\}$. Finally, in the bottom panel, the alleles on each haplotype have been simulated according to (4.4). In particular, both h_{4A} and h_{4B} copied h_2 at locus 3, but the copy is incorrect. That is, a mutation event has occurred at locus 3. However, at all other loci, no mutation event has occurred, and hence the copy is correct.

We refer the interested reader to Li and Stephens (2003) for more details regarding their choice of parameters, and an alternative conditional distribution that performs better in their simulation study. However, we are not concerned

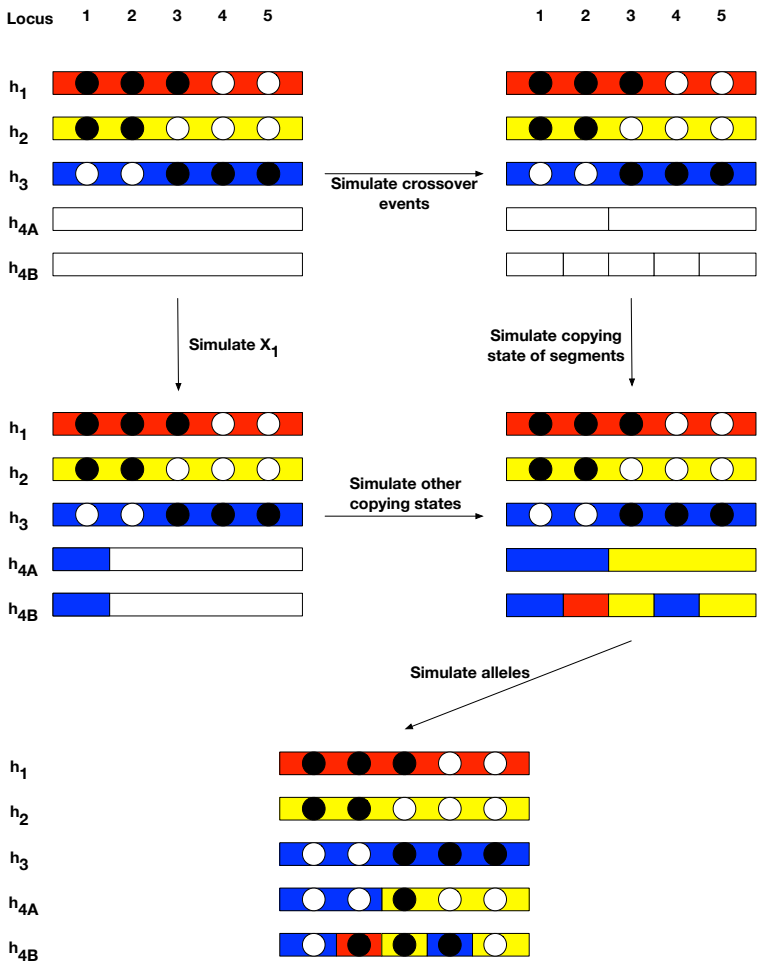

Figure 4.1: Example – using the LS model to simulate two new haplotypes from three existing haplotypes. Each existing haplotype is represented by a different background colour, and the colour of each circle represents one of two possible alleles. We illustrate two perspectives of the LS model. In the first perspective, the copying state at locus 1 for each new haplotype has been simulated in the centre-left panel, and all other copying states have been simulated in the centre-right panel. In the second perspective, crossover events have been simulated on each new haplotype in the top-right panel, and the copying state of each resultant segment has been simulated in the centre-right panel. Finally, the copying state on each new haplotype at each allele has been simulated in the bottom panel.

with the alternative distribution in this thesis, because HAPGEN and HAPGEN2 build on the conditional distribution we have already discussed.

4.1.2 Simulating Genotype Data with One Disease SNP

We now discuss HAPGEN, an extension of the LS model by Spencer *et al.* (2009) designed to include the presence of a disease SNP. HAPGEN can simulate an arbitrary number of controls and cases.

HAPGEN uses the same framework as the LS model to simulate new haplotypes from known haplotypes $h_1, \ldots, h_k$ using copying states. The known haplotypes may include both real and simulated haplotypes. However, unlike the LS model, HAPGEN simulates haplotypes h_{k+1} and h_{k+2} simultaneously, which are assumed to come from a diploid organism. Also, the copying state is initialised at the disease locus $d \in \{1, \ldots, S\}$, *not* the first locus, and it is initialised after simulating the alleles on the two haplotypes at the disease locus, $h_{k+1,d}$ and $h_{k+2,d}$. We now detail the methodology of HAPGEN.

To simulate the alleles at the disease locus, denote by a or 0 the wild-type allele, and denote by A or 1 the minor or rare allele. Assume that the presence of one or two rare alleles at the disease locus in an individual increases the probability of disease. Thus the three possible genotypes G are aa, Aa and AA. Write

$$(h_{k+1,d}, h_{k+2,d}) = \begin{cases} (1,1) & \text{if } G = AA \\ (1,0) & \text{if } G = Aa \\ (0,0) & \text{if } G = aa. \end{cases}$$

Also, denote by p the sample minor allele frequency at this locus, that is, the proportion of haplotypes where the allele at this locus is the minor allele. Denote the event that an individual is a case by D. Assume that controls are random individuals from the population, *not* individuals from the population specifically selected to not have the disease. The procedure that HAPGEN follows to simulate alleles $h_{k+1,d}$ and $h_{k+2,d}$ depends on the disease status of the subject.

Suppose that the simulated individual is a control. Consequently, $h_{k+1,d}$ and $h_{k+2,d}$ are simulated independently, according to the sample allele frequencies. Thus for $b \in \{1, 2\}$,

$$\Pr(h_{k+b,d} = c) = \begin{cases} p & \text{if } c = 1 \\ 1 - p & \text{if } c = 0. \end{cases}$$

Consequently, the genotype probabilities are

$$\Pr(G = g) = \begin{cases} p^2 & \text{if } g = AA \\ 2p(1-p) & \text{if } g = Aa \\ (1-p)^2 & \text{if } g = aa, \end{cases} \tag{4.8}$$

which are used to simulate alleles $h_{k+1,d}$ and $h_{k+2,d}$ for a control.

Now, suppose that the simulated individual is a case. Denote by α and β respectively the penetrance of having one or two disease alleles compared to having no disease alleles. The parameters α and β are also known as the *heterozygote relative risk* and the *homozygote relative risk* respectively. Thus

$$\alpha = \frac{\Pr(D|G = Aa)}{\Pr(D|G = aa)} \tag{4.9}$$

and

$$\beta = \frac{\Pr(D|G = AA)}{\Pr(D|G = aa)}.$$

Then for each genotype $g \in \{AA, Aa, aa\}$, by Bayes' theorem,

$$\Pr(G = g|D) = \frac{\Pr(D|G = g)\Pr(G = g)}{\Pr(D)}. \tag{4.10}$$

Using the law of total probability,

$$\begin{aligned}
\Pr(D) &= \sum_g \Pr(D|G=g)\Pr(G=g) \\
&= \Pr(D|G=aa)\Pr(G=aa) + \Pr(D|G=Aa)\Pr(G=Aa) \\
&\quad + \Pr(D|G=AA)\Pr(G=AA) \\
&= \Pr(D|G=aa)(1-p)^2 + \alpha\Pr(D|G=aa)2p(1-p) \\
&\quad + \beta\Pr(D|G=aa)p^2 \\
&= \Pr(D|G=aa)[(1-p)^2 + 2\alpha p(1-p) + \beta p^2] \\
&= \Pr(D|G=aa)\gamma, \qquad (4.11)
\end{aligned}$$

where $\gamma = (1-p)^2 + 2\alpha p(1-p) + \beta p^2$ for notational brevity. Hence, from (4.10) and (4.11),

$$\Pr(G=g|D) = \frac{\Pr(D|G=g)\Pr(G=g)}{\Pr(D|G=aa)\gamma}.$$

Thus we can calculate the probability $\Pr(G=g|D)$ for $g \in \{AA, Aa, aa\}$. For example,

$$\begin{aligned}
\Pr(G=Aa|D) &= \frac{\Pr(D|G=Aa)\Pr(G=Aa)}{\Pr(D|G=aa)\gamma} \\
&= \frac{\Pr(D|G=Aa)}{\Pr(D|G=aa)}\frac{2p(1-p)}{\gamma} && \text{(from (4.8))} \\
&= \alpha[2p(1-p)]\gamma^{-1} && \text{(from (4.9))} \\
&= 2\gamma^{-1}\alpha p(1-p).
\end{aligned}$$

Similarly,

$$\Pr(G=aa|D) = \gamma^{-1}(1-p)^2$$

and

$$\Pr(G=AA|D) = \gamma^{-1}\beta p^2.$$

Combining these probabilities, we have that

$$\Pr(G=g|D) = \begin{cases} \gamma^{-1}\beta p^2 & \text{if } g = AA \\ 2\gamma^{-1}\alpha p(1-p) & \text{if } g = Aa \\ \gamma^{-1}(1-p)^2 & \text{if } g = aa, \end{cases} \qquad (4.12)$$

which are used to simulate alleles $h_{k+1,d}$ and $h_{k+2,d}$ for a case. In particular, note that when $\alpha = \beta = 1$, $\gamma = (1-p)^2 + 2p(1-p) + p^2 = 1$. Thus from (4.8) and (4.12), the probability distribution for a case reduces to that for a control.

HAPGEN uses the alleles $h_{k+1,d}$ and $h_{k+2,d}$ to initialise the copying state X_d at the disease loci on both haplotypes. For $b \in \{1, 2\}$, X_d is simulated according to

$$\Pr(X_d = x) \propto \begin{cases} 1 - \nu_k & \text{if } h_{x,d} = h_{k+b,d} \\ \nu_k & \text{if } h_{x,d} \neq h_{k+b,d}, \end{cases} \tag{4.13}$$

where ν_k is given in (4.6). In other words, with probability $1 - \nu_k$, the copying state simulated corresponds to the event that $h_{k+b,d}$ is a correct copy, and with probability ν_k, the copying state corresponds to the event that $h_{k+b,d}$ is an incorrect copy. These are the same probabilities as (4.6) and (4.7) in the LS model. However, in the LS model, these probabilities are used to sample the haplotypes from the copying states; here, these probabilities are used to sample the copying states from the haplotypes. Also, more than one copying state x may satisfy the events $h_{x,d} = h_{k+b,d}$ and $h_{x,d} \neq h_{k+b,d}$. Consequently, for each event, the copying state is sampled uniformly from the set of copying states satisfying the event.

Once the copying state at the disease allele is initialised, the copying states to the right of the disease locus are simulated according to (4.2) for $j \in \{d, d+1, \ldots, S-1\}$. Similarly, the copying states to the left of the disease locus are simulated according to

$$\Pr(X_{j-1} = x' | X_j = x) = k^{-1}(1 - \alpha_{j-1}) + \alpha_{j-1} I\{x' = x\}, \tag{4.14}$$

where $j \in \{2, 3, \ldots, d\}$, and $x, x' \in \{1, \ldots, k\}$. Recall that $\alpha_{k,j}$ is the probability that a crossover event does not occur between alleles j and $j+1$. Finally, (4.4) is used to simulate the alleles on h_{k+1} and h_{k+2} at all loci except the disease locus.

We now detail an example of using HAPGEN to simulate two haplotypes on a diploid organism for a case or a control. Suppose that we simulate haplotypes h_4 and h_5 in a diploid organism from haplotypes h_1, h_2 and h_3, as in the top left

panel of Figure 4.2. In this example, the disease locus $d = 3$, with relative risks $\alpha = 2$ and $\beta = 4$. We use the same notation in this figure as in Figure 4.1, except the alleles at the disease locus are indicated by stars.

In the top right panel, we simulate the alleles at the disease locus. Suppose that h_4 and h_5 come from a control. Then we simulate the disease alleles according to (4.8), where p is the frequency of the minor allele, which we denote by A. In the figure, the minor allele is indicated by the white star, so $p = \frac{1}{3}$. Thus, from (4.8), we simulate the disease alleles on h_4 and h_5 according to the probability distribution

$$\Pr(G = g) = \begin{cases} \frac{1}{9} & \text{if } g = AA \\ \frac{4}{9} & \text{if } g = Aa \\ \frac{4}{9} & \text{if } g = aa. \end{cases}$$

However, if haplotypes h_4 and h_5 come from a case, then we simulate the disease alleles according to (4.12), where $\gamma = (1-p)^2 + 2\alpha p(1-p) + \beta p^2 = \frac{16}{9}$. Hence these probabilities are

$$\Pr(G = g|D) = \begin{cases} \frac{1}{4} & \text{if } g = AA \\ \frac{1}{2} & \text{if } g = Aa \\ \frac{1}{4} & \text{if } g = aa. \end{cases}$$

Suppose that the alleles simulated on haplotype h_4 and h_5 are A and a respectively, as in the top right panel of Figure 4.2, and denoted by a white star and a black star respectively. We then simulate the copying states at these alleles in the centre-right panel of this figure. These copying states are simulated according to (4.13). For example, for $b = 1$, which corresponds to haplotype h_4, $h_{x,3} = h_{4,3}$ for $x = 2$, and $h_{x,3} \neq h_{4,3}$ for $x \in \{1, 3\}$. In other words, if h_4 copies h_2 at locus 3, then the copy is correct, but if h_4 copies h_1 or h_3, then the copy is incorrect. Thus the simulation probabilities are

$$\Pr(X_3 = x) \propto \begin{cases} 1 - \nu_3 & \text{if } x = 2 \\ \nu_3 & \text{if } x \in \{1, 3\}. \end{cases}$$

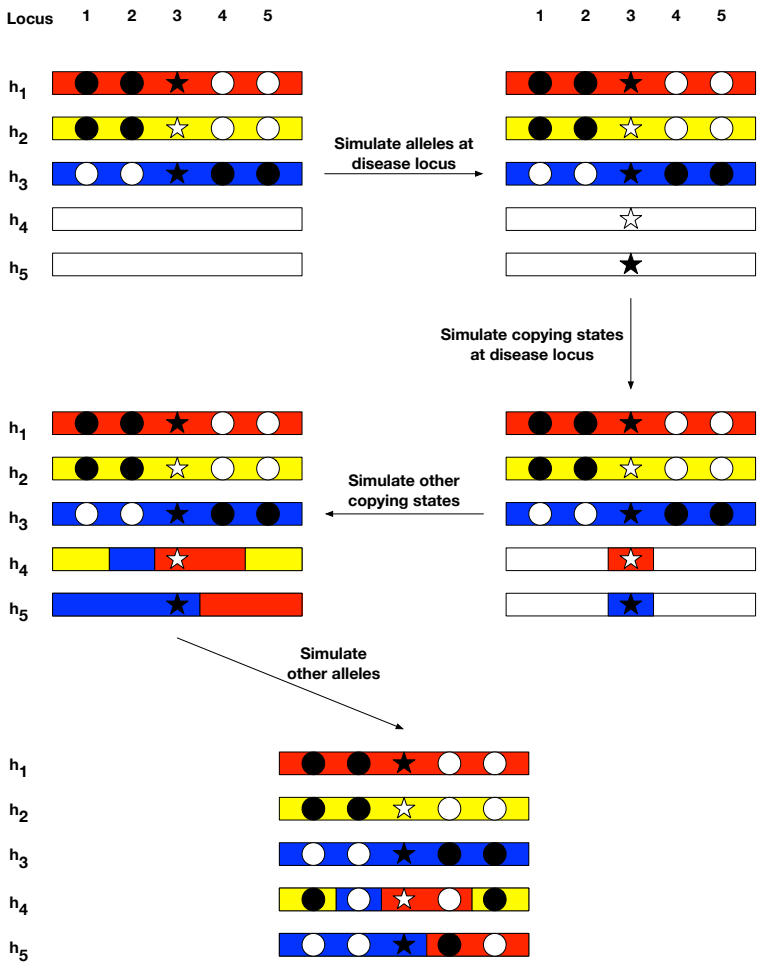

Figure 4.2: Example – using HAPGEN to simulate two new haplotypes from three existing haplotypes. Each existing haplotype is represented by a different background colour, and the colour of each circle and star represents one of two possible alleles. Circles indicate alleles at the non-disease loci and stars indicate alleles at the disease locus. In the top-right panel of the figure, the alleles at the disease locus have been simulated for both new haplotypes. In the centre-right panel, the simulated alleles are used to simulate copying states, which are used to simulate the copying states at the rest of the loci in the centre-left panel. Finally, in the bottom panel, the alleles at the rest of the loci are simulated for both new haplotypes.

Normalising, we have that

$$\Pr(X_3 = x) = \begin{cases} \frac{1}{2}\nu_3 & \text{if } x = 1 \\ 1 - \nu_3 & \text{if } x = 2 \\ \frac{1}{2}\nu_3 & \text{if } x = 3. \end{cases}$$

In the centre-right panel of Figure 4.2, $X_3 = 1$ for h_4 and $X_3 = 3$ for h_5.

In the centre-left panel of the figure, the copying states to the right of X_3 are simulated according to (4.2), and the copying states to the left of X_3 are simulated according to (4.14). Finally, in the bottom panel of this figure, the alleles at the non-disease loci are simulated according to (4.4).

4.1.3 Simulating Genotype Data with Many Disease SNPs

In this subsection, we detail HAPGEN2 (Su *et al.*, 2011), the method that we used to simulate genotype data. Su *et al.* (2011) use somewhat different notation to Li and Stephens (2003) and Spencer *et al.* (2009). Due to the complexity of HAPGEN2, we will use the notation from Su *et al.* (2011) in our description of it.

Denote the set of haplotypes in the real data by

$$H^R = \{h_1, \ldots, h_r\},$$

the set of haplotypes to be simulated for controls by

$$H^P = \{h_{r+1}, \ldots, h_p\},$$

and the set of haplotypes for cases by

$$H^Q = \{h_{p+1}, \ldots, h_q\}.$$

Each haplotype consists of alleles at L genetic loci, so write $h_i = (h_{(i,1)}, \ldots, h_{(i,L)})$. Denote by 0 the wild-type allele and 1 the rare allele.

Suppose that K genetic loci harbor disease SNPs. Denote these loci by $d_k \in \{1, \ldots, L\}$, where $k \in \{1, \ldots, K\}$. Let $D = \{d_k : k \in \{1, \ldots, L\}\}$ be the set of

such loci. Assume that the presence of one or two rare alleles at locus d_k increases the probability of disease, and denote the respective penetrances by rr_k^1 and rr_k^2 respectively. Also, for notational simplicity, let $rr_k^0 = 1$ be the trivial penetrance of having no disease alleles.

HAPGEN2 simulates control haplotypes H^P first, and then case haplotypes H^Q. The simulation of control haplotypes occurs in exactly the same way as the LS Model (Li and Stephens, 2003). We recall the method here, because it will aid our explanation of the way that HAPGEN2 simulates case haplotypes. Consider simulating haplotype $h_{i+1} \in H^P$ from $h_1, \ldots, h_i$. As in Li and Stephens (2003) and Spencer *et al.* (2009), the copying state of haplotype h_{i+1} at locus $j \in \{1, \ldots, L\}$ is the haplotype in $\{h_1, \ldots, h_i\}$ that h_{i+1} copies at locus j. However, Su *et al.* (2011) denote this copying state by $z_{i+1,j}$. The first copying state is simulated uniformly from $\{1, \ldots, i\}$. The rest of the copying states are then simulated using the probabilities

$$\Pr(z_{i+1,j} = z' | z_{i+1,j-1} = z) = \frac{1 - \alpha_{i,j}}{i} + \alpha_{i,j} I\{z = z'\},$$

where $\alpha_{i,j}$ is given in (4.3). However, in this equation, i replaces k. Also, Su *et al.* (2011) define c_j and d_j using loci $j-1$ and j. In contrast, Li and Stephens (2003) and Spencer *et al.* (2009) define these quantities in terms of loci j and $j+1$.

In other words, with probability $\alpha_{i,j}$ no crossover event occurs between $j-1$ and j. If the crossover event does occur, then the new copying state $z_{i+1,j}$ is simulated uniformly from $\{1, \ldots, i\}$. The process of simulating copying states can be considered as the splitting up of haplotype h_{i+1} at each crossover event into segments, $h_{i+1,s_1}, \ldots, h_{i+1,s_n}$, where the copying state of each allele on each segment is the same.

Finally, the allele at each locus is simulated according to the probability

$$\Pr(h_{i+1,j} = h_{z,j} | z_{i+1,j} = z) = 1 - \nu_i,$$

where $\nu_i = \frac{\tilde{\theta}}{2(i+\tilde{\theta})}$ to account for the possibility of a mutation. That is, with probability ν_i allele $h_{i+1,j}$ is copied from allele $h_{z,j}$ incorrectly, where z is the

haplotype being copied at locus j. As in Li and Stephens (2003) and Spencer *et al.* (2009), $\tilde{\theta}$ is Watterson's point estimate (4.5).

We now discuss simulating a pair of haplotypes on a case individual with HAPGEN2. For $b \in \{1, 2\}$, crossover events between alleles $h_{i+b,j-1}$ and $h_{i+b,j}$ on haplotype h_{i+b} are simulated using the probability $1 - \alpha_{i,j}$. Let

$$h_D^b = \{h_{i+b,d_1}, \ldots, h_{i+b,d_k}\}$$

be the set of disease alleles on haplotypes h_{i+1} and h_{i+2} respectively. The crossover events split up h_D^b into n_b segments

$$\left\{h^b_{s_1^b}, \ldots, h^b_{s_{n_b}^b}\right\}.$$

Bayes' theorem is then used to calculate the probability of observing alleles h_D^1 and h_D^2 on the disease loci of haplotypes h_{i+1}, h_{i+2} for a case:

$$\begin{aligned}
\Pr[(h_D^1, h_D^2)|\text{case}] &\propto \Pr[\text{case}|(h_D^1, h_D^2)] \Pr(h_D^1, h_D^2) && (4.15)\\
&= \left[\prod_{k=1}^{K} \Pr(\text{case}|h_{i+1,d_k}, h_{i+2,d_k})\right] \Pr(h_D^1, h_D^2) && (4.16)\\
&= \left[\prod_{k=1}^{K} \Pr(\text{case}|h_{i+1,d_k}, h_{i+2,d_k})\right] \Pr(h_D^1) \Pr(h_D^2) && (4.17)\\
&\propto \left[\prod_{k=1}^{K} rr_k^{h_{i+1,d_k}+h_{i+2,d_k}}\right] \Pr(h_D^1) \Pr(h_D^2) && (4.18)\\
&= \left[\prod_{k=1}^{K} rr_k^{h_{i+1,d_k}+h_{i+2,d_k}}\right] \prod_{j_1=1}^{n_1} \Pr\left(h^1_{s^1_{j_1}}\right) \prod_{j_2=1}^{n_2} \Pr\left(h^2_{s^2_{j_2}}\right), && (4.19)
\end{aligned}$$

where $\Pr(h_s)$ is the marginal frequency of the haplotype segment h_s in the original haplotype data H^R and the controls H^P. Note that the probabilities will be normalised by considering all possible sets of alleles (h_D^1, h_D^2).

We now justify this calculation. (4.15) is equal to (4.16) because HAPGEN2 assumes that the penetrances at multiple disease loci combine multiplicatively. (4.16) is equal to (4.17) because $\Pr(h_D^1, h_D^2) = \Pr(h_D^1) \Pr(h_D^2)$, a consequence of

the fact that the haplotypes on the two chromosomes are independent. (4.17) is proportional to (4.18) because

$$\Pr(\text{case}|h_{i+1,d_k}, h_{i+2,d_k}) \propto rr_k^{h_{i+1,d_k}+h_{i+2,d_k}},$$

for all k, by definition of the penetrances. Finally, for $b \in \{1,2\}$, the set of segments

$$\left\{h^b_{s^b_b}, \ldots, h^b_{s^b_{n_b}}\right\}$$

are separated by crossover events, which means that they are independent. Thus (4.18) is equal to (4.19).

The simulated alleles at the disease loci are used to simulate the copying state for each segment s on haplotypes h_{i+1} and h_{i+2} separately. Consider haplotype h_{i+b}, where $b \in \{1,2\}$. If a segment includes a disease allele, then the copying state is simulated using the probability

$$\Pr(z_{i+b,j} = z) \propto \prod_{d_k : d_k \in s} \nu_{i+b}^{1-I_{i+b,d_k}} (1-\nu_{i+b})^{1-I_{i+b,d_k}}, \tag{4.20}$$

for all $j \in s$ and $z \in \{1, \ldots, i\}$. In this probability, $I_{i+b,d_k} = I\{h_{i+b,d_k} = h_{z,d_k}\}$. In other words, if there are disease alleles present on the segment, then for each disease allele, weight the probability of a given copying state by ν_{i+b} for each incorrect copy corresponding to the copying state, and weight the probability by $1-\nu_{i+b}$ for each correct copy corresponding to the copying state.

Otherwise, if a segment does not include a disease allele, then the copying state is sampled uniformly from $\{1, \ldots, i\}$. Finally, alleles that are not on disease loci are simulated from the respective copying states in the same way as the controls.

We demonstrate HAPGEN2 with an example, illustrated in Figure 4.3. In particular, suppose that we haplotypes h_1, h_2 and h_3 are the real data, and that we simulated h_4 and h_5 for a control. We demonstrate simulating a case using HAPGEN2. As in Figure 4.2, we display alleles at disease loci with stars. In this example, the disease loci are $d_1 = 2$ and $d_2 = 3$. Let the penetrances at $d_1 = 2$ be $rr_1^1 = rr_1^2 = 3$, and let the penetrances at $d_2 = 3$ be $rr_2^1 = 2$ and $rr_2^2 = 4$.

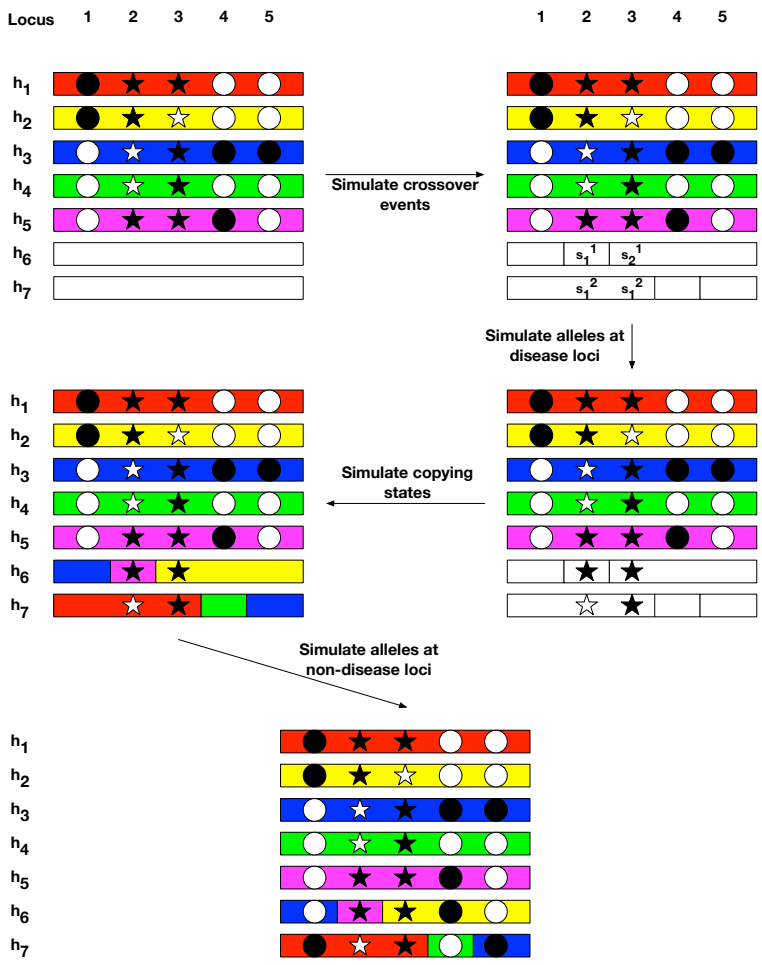

Figure 4.3: Example – using HAPGEN2 to simulate two new haplotypes from five existing haplotypes. Each existing haplotype is represented by a different background colour, and the colour of each circle and star represents one of two possible alleles. Circles indicate alleles at the non-disease loci and stars indicate alleles at the disease loci. In the top-right panel, crossover events are simulated which break up the new haplotypes into segments. In the centre-right panel, the alleles at the disease loci are simulated, and these are used in the centre-left panel to simulate coping states on all loci of the new haplotypes. Finally, the alleles at the non-disease loci are simulated in the bottom panel.

$(h_{6,2}, h_{6,3}, h_{7,2}, h_{7,3})$	$rr_1^{h_{6,2}+h_{7,2}} rr_2^{h_{6,3}+h_{7,3}} \Pr(h_{6,2}) \Pr(h_{6,3}) \Pr(h_{7,2}, h_{7,3})$
$(0,0,0,0)$	$1 \times 1 \times \frac{3}{5} \times \frac{4}{5} \times \frac{2}{5} = \frac{24}{125}$
$(1,0,0,0)$	$3 \times 1 \times \frac{2}{5} \times \frac{4}{5} \times \frac{2}{5} = \frac{48}{125}$
$(0,1,0,0)$	$1 \times 2 \times \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} = \frac{12}{125}$
$(0,0,1,0)$	$3 \times 1 \times \frac{3}{5} \times \frac{4}{5} \times \frac{2}{5} = \frac{72}{125}$
$(0,0,0,1)$	$1 \times 2 \times \frac{3}{5} \times \frac{4}{5} \times \frac{1}{5} = \frac{24}{125}$
$(1,1,0,0)$	$3 \times 2 \times \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} = \frac{36}{125}$
$(0,1,0,1)$	$1 \times 4 \times \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} = \frac{12}{125}$
$(0,0,1,1)$	$3 \times 2 \times \frac{3}{5} \times \frac{4}{5} \times 0 = 0$

Table 4.1: Example – calculating simulation probabilities in HAPGEN2

Simulate crossover events on the top right panel of the figure, where a crossover event occurs between alleles $h_{k+b,j-1}$ and $h_{k+b,j}$ with probability $1 - \epsilon_{k,j}$, for $b \in \{1, 2\}$. These crossover events break up the alleles at the disease loci into segments. In particular, on haplotype h_6, the disease loci are broken up into $n_1 = 2$ segments, $h^1_{s^1_1} = \{h_{6,2}\}$ and $h^1_{s^1_2} = \{h_{6,3}\}$. However, on haplotype h_7, the disease loci lie on $n_2 = 1$ segment $h^2_{s^2_1} = \{h_{7,2}, h_{7,3}\}$. Consequently, we can reduce (4.19) to

$$\begin{aligned} &\Pr(h_{6,2}, h_{6,3}, h_{7,2}, h_{7,3} | \text{case}) \\ &\quad \propto rr_1^{h_{6,2}+h_{7,2}} rr_2^{h_{6,3}+h_{7,3}} \Pr(h_{6,2}) \Pr(h_{6,3}) \Pr(h_{7,2}, h_{7,3}). \end{aligned} \tag{4.21}$$

Recall that $\Pr(h_s)$ is the frequency of haplotype segment h_s in the real haplotypes and the simulated control haplotypes. Denote by 0 the wild-type allele, which is displayed by a black shape in Figure 4.3, and denote by 1 the rare allele, which is displayed by a white shape in the figure. There are $2^4 = 16$ possible values for $(h_{6,2}, h_{6,3}, h_{7,2}, h_{7,3})$, and we display the product of the right-hand side of (4.21) for some of these values in Table 4.1. Of course, the products must be normalised to ensure that (4.21) is a valid probability distribution. We then simulate the alleles on h_6 and h_7 at the disease loci using this probability distribution. We display a possible realisation of this simulation in the centre right panel of Figure 4.3.

In the centre left panel of Figure 4.3, we simulate the copying state of each segment of h_6 and h_7. We only detail some of the simulation probabilities for these haplotypes. From the centre right panel of Figure 4.3, the leftmost segment contains no disease alleles, so the copying state is simulated uniformly from $\{1, \ldots, 5\}$. However, the second segment from the left only contains locus 2. Let $I_{6,2} = I\{h_{6,2} = h_{z,2}\}$ be the indicator of the event that the copy at the disease allele on the segment is correct for a given copying state. Thus

$$I_{6,2} = \begin{cases} 0 & \text{if } z \in \{3,4\} \\ 1 & \text{if } z \in \{1,2,5\}. \end{cases}$$

Hence, (4.20) reduces to

$$\begin{aligned} \Pr(z_{6,2} = z) &\propto \nu_6^{1-I_{6,2}}(1-\nu_6)^{I_{6,2}} \\ &= \begin{cases} \nu_6 & \text{if } z \in \{3,4\} \\ 1-\nu_6 & \text{if } z \in \{1,2,5\} \end{cases} \end{aligned}$$

Thus the simulation probabilities are

$$\Pr(z_{6,2} = z) = \begin{cases} \frac{1}{2}\nu_6 & \text{if } z \in \{3,4\} \\ \frac{1}{3}(1-\nu_6) & \text{if } z \in \{1,2,5\}. \end{cases}$$

We also give an example of simulating the copying state of the first segment of h_7, which we denote by s_2. Let $I_{7,2} = I_{\{h_{7,2}=h_{z,2}\}}$ and $I_{7,3} = I\{h_{7,3} = h_{z,3}\}$ be the indicators of the events that the copies at the disease alleles on the segment are correct. Thus,

$$I_{7,2} = \begin{cases} 0 & \text{if } z \in \{1,2,5\} \\ 1 & \text{if } z \in \{3,4\} \end{cases}$$

and

$$I_{7,3} = \begin{cases} 0 & \text{if } z = 2 \\ 1 & \text{if } z \neq 2. \end{cases}$$

Hence, for all $j \in s_2 = \{1, 2, 3\}$, (4.20) reduces to

$$\Pr(z_{7,j} = z) \propto \nu_7^{1-I_{7,2}}(1-\nu_7)^{I_{7,2}}\nu_7^{1-I_{7,3}}(1-\nu_7)^{I_{7,3}}$$
$$= \begin{cases} \nu_7^2 & \text{if } z = 2 \\ \nu_7(1-\nu_7) & \text{if } z \in \{1, 5\} \\ (1-\nu_7)^2 & \text{if } z \in \{3, 4\}. \end{cases}$$

The simulation probabilities are then normalised as before.

Finally, the alleles on the non-disease loci are simulated from the copying states in the same way as they were for the controls. We illustrate this in the bottom panel of Figure 4.3.

4.2 Implementing HAPGEN2 and the GSA Methods

We now discuss the procedures that we used to simulate gene sets and implement the GSA methods. In particular, we selected realistic values of each parameter that we could vary in our study. We discuss our choices for the values that we used for each parameter, based on the relevant literature.

4.2.1 Some Comments on Implementation

To run HAPGEN2, users need to specify three input files:

1. a file containing real haplotype data on a given chromosome for a number of individuals;

2. a legend file containing *reference SNP (rs)* identifiers from the Database of Short Genetic Variations (dbSNP) (Sherry *et al.*, 2001), position numbers and the two possible alleles at each locus; and

3. a file containing the genetic distance between each SNP.

We used genomic data collected from CEPH (Utah residents with ancestry from northern and western Europe) (abbreviation: CEU) as part of the International HapMap Project (Gibbs *et al.*, 2003). In particular, we used the HapMap 3 (release 2) haplotypes, and we refer these data hereafter as the HapMap CEU Data. To simplify our simulations, we only simulated genotype data from Chromosome three. 96537 SNPs on Chromosome three are included in the HapMap CEU Data.

For each of the six GSA methods, we investigated whether we should use the author's implementation or implement it ourselves. For example PARIS and ProxyGeneLD use data that describes the LD structure of the genome. Consequently, it was easier to try using the implementations of these methods by Yaspan *et al.* (2011) and Hong *et al.* (2009) respectively. We used the implementation of PARIS by Yaspan *et al.* (2011), but the link to the implementation of ProxyGeneLD by Hong *et al.* (2009) in the literature was broken, so we could not use it. The other methods do not require additional input data, so we sought to implement them ourselves in R (R Core Team, 2013). Due to time constraints, we only implemented the SRT, MPEVA, and MGSEA ourselves.

PARIS maps SNPs to genes using SNP positions and gene position ranges from the Ensembl database (Cunningham *et al.*, 2015). In particular, PARIS uses the National Center for Biotechnology Information (NCBI) build 37a, Ensembl release 56, and dbSNP build 130 (Sherry *et al.*, 2001). The dbSNP database defines the *rs* ID for each SNP. However, the HapMap CEU Data that we used is based on NCBI build 36 and dbSNP build 126 (Sherry *et al.*, 2001). Consequently, to compare each method, we converted the positions and *rs* identifiers of each SNP from the databases used in the HapMap CEU Data (which we use in HAPGEN2) to positions and *rs* identifiers in the databases used in PARIS. We performed the necessary data conversions using the following procedure.

We used data from dbSNP to convert the *rs* identifiers on the HapMap CEU Data from dbSNP build 126 to dbSNP build 130. We also obtained a map between SNP *rs* IDs and positions that was consistent with Ensembl release 56 and dbSNP build 130 by running PARIS with a list of gene symbols from Chromosome three. We obtained the list of gene symbols from the Ensembl database using the R package `biomaRt` (Durinck *et al.*, 2005; Durinck *et al.*, 2009). We then compared the position of each SNP that was consistent with Ensembl release 56 with the position given by the HAPGEN2 legend. While the position of each SNP was different, the order of the SNPs using both databases was almost identical. However, the order of three SNPs (*rs*11130263, *rs*11928389 and *rs*13093798) had reversed between databases, and the position of two SNPs had changed completely (*rs*1979334 and *rs*9837104). Consequently, we tagged these SNPs to ensure that we did not include them in any gene sets that we analysed. When we removed these SNPs and reordered the SNPs by position using both databases, the ordering was exactly the same. We then ran PARIS again using the list of gene symbols and "good" SNPs to generate SNP to gene maps that were consistent with the databases that PARIS uses.

The other database that PARIS uses is version 27 of the HapMap LD data. Unfortunately, the corresponding data that provides the genetic distance between SNPs is not available for this LD data, so we decided to use the genetic distance data provided with the HapMap CEU Data. We recognise that this decision may reduce the accuracy and precision of our analyses that use PARIS.

We used the SNP to gene map data provided by PARIS and genotype data simulated using HAPGEN2 to implement each GSA method. Since we only simulated data from Chromosome three, we implemented each method on arbitrary gene sets. We then used R to implement the SRT, MPEVA and MGSEA on our simulated data, and we used the implementation of PARIS by Yaspan *et al.* (2011). For each method, we performed GWA studies by arranging the genotype data at each genetic locus into a 3×2 contingency table. We then used Pearson's

chi-squared test at each locus if the Cochran conditions were satisfied, and FET otherwise. Recall that a contingency table satisfies the Cochran conditions if no cell contains a zero, and if more than 80 per cent of the cells contain a number no less than five. Since PARIS and the SRT do not include a correction for testing multiple gene sets simultaneously, we decided to compare the results obtained by each method without correcting them for multiple testing. For the interested reader, the code is available on request on GitHub.

4.2.2 Selecting Simulation Parameters

As discussed earlier, it is important to select realistic values for parameters in a simulation study. Consequently, we reviewed the necessary literature to inform our choices, and we summarise these values in Table 4.2. In this table, sets of values in braces indicate that we simulated data using all of the values in braces. In particular, we varied the homozygote relative risk r (three values), the number of base pairs b to extend gene boundaries (four values), and the size of gene sets m in terms of the number of genes (four values). Consequently, we simulated $3 \times 4 \times 4 = 48$ data sets for our simulation study. In each simulation, we simulated 2000 gene sets in total: 500 with k disease genes for $k \in \{0, 1, 2, 3\}$. We now justify the choices that we made for the value(s) of each parameter.

The Relative Risks and Disease Model

Spencer *et al.* (2009) conducted a simulation study that used a multiplicative disease model and heterozygote relative risks of 1.3, 1.5 and 1.7. Su *et al.* (2011) also used a heterozygote relative risk of 1.3 in demonstrating HAPGEN2. We decided to simulate genotype data using multiplicative disease models and heterozygote relative risks of 1.2, 1.5 and 1.8, to explore a slightly wider area of the parameter space and to see what effect changing the relative risk had on the performance

Parameter	Values
Homozygote relative risk r	$\{1.44, 2.25, 3.24\}$
Disease model at each genetic locus	Multiplicative
Disease SNPs per disease gene	1
Size of genes in gene sets	$[5, 12]$
Random seed used to select genes to be disease genes	1
Random seed used to select SNPs to be disease SNPs	1
Number of cases and controls	500 each
Number of base pairs b to extend gene boundaries	$\{0, 1000, 20000, 50000\}$
Size of gene sets m in terms of the number of genes	$\{5, 10, 20, 50\}$
Random seed used to select gene sets	1
Number of disease genes k in each gene set	$\{0, 1, 2, 3\}$
Number of gene sets with each level of disease genes	500

Table 4.2: Parameters used to simulate gene sets

of each GSA method. We parameterised our simulations using the homozygote relative risk r and disease model. Thus $r \in \{1.44, 2.25, 3.24\}$.

The Number of Disease SNPs and Disease Genes

We reviewed the literature to select realistic values for the number of disease SNPs and disease genes to include in our study. However, HAPGEN2 often returned segmentation fault errors when we attempted to use it with more than ten disease SNPs. Consequently, to ensure that the code did not return errors, we limited the number of disease SNPs that we used in our simulations. In particular, we simulated genotype data using three disease genes, where each disease gene had one disease SNP in it.

The Number of Base Pairs to Extend Gene Boundaries and the Range of Gene Sizes

As we saw in Chapter 3, the authors of the GSA methods gave quite different recommendations regarding the number of base pairs b to extend gene boundaries to account for possible promoter regions. For example, Holmans *et al.* (2009) used both $b = 0$ and $b = 20000$, Hong *et al.* (2009) used $b = 1000$, Yaspan *et al.* (2011) used $b = 50000$ and Wang *et al.* (2007) used $b = 500000$.

In our simulation study, we measured the size of a gene in terms of the number of SNPs that lie within the positional boundaries of the gene, without taking into account the gene boundary extension detailed above. We adopted this convention to ensure that we could change the gene size range and gene boundary extension parameters separately. Furthermore, we kept the gene size range fixed at $[5, 12]$. That is, we simulated gene sets using genes that contained 5 to 12 SNPs in them, not taking into account the number of base pairs to extend gene boundaries. However, we used gene boundary extensions $b \in \{0, 1000, 20000, 50000\}$ to reflect the choices made by most of the authors of the GSA methods. Nonetheless, when we varied b, we used the same genes in each simulation. Consequently, we were able to vary the effective size of each gene (in terms of the number of SNPs, taking b into account) in the gene set, while preventing our results from being effected by using entirely different genes.

Number of Cases and Controls

The number of cases and controls analysed in a GWA study can vary considerably. For example, O'Dushlaine *et al.* (2009) applied the SRT to a GWA study that genotyped 900 cases and 867 controls, Holmans *et al.* (2009) applied ALIGATOR to a GWA study that genotyped 1748 cases and 2953 controls, and Wang *et al.* (2007) applied MGSEA to a GWA study that genotyped 267 cases and 270

Method	Parameter	Values
PARIS	SNP significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
PARIS	Approximate bin size B	$\{500, 1000, 2000\}$
PARIS	Random seed to sample block collections	$\{1, 2, 3, 4, 5\}$
PARIS	Number of block collections to sample	1000
The SRT	SNP significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
The SRT	Number of data sets to simulate	1000
MPEVA	Gene significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
MGSEA	GSEA parameter p	1
MGSEA	Number of data sets to simulate	1000

Table 4.3: Parameters used in GSA methods

controls. We decided to simulate genotype data for 500 cases and 500 controls in all of our simulations.

The Size of Gene Sets in Terms of the Number of Genes

The size of gene sets analysed in a GWA study can vary considerably, too. For example, one of the gene sets that Yaspan *et al.* (2011) analysed, KEGG hsa:00072, has nine genes in it, and the GO categories identified by Holmans *et al.* (2009) as associated with Crohn's disease contained between 3 and 492 genes. We analysed gene sets containing $m \in \{5, 10, 20, 50\}$ genes in them, to investigate the effect of changing gene set size on the performance of each method.

4.2.3 Selecting Method Parameters

Recall that there are parameters in PARIS, the SRT, MPEVA and MGSEA that can be varied. We display in Table 4.3 the parameters in each method and the values that we chose for them. We make some brief comments about these parameters.

For the methods that use a significance criterion, we saw in Chapter 3 that choosing an appropriate criterion is important. Consequently, we decided to use PARIS, the SRT and MPEVA with a range of significance criteria. For simplicity, we used the same criteria for all of these methods. We followed the recommendation of O'Dushlaine *et al.* (2009) in using $\alpha \in \{0.001, 0.005, 0.01, 0.05\}$. Similarly, PARIS, the SRT and MGSEA all use a simulation procedure to calculate an empirical p-value for each gene set. Each author used 1000 simulations, so we did the same.

Yaspan *et al.* (2011) recommend using an approximate bin size $B = 10000$ in PARIS. However, the recommendation is based on analysing data from all chromosomes. In contrast, our simulation study only uses genotype data from Chromosome three, so we tried using smaller bin sizes. We also varied the random seed used in PARIS to sample the random block collections, even though we expect changing the seed to have a negligable effect on the results obtained by PARIS. Finally, Subramanian *et al.* (2005) recommend using $p = 1$ in GSEA, so we did the same in MGSEA.

4.3 Comparing the GSA Methods

We now discuss the methods that we used to compare PARIS, the SRT, MPEVA and MGSEA. We display and discuss our results in Chapter 5.

To analyse the performance of the GSA methods, we classed gene sets as *non-disease* or *disease*, according to the number of disease genes in them. Recall that we analysed 2000 gene sets in each simulation, where 500 had k disease genes in them for $k \in \{0, 1, 2, 3\}$. Consequently, in each simulation, we classed a gene set as *disease* if it contained at least d disease genes, where $d \in \{1, 2, 3\}$.

We can interpret d in the context of biological pathways as follows. Recall that biological pathways are sets of genes that together perform a specific biological function. Furthermore, a biological pathway has redundancy if it can perform its

function even when a protein that was produced by a gene in the pathway does not function correctly. Consequently, a pathway with a high level of redundancy corresponds to a large value of d.

4.3.1 ROC Curves

We compared the performance of each method at classifying gene sets using empirical ROC curves. We briefly define and discuss ROC curves. For the interested reader, a more detailed discussion of ROC curves can be found in literature such as Metz (1978).

Many hypothesis tests reject a null hypothesis H_0 if and only if the p-value is no greater than a significance level $\alpha \in [0, 1]$. Let

$$F_0(\alpha) = \Pr(\text{reject } H_0 \text{ at level } \alpha | H_0 \text{ true}), \text{ and}$$
$$F_1(\alpha) = \Pr(\text{reject } H_0 \text{ at level } \alpha | H_0 \text{ false}).$$

An *ROC curve* plots $F_1(\alpha)$ on the y-axis against $F_0(\alpha)$ on the x-axis for all $\alpha \in [0, 1]$.

The probabilities F_0 and F_1 are often unknown. However, in a simulation study, these probabilities can be estimated empirically by repeating the test many times when the null hypothesis is true and when the null hypothesis is false. We say that the *false positive rate (FPR)* is the proportion of true null hypotheses that are incorrectly rejected, and the *true positive rate (TPR)* is the proportion of false null hypotheses that are correctly rejected. An *empirical ROC curve* plots the TPR against the FPR for each α. However, since there are a finite number of points on an empirical ROC curve, they are often connected with straight lines. These straight lines interpolate the TPR and FPR, but it is important to remember that the interpolated points on an empirical ROC curve cannot be reached by any significance level.

Tests that perform well have empirical ROC curves that contain points in the upper left-hand corner of the unit square, close to the point $(0, 1)$, and the

area under the ROC curve (AUC) is close to 1. In contrast, mediocre tests have empirical ROC curves that lie close to the line FPR = TPR, which corresponds to a test that classifies by random chance. The AUC for a mediocre test is close to $\frac{1}{2}$. In Figure 4.4, we display examples of ROC curves corresponding to a test with good performance and a test with mediocre performance. In this figure, the x-coordinate is the FPR and the y-coordinate is the TPR. The colour of the empirical ROC curve corresponds with the performance of the test.

4.3.2 Using ROC curves to Compare GSA methods

To compare PARIS, the SRT, MPEVA and MGSEA, we first used empirical ROC curves to determine the method parameters, listed in Table 4.3, that only changed the performance of each method negligibly. To reduce the number of analyses that we needed to perform, we fixed the gene set size $m = 5$ and gene boundary extension $b = 0$ in these simulations. Consequently, when we compared the methods with each other, we only varied the parameters that markedly affected the performance of each method. We display and discuss these results in Section 5.1.

We then compared PARIS, the SRT, MPEVA and MGSEA by producing scatterplots of the AUC obtained by each method. We implemented these methods on gene sets simulated using the parameters in Table 4.2 to elucidate the relationship between the performance of each method and the parameters gene set size m, gene boundary extension b and homozygote relative risk r. We display and discuss scatterplots of the AUC in Section 5.2.

4.3.3 Other Procedures to Compare GSA Methods

We also tested the sensitivity of PARIS, the SRT and MGSEA to SNPs with very small p-values. To conduct these tests, we identified gene sets that were assigned disparate p-values by the GSA methods. We then produced Manhattan plots to display the p-values of SNPs in these gene sets. Manhattan plots display

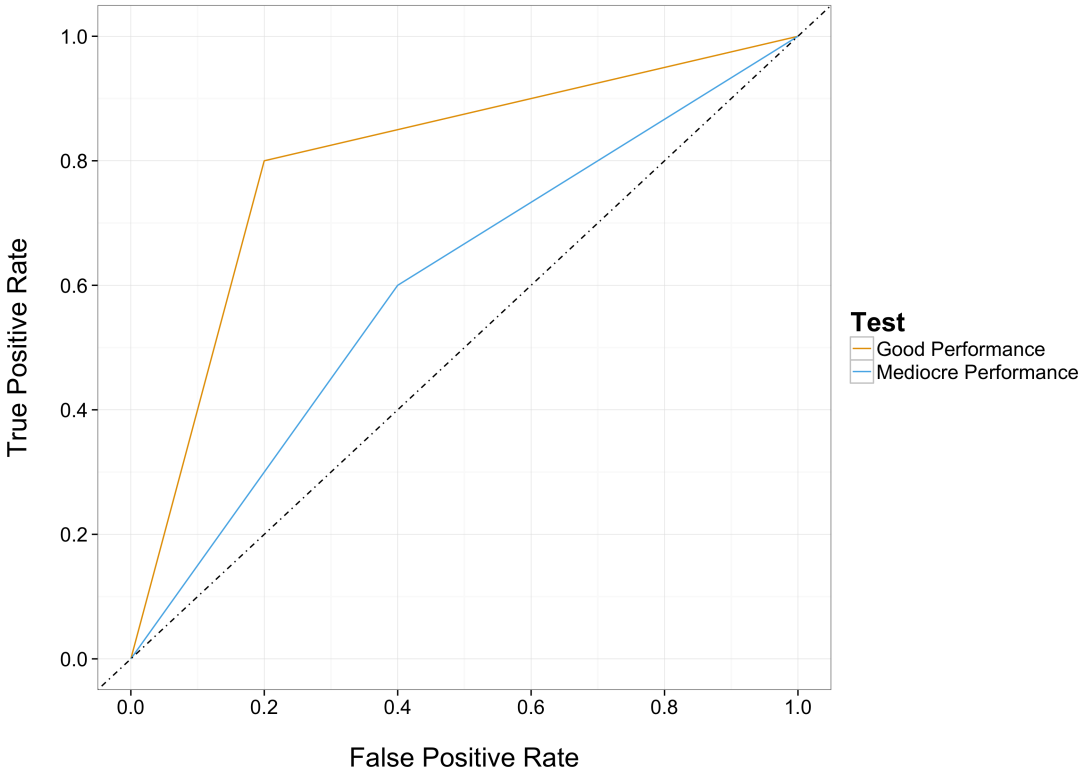

Figure 4.4: Example – empirical ROC curves corresponding to tests with good performance and mediocre performance. The x-coordinate is the FPR and the y-coordinate is the TPR. The colour of the empirical ROC curve corresponds with the performance of the test. These empirical ROC curves are made up of only two straight lines to emphasise the property that all empirical ROC curves are made up of straight lines. Furthermore, in some of our results, the empirical ROC curves are also made up of a small number of straight lines. However, many empirical ROC curves are made up of a larger number of straight lines.

$-\log_{10}(p)$ on the y-axis, where p is the SNP p-value, and SNP position on the x-axis. We display and discuss these Manhattan plots in Section 5.3.

4.4 Chapter Summary

In Section 4.1, we discussed three methods of simulating genotype data: the LS model (Li and Stephens, 2003), HAPGEN (Spencer *et al.*, 2009) and HAPGEN2 (Su *et al.*, 2011). We used HAPGEN2 in our simulation study, because it can simulate genotype data with multiple disease loci. However, we deferred the details of HAPGEN2 to Subsection 4.1.3. In a simulation study, it is important to select values of parameters to ensure that the simulated data are realistic. Consequently, in Section 4.2, we discussed the values that we chose for each parameter. We also detailed the procedures that we used to implement the simulation methods and the GSA methods. Finally, we discussed the procedures that we used to compare the performance of the GSA methods in Section 4.3. We display and discuss the results that we obtained in Chapter 5.

Chapter 5

Results

In this chapter, we display and discuss results about the performance of PARIS, the SRT , MPEVA and MGSEA. In Section 5.1, we display empirical ROC curves to analyse the effect of changing the value of each method parameter on the performance of the method. In particular, we present the method parameters that markedly affected the performance of the methods. We display the method parameters that we used in Table 5.1.

We then display scatterplots of the AUC obtained by each GSA method in Section 5.2 for various values of the gene set size m, gene boundary extension b and homozygote relative risk r, as detailed in Table 5.2. We use these scatterplots to recommend the GSA method that gives the best overall performance.

Finally, in Section 5.3, we display Manhattan plots to investigate gene sets that were assigned disparate p-values by different methods. In particular, we test the sensitivity of different methods to SNPs with very small p-values.

To assist our discussion of the results, we display in Figure 5.1 Manhattan plots for the genotype data simulated using the parameters in Table 5.2. In each plot in this figure, each point represents a SNP. The x-coordinate of each point represents the position of the SNP on chromosome three, and the y-coordinate of each point is $-\log_{10}(p)$, where p is the SNP p-value obtained in the GWA study.

Method	Parameter	Values
PARIS	SNP significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
PARIS	Approximate bin size B	$\{500, 1000, 2000\}$
PARIS	Random seed to sample block collections	$\{1, 2, 3, 4, 5\}$
PARIS	Number of block collections to sample	1000
The SRT	SNP significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
The SRT	Number of data sets to simulate	1000
MPEVA	Gene significance level α	$\{0.001, 0.005, 0.01, 0.05\}$
MGSEA	GSEA parameter p	1
MGSEA	Number of data sets to simulate	1000

Table 5.1: Parameters used in GSA methods

Parameter	Values
Homozygote relative risk r	$\{1.44, 2.25, 3.24\}$
Disease model at each genetic locus	Multiplicative
Disease SNPs per disease gene	1
Size of genes in gene sets	$[5, 12]$
Random seed used to select genes to be disease genes	1
Random seed used to select SNPs to be disease SNPs	1
Number of cases and controls	500 each
Number of base pairs b to extend gene boundaries	$\{0, 1000, 20000, 50000\}$
Size of gene sets m in terms of the number of genes	$\{5, 10, 20, 50\}$
Random seed used to select gene sets	1
Number of disease genes k in each gene set	$\{0, 1, 2, 3\}$
Number of gene sets with each level of disease genes	500

Table 5.2: Parameters used to simulate gene sets

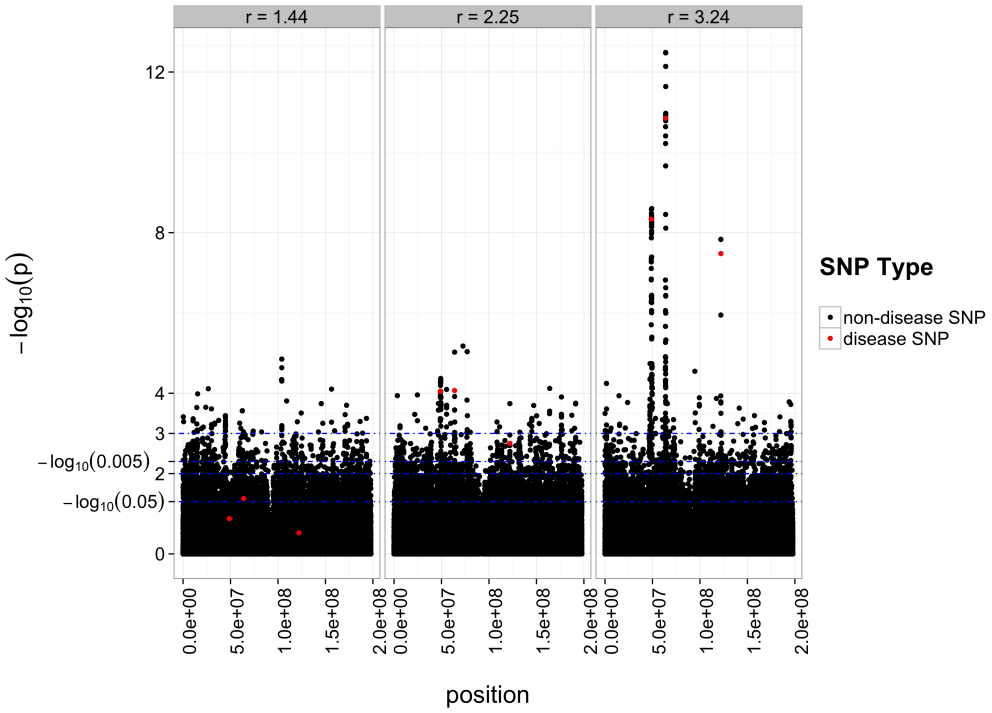

Figure 5.1: Manhattan plots obtained from genotype data simulated using HAPGEN2. The three panels correspond to different values of the homozygote relative risk r. The x-coordinate is the position of each SNP on the chromosome, and the y-coordinate is $-\log_{10}(p)$ on the y-axis, where p is the SNP p-value. Non-disease SNPs are represented by black circles, and disease SNPs are represented by red circles.

Furthermore, we use black circles to indicate non-disease SNPs and red circles to indicate disease SNPs. Each panel in the figure contains a Manhattan plot for the genotype data simulated using a different value of the homozygote relative risk r.

In the Manhattan plots, we expect p to be close to zero for disease SNPs, and uniformly distributed on $(0, 1)$ for non-disease SNPs. Consequently, we expect that $-\log_{10}(p)$ to be large for disease SNPs, and between 0 and 2 for most non-disease SNPs. From Figure 5.1, for $r \in \{1.44, 2.25\}$, the p-values of the non-disease SNPs are approximately uniformly distributed. However, for $r = 1.44$, $-\log_{10}(p)$ is not large at all for the disease SNPs. Also, for $r = 3.24$, there are a considerable number of non-disease SNPs very close to the disease SNPs with very small p-

values, which are not following a uniform distribution. This is probably a result of LD between SNPs that are close to each other on the chromosome. Consequently, we can consider these non-disease SNPs as markers for the respective disease SNPs. However, in the following results, we assume that markers are false positive. If the aim of the analysis is to identify genetic factors that cause the disease, then this assumption makes sense, because markers are merely correlated with the disease. However, if the aim of the analysis is diagnosis or prediction, then our assumption is invalid, because genetic markers are critically important in diagnosis or prediction.

As we discuss, these properties of the simulated genotype data impacted our analyses.

5.1 Analysing the Effect of GSA Method Parameters on Performance

In our simulation study, we analysed the effect of changing method parameters on the performance of each GSA method. To conduct this part of the study, we used the parameters in Table 5.2, although we fixed the gene set size $m = 5$ and the gene boundary extension $b = 0$.

5.1.1 PARIS: Varying the Seed and Bin Size

In our simulation study, we found that changing the approximate bin size B and seed used by PARIS only negligibly changed its performance. This is consistent with the comment by Yaspan *et al.* (2011) that the results of PARIS are robust to the choice of bin size. We display in Figure 5.2 empirical ROC curves that indicate the performance of using PARIS with approximate bin sizes $B \in \{500, 1000, 2000\}$ on simulated gene sets. In this figure, we fixed the seed to 1 and the SNP significance level $\alpha = 0.05$. Each column of panels corresponds to changing the value

of r, the homozygote relative risk. Each row of panels corresponds to a different value of d, the minimum number of disease genes in a disease gene set. And in each panel, we display empirical ROC curves corresponding to using PARIS with different bin sizes.

Similarly, in Figure A.1, we plot empirical ROC curves that display the performance of using PARIS with random seeds of 1, 2, 3, 4 and 5 on simulated gene sets. In this figure, we fixed the SNP significance level $\alpha = 0.05$ and the approximate bin size $B = 1000$. This figure demonstrates that changing the random seed also had a negligible effect on the performance of PARIS, as we expect.

Consequently, we used PARIS with a seed of 1 and approximate bin size $B = 1000$ in the rest of our analyses.

5.1.2 PARIS: Changing the SNP Significance Level

Empirical ROC Curves

Unlike changing the seed and bin size, the performance of PARIS changed markedly when we changed the SNP significance level α. We display empirical ROC curves that illustrate these differences in Figure 5.3 and make several comments about them. In these comments, we measure performance using the AUC.

From Figure 5.3, no significance level α gives the best performance overall. For example, from the left column of the figure, using $a = 0.05$ gives the best performance when $r = 1.44$. Also, from the middle and right columns of Figure 5.3, when $r \in \{2.25, 3.24\}$ and $d = 1$, the performance of PARIS with $\alpha \in \{0.001, 0.005\}$ is generally better than the performance of PARIS with $\alpha \in \{0.01, 0.05\}$. However, this relationship reverses as d increases.

The result for $r = 1.44$ is consistent with the left panel of Figure 5.1: when $r = 1.44$, two disease SNPs have p-values greater than 0.05, and the other has

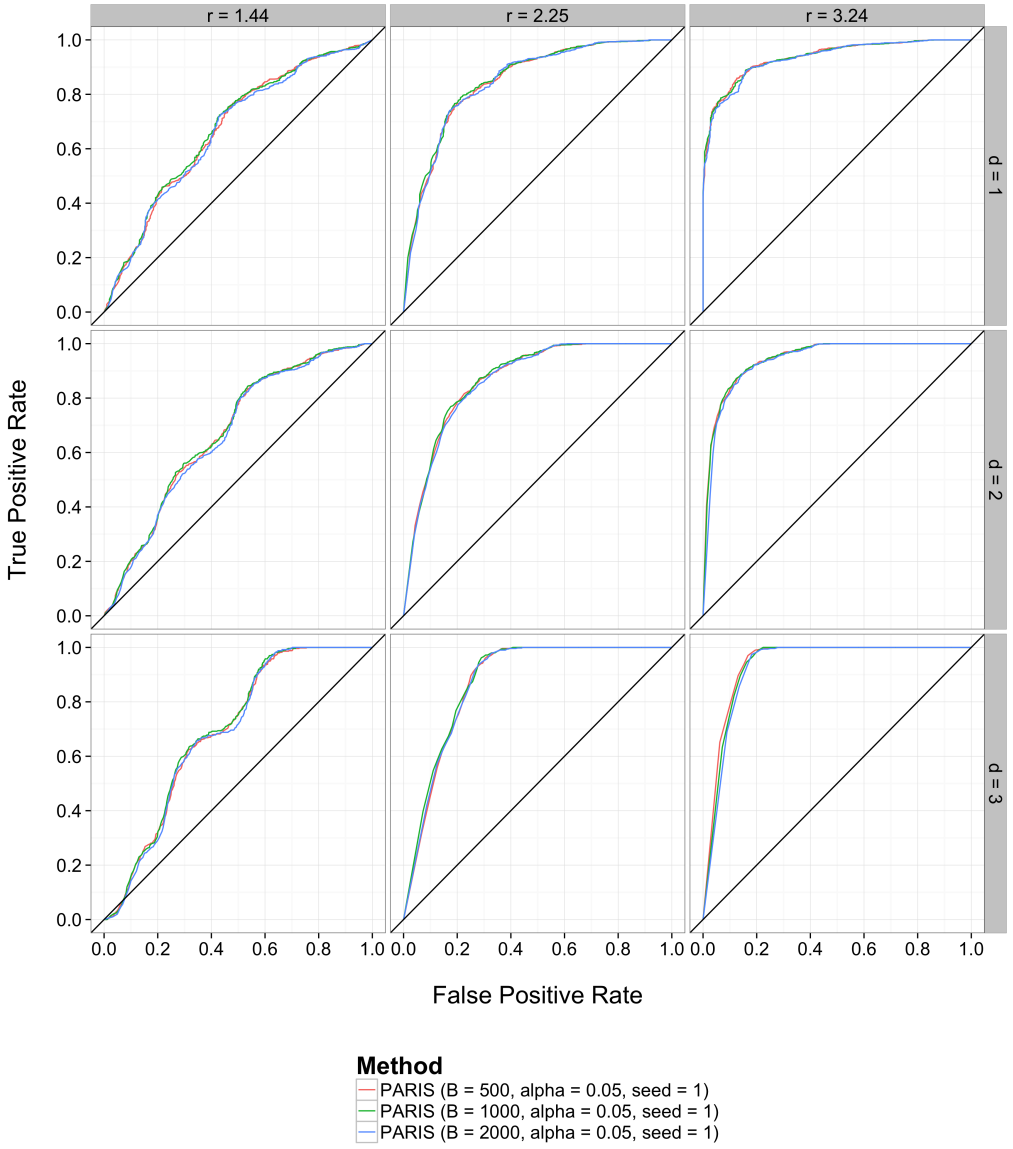

Figure 5.2: Empirical ROC curves obtained from using PARIS with a seed of 1, SNP significance level $\alpha = 0.05$ and different bin sizes on gene sets simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and each column corresponds to a different value of r. The approximate bin size B is indicated by the colour used to display the empirical ROC curve.

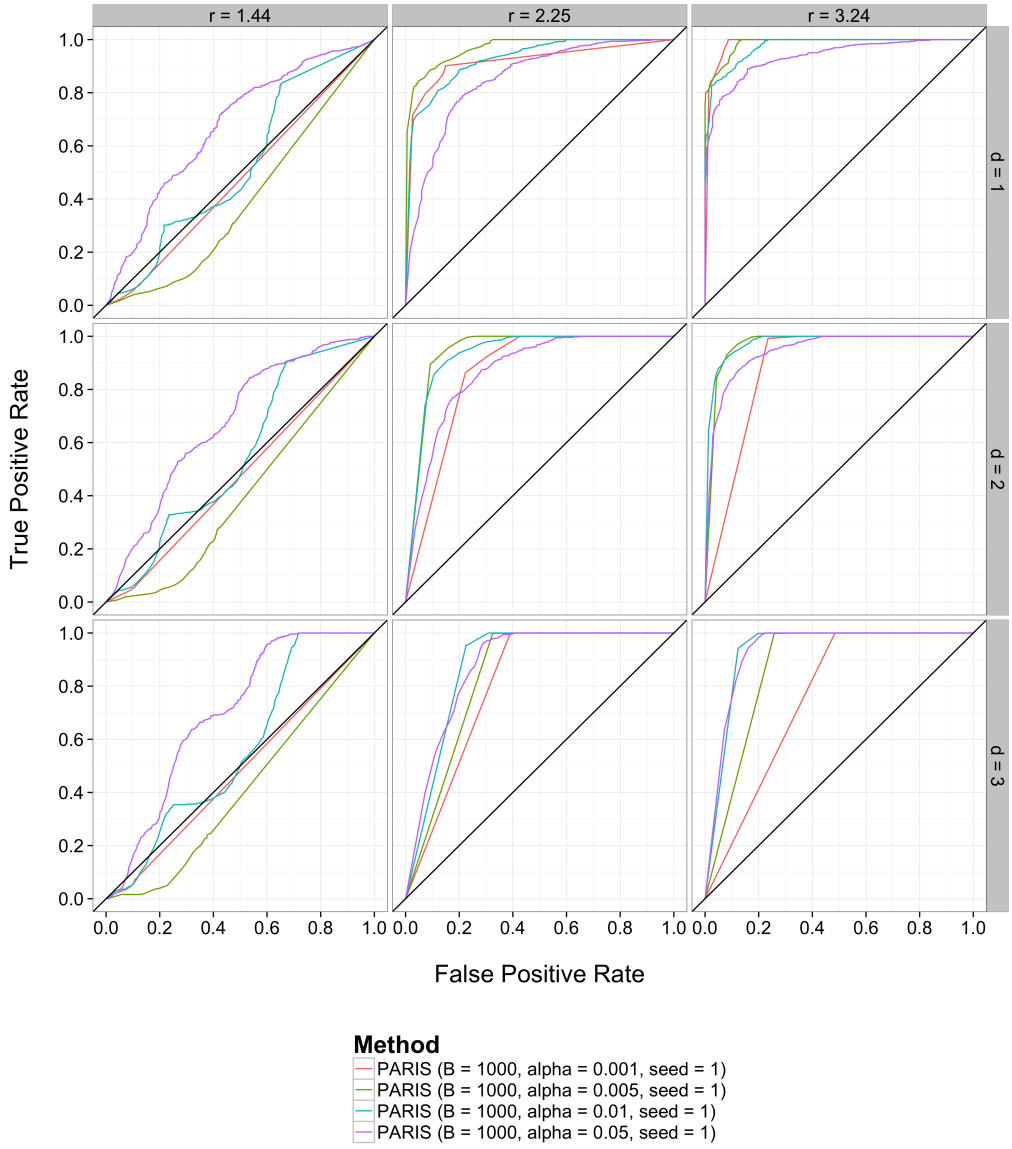

Figure 5.3: Empirical ROC curves obtained from using PARIS with a seed of 1, approximate bin size $B = 1000$ and different SNP significance levels on gene sets simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and each column corresponds to a different value of r. The SNP significance level α is indicated by the colour used to display the empirical ROC curve.

a p-value between 0.01 and 0.05. Thus the only significance level that correctly classifies a disease SNP as significant is 0.05.

Frequency Polygons

Some of the empirical ROC curves in Figure 5.3 have the notable property that they are made up of obvious straight lines. We explored this feature further by producing histograms and frequency polygons of the p-values of the gene sets, which we display in Figure 5.4. When we display such plots, each column of panels corresponds to a different value of r, and each row corresponds to the p-values obtained using a different method. The histograms display the distribution of p-values in all gene sets, and the frequency polygons display the distribution of p-values in the gene sets with k disease genes in them, for $k \in \{0, 1, 2, 3\}$. We use a bin width of 0.05 to display the histograms and frequency polygons. We also use a logarithmic scale to display the frequencies in the plots, because it made the distributions of the p-values for each set of gene sets clearer.

In the plots in Figure 5.4, the line corresponding to $k = 0$ displays the distribution of the p-values assigned to gene sets with no disease genes in them. Consequently, we may expect this distribution to be approximately uniform. However, as k increases, we expect the distribution of the p-values of the gene sets with k disease genes in them to become more and more positively skewed. Consequently, we expect the histograms, which display the sum of the counts displayed by the frequency polygons, to be positively skewed as well.

The frequency polygons in Figure 5.4 have a number of features that do not follow our expectations. For example, when $\alpha \neq 0.05$, the p-values of the gene sets are either less than 0.15, or greater than 0.95. This property severely violates our expectation that the p-values of the gene sets with no disease genes in them should be close to a uniform distribution.

Furthermore, consider the distribution of the p-values of gene sets with no disease genes in them obtained by PARIS with $\alpha = 0.05$. For $r \in \{1.44, 2.25\}$,

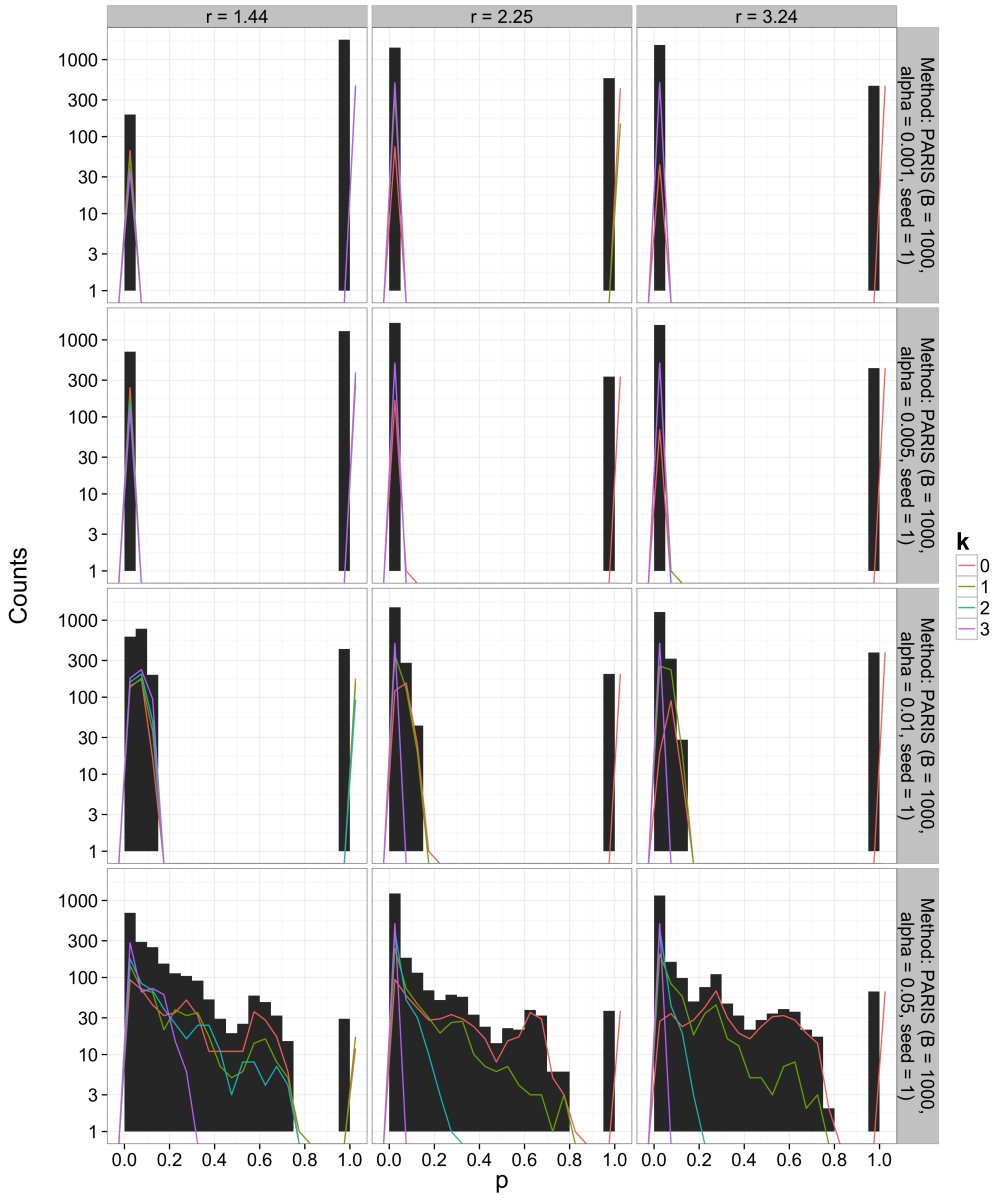

Figure 5.4: Frequency polygons and histograms obtained from using PARIS with a seed of 1, approximate bin size $B = 1000$ and different SNP significance levels on gene sets simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different SNP significance level α and each column corresponds to a different value of r. In each panel, the histogram displays the distribution of the p-values of all simulated gene sets, and the frequency polygons display the distribution of the p-values of the gene sets according to the number of disease genes in them. The colour of the frequency polygons indicates the number of disease genes in the gene sets.

the distribution is still not uniform, even though the gap in the distribution is far smaller than the distributions corresponding to the other significance levels. For example, there are nearly 100 gene sets in the $p \in [0, 0.05]$ bin. Suppose that the minimum number of disease genes in a disease gene set $d = 1$, so that the non-disease gene sets are those with no disease genes in them. If a gene set significance level of 0.05 is used, then the FPR is nearly $\frac{100}{500} = 0.2$. This is unusually large for a test with a significance level of 0.05.

However, these results may be consequences of parameter values that we used in this simulation, such as the number of disease SNPs (3) and the number of genes in gene sets (5). If we simulated genotype data with more disease SNPs, or if we simulated gene sets with more genes in them, then these anomalies may disappear. However, we do not investigate this problem further, because the aim of this section is to analyse the effect of changing the SNP significance level on the performance of PARIS.

Summary

In summary, the performance of PARIS was affected by changing the SNP significance level. Consequently, to compare PARIS with the other GSA methods, we implemented it with multiple significance levels.

5.1.3 The SRT and EVA: Changing the Significance Levels

Similarly to PARIS, changing the SNP significance level α in the SRT and the gene significance level α in MPEVA changed the performance of these methods considerably. We display empirical ROC curves for the SRT and MPEVA in Figures A.2 and A.3 respectively.

From the empirical ROC curves, no significance level α gives the best performance for the SRT or MPEVA. For example, from the left column of the figures, using the SRT and EVA with $\alpha = 0.01$ gives the best performance when $r = 1.44$.

However, consider using the SRT on the gene sets simulated using $r \in \{2.25, 3.24\}$. From Figure A.2, there is no clear relationship between the significance level that yields the best performance and d. In contrast, from Figure A.3, the relationship between the optimal significance level to use with EVA and the relative risk is clearer. When $r = 2.25$, the optimal significance level $\alpha = 0.005$, and when $r = 3.24$, the optimal significance level $\alpha = 0.001$.

In summary, the performance of the SRT and MPEVA was affected by changing the significance levels. Consequently, to compare these methods with the other GSA methods, we implemented them with multiple significance levels.

5.2 Comparing the GSA Methods

In the previous section, we concluded that the performance of PARIS, the SRT and MPEVA was markedly affected by the changing the SNP or gene significance level α. Consequently, we compared the AUC obtained by applying each method to simulated gene sets using different significance levels $\alpha \in \{0.001, 0.005, 0.01, 0.05\}$ to determine which level gives the best performance for each method and each simulation. We also compared the methods with each other, to determine which method gives the best overall performance.

From the empirical ROC curves that we displayed in Section 5.1, the homozygote relative risk r had a significant effect on the performance of PARIS, the SRT and MPEVA. However, the performance of each method was very similar for $r = 2.25$ and $r = 3.24$. Consequently, we compared the performance of each method separately for $r = 1.44$ and $r = 2.25$.

In Figures 5.5 and 5.6, we display scatterplots of the AUC obtained by using each method on gene sets simulated using different values of b and m, where $r = 1.44$ and $r = 2.25$ respectively. In each panel in these figures, we plot the method on the x-axis and the AUC on the y-axis. We have also drawn dotted lines where AUC = 0.5, which corresponds to the method that classifies gene sets

by random chance. The colour of each element in each scatterplot corresponds to the size m of the gene sets analysed, and the shape corresponds to a different value of d, the minimum number of disease genes in a disease gene set. Each row in this figure corresponds to a different value of the gene boundary extension b.

We expect the points in the scatterplots to be closer to the line AUC = 0.5 in Figure 5.5 compared to Figure 5.6, because the empirical ROC curves in Section 5.1 are closer to the line FPR = TPR for the simulations where $r = 1.44$ compared with the simulations where $r = 2.25$. This result is apparent from Figure 5.1: overall, the p-values of the disease SNPs are smaller than the p-values of the non-disease SNPs when $r = 2.25$, but this is not the case when $r = 1.44$. Consequently, classifying gene sets is more difficult when $r = 1.44$.

We also expect the performance of each method to decrease as b and m increase, because the proportion of disease SNPs in gene sets decreases as these parameters increase. However, as we saw from the empirical ROC curves, the relationship between performance and d is complex.

It is also important to remember that the values of the AUC displayed in Figures 5.5 and 5.6 are based on *empirical* ROC curves, which are only estimates of the true ROC curves. Consequently, there is a measure of uncertainty to the values of the AUC that we display in these figures.

We use each figure to comment on the significance levels α that gives the best results for each method, and we also comment on the methods that gives the best overall performance.

5.2.1 The Performance of Each Method when $r = 1.44$

Recall that for tests whose performance is comparable to random chance, the AUC is approximately 0.5. Consequently, from Figure 5.5, the performance of all methods is not markedly better than random chance, regardless of the significance level used. However, the performance of MPEVA, PARIS and the SRT is sensitive

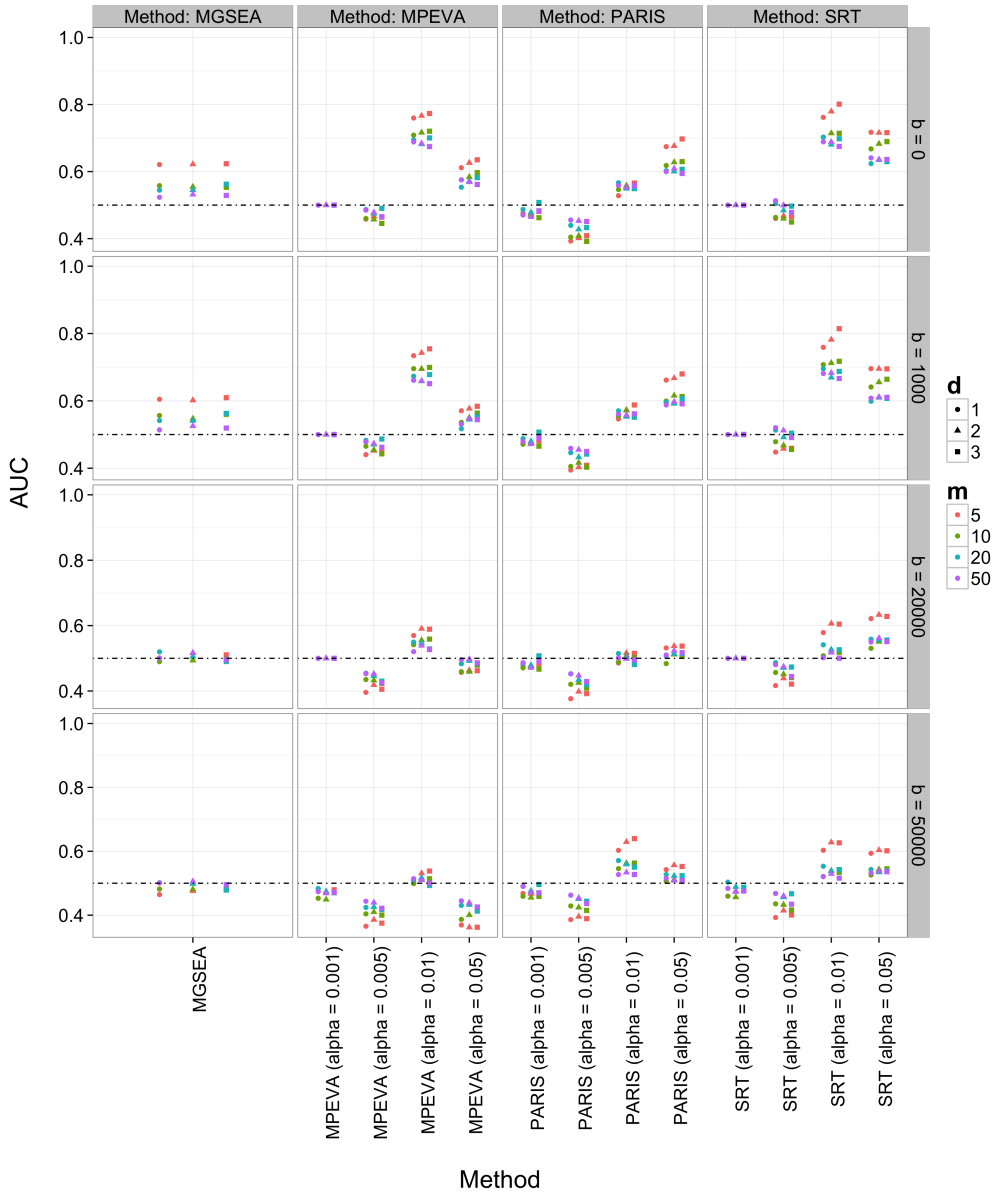

Figure 5.5: The AUC obtained by using each GSA method on gene sets simulated using homozygote relative risk $r = 1.44$. Each row corresponds to a different value of the gene boundary extension b, and each column corresponds to a different method. The AUC obtained by each method is given on the y-axis, and the method (with a significance level, if appropriate) is given on the x-axis. The shape of each object corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and the colour of each object corresponds to a different value of the gene set size m.

to the significance level α. For example, the optimal significance level $\alpha = 0.01$ for MPEVA and the SRT, whereas the optimal significance level $\alpha = 0.05$ for PARIS. We do not expect the optimal significance level for these methods to be the same, because they use the significance level in different ways.

As we expect, the overall performance of the methods decreases as the gene boundary extension b and gene set size m increase. However, there is no obvious relationship between the AUC and d, the minimum number of disease genes in a disease gene set.

In this set of simulations none of the methods were effective and it is not possible to draw meaningful conclusions about their relative performance.

5.2.2 The Performance of Each Method when $r = 2.25$

From Figure 5.6, the AUC obtained by all methods is better when $r = 2.25$ compared with $r = 1.44$, as we expect. Overall, the method that obtained the highest AUC is the SRT. Furthermore, from the far-right column of Figure 5.6, the performance of the SRT is robust to the choice of significance level, excluding the case where $\alpha = 0.05$. In contrast, the performance of PARIS and EVA is more sensitive to changes in α.

From the Figure 5.6, there is no obvious relationship between the AUC and d or b. The lack of relationship between the AUC and b is unexpected. For example, increasing b reduces the performance of EVA with $\alpha \neq 0.001$ and the SRT with $\alpha = 0.05$. In contrast, increasing b increases the performance of EVA with $\alpha = 0.001$, and it does not affect the performance of MGSEA, PARIS with $\alpha = 0.05$ and the SRT with $\alpha = 0.01$.

As we expect, the AUC obtained by MGSEA and MPEVA decreases as m increases. However, for PARIS and the SRT with $\alpha \in \{0.001, 0.005\}$ and $d = 3$, the AUC decreases as m *decreases*. Recall that if $d = 3$, then gene sets with one or two disease genes in them are non-disease gene sets. If the number of disease

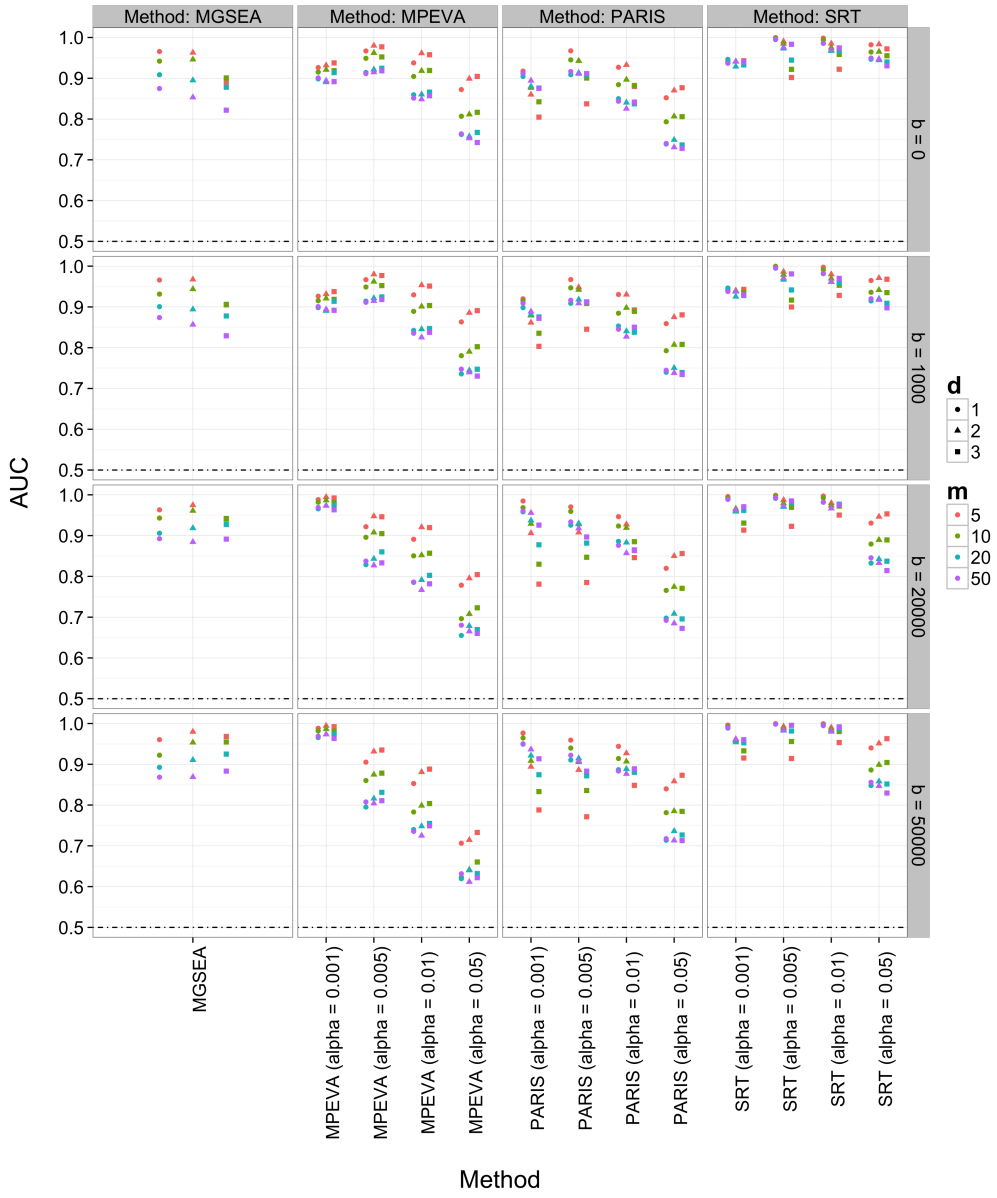

Figure 5.6: The AUC obtained by using each GSA method on gene sets simulated using homozygote relative risk $r = 2.25$. Each row corresponds to a different value of the gene boundary extension b, and each column corresponds to a different method. The AUC obtained by each method is given on the y-axis, and the method (with a significance level, if appropriate) is given on the x-axis. The shape of each object corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and the colour of each object corresponds to a different value of the gene set size m.

genes is held constant, then the proportion of disease genes in a gene set is larger for smaller gene sets. Consequently, for smaller m, gene sets with one or two disease genes in them have smaller p-values more often than larger gene sets, and hence they are more likely to be false positives. This may be why the performance of PARIS and the SRT decreases as m decreases.

We summarise the results obtained by each method when $r = 2.25$ using Figure 5.7. In this figure, we display boxplots of the AUC obtained by each GSA method. Each boxplot displays the distribution of the AUC obtained by each GSA method, with different significance levels for all methods apart from MGSEA, and different values of d, on gene sets simulated using $m \in \{5, 10, 20, 50\}$. The AUC is given on the y-axis, and the method is given on the x-axis. The colour of each boxplot corresponds to a different value of b. From the middle two columns of Figure 5.7, the methods whose performance varies the most across our simulations are MPEVA and PARIS. Also, the boxplots corresponding to the SRT are closer to AUC = 1 than the boxplots corresponding to any other method. Consequently, for the range of scenarios that we considered, the SRT often gives the best performance, and it is unlikely to perform markedly worse than any other method.

5.2.3 Summary of the Performance of Each Method

The AUC obtained by all methods was uniformly greater when $r = 2.25$ compared with $r = 1.44$, as we expect. For MPEVA, PARIS and the SRT, we display the significance level α that gives the best performance when $r \in \{1.44, 2.25\}$ in Table 5.3.

The methods that gives the best overall performance are the SRT with $\alpha \in \{0.001, 0.005, 0.01\}$. We recommend using $\alpha = 0.01$, since one of the aims of GSA is to detect gene sets containing multiple moderately significant SNPs that may not be detected by classical GWA studies.

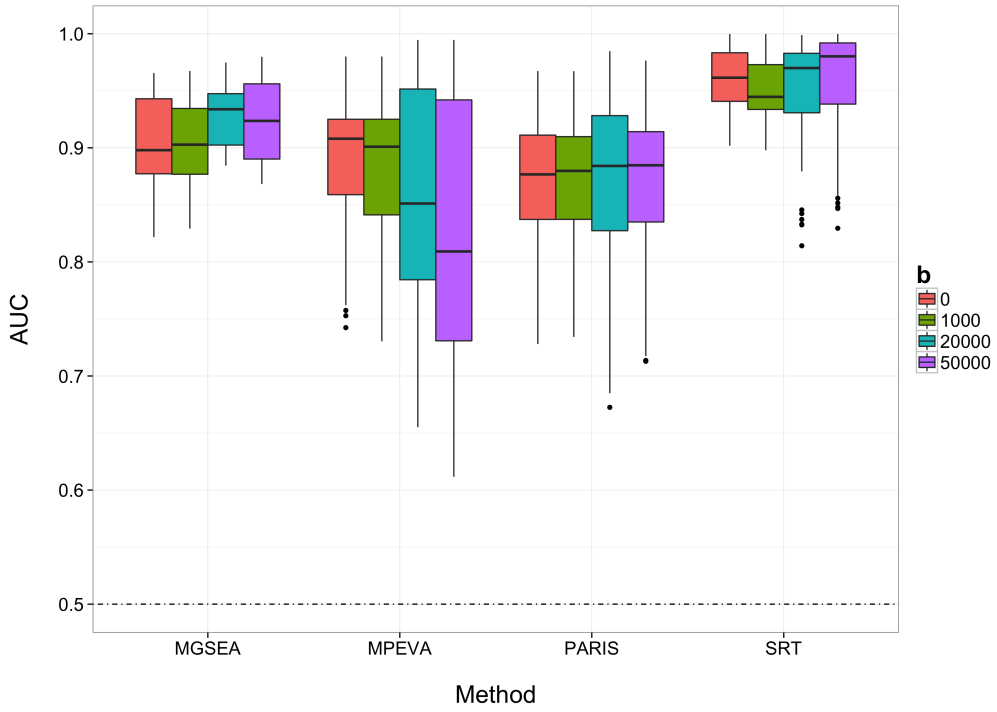

Figure 5.7: Boxplots of the AUC obtained by using each GSA method on gene sets simulated using homozygote relative risk $r = 2.25$. Each boxplot displays the distribution of the AUC obtained by each GSA method, with different significance levels for all methods apart from MGSEA, and different values of d (the minimum number of disease genes in a disease gene set), on gene sets simulated using gene set size $m \in \{5, 10, 20, 50\}$. The AUC is given on the y-axis, and the method is given on the x-axis. The colour of each boxplot corresponds to a different value of b, the gene boundary extension.

r	1.44	2.25
MPEVA	0.01	0.005 (small b) or 0.001 (large b)
PARIS	0.05	Varies
The SRT	0.01	0.01

Table 5.3: Optimal significance levels for each method, where the homozygote relative risk $r \in \{1.44, 2.25\}$.

5.3 Gene Sets Assigned Disparate p-values

In this section, we highlight gene sets that were assigned disparate p-values by different GSA methods. In particular, we tested the sensitivity of each method to SNPs with very small p-values. To perform this sensitivity analysis, analysed gene sets containing

1. a highly significant SNP but few significant SNPs otherwise, and
2. many SNPs that are significant, but not highly significant.

In these categories, a highly significant SNP is a SNP with a very small p-value, and a significant SNP is a SNP with a p-value less than some significance level α.

Recall that MGSEA is a GSE method, which means that it directly uses SNP and gene test statistics without using a significance criterion. Consequently, we expect MGSEA to assign a small p-value to gene sets in category one, and a larger p-value to gene sets in category two. In contrast, we expect the reverse relationship for PARIS and the SRT, which only use the p-values of SNPs to categorise them as significant or nonsignificant.

For simplicity, we simulated gene sets using $r = 2.25, m = 5$ and $b = 0$. We divide the results that we display in this section into results that followed our expectations, and results that were contrary to our expectations.

5.3.1 Gene Sets that Followed our Expectations

Gene Sets in Category One

As we expect, many gene sets in category one were assigned small p-values by MGSEA and larger p-values by PARIS and the SRT. We highlight two examples.

In Figure 5.8, we display a Manhattan plot highlighting a gene set that was assigned p-values of 0.002 and 0.2867 by MGSEA and the SRT ($\alpha = 0.05$) respectively. In this plot, the grey circles represent non-disease SNPs that are not in the

gene set. Coloured circles represent SNPs that are in the gene set, and the colour indicates the gene that the SNP is in. Black squares indicate disease SNPs that are not in the gene set. As in Figure 5.1, we plot the position of the SNP on the x axis, and $-\log_{10}(p)$ on the y-axis, where p is the SNP p-value. We also draw horizontal dotted lines where $p \in \{0.0001, 0.001, 0.01, 0.05\}$, to aid our discussion of the results.

The reason for the disparate results is clear: the gene set in Figure 5.8 contains a SNP with $p < 0.001$ in the ZIC4 gene, which is markedly affecting the result obtained by MGSEA. However, this SNP is not a disease SNP, and it is not in LD with any of the disease SNPs. Consequently, this is a false positive gene set for MGSEA, and a true negative gene set for the SRT.

We also display in Figure 5.9 a gene set assigned p-values of 0.000 and 0.659 by MGSEA and PARIS ($\alpha = 0.05$) respectively. This gene set contains a disease SNP with $p < 0.0001$ in the SLC25A20 gene, which we display using a coloured square, and two other SNPs with $p < 0.05$. However, all of the other SNPs in this gene set have p-values greater than 0.05. Consequently, this gene set is also consistent with our expectations. If $d = 1$, then this gene set is a disease gene set; it is a true positive gene set for MGSEA and a false negative gene set for PARIS.

Gene Sets Containing Many Significant SNPs

As we expect, many gene sets that contained numerous significant SNPs, but no highly significant SNPs, were assigned small p-values by PARIS and the SRT, and larger p-values by MGSEA.

For example, consider the gene set highlighted in the Manhattan plot in Figure 5.10, which was assigned p-values of 0.043 and 0.198 by the SRT ($\alpha = 0.05$) and MGSEA respectively. This gene set has five SNPs with $p < 0.05$, but none of these SNPs are highly significant.

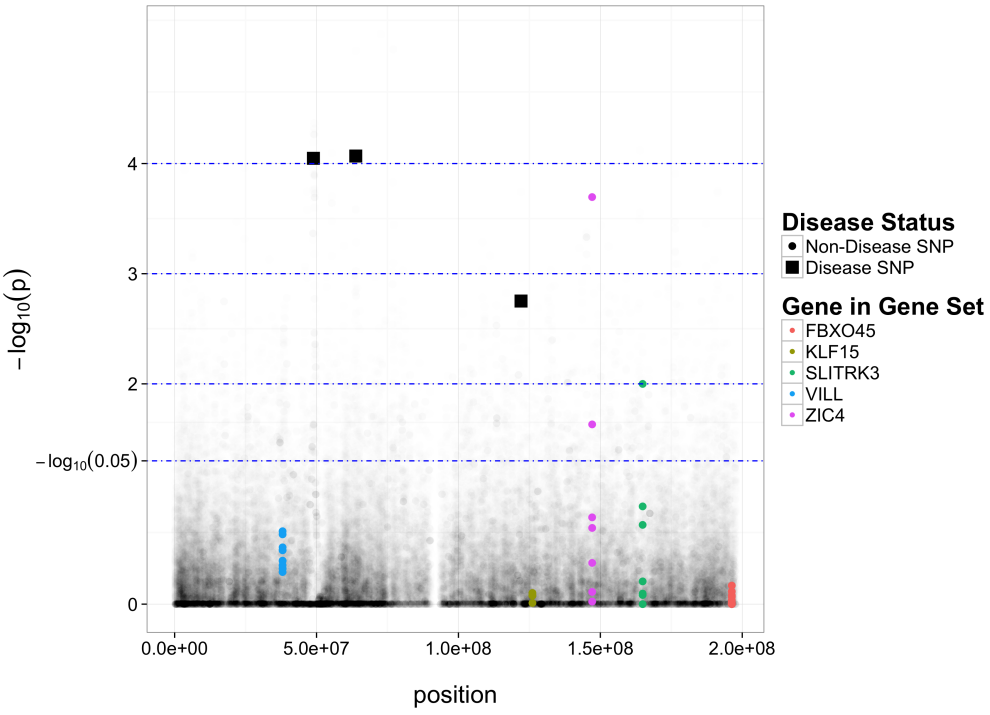

Figure 5.8: Manhattan plot highlighting a gene set in category one that follows our expectations. This gene set was assigned p-values of 0.002 and 0.2867 by MGSEA and the SRT ($\alpha = 0.05$) respectively. The grey circles represent non-disease SNPs that are not in the gene set. Coloured circles represent SNPs that are in the gene set, and the colour indicates the gene that the SNP is in. Black squares indicate disease SNPs that are not in the gene set. The x-coordinate is the position of the SNP on the chromosome, and the y-coordinate is $-\log_{10}(p)$, where p is the SNP p-value. Since this gene set contains no disease genes, this is a false positive gene set for MGSEA and a true negative gene set for the SRT. Furthermore, since this gene set is in category one, the results obtained by the SRT and MGSEA for this gene set are expected.

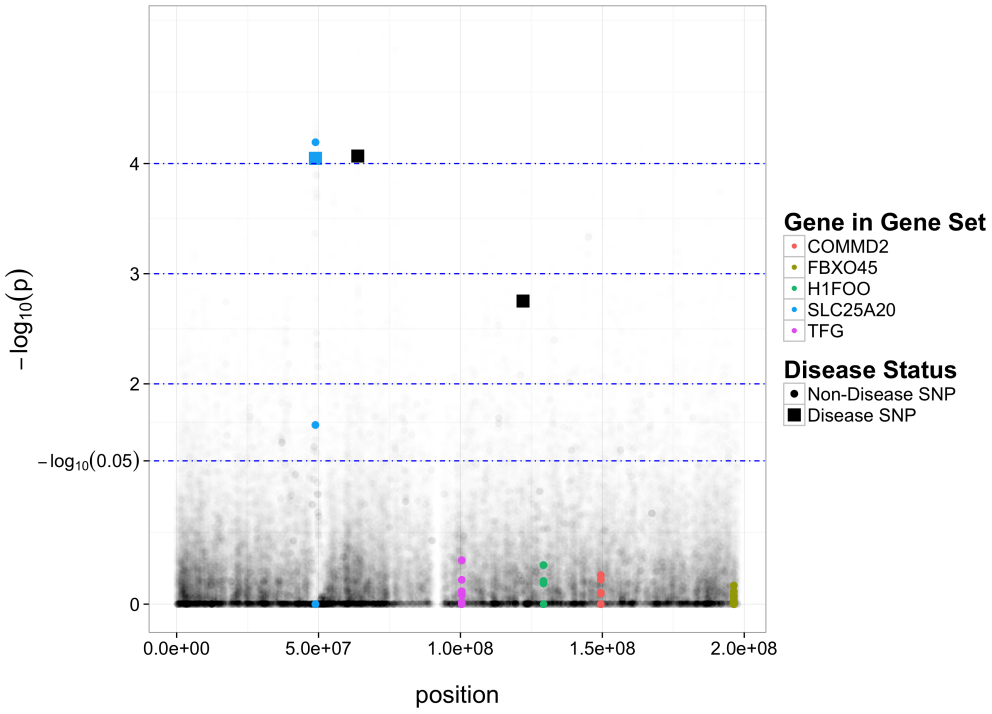

Figure 5.9: Manhattan plot highlighting a gene set in category one that follows our expectations. This gene set was assigned p-values of 0.000 and 0.659 by MGSEA and PARIS ($\alpha = 0.05$) respectively. The grey circles represent non-disease SNPs that are not in the gene set. Coloured circles represent SNPs that are in the gene set, and the colour indicates the gene that the SNP is in. Black squares indicate disease SNPs that are not in the gene set, and coloured squares indicate disease SNPs that are in the gene set. The x-coordinate is the position of the SNP on the chromosome, and the y-coordinate is $-\log_{10}(p)$, where p is the SNP p-value. This gene set contains one disease gene. Consequently, if $d = 1$, this is a true positive gene set for MGSEA and a false negative finding for PARIS. Since this gene set is in category one, the results obtained by PARIS and MGSEA for this gene set are expected.

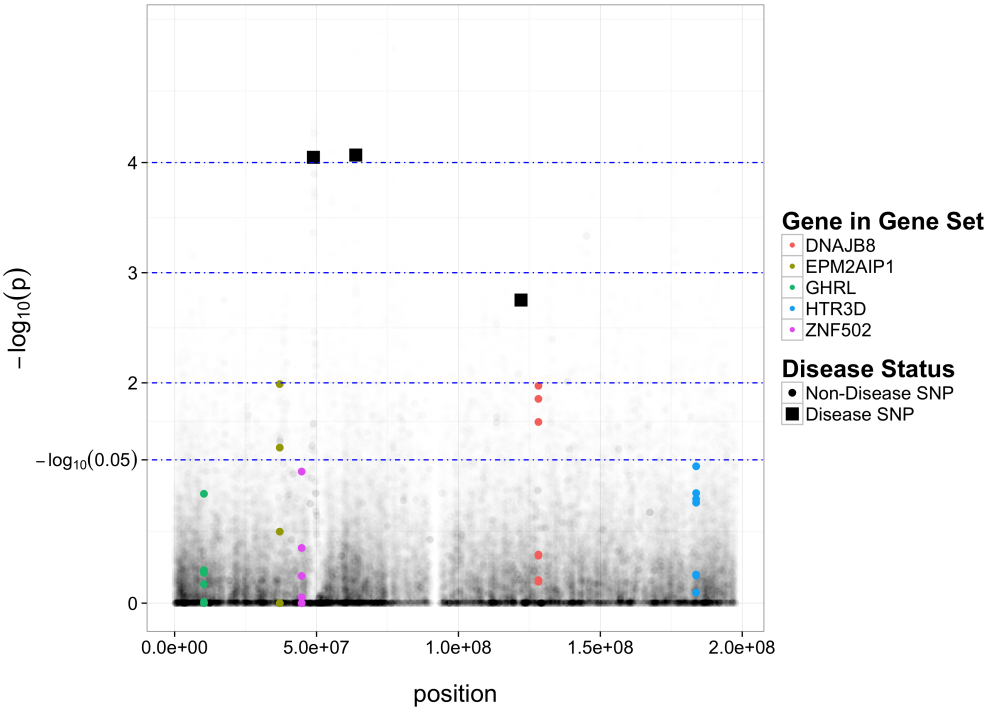

Figure 5.10: Manhattan plot highlighting a gene set in category two that follows our expectations. This gene set was assigned p-values of 0.043 and 0.198 by the SRT ($\alpha = 0.05$) and MGSEA respectively. The grey circles represent non-disease SNPs that are not in the gene set. Coloured circles represent SNPs that are in the gene set, and the colour indicates the gene that the SNP is in. Black squares indicate disease SNPs that are not in the gene set. The x-coordinate is the position of the SNP on the chromosome, and the y-coordinate is $-\log_{10}(p)$, where p is the SNP p-value. Since this gene set is in category two, the results obtained by the SRT and MGSEA for this gene set are expected.

5.3.2 Gene Sets that did not Follow our Expectations

In contrast to the previous results, numerous gene sets did not follow our expectations regarding the sensitivity of the GSA methods to SNPs with very small p-values. In particular, some gene sets in category one were assigned smaller p-values by PARIS or the SRT. We illustrate this result using the gene set highlighted in Figure 5.11. This gene set was assigned p-values of 0.000 and 0.158 by PARIS ($\alpha = 0.05$) and MGSEA respectively. This gene set contains one SNP with $p < 0.001$ and two with $p < 0.05$, but no other significant SNPs ($\alpha = 0.05$). Consequently, the results obtained by PARIS and MGSEA are unexpected.

We speculate about the reason that this unexpected result occurred. Recall that since we only used three disease SNPs in our analysis and since this gene set only has five genes in it, the distribution of p-values obtained by each method may have been distorted. This could be why PARIS is assigning this gene set a p-value of zero, even though it only has three significant SNPs ($\alpha = 0.05$). This result also highlights the fact that the results obtained by GSE methods are affected by factors other than SNPs with very small p-values, and the results obtained by methods that use a significance criterion are affected by factors other than the number of significant SNPs in the gene set.

5.4 Chapter Summary

In Section 5.1, we displayed ROC curves to demonstrate that the only method-specific parameters that affected the performance of PARIS, the SRT and MPEVA non-negligibly are the SNP and gene significance levels α. Consequently, when we compared PARIS, the SRT, MPEVA and MGSEA in Section 5.2, we varied the significance levels. In this section, we used scatterplots of AUC to demonstrate that the SRT performed best in our simulation study. In Section 5.3, we displayed Manhattan plots to demonstrate that MGSEA is often affected by the presence of single, highly significant SNPs, while PARIS and the SRT are often robust to such

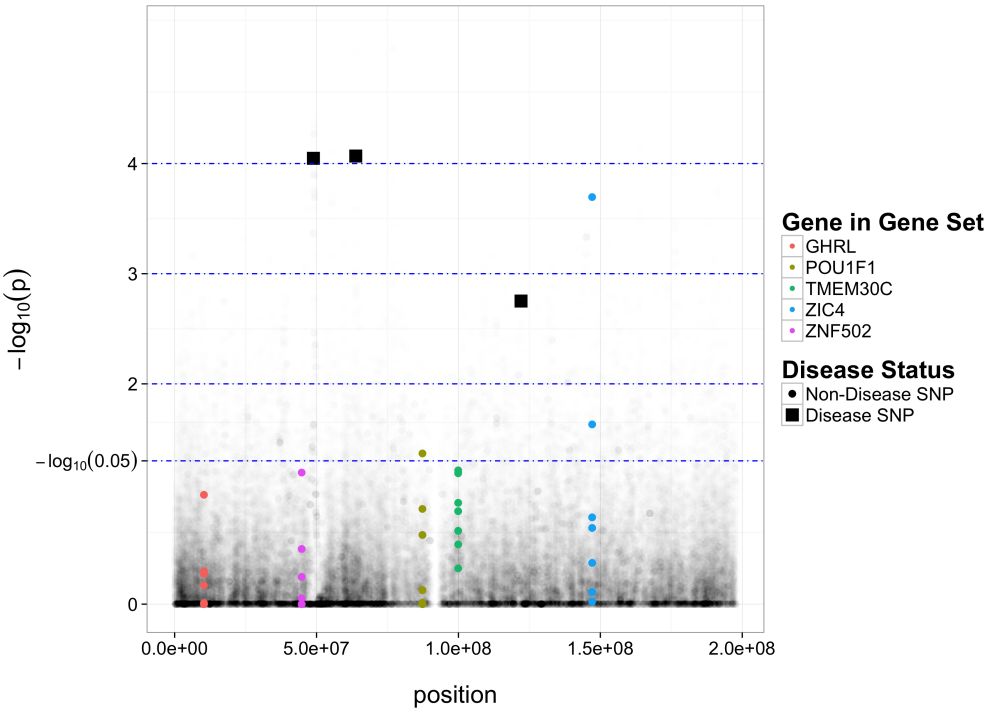

Figure 5.11: Manhattan plot highlighting a gene set in category one that does not follow our expectations. This gene set was assigned p-values of 0.000 and 0.158 by PARIS ($\alpha = 0.05$) and MGSEA respectively. The grey circles represent non-disease SNPs that are not in the gene set. Coloured circles represent SNPs that are in the gene set, and the colour indicates the gene that the SNP is in. Black squares indicate disease SNPs that are not in the gene set, and coloured squares indicate disease SNPs that are in the gene set. The x-coordinate is the position of the SNP on the chromosome, and the y-coordinate is $-\log_{10}(p)$, where p is the SNP p-value. This gene set contains one SNP with $p < 0.001$ and two with $p < 0.05$, but no other significant SNPs ($\alpha = 0.05$). Consequently, the results obtained by PARIS and MGSEA for this gene set are unexpected.

SNPs. However, not all gene sets followed these trends. This result highlights the fact that other factors impact the results obtained by GSE methods and methods that use a significance criterion.

Chapter 6

Conclusion

GWA studies are designed to gain further information about the SNPs that are associated with a given disease. One of the purposes of GWA studies is to elucidate the way that genetic variation as a whole causes diseases, so that treatments for them can be improved. We detailed the necessary biological and statistical background for GWA study in Chapter 2.

However, GWA studies have a number of shortcomings. For example, they lack power to detect SNPs with a small effect size, and they cannot account for epistasis. Consequently, many novel methods have been developed to identify gene sets that are associated with a disease. These methods are known as GSA methods. A number of reviews of GSA methods exist in the literature, such as Fridley and Biernacka (2011), Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011). However, none of these reviews compare specific GSA methods in detail or test their conclusions by implementing GSA methods on data.

In Chapter 3, we detailed six GSA methods: PARIS, the SRT, MPEVA, ProxyGeneLD, ALIGATOR and MGSEA. We then compared these methods theoretically. We structured our comparison around seven important issues that need to be considered in GSA, as detailed in the reviews by Fridley and Biernacka (2011),

Holmans (2009), Mooney *et al.* (2014), Ramanan *et al.* (2012), Wang *et al.* (2010), and Wang *et al.* (2011).

In particular, the SRT is the only method that tests a self-contained null hypothesis; all of the other methods test a competitive null hypothesis. If a self-contained method is required, then the SRT is a good choice, because it accounts for LD and gene size. The only competitive method that accounts for LD and gene size is MGSEA. However, unlike the other methods, it is sensitive to SNPs with very small p-values. PARIS and ProxyGeneLD attempt to account for LD by partitioning SNPs into LD and LE blocks, however, current partitioning methods do not guarantee robustness to LD, and more research into this problem is required. In contrast, MPEVA and ALIGATOR do not account for LD at all. Regarding gene size, PARIS and ALIGATOR both account for it, whereas MPEVA and ProxyGeneLD do not.

In Chapter 4, we detailed the advantages and disadvantages of comparing these methods using a simulation study, and we discussed three methods of simulating genetic data: the LS method (Li and Stephens, 2003), HAPGEN (Spencer *et al.*, 2009) and HAPGEN2 (Su *et al.*, 2011). We used HAPGEN2, because it can simulate genetic data in the presence of multiple disease SNPs. We then detailed the procedures that we used to simulate genetic data using HAPGEN2, implement the GSA methods, and compare their performance. Due to time constraints and the availability of software and data, we only implemented PARIS, the SRT, MPEVA and MGSEA. We also reviewed the literature to ensure that we selected realistic values for each parameter that we could vary in our simulation study.

Finally, in Chapter 5, we displayed and discussed our results. While no method gave the best performance in all simulations, we recommend using the SRT, because it consistently performed well in our study. Furthermore, we recommend using a SNP significance level $\alpha = 0.01$, because the aim of GSA is to detect gene sets containing many moderately significant SNPs that may not be detected by traditional GWA study. Unfortunately, the performance of the competitive

methods was not as robust to parameters such as gene set size and gene boundary extension. This lack of robustness was expected, because we allowed these parameters to vary while we kept constant the number of disease SNPs in the gene sets. Consequently, the *proportion* of disease SNPs in the gene sets decreased, which reduced the performance of these methods.

Much future work still needs to be done in the area of GSA. Due to the time constraints in writing this thesis, we only compared six GSA methods. However, Mooney *et al.* (2014) detail 55 GSA methods in Table 1 that can be used on GWA study data. These methods use a wide range of statistical techniques, such as Bayesian methods (Shahbaba *et al.*, 2012), ridge regression (Chen *et al.*, 2010) and principle component analysis (Lu *et al.*, 2014). Consequently, a future simulation study could implement and compare more of these methods. Furthermore, we only simulated genetic data on chromosome three, and while our choices for the values of the parameters in our simulation study were realistic, we only performed a small number of simulations. Consequently, a future simulation study could simulate data from more chromosomes, use more simulations, and vary the values of parameters more. Alternatively, the GSA methods could be implemented on real genetic data.

Another area of future work concerns measuring the performance of a method. For example, we focused on using ROC curves and the AUC, noting that other simulation studies had also used the AUC to measure performance (Lu *et al.*, 2014; Zhang *et al.*, 2014). However, when gene sets are analysed using GSA methods, a significance level $\alpha \leq 0.05$ is commonly used. Consequently, the section of ROC curves where FPR > 0.05 lacks meaning. A common approach is to calculate the statistical power obtained by methods at given significance levels (Jia *et al.*, 2011; Lu *et al.*, 2014). However, we could also consider the AUC between FPR $= 0$ and FPR $= 0.05$. For example, suppose that the ROC curve of a method follows the line FPR $=$ TPR until FPR $= 0.05$, and then the TPR increases rapidly. Then the set of gene sets with p-values less than 0.05 would contain a similar number

of true positives and false positives ($\alpha = 0.05$). Thus the performance of using the method with gene set significance level $\alpha = 0.05$ is no better than random chance. In this case, the AUC between FPR = 0 and FPR = 1 is high, but the AUC between FPR = 0 and FPR = 0.05 is the same as the method that classifies by random chance.

Finally, we were unable to test our theoretical conclusions about how robust each GSA method is to parameters such as LD and gene size. Regarding gene size, this occurred because we used the *number* of disease SNPs in a disease gene as a parameter in our simulation study. Consequently, increasing the number of SNPs in the gene sets (by increasing the gene boundary extension) decreased the proportion of disease SNPs in the gene sets, which may have confounded our results. In future simulation studies, we recommend keeping the proportion of disease SNPs or disease genes fixed.

Appendix A

Additional Figures to Compare GSA Methods

A.1 Varying the Approximate Bin Size in PARIS

In Figure A.1, we plot empirical ROC curves that display the performance of using PARIS with random seeds of 1, 2, 3, 4 and 5 on simulated gene sets. In this figure, we fixed the SNP significance level $\alpha = 0.05$ and the approximate bin size $B = 1000$. This figure demonstrates that changing the random seed also had a negligible effect on the performance of PARIS, as we expected.

A.2 Varying the Significance Levels in the SRT and MPEVA

Similarly to PARIS, changing the SNP significance level α in the SRT and the gene significance level α in MPEVA changed the performance of these methods considerably. We display empirical ROC curves for the SRT and MPEVA in Figures A.2 and A.3 respectively.

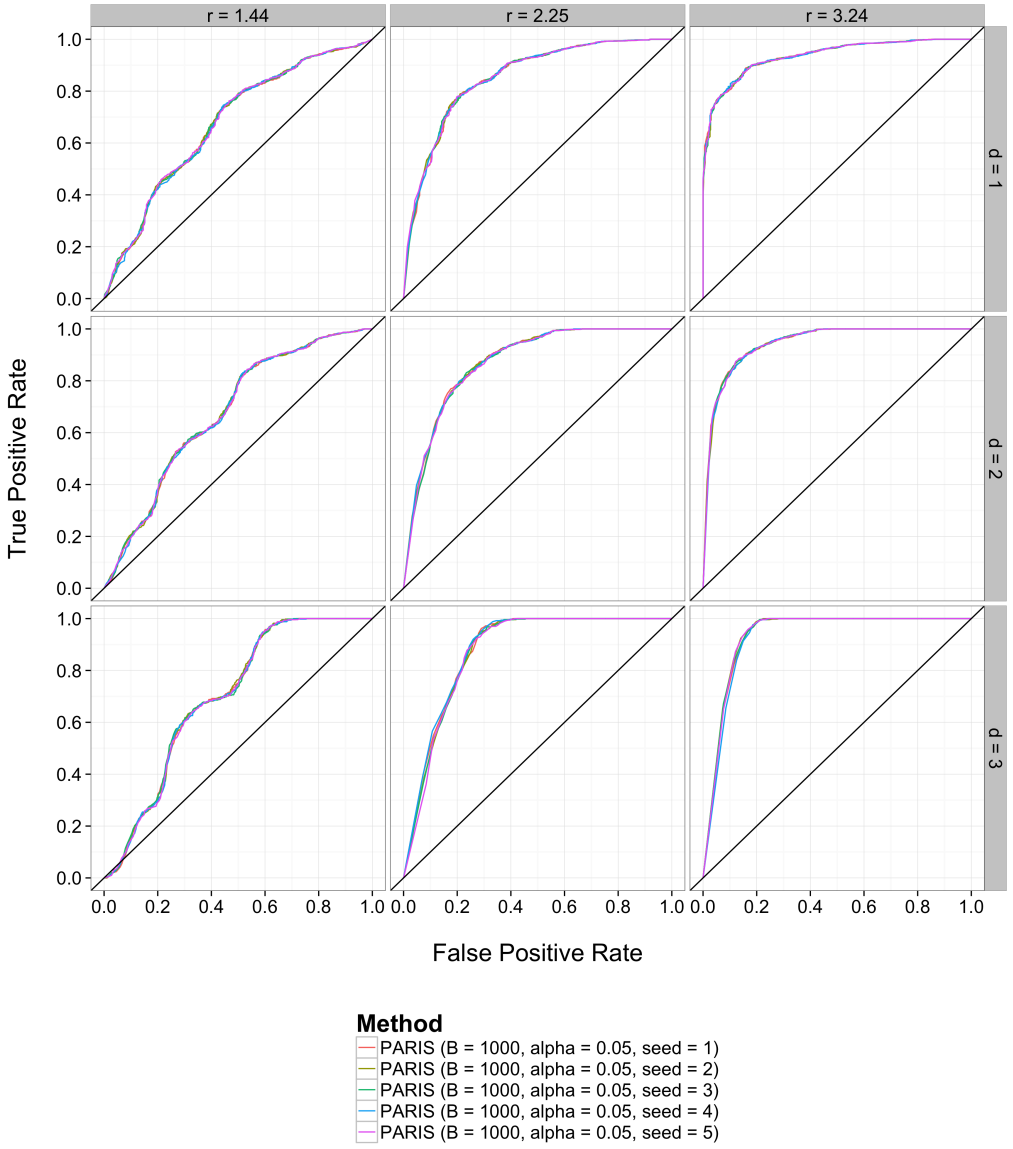

Figure A.1: Empirical ROC curves obtained from using PARIS with SNP significance level $\alpha = 0.05$, approximate bin size $B = 1000$ and different random seeds on genetic data simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and each column corresponds to a different value of r. The seed is indicated by the colour used to display the empirical ROC curve.

From the empirical ROC curves, no significance level α gives the best performance for the SRT or MPEVA. For example, from the left column of the figures, using the SRT and EVA with $\alpha = 0.01$ gives the best performance when $r = 1.44$. However, consider using the SRT on the gene sets simulated using $r \in \{2.25, 3.24\}$. From Figure A.2, there is no clear relationship between the significance level that yields the best performance and d. In contrast, from Figure A.3, the relationship between the optimal significance level to use with EVA and the relative risk is clearer. When $r = 2.25$, the optimal significance level $\alpha = 0.005$, and when $r = 3.24$, the optimal significance level $\alpha = 0.001$.

In summary, the performance of the SRT and MPEVA was affected by changing the significance level. Consequently, to compare these methods with the other GSA methods, we implemented them with multiple significance levels.

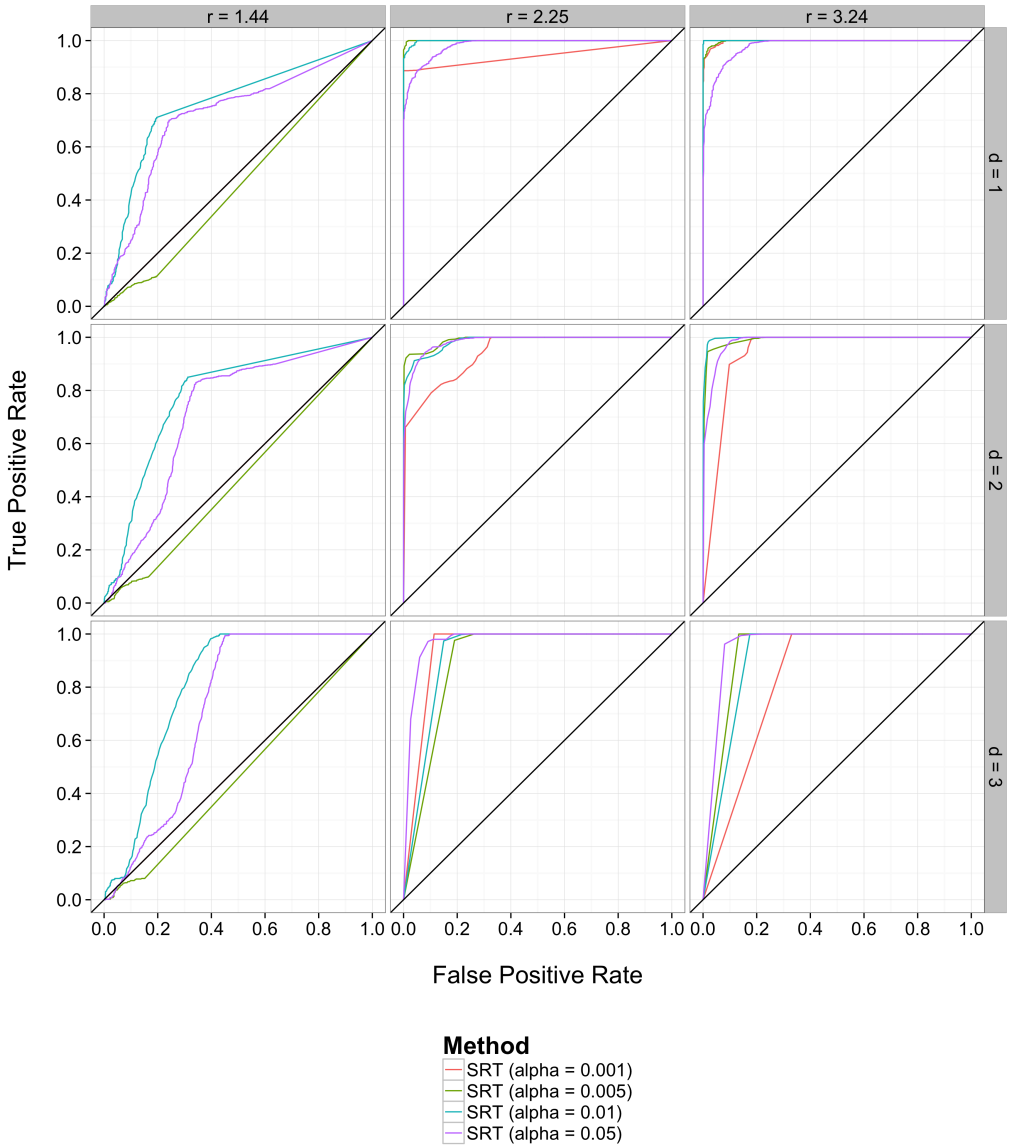

Figure A.2: Empirical ROC curves obtained from using the SRT with different SNP significance levels on gene sets simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and each column corresponds to a different value of r. The significance level α is indicated by the colour used to display the empirical ROC curve.

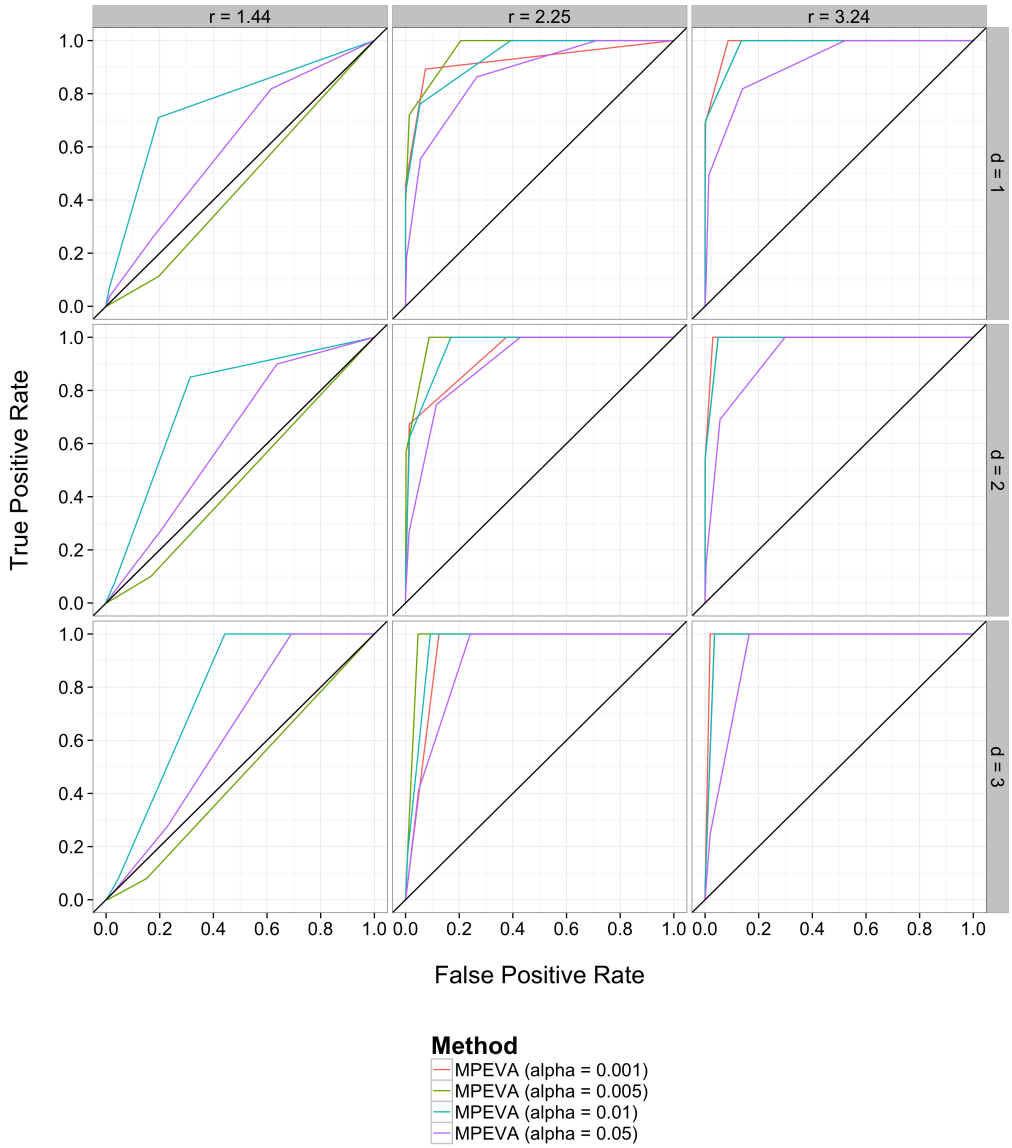

Figure A.3: Empirical ROC curves obtained from using MPEVA with different gene significance levels on gene sets simulated using gene set size $m = 5$, gene boundary extension $b = 0$ and different values of the homozygote relative risk r. Each row corresponds to a different value of d, the minimum number of disease genes in a disease gene set, and each column corresponds to a different value of r. The significance level α is indicated by the colour used to display the empirical ROC curve.

Bibliography

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* 11.3, pp. 375–386.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25.1, pp. 25–29.

Askland, K., Read, C., and Moore, J. (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.* 125.1, pp. 63–79.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1, pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29.4, pp. 1165–1188.

Bonferroni, C. E. (1936). *Statistical Theory of Classes and Calculating Odds*. Library International Seeber.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Vol. 2. Duxbury Pacific Grove, CA.

Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., and Hsu, L. (2010). Insights into colon cancer etiology via a regularized approach

to gene set analysis of GWAS data. *The Am. J. Hum. Genet.* 86.6, pp. 860–871.

Cochran, W. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* 10, pp. 417–451.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* (2015). Ensembl 2015. *Nucleic Acids Res.* 43.D1, pp. D662–D669.

Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine (2015a). *dbSNP accession: {rs1979334, rs9837104, rs11130263, rs11928389, rs13093798} (dbSNP Build ID: 130).* URL: `http://www.ncbi.nlm.nih.gov/SNP/` (visited on June 10, 2015).

Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine (2015b). *(dbSNP Build ID: 126 - 130).* URL: `http://www.ncbi.nlm.nih.gov/SNP/` (visited on June 10, 2015).

Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine (2015c). *RsMergeArch table.* URL: `ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/RsMergeArch.bcp.gz` (visited on June 10, 2015).

Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A., *et al.* (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4.5, P3.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21.16, pp. 3439–3440.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4.8, pp. 1184–1191.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3.1, pp. 87–112.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Series B Stat. Methodol.* 64.4, pp. 657–680.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* 85.1, pp. 87–94.

Foulkes, A. S. (2009). *Applied Statistical Genetics with R: for Population-based Association Studies*. Springer.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42.12, pp. 1118–1125.

Fridley, B. L. and Biernacka, J. M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur. J. Hum. Genet.* 19.8, pp. 837–843.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* 12.1, pp. 1–77.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., *et al.* (2003). The international HapMap project. *Nature* 426.6968, pp. 789–796.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23.8, pp. 980–987.

Gonick, L. (1991). *Cartoon Guide to Genetics*. HarperCollins.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* 8.29, pp. 299–309.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, pp. 49–50.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6.2, pp. 65–70.

Holmans, P. (2009). Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.* 72, pp. 141–179.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., Owen, M. J., O'Donovan, M. C., and Craddock, N. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* 85.1, pp. 13–24.

Hong, M.-G., Pawitan, Y., Magnusson, P. K., and Prince, J. A. (2009). Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* 126.2, pp. 289–301.

Hosack, D. A., Dennis Jr, G., Sherman, B. T., Lane, H. C., Lempicki, R. A., *et al.* (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4.10, R70.

International HapMap Consortium (2014). *HapMap data: CEPH (Utah residents with ancestry from northern and western Europe) (abbreviation: CEU)*. URL: `https://mathgen.stats.ox.ac.uk/wtccc-software/HM3.tgz` (visited on Apr. 28, 2014).

International Multiple Sclerosis Genetics Consortium and Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 476.7359, pp. 214–219.

Jia, P., Wang, L., Meltzer, H. Y., and Zhao, Z. (2011). Pathway-based analysis of GWAS datasets: effective but caution required. *Int. J. Neuropsychopharmacol.* 14.4, pp. 567–572.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28.1, pp. 27–30.

Laird, N. M. and Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics*. Springer.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165.4, pp. 2213–2233.

Lu, M., Lee, H.-S., Hadley, D., Huang, J. Z., and Qian, X. (2014). Supervised categorical principal component analysis for genome-wide association analyses. *BMC Genomics* 15.Suppl 1, S10.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456.7218, pp. 18–21.

Metz, C. E. (1978). "Basic principles of ROC analysis". *Seminars in nuclear medicine*. Vol. 8. 4. Elsevier, pp. 283–298.

Mi, H., Muruganujan, A., and Thomas, P. D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* 41.D1, pp. D377–D386.

Mooney, M. A., Nigg, J. T., McWeeney, S. K., and Wilmot, B. (2014). Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* 30.9, pp. 390–400.

O'Dushlaine, C., Kenny, E., Heron, E. A., Segurado, R., Gill, M., Morris, D. W., and Corvin, A. (2009). The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* 25.20, pp. 2762–2763.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dubl. Phil. Mag. J. Sci.* 50.302, pp. 157–175.

Polychronakos, C. and Li, Q. (2011). Understanding type 1 diabetes through genetics: advances and prospects. *Nat. Rev. Genet.* 12.11, pp. 781–792.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* 66.3, pp. 403–411.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: `http://www.R-project.org/` (visited on Nov. 16, 2015).

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* 28.7, pp. 323–332.

Reif, D. M., Dudek, S. M., Shaffer, C. M., Wang, J., and Moore, J. H. (2005). "Exploratory Visual Analysis of Pharmacogenomic Results". *Pacific Symposium on Biocomputing*. Vol. 10. Citeseer, pp. 296–307.

Shahbaba, B., Shachaf, C. M., and Yu, Z. (2012). A pathway analysis method for genome-wide association studies. *Stat. Med.* 31.10, pp. 988–1000.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29.1, pp. 308–311.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* 5.5, e1000477.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. Series B Stat. Methodol.* 62.4, pp. 605–635.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat.* 31.6, pp. 2013–2035.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27.16, pp. 2304–2305.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102.43, pp. 15545–15550.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U. S. A.* 102.38, pp. 13544–13549.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90.1, pp. 7–24.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81.6, pp. 1278–1283.

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11.12, pp. 843–854.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* 98.1, pp. 1–8.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7.2, pp. 256–276.

Weinberg, W. (1908). Über vererbungsgesetze beim menschen. *Mol. Gen. Genet.* 1.1, pp. 440–460.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* 16.2, p. 97.

Wright, S. (1938). Size of population and breeding structure in relation to evolution. *Science* 87.2263, pp. 430–431.

Yaspan, B. L., Bush, W. S., Torstenson, E. S., Ma, D., Pericak-Vance, M. A., Ritchie, M. D., Sutcliffe, J. S., and Haines, J. L. (2011). Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.* 129.5, pp. 563–571.

Zhang, X., Xue, F., Liu, H., Zhu, D., Peng, B., Wiemels, J. L., and Yang, X. (2014). Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies. *BMC Genet.* 15.1, p. 130.

Zheng, G., Yang, Y., Zhu, X., and Elston, R. (2012). *Analysis of Genetic Association Studies*. Springer.