Gene x Gene Interactions in Genome Wide Association Studies

Kanishka Bhattacharya

Wellcome Trust Centre for Human Genetics,

Medical Sciences Division, University of Oxford.

Thesis submitted for the degree of Doctor of Philosophy

Wolfson College

Michaelmas Term 2013

To Sidney, the guide dog ...

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Andrew P. Morris for his patient guidance and support throughout, particularly when I was on a medical break and found it difficult to work. A big note of thanks to Reedik Magi, Andrew Rimmer and Shamal Faily for their help with my C/C++ programming. I would also like to thank Mihai Duta at the Oxford Supercomputing Services, John Broxholme, Anuj Goel and Ashish Kumar for their programming related help at various stages of my work.

I owe a great deal to the funding from Wellcome Trust, without which I would not have been able to start my DPhil. I am very thankful to Wolfson College for providing non-academic support, which helped give a more complete feel to my graduate life. A big thank you to my partner, Imogen Wade, for being very supportive through the difficult days after my accident and the long, painful days of recovery.

Last but not least, I should thank my dog Sidney for keeping my feet warm while I got on with the corrections. He was always ready for a play when I needed a distraction.

Abstract

Genome wide association studies (GWAS) have revolutionized our approach to mapping genetic determinants of complex human diseases. However, even with success from recent studies, we have typically been able to explain only a fraction of the trait heritability. GWAS are typically analysed by testing for the marginal effects of single variants. Consequently, it has been suggested that gene-gene interactions might contribute to the "missing heritability" of complex diseases.

GWAS incorporating interaction effects have not been routinely applied because of statistical and computational challenges relating to the number of tests performed, genome-wide. To overcome this issue, I have developed novel methodology to allow rapid testing of pairwise interactions in GWAS of complex traits, implemented in the IntRapid software. Simulations demonstrated that the power of this approach was equivalent to computationally demanding exhaustive searches of the genome, but required only a fraction of the computing time. Application of IntRapid to GWAS of a range of complex human traits undertaken by the Wellcome Trust Case Control Consortium (WTCCC) identified several interaction effects at nominal significance, which warrant further investigation in independent studies.

In an attempt to fine-map the identified interacting loci, I undertook imputation of the WTCCC genotype data up to the 1000 Genomes Project reference panel (Phase 1 integrated release, March 2012) in the neighbourhood of the lead SNPs. I modified the IntRapid software to take account of imputed genotypes, and identified stronger signals

of interaction after imputation at the majority of loci, where the lead SNP often had moved by hundreds of kilobases.

The X-chromosome is often overlooked in GWAS of complex human traits, primarily because of the difference in the distribution of genotypes in males and females. I have extended IntRapid to allow for interactions with the X chromosome by considering males and females separately, and combining effect estimates across the sexes in a fixed-effects meta-analysis. Application to genotype data from the WTCCC failed to identify any strong signals of association with the X-chromosome, despite known epidemiological differences between the sexes for the traits considered.

The novel methods developed as part of this doctoral work enable a user friendly, computationally efficient and powerful way of implementing genome-wide gene-gene interaction studies. Further work would be required to allow for more complex interaction modelling and deal with the associated computational burden, particularly when using "next-generation" sequencing (NGS) data which includes a much larger set of SNPs. However, IntRapid is demonstrably efficient in exhaustively searching for pairwise interactions in GWAS of complex traits, potentially leading to novel insights into the genetic architecture and biology of human disease.

Contents

Nomenclature **xiii**

1 Introduction **1**

1.1 Mendelian and complex genetic diseases 4

1.2 Structure of the Human Genome 7

1.2.1 Polymorphism 7

1.2.2 Haplotype vs Genotype data 8

1.2.3 Linkage Disequilibrium 9

1.3 Association Analysis 12

1.4 Genome Wide Association Studies 13

1.4.1 Study Design : Choice of SNPs 14

1.4.2 Study Design : Choice of samples 15

1.4.3 Statistical Quality Control 15

1.4.3.1 SNP Quality Control 15

1.4.3.2 Sample Quality Control 17

1.5 Testing marginal effects in GWAS 20

1.5.1 Case-control design 20

1.5.1.1 Genotypic test of independence 20

1.5.1.2 Allelic test of independence 21

1.5.1.3 Cochran-Armitage Trend Test 23

1.5.1.4 Generalized linear models 24

1.5.2 Quantitative trait design 25

1.5.3 Software for GWAS 25

1.5.4 Imputation 26

1.5.5 Significance threshold and Multiple Testing 26

1.6 Success and Potential . . . 28
1.7 Limitations of GWAS . . . 28
1.8 Aims of thesis . . . 30

2 Gene-gene Interactions **33**
2.1 Introduction . . . 33
2.2 Challenges and strategies for gene-gene interaction testing . . . 39
2.2.1 Generalised Linear models (GLM) . . . 40
2.2.1.1 Additive-additive interaction test . . . 41
2.2.1.2 General interaction test . . . 41
2.2.1.3 Case only study . . . 42
2.2.1.4 Testing for association allowing for interaction . . . 44
2.2.1.5 Two stage two locus methods . . . 45
2.2.1.6 Higher-order interactions . . . 46
2.2.1.7 Other approaches . . . 47
2.2.2 Data mining and related methods . . . 47
2.2.2.1 Recursive partitioning methods . . . 48
2.2.2.2 Multifactor Dimensionality Reduction method . . . 50
2.2.2.3 ReliefF, Tuned ReliefF and evaporative cooling . . . 51
2.2.3 Bayesian model selection approaches . . . 52
2.2.3.1 Bayesian epistasis association mapping . . . 53
2.3 Chapter Summary . . . 53

3 Coverage and Power of gene-gene interaction tests **55**
3.1 Introduction . . . 55
3.2 Models of interaction . . . 57
3.2.1 Data Simulation . . . 58
3.2.2 Incorporating LD . . . 60
3.2.3 Tests for interaction . . . 62
3.3 Results . . . 62
3.4 Discussion . . . 72
3.5 Chapter Summary . . . 74

4 Genome wide scan for gene-gene interactions **75**
4.1 Introduction . . . 75
4.2 Rapid interaction testing . . . 76
4.2.1 Stage 1 . . . 78
4.2.2 Stage 2 . . . 80
4.3 Software development . . . 81
4.4 Simulations . . . 83
4.5 Application to GWAS data from the WTCCC . . . 87
4.5.1 Quality Control . . . 89
4.5.2 Implementation of interaction scans using IntRapid . . . 90
4.5.3 Results . . . 91
4.5.3.1 Disease phenotypes . . . 91
4.5.3.2 Quantitative trait - BMI . . . 99
4.6 Summary . . . 101

5 Interaction between Imputed signals **103**
5.1 Introduction . . . 103
5.2 Imputation . . . 104
5.2.1 Impute . . . 105
5.2.2 Probability distribution of imputed genotypes . . . 107
5.3 IntRapid for imputed signals . . . 108
5.3.1 Stage 1 . . . 108
5.3.2 Stage 2 . . . 109
5.4 Application to diseases from WTCCC . . . 110
5.4.1 Imputation of signal neighbourhoods . . . 110
5.4.2 Quality control and imputation . . . 110
5.4.3 Results . . . 113
5.4.4 Permutation tests . . . 118
5.5 Discussion . . . 119
5.6 Chapter Summary . . . 120

6 X Chromosome interaction **122**
6.1 Introduction . . . 122
6.2 Methods . . . 123
6.3 Stage 1 of IntRapid for X - Chromosomes . . . 124
6.3.1 Within X-Chromosome : Females . . . 125
6.3.2 Within X-Chromosome : Males . . . 126
6.3.3 Autosomal SNP and X-Chromosome : Females . . . 128
6.3.4 Autosomal SNP and X-Chromosome : Males . . . 128
6.3.5 Odds ratios . . . 129
6.3.6 Meta analysis . . . 130
6.4 Stage 2 - Generalised linear model . . . 132
6.5 Meta-analysis of fixed effects . . . 133
6.6 Implementation in IntRapid software . . . 134
6.7 Application to WTCCC . . . 134
6.8 Chapter Summary . . . 135

7 Discussion **137**
7.1 Limitations and possible solutions . . . 141
7.2 Conclusion . . . 142

References **185**

List of Figures

2.1 Epistasis causing change in fur colour in the Labrador breed of dog. Here "B" stands for black, "b" stands for chocolate, "E" stands for ability to express coat colour and "e" stands for inability to express dark coat colour. 34

2.2 Reproduced from [Moore, 2005] 36

3.1 Power of additive-additive and general interaction tests against MAF and λ for "pure additive" and "pure dominance" theoretical interaction models. 64

3.2 Power of additive-additive and general interaction tests against MAF and λ for "dominant-dominant" and "additive-additive" theoretical interaction models. 65

3.3 Power of additive-additive and general interaction tests against MAF and λ for a theoretical "recessive-recessive" interaction model. 66

3.4 False positive rate (FPR) of additive-additive and general interaction tests against MAF and λ for theoretical "simple additive" and "simple dominant" interaction models. These models have no epistasis and only have additive main effects and dominance main effects respectively. 68

3.5 **Impact of LD**: Power against LD (r^2), for specific cases of MAF and λ where either of the tests have achieved a reasonably large statistical power. We have particularly focussed on cases where the general interaction test appeared to be significantly more powerful than the additive-additive interaction test. 69

3.6 **Impact of LD in a "recessive-recessive" model**: Power against LD (r^2), for MAF of 0.02 at both loci and λ of 2%. Here, the general interaction is significantly more powerful. 71

4.1 **Additive-additive model**: Power levels for various cut off p-values in stage 1 of "IntRapid". Power at a nominal significance threshold of the interaction p-value $p_{GLM} < 10^{-10}$, of the rapid interaction testing strategy for an additive-additive two-locus model of association, with causal allele frequency (CAF) of 50% at one SNP and 10% or 50% at the other. Power is presented as a function of the additive-additive interaction component for a range of thresholds, α_{RAPID}, for carrying forward pairs of SNPs from the rapid testing stage to the GLM analysis framework. 85

4.2 **Dominant-dominant model**: Power levels for various cut off p-values in stage 1 of "IntRapid". Power at a nominal significance threshold of the interaction p-value $p_{GLM} < 10^{-10}$, of the rapid interaction testing strategy for an additive-additive two-locus model of association, with causal allele frequency (CAF) of 50% at one SNP and 10% or 50% at the other. Power is presented as a function of the additive-additive interaction component for a range of thresholds, α_{RAPID}, for carrying forward pairs of SNPs from the rapid testing stage to the GLM analysis framework. 86

4.3 Genome wide plot of interaction p-values from Stage 1 of IntRapid in BD. 93

4.4 Genome wide plot of interaction p-values from a GLM in BD. . . 94

4.5 Plot shows the -log10 p-values in Stage-1 and Stage-2 mostly agree at the smaller end of the spectrum. The level of agreement is much lower in a plot for all -log10 p-values. The horizontal line in plot (b) helps identify cases where the stage-1 p-values were large but the stage-2 p-values were lower than 10^{-4}. 95

4.6 QQ plot of -log10 p-values from stage 2 of applying IntRapid to the WTCCC dataset of BD 96

5.1 An outline of imputation reproduced from Marchini and Howie [2010] 106

5.2 Signal plots for interaction signals found in BD, using LocusZoom 116
5.3 Signal plots for interaction signals found in CHD, using LocusZoom 117

List of Tables

1.1 Possible diplotypes 8
1.2 Contingency table for Score Tests 20
1.3 Contingency table for a pair of alleles 21
1.4 Additive and dominance coding for genotypes 24

3.1 Theoretical models of interaction between two loci. For a given model and combination of genotypes, the mean of the corresponding quantitative phenotypic trait are represented in the matrices (as explained in Table 3.2). 58
3.2 Mean phenotypic responses for combination of genotypes 59

4.1 Contingency table for a pair of SNPs 79
4.2 Contingency table for a pair of alleles 79
4.3 Population mean trait values of theoretical epistasis models 84
4.4 Disease cohorts, their abbreviations and calculated heritability in the WTCCC study 88
4.5 Time required to complete analysis on the Type 2 diabetes and Obesity (as measured by BMI) samples depending on threshold p-value at stage 1 of the "IntRapid" method. 90
4.6 Sample sizes and number of SNPs used in the seven diseases . . . 91
4.7 Gene-gene interaction signals obtained after applying IntRapid on all seven WTCCC diseases. Pairs of SNPs where the p-value corresponding to the interaction term in the GLM satisfies a minimum threshold of 10^{-10} 98

4.8 In obesity as measured by BMI, pairs of SNPs where the p-value corresponding to the interaction term in the GLM satisfies a minimum threshold of 10^{-10} 100

5.1 Expected genotype calculation from imputed data 107
5.2 Contingency table for a pair of imputed SNPs 108
5.3 Contingency table with probability of a pair of alleles 109
5.4 These SNPs were the top hits in their respective 1Mb neighbourhoods after using IntRapid to conduct genome wide interaction scans for all WTCCC1 diseases. An immediate 1Mb region was imputed around these signals using the 1000 genomes reference panel. The only exceptions were the HLA regions in RA and T1D, where a 4Mb region was imputed. 111
5.5 Gene-gene interaction signals obtained after applying IntRapid to imputed neighbourhoods of SNPs previously identified as part of two loci interactions. The SNPs corresponding to the strongest signals in each neighbourhood are reported and thus not all SNPs that appear in the table are imputed. 'Imp.' stands for 'Imputation' 115

6.1 Contingency table for a pair of SNPs in the X-Chromosome of females. 125
6.2 Allelic contingency table for a pair of SNPs in the X-Chromosome for females. 126
6.3 Contingency table for a pair of SNPs in the X Chromosome of males. 127
6.4 Allelic contingency table for a pair of SNPs in males. 127
6.5 Contingency table for an autosomal SNP and an X-Chromosome SNP in males. 128
6.6 Allelic contingency table for an autosomal SNP and an X-Chromosome SNP for males. 129

Chapter 1

Introduction

Genome wide association studies (GWAS) have successfully mapped many genes associated with a variety of complex diseases [Barrett et al., 2008, Wellcome Trust Case Control Consortium, 2007, Wellcome Trust Case Control Consortium and Australo-Anglo-American Spondylitis Consortium (TASC), 2007]. These carefully designed studies have made significant contributions to our understanding of the genetics of complex diseases. Recently conducted large scale meta-analyses (where multiple independent GWAS are combined in a statistically rigorous way) involving thousands of individuals have uncovered several moderate to small sized effects for a variety of complex traits [Barrett et al., 2009, Cooper et al., 2008, Lohmueller et al., 2003, Zeggini et al., 2008]. Further, expanding existing datasets via imputation with respect to dense reference panels like the HapMap [1000 Genomes Project Consortium, 2010, 2012, The International HapMap Consortium, 2003, 2005, 2007] has improved the chances of identifying additional signals. These efforts have helped improve power and coverage of GWAS. However, the primary analysis performed on GWAS usually pertains to tests for single variant effects, otherwise known as main effects. Depending on the sample size, single variant analysis is typically able to identify only those common variants which

have at least a moderate sized *marginal* (i.e. genotypes at the variant have different disease risk or mean trait values). For most complex traits, even with all of these signals put together, we are still only able to explain a small fraction of the estimated "heritability" [Manolio, 2010, Manolio et al., 2009] (i.e. the amount of variance in the trait that is explained by genetic factors).

One possible explanation for this gap in our understanding is interaction between genes, whereby the effect of one is modified by another. It has been suggested that looking for pairwise and other higher order statistical interactions could be a potential way forward to try to understand the genetic framework underlying complex diseases. It has been demonstrated that two or multi-gene interactions exist in several different species [Albrechtsen et al., 2007, Coffey et al., 2004, Flint and Mackay, 2009, Zee et al., 2002]. When these interaction pathways are combined with environmental factors, the disease risk or distribution of the quantitative trait changes. However, Coffey et al. [2004] could not cross-validate the initially detected interaction signals for risk of myocardial infarction in humans. They emphasise the need for well validated interaction signals, given the risk of finding signals which in reality do not exist (i.e. false positive signals). Several obesity related gene-gene interactions have been identified in mice [Stone et al., 2006, Stylianou et al., 2006, Warden et al., 2004]. Yu et al. [2008] produced a database of all reported interactions in Drosophila. Dong et al. [2003] and Dong et al. [2005] have reported the existence of gene-gene interactions related to extreme obesity in humans. Strange et al. [2010] found evidence of interaction between the *HLA-C* and *ERAP1* loci in psoriasis. The signal was replicated in independent samples.

The interplay between genes, often termed "epistasis", could mean the alter-

ation (suppression or magnification) of the action of one gene in the presence or absence of another [Cordell, 2002]. It is possible that some of these interactions could have very small or no marginal effects at all. It has been suggested [Carlborg and Haley, 2004, Williams et al., 2004] that similar signals could exist in humans and thus could help in explaining a larger proportion of the genetic contribution to phenotypic variance, which remains unexplained at this stage.

However, so far, gene-gene interactions have not been routinely evaluated at a genome wide level in humans. Besides the lack of statistical power in most GWAS datasets, even a two locus exhaustive scan poses a massive computational burden. These problems may have restricted researchers from considering this as an avenue for exploring further genetic complexities. However, with substantially cheaper genotyping costs, and merging of several big studies (meta-analyses), we now have larger sample sizes than ever, with some present day studies having sample sizes of more than $250,000$ individuals [Heard-Costa et al., 2009, Lindgren et al., 2009]. We also have more dense genotyping arrays (often with more than 1 million markers) leading to better coverage of genetic variation. With better computational power and availability of sophisticated computing clusters, we are now able to run complex statistical algorithms on datasets larger than ever. But the sheer increase in sample size comes with several statistical challenges.

This thesis aims to look into new methods to scan for gene-gene interactions on a genome wide scale. The new methods need to be computationally efficient so they can be implemented on ever larger datasets and perform exhaustive interaction scans. The methods will also be implemented in user friendly software and made freely available to the wider research community. Simulation studies will also need to be performed at each stage to compare the power achieved with

such new methods, compared to more traditional methods.

This chapter gives a brief introduction to issues arising from present day approaches to identify genetic markers which affect common complex disorders. It also includes a discussion of some advanced modelling techniques in the field. It includes a brief overview of GWAS, its development over the years and its success in finding disease associated genetic markers in the human genome.

1.1 Mendelian and complex genetic diseases

Genetics has come a long way since the theories of discrete inheritance suggested by Gregor Mendel [Hutchinson et al., 2005] in the mid-nineteenth century. From what started as an observation of traits through generations, genetics has now been transformed into a much wider subject group, with much more detailed knowledge of the physical basis of inheritance. The discrete units of inheritance noted by Mendel would be analogous to the present day idea of genes. Although it started with the Mendelian theory of inheritance, modern genetics perhaps owes its beginning to the discovery of the double helical structure by Watson and Crick [Watson and Crick, 2003]. It is only after their discovery of the structure of Deoxyribonucleic acid (DNA) in 1953 that we started making real inroads into the detailed framework of the human genome. More recently, with the completion of the Human Genome Project, and massive improvements in the efficiency of genotyping technology, we were able to genotype several hundreds of thousands of genetic variants over thousands of individuals [Greenhalgh, 2005, Platzer, 2006]. It has also been pointed out that, more often than not, inherited traits are not transmitted as discrete units as Mendel had described. Much rather they are

passed on as a complex genetic network, most of which we are yet to unravel. It seems likely that common complex diseases are triggered by several genes acting together in different parts of the genome and even interact with each other and the environment at different levels of complexity. Here, interaction between multiple genes could be defined as the phenomenon under which the presence of these genes causes an amplified or suppressed effect on the phenotype, compared to the otherwise expected additive effect. Thus a significant interaction effect between multiple genes could be tracked measuring the deviation from the sum of the independent effects in question (Refer to Chapter 2 for further details).

In Mendelian disorders, such as sickle-cell anaemia [Christensen and Murray, 2007], Tay-Sachs disease, cystic fibrosis and xeroderma pigmentosa, disease status is typically determined by independent single gene mutations which are passed on from one generation to another as discrete genetic units. However, complex diseases can be caused by hundreds of variants, and many of these may work together in a complex manner. The approach in single gene disorders relies on a strong association between the phenotype and the genetic variant in question, besides a detailed knowledge of the mode of inheritance of the genes in question [Hamosh et al., 2005].

For Mendelian diseases, mutations are "highly penetrant", i.e. have a high probability of resulting in affection in a carrier, depending on the mode of inheritance. In complex diseases, the traits are quite often affected by different sets of genes in different biological 'pathways'. Usually, most genes have a moderate to small sized effect, so that each individual disease causing mutation has only a small probability of resulting in affection in a carrier. It thus requires huge sample sizes to successfully map them. The levels of penetrance also differ

based on environmental interactions and other epigenetic causes i.e., effects over and above the effect of underlying DNA sequences. For example, it is common knowledge that smoking is associated with cancer. However, if an individual had genes associated with cancer, besides being a regular smoker, he/she might have a magnified probability of developing the disease. This is referred to as "gene-environment" interaction. Many complex diseases may be heterogeneous, such that affected individuals may have very different underlying genetic make-up. Studies like those of Barrett et al. [2008], Scott et al. [2007], Wellcome Trust Case Control Consortium [2007] are some of the many studies performed recently to track down the multiple variants associated with complex diseases. A detailed map of disease linked (or associated) variants is still unavailable for all complex diseases. Much of the heritability of major complex traits remains unexplained. Consequently, it has been proposed that, the more complex gene-gene or gene-environment interactions need to be investigated for all diseases. Method development for understanding genetic interactions have seen rapid progress in recent years and there is now more interest that ever to understand the more complex patterns in the human genome.

Modern techniques, like high-throughput genotyping, have brought down the costs of assaying genetic variation in humans. With more precise methods to read data, we are now able to have very small error rates at the order of millions of variants from many thousands of individuals. Moreover, advanced statistical algorithms, better computing power, and much larger sample sizes have enabled us to identify some of these moderate or small effects with very low false positive rates and, adequate levels of power for marginal effects. We are now in a situation where we know the amounts of data will increase exponentially from here onwards,

but the kind of statistical tools we have are quite limited, and certainly not equipped to decipher more complex statistical relationships in the datasets.

1.2 Structure of the Human Genome

Human beings, like many multicellular organisms, are made of billions of small units of life called cells. A cell is the basic structural and functional unit of any living organism. Eukaryotic cells hold membrane enclosed organelle called the nucleus. The nucleus contains multiple long linear dioxyribonucleic acid (DNA) molecules along with proteins, ribonucleic acid (RNA) molecules, and surrounding structures to form chromosomes. Humans have 23 pairs of chromosomes, out of which 22 pairs are autosomal and the remaining two form the sex chromosomes - XY for men and XX for women.

1.2.1 Polymorphism

There are four different nucleotide bases found in DNA, which are adenine (A), guanine (G), cytosine (C), and thymine (T). These hold the key to inherited human traits and diseases. At a particular site, or a short interval of sites, the base might vary between individuals, so it is referred to as polymorphic (also known as **polymorphism**) [Hartl and Clark, 2007]. Each base observed at the site is referred to as an **allele**, and for a single nucleotide site, or **base pair** (bp), the polymorphism is referred to as a **Single Nucleotide Polymorphism** (short form is SNP - pronounced 'snip'). Human beings inherently have a very low mutation rate (10^{-8} per base pair per generation), and hence most SNPs in humans are diallelic. Most polymorphisms in a human are SNPs and analysis of

this type of variation is the foundation of association studies.

Locus (plural loci) is a specific site in the genome. Because SNPs are typically di-allelic, humans usually have three possible genotypes at a locus. These three genotypes are the combinations of the two possible alleles at the site.

1.2.2 Haplotype vs Genotype data

A haplotype can be defined as a set of variants or polymorphisms in the same strand of a chromosome that are inherited together from the same parent. It can be a combination of alleles or a set of SNPs on the same chromosome [Hartl et al., 1997]. For a pair of di-allelic SNPs in a population there can be up to four possible haplotypes. However, there can be ten possible haplotype pairs for any diploid individual. A pair of haplotypes is referred to as **diplotypes**. For example, let us assume that in one particular location for alleles A and G, we have possible genotypes AA, AG and GG, and at a neighbouring location for alleles C and T, we have possible genotypes CC, CT and TT. This could mean that for an individual there could be ten possible assignments at the two loci (Table: 1.1). However, it is not possible determine the phase of the double heterozygote.

	AA	**AG**	**GG**
CC	AC/AC	AC/GC	GC/GC
CT	AC/AT	AC/GT or AT/GC	GC/GT
TT	AT/AT	AT/GT	GT/GT

Table 1.1: Possible diplotypes

1.2.3 Linkage Disequilibrium

The closer a pair of loci are, the higher their probability of being inherited together [Hartl and Clark, 2007]. The two alleles of a given genotype are inherited from the two parents. For variants that are located close to each other, the alleles are in gametic phase and hence are usually transmitted together. For loci which are further away from each other it is more likely that recombination events could split up the alleles during meiosis. Based, on this principle, the recombination rate (θ) is defined as the probability of recombination between two loci.

Since blocks of SNPs that are close to each other are typically inherited together, the genotypes within an individual are statistically correlated at these variants. This allows us to select only a few SNPs from a block, and yet represent a majority of the genetic variation contained in the region. If the SNPs were inherited completely at random, for a given SNP, we would expect the haplotype frequency to be equal to the product of the constituent allele frequencies. Any deviation from this equilibrium is used to represent correlation between SNPs. This is also known as linkage disequilibrium (LD).

Mathematical definition Let p_{ij} denote the population proportion of ij haplotypes in a population, while $p_{i.}$ and $p_{.j}$ denote the marginal allele proportions, which could be assumed to be constant over generations. Under linkage equilibrium, population haplotype proportions are constant over time, and using the definition of statistical independence, it is equal to the product of the constituent allele proportions:

$$p_{ij} = p_{i.}p_{.j}, \quad \text{for } i, j = 0 \ \& \ 1 \tag{1.1}$$

Deviation from this equilibrium state is referred to as LD. There are several popular measures of LD.

1. The difference between each haplotype proportion and its value under linkage equilibrium:

$$D_{ij} = p_{ij} - p_{i.}p_{.j} \tag{1.2}$$

can be estimated by replacing the population proportions with the corresponding sample proportions. When both loci are diallelic,

$D_{00} = D_{11} = -D_{01} = -D_{10}$.

2. D_{ij} has several limitations as a measure of LD. In particular, the range of possible values of D_{ij} depends on the allele proportions. For example, in the diallelic case we have

$-D^- \leq D_{ij} \leq D^+$ where,

$$\begin{aligned} D^- &\equiv \min\{p_{0.}p_{.0}, p_{1.}p_{.1}\} \\ D^+ &\equiv \min\{p_{0.}p_{.1}, p_{1.}p_{.0}\} \end{aligned}$$

Thus, a value of D_{ij} that arises for one pair of loci may be impossible at another pair given their allele proportions, making cross-locus comparisons difficult. To overcome this problem we can use,

$$D'_{ij} = \begin{cases} D_{ij}/D^{+} & \text{if } D_{ij} \geq 0 \\ D_{ij}/D^{-} & \text{if } D_{ij} < 0 \end{cases} \tag{1.3}$$

which can take any value between -1 and 1, irrespective of the allele proportions.

3. D_{ij}^2 can be scaled to obtain the squared correlation coefficient which in this context is usually called r^2:

$$r_{ij}^2 = \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}(1-p_{i.})(1-p_{.j})}. \tag{1.4}$$

A value of 1 for r^2 represents perfectly correlated SNPs where only two possible haplotypes exist. A value of 1 for D' corresponds to "complete LD" when only three of the four possible haplotypes between a pair of SNPs.

Importance of studying LD The study of LD plays a central role in several aspects of genetics. Its importance has been long recognized, primarily because of the information it provides regarding the inheritance of haplotypes and hence identification of disease genes.

LD enables design of efficient GWAS genotyping arrays by allowing genotyping of only a subset of variants, and yet make inferences about correlated SNPs (refer to 1.4.1 for details). The International HapMap Project [The International HapMap Consortium, 2003, 2005, 2007] and more recently the 1000 Genomes Project [1000 Genomes Project Consortium, 2010, 2012] have played key roles in studying LD structure in multiple populations from different ethnic groups, and has also helped in providing easy access to such information online. These

reference panels of human genetic variation have enabled researchers to work towards the identification of disease associated genes, as described below.

1.3 Association Analysis

For most complex diseases, many variants are believed to influence the disease phenotype. Linkage studies focussing on the inheritance of genetic variants with disease in families, have had limited success in identifying loci for complex diseases [Gulcher et al., 2001]. Association studies may be more powerful when the disease is multi-factorial and the genetic effect of each locus is small [Long and Langley, 1999, Risch and Merikangas, 1996, Schork et al., 1998, Tu and Whittemore, 1999]. Association studies typically use independent samples from a homogeneous population, collecting cases and controls for a binary disease phenotype. This makes it much easier for association studies to collect reasonable sample sizes since large families with multiple affected members needed for linkage analysis are not required.

In an association analysis, genotype frequencies are tested for correlation with a disease trait, which is already known to be genetic. However early association studies focused on variants in pre-specified "candidate" genes of interest. However, these methods had limited success in finding variants which explained variability in complex traits. Quite often, these studies were not very well designed. For example, they had small sample sizes, the cases and controls were not well matched for demographic features or potential confounding factors, the results were not properly adjusted for multiple comparisons, or were obtained from heterogeneous populations, and sometimes the candidate genes were not definitive

[Cardon and Bell, 2001].

For binary traits where, for example, individuals can be characterised as 'affected' or 'unaffected', it is typical to make use of a 'case-control' design. With cheaper genotyping costs, we can now afford to genotype millions of variants across the genome of thousands of individuals, enabling us to implement association studies on large sample sizes. Further, the coverage and power of these studies could be increased by accounting for the LD between genotyped and non-genotyped SNPs. Information from standard SNP reference panels [The International HapMap Consortium, 2003] can be used to expand the datasets so they contain information on many more variants than were typed in the study. This process, also known as imputation [Halperin and Stephan, 2009, Howie et al., 2009], enables the inclusion of SNPs which were otherwise not present in the genotyping chip of the study (further details in Chapter 5). With better study design and quality control approaches, GWAS have proved to be more powerful than candidate gene based association studies. The next section explains GWAS in greater detail.

1.4 Genome Wide Association Studies

GWAS have established themselves as the most popular means of exploring the human genome for variants contributing effects to complex human traits. However, this involves a whole range of statistical procedures to optimize the power of study. Before testing for association with the trait many statistical quality control measures are carried out. This minimizes the chance of coming across false positive associations due to poor quality genotyping or study design.

1.4.1 Study Design : Choice of SNPs

A popular hypothesis is that genetic effects of complex diseases may be embedded in commonly varying SNPs - defined to have minor allele frequency (MAF) of at least 5%. This hypothesis has led us to investigate complex trait associations by genotyping only the commonly varying SNPs. The SNPs used for a study in a given population, would depend on the choice of the product used to genotype variants. As some modern studies have more than a million SNPs, it is quite likely that a common causal variant will be in high LD with one included on the genotyping chip used. This is often called the 'common disease, common variant' hypothesis.

Most modern genotyping chips can contain hundreds of thousands of commonly varying SNPs, which covers a majority of the common variation in the human genome because of the structure of LD. Information about LD values across the genome helps us identify regions where a higher density of SNPs need to be genotyped to capture variation. The Hap Map Project [The International HapMap Consortium, 2003] has a detailed scan of SNPs which represent blocks of the genome with high LD, also known as tagSNPs [Botstein and Risch, 2003]. There are selections of tagSNPs for different populations and these are quite often used to design genotyping chips for large scale GWAS. This acts as a massive cost saving technique [Daly et al., 2001, Gabriel et al., 2002, Patil et al., 2001], as well as increasing the possibility of covering the 'causal' SNP in the analysis, assuming it is common. The process of selection and prioritization of tag SNPs is key to a successful association study [Carlson et al., 2004, Ke and Cardon, 2003, Meng et al., 2003, Stram et al., 2003, Weale et al., 2003].

1.4.2 Study Design : Choice of samples

For any GWAS, it is vital to make sure that all individuals are unrelated and come from the same population. Besides the genotype data, information on possible confounding variables are also collected. These include variables like genotyping batch, ethnicity, non-genetic risk factors. These potentially secondary variables are used in downstream analysis to ensure the genetic signals are not owing to confounding factors. Further, for case-control studies, samples should be as homogeneous as possible to improve statistical power of the study - both in terms of ancestry and phenotype.

1.4.3 Statistical Quality Control

The key to a successful GWAS is to ensure statistical quality control (QC) of the genotype data is conducted robustly and genotypes or individuals are removed when the quality is not acceptable [Turner et al., 2011, Weale, 2010]. Genotyping error can create noise in the data, which might cloud downstream analyses and result in false positive or false negative association signals. Thus, the interpretation of the study results would be severely compromised if the initial quality control is not carried out effectively.

1.4.3.1 SNP Quality Control

Marker Genotyping Call Rate: Genotype calling forms a fundamental process to determine the genotypes of a given variant in a sample of individuals and are typically estimated by plotting the signal intensities of the two alleles. For looking at large number of variants, efficient calling algorithms are required to

determine genotypes. There are a variety of genotype calling algorithms [Marchini and Howie, 2008] available, and quite often the choice is dependent on the nature of the genotyping chip used. The output usually gives a probability for each of the three possible genotypes at a given variant for each individual. For the calling algorithm Chiamo [Marchini and Howie, 2008], a genotype is "called" if the probability of the most likely genotype, is at least 90%, and a missing genotype otherwise.

SNPs with high levels of missingness are more likely to have occurred in situations where an automated algorithm struggles to make a definitive decision above genotype calls. SNPs with large rates of missingness often reflect poor quality DNA or difficulties in genotype calling because of other technical issues. The threshold for missingness will vary from one study to another. The Wellcome Trust Case Control Consortium (WTCCC) used a missing rate of 5% [Wellcome Trust Case Control Consortium, 2007]. The choice of threshold is usually made by considering the distribution of missing genotype rates across SNPs to identify clear outliers in the distribution.

For disease phenotypes, missingness rates of all SNPs are typically compared between the cases and controls and those with a significant difference are eliminated from the study [Clayton et al., 2005].

Deviations from HWE: Hardy Weinberg equilibrium (HWE) is a theoretical idea stating that, under random mating, the overall allele and genotype frequencies in a population remain constant. Any substantial deviations from this equilibrium can be indicative of genotyping error [Teo et al., 2007]. However, under a recessive disease model, and in particular for Mendelian diseases, one

might genuinely expect significant deviations from the HWE. Hence very stringent thresholds (p-values have to be very small) are used to judge significance for deviation from HWE. Several different procedures have been proposed to measure HWE, however in principle it can be tested by comparing the genotype and allele frequencies in a sample, typically by means of an exact test.

Minor Allele Frequency (MAF): In association studies, common SNPs tend to have relatively smaller false positive error rates than rare SNPs [Gorlov et al., 2008, Moskvina et al., 2006, Tabangin et al., 2009]. For low MAF (less than 5% or 1% for most studies depending on sample size), the power to detect any associated signals would typically be very low under the common disease, common variant hypothesis. This has prompted standard QC procedures to exclude SNPs with low MAF from downstream analysis.

1.4.3.2 Sample Quality Control

Sample Genotyping Call Rate: When a large fraction of genotype calls for a sample are missing, it reflects poor quality DNA [Turner et al., 2011]. This could lead to noisy data in the genotype calling process. Thus samples with low genotyping call rate are eliminated from downstream analysis. As with SNP QC, the threshold further varies from one study to another, depending on the genotyping platform used, quality of DNA samples, and other human or equipment errors. The appropriate threshold for sample call rate will vary from one study to another. In the WTCCC study 250 samples with more than 3% missingness across all SNPs were removed [Wellcome Trust Case Control Consortium, 2007].

Heterozygosity: Cross sample contamination or poor quality DNA can lead to higher levels of heterozygous genotypes than expected. Since poor quality DNA affects both heterozygosity and missingness, a usual practice plots these parameters against each other. The appropriate threshold will vary from one study to another and will depend on the allele frequencies of SNPs on the genotyping array. The WTCCC removed six samples with heterozygosity of more than 30% and another three with heterozygosity of less than 23% [Wellcome Trust Case Control Consortium, 2007].

Relatedness: A design for large scale association studies requires the samples to be unrelated to each other. However, at times, unknowingly, the sample ends up including relatives or perhaps erroneously includes the same DNA sample of the same person twice. The extent of alleles shared between every pair of individuals can be calculated, a measure referred to as Identity-by-State (IBS). Given IBS information for a homogeneous sample, extent of DNA shared between pairs of individuals can be calculated genome-wide, also known as Identity-by-Descent (IBD). Siblings are expected to have an IBD of around 50% and identical twins and duplicated samples should have an IBD of nearly 100% (allowing for genotyping error). When a pair of related individuals are identified, the sample with the highest rate of missingness is typically removed from downstream analysis.

Population Substructure: Population substructure is observed when subsets of the sample exhibit systematic differences in their population ancestry and phenotype of interest. It is critical to account for these differences to avoid spurious associations [Cardon and Palmer, 2003]. This can be achieved by en-

suring the samples come from a homogeneous population. Self reported ancestry can be matched against genetically inferred ancestry to ensure homogeneity. Often combining data from multiple sites in a joint analysis can lead to population stratification if both allele frequencies and phenotypes differ significantly between groups.

Statistical methods have been developed to account for population stratification in GWAS. **Genomic Control** estimates an 'inflation' factor and corrects all test statistic estimates downwards [Devlin and Roeder, 1999]. **Structured association** uses genotype data to infer population structure and conducts association tests in each sub-sample [Falush et al., 2003, 2007, Pritchard et al., 2000]. It is also used to detect outliers in an otherwise homogeneous sample which can then be eliminated from further analysis. The algorithm was implemented in a software called STRUCTURE (`pritchardlab.stanford.edu/structure.html`).

The risk of false positive signals caused by confounding population stratification increases with large sample sizes [Marchini et al., 2004]. The EIGENSTRAT approach was developed to overcome these concerns [Patterson et al., 2006, Price et al., 2006]. It uses principal component analysis (PCA) to explicitly detect and adjust for population structure for very large data sets in a computationally efficient manner. PCA reduces a set of correlated variables to independent variables, called principal components (PC), such that the first PC captures as much of the variance as possible and the subsequent PCs capture the most variance possible subject to a orthogonality condition with the preceding PCs. Adjusting for PCs as covariates in downstream association analyses can thus be used to adjust for population substructure.

1.5 Testing marginal effects in GWAS

1.5.1 Case-control design

In a case-control design, a null hypothesis of no association (or complete independence) between the genotypes of a given SNP and the disease status of individuals is tested. For a sample of n individuals, with $n_{1.}$ cases and $n_{2.}$ controls, a single SNP with three genotypes (for example AA, AG and GG) could be considered. If, n_{ij} were the number of genotypes, j, observed for a phenotype, i (phenotype 1 being cases and 2 being controls), then Table 1.2 gives the contingency table for the SNP in question.

Table 1.2: Contingency table for Score Tests

Disease Status	Genotypes			
	AA	AG	GG	Total
Case	n_{11}	n_{12}	n_{13}	$n_{1.}$
Control	n_{21}	n_{22}	n_{23}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

1.5.1.1 Genotypic test of independence

Under assumptions of complete independence between the disease status and genotypes, the expected frequencies are calculated (E_{ij}).

$$E_{ij} = \frac{n_{i.}n_{.j}}{n} \tag{1.5}$$

The following statistic, equation (1.6), is then used to test for the extent of deviation from this null expectation:

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1.6}$$

where, $\chi^2_{(r-1)(c-1)}$ is the test statistic, n_{ij} are the observed frequencies or n_{ij}, E_{ij} are the expected frequencies under a null hypothesis of no association, r and c stand for the number of rows (two for a case-control design) and columns (three for diallelic genotypes) respectively. The number of degrees of freedom is calculated as $(r-1)(c-1)$.

For a simple test of association, when there are three genotypes (three columns) and a binary phenotype (two rows) the degrees of freedom is two. When the genotype counts are small, and the asymptotic assumptions of a chi-squared test of association are not valid, one may use a Fisher's exact test [Agresti, 2007].

1.5.1.2 Allelic test of independence

It is typical to assume a linear relationship between the number of minor alleles in a genotype and disease risk (or more precisely log-risk). Under this model, a 2×2 contingency table is created for the case-control status of a disease against the alleles (A and G say). We may create the following contingency table for the disease status against the allele counts.

Table 1.3: Contingency table for a pair of alleles

Disease status	Alleles	
	A	G
Cases	w	x
Controls	y	z

where (from table 1.2),

$$\begin{aligned} w &= 2n_{11} + n_{12}, \\ x &= 2n_{13} + n_{12}, \\ y &= 2n_{21} + n_{22}, \\ z &= 2n_{23} + n_{22} \end{aligned}$$

A chi-squared test of independence may then be applied on this table (as described in 1.5.1.1). An assessment of the direction and magnitude of the effect of the SNP on disease risk can be provided by the odds ratio (OR) which, for a 2×2 table of this kind, is calculated as follows.

$$OR = \frac{w \times z}{x \times y} \tag{1.7}$$

and the corresponding variance would be,

$$Var(log(OR)) = \left(\frac{1}{w} + \frac{1}{x} + \frac{1}{y} + \frac{1}{z}\right) \tag{1.8}$$

This method, however, requires an assumption of HWE in cases and controls and 'double counts' individuals by focussing on alleles. A more appropriate approach, under a linear model in the allelic log-OR ($\Psi_{GG} = 2 \times \Psi_{AG}$), is performed by means of the Cochran-Armitage trend test. Unlike the allelic association test the Cochran-Armitage trend test remains unbiased when there is significant deviance from the HWE [Guedj et al., 2008].

1.5.1.3 Cochran-Armitage Trend Test

The Cochran-Armitage trend test is a method to detect association between two categorical variables which have 2 and k categories respectively [Freidlin et al., 2009]. A cross-tabulation results in a $2 \times k$ table. The test is particularly useful if there are reasons to believe the existence of an ordinal trend in the k categories of the second variable.

In this case, k represents the number of genotypes ($k = 3$) and we assume the presence of each additional 'causal' allele in the genotype results in a linear increase of log-OR. Using notation from Table 1.2, the test statistic is,

$$T \equiv \sum_{i=1}^{3} t_i(n_{1i}n_{1.} - n_{2i}n_{2.}), \tag{1.9}$$

where t_i are weights depending on the nature of underlying assumption. We would use $t = (0, 1, 2)$ for an additive model [Sasieni, 1997].

Under the null hypothesis of no association,

$$Z = \frac{T}{\sqrt{\mathrm{Var}(T)}} \sim \mathrm{N}(0, 1) \tag{1.10}$$

where,

$$\mathrm{V}ar(T) = \frac{n_{1.}n_{2.}}{n}\left(\sum_{i=1}^{3} t_i^2 n_{.i}(n - n_{.i}) - 2\sum_{i=1}^{2}\sum_{j=i+1}^{3} t_i t_j n_{.i} n_{.j}\right), \tag{1.11}$$

1.5.1.4 Generalized linear models

A more flexible approach for testing association is to make use of a generalized linear modelling (GLM) framework which, in the context of a binary trait, models the logarithm of the odds of disease as a linear function of the genotypes (equation (1.12)). Specifically,

$$g(y) = \beta_0 + \beta_{A_1} x_{A_1} + \beta_{D_1} x_{D_1} \tag{1.12}$$

where, $g(y)$ is the link function for the disease phenotype (logistic for a case control design), β_{A_1} and β_{D_1} correspond to the "additive" and "dominance" effects of the SNP. The terms x_{A_1} and x_{D_1} are indicator variables representing the genotype of the individual summarized in Table 1.4. Typically, we focus on the additive component and assume $\beta_{D_1} = 0$. The resulting test is equivalent to a Cochran-Armitage trend test and is usually used for single variant analysis of GWAS. The association test is performed by comparing model (1.12) with a null model, where β_{A1} and β_{D1} are zero. Under the null hypothesis, twice the difference in the log-likelihoods of these two models has a chi-squared distribution, with degrees of freedom equal to the number of parameters set to zero.

Table 1.4: Additive and dominance coding for genotypes

Genotype	Additive Component x_{A_1}	Dominance Component x_{D_1}
AA	0	0
AG	1	1
GG	2	0

The flexibility of modelling offered by logistic regression can be used to account for confounding factors, such as gender and age, by including them as

covariates in equation 1.12. In the presence of population substructure, PCs from EIGENSTRAT can be used as covariates to account for the stratification.

1.5.2 Quantitative trait design

When the phenotype is continuous in nature, an ANOVA or a linear regression framework could be used. The ANOVA, like the χ^2 test used for case control designs, checks for deviation from a state of no association (or complete independence) between the SNP and the phenotype (in this case the mean trait values of each genotype will be equal). When the alleles are, however, assumed to be acting additively, we assume a linear relationship between the number of alleles in the genotype and the trait. This assumption reduces the degrees of freedom of the test to one, making it more powerful when the assumption of additive effect of the alleles holds. A suitable transformation of the continuous phenotype may be required, to have a resultant variable following normality. In this case, the link function $g(y)$ (from equation 1.12) is identity.

1.5.3 Software for GWAS

The afore described statistical tests could be performed in standard statistical software such as R, Matlab and SAS. However, specialist software, such as PLINK and SNPTEST, has been designed that allow for quality control and handling of large data sets. PLINK and SNPTEST are freely available and widely used by the statistical genetics community. Many well established methods in statistical genetics are readily available in these packages and it helps many researchers focus on the analysis and the biological implications of the results, rather than

programming.

1.5.4 Imputation

Most modern genotyping arrays contain in the order of 500,000 to one million SNPs, and cover more than 80% of the common variation found in European ancestry populations [Spencer et al., 2009]. However, this coverage could be improved by using more detailed genotyping reference panels like HapMap or 1000 Genomes. SNPs not present in the genotyping chip could be imputed by exploiting the LD between 'observed' and 'non-observed' SNPs in more dense reference panels. Details of this method are described in [Howie et al., 2009]. This improves the coverage of the study and increases the chances of finding signals which, owing to LD, would have been too weak to find otherwise. Software like MACH and IMPUTE are available to carry out genotypic imputations on the scale of the whole genome, and work seamlessly with reference panels from HapMap and the 1000 Genomes Project [1000 Genomes Project Consortium, 2012, The International HapMap Consortium, 2007]. One might either threshold imputed probabilities at 90% (say) to decide on the most likely genotype at an imputed location or compute the expected number of minor alleles at a given location using imputation probabilities. Imputation will be discussed in much greater detail in Chapter 5.

1.5.5 Significance threshold and Multiple Testing

A statistical test is typically considered 'significant' if the p-value is less than 0.05. This significance threshold ,α, is the probability of observing a false positive

given the data, in this case 1/20. If, however, multiple (n) independent tests are carried out in the same dataset, the number of expected false positives is $\alpha \times n$. In the GWAS context, for a genotyping chip with a million independent or weakly correlated markers, and hence a million statistical tests, the expected false positives are around $50,000$, at a 5% significance threshold. To correct for this, one may use the Bonferroni's correction where the usual significance threshold of 5% is divided by the number of tests performed to yield a new much lower α level. The Bonferroni test is conservative because it does not take account of correlation between tests due to LD. Hence, other methods have been suggested, including the permutation test which can be computationally demanding [Pe'er et al., 2008]. A wide range of analytical and empirical approaches have been used to determine the appropriate level of significance for GWAS of common variants. Most approaches point to a significant threshold of the order of $p \leq 5 \times 10^{-8}$ and this is now widely accepted as the "gold standard" threshold for genome-wide significance [Pe'er et al., 2008].

Once an association study is conducted with due precautions, signals should ideally be replicated in different samples from the same population (or in other). In order to get the most out of existing data sets one might combine them together in a statistically sound manner to identify genetic signals with a relatively small effect size.

However, individual level genotype data are typically not available to researchers interested in conducting combined analysis. However, meta-analysis can be used to combine association summary statistics (log OR, variance, p-value) across studies, without the need to exchange individual level data. Summaries of approaches to perform meta-analysis are provided by Evangelou and Ioannidis

[2013], Han and Eskin [2012]. Meta-analysis has also been enabled in the context of GWAS through imputation, allowing studies typed with different genotyping arrays to be combined at the same set of SNPs in a reference panel.

1.6 Success and Potential

GWAS have successfully uncovered many novel loci which hold association with complex diseases including Crohn's, type 2 diabetes and macular degeneration [Barrett et al., 2008, 2009, Neale et al., 2010, Scott et al., 2007]. It has helped us obtain insights into the biological pathways of these diseases, a key objective of conducting these studies. Even when we are not completely sure of the precise functional role of the discovered variants, with more and more data coming in, and with systematic investigations into the underlying biological pathways, we should be able to understand the complex pathophysiology much better. With cheaper genotyping and larger arrays, we now have larger sample sizes and better coverage of common SNPs than ever before. Furthermore, with reference panels from the 1000 Genomes Project, we are able to impute more than 30 million variants genome wide including many at much lower frequency. This has helped us in discovering many loci, each with moderate effects on disease, and is expected to increase the amount of variance explained in the phenotype [Morris et al., 2012].

1.7 Limitations of GWAS

Despite the promise that GWAS bring, there are limitations to this approach [Ward and Kellis, 2012]. First phenotypes are not always very well defined. This

ambiguity would naturally collude any findings achieved in downstream analyses. Once we are able to decide on a coherent definition of the phenotype, there are issues of multiple possibilities of underlying genetics for the same phenotype.

The thousands of discovered and replicated disease variants have very modest effects and account for a very small fraction of the observed phenotypic variance or estimated 'heritability' [Manolio et al., 2009]. Thus these loci are not very useful for disease risk prediction.

Most discovered loci are characterised by common variants with association signals mapping to hundreds of kb because of LD. This makes fine mapping and discovery of the specific causal variant in each locus difficult. Most of the identified loci are non-coding and many are quite far from discovered genes. Thanks to LD it is impossible to be sure about the exact variant which is 'causal' and we can only be imprecise about its likely location. This makes downstream functional analysis somewhat difficult [Cooper and Shendure, 2011].

GWAS are not designed to evaluate the impact of low-frequency and rare genetic variation on complex traits. Some evidence has been produced to show that rare variants and functional variations could have a role to play in explaining phenotypic variance [Cirulli and Goldstein, 2010]. This issue is certainly being addressed at the moment, with studies like the 1000 Genomes Project, large-scale whole-genome and whole exome sequencing projects in process. There have been other studies to show that copy number variants may also be able to explain significant amounts of the 'missing heritability'.

GWAS analyses have typically focussed on simple, single variant tests of association. It is likely that the actual genetic make up of these diseases is very complex and so far we have been over simplifying our hypothesis. It is probably

quite unrealistic to assume that there would be marginal effects of genes and, in general, little or no statistical interaction between them. In the presence of statistical interaction, the sum of the effects of individual genes on the phenotype is diminished or magnified (further details in Chapter 2). One of the major reasons why we have explained so little of the phenotypic variance could be because we typically do not account for interactions. As described earlier, evidence from animal models suggest that higher order interactions of genes could explain more of the variance. Quite often, even when there are no marginal effects for some genes, they may still have interaction effects. In other words, functions of some genes can be triggered by another gene, which could possibly lie in a completely different part of the genome. There could also be environmental effects, which could silence genes and hence have a massive effect on the phenotype. Thus, more complex statistical methods are required to model gene-gene or gene-environment interactions, multiple causal variants at the same locus (conditional analysis), etc.

1.8 Aims of thesis

This thesis highlights the importance of studying gene-gene interactions in GWAS and suggests methods to make such studies computationally feasible without compromising on statistical power, allowing for imputed genotype data and analysis of SNPs on chromosome X.

Chapter 2 introduces the concept of epistasis. It explains the different kinds of epistatic models one could expect. It then moves on to consider the link between statistical interaction of genes and epistasis. After explaining the different ways of identifying significant statistical interactions, it discusses existing methods, and

their advantages and limitations.

Chapter 3 introduces two tests for identifying statistical interaction between genetic variants. The first one assumes independence of the alleles at a SNP and thus each variant is represented by the number of 'minor' alleles present under an additive model. This is also known as the additive test of interaction. The second test of interaction does not assume independence of the alleles and considers the three genotypes as distinct categories. This is called the general test for interaction. This chapter compares the performance of these two tests by simulation by considering various scenarios underlying interaction models, MAF and effect sizes.

Chapter 4 describes a novel algorithm that was designed to scan for gene-gene interactions on the scale of the whole genome. It then briefly describes software that was developed to implement this novel algorithm. Simulations were performed to optimise the performance of the algorithm, and the method was applied to GWAS data for seven diseases studied by the WTCCC [Wellcome Trust Case Control Consortium, 2007], as well as obesity. The diseases are described and the significant interaction results discussed.

Chapter 5 discusses the role of imputation in GWAS and gene-gene interaction studies. It discusses in some details the methods involved to impute data for a genome wide interaction scan. It then discusses the results obtained and the fine-mapping of regions by using imputed data. It also discusses the presence of multiple signals in the HLA region of Chromosome 6 for Rheumatoid Arthritis and Type 1 diabetes.

Chapter 6 discusses the complexities in studying the sex chromosomes, and develops a novel way of modelling interactions with SNPs on the X-Chromosome

by combining estimates from males and females in a fixed-effects meta-analysis. This algorithm is then applied to the seven WTCCC diseases resulting in an interaction scan of chromosome X with each autosomal chromosome and interactions within chromosome X.

Chapter 7 provides an overall summary of findings from all chapters and links them to the original objectives. It then discusses limitations of methods used and possible solutions. The chapter concludes with possible future research proposals.

Chapter 2

Gene-gene Interactions

2.1 Introduction

Epistasis is the phenomenon under which the action of one gene is suppressed by another gene. Knowledge of biochemistry and physiological genetics suggests that epistasis, in some form or the other, definitely exists [Wright, 1980]. An example of epistasis where the resultant phenotypes are obvious is the coat colour of Labrador Retrievers. It is known that fur colour is a polygenic trait, and is controlled by two different genes, or four different alleles. While one of the genes controls fur colour, the other (epistatic gene) controls the expression of the former. If an offspring receives at least one dominant allele (Black or "B") from the epistatic gene, then the fur colour is influenced by the dominant allele of the former gene. The colour corresponds to the recessive allele (Brown or "b") otherwise. However, if the offspring receives homozygous recessive alleles ("ee") from the epistatic gene, then irrespective of alleles passed on from the first gene, the coat colour is golden (Figure: 2.1).

Figure 2.1: Epistasis causing change in fur colour in the Labrador breed of dog. Here "B" stands for black, "b" stands for chocolate, "E" stands for ability to express coat colour and "e" stands for inability to express dark coat colour.

The precise understanding of epistasis differs from one subject field to another. William Bateson [William and Gregor, 1909] described epistasis as the phenomenon which might be responsible for distortions from Mendelian segregation ratios. Soon after, Sir Ronald Fisher [Fisher, 1918] described epistasis as deviations from linear statistical models. Thus, it started being understood as a phenomenon where one gene masks the effect of another [Phillips, 2008]. In more modern times, with growing efforts to track epistatic effects, scientists often look for statistical interactions and assume that it would imitate a real epistatic effect. However in reality, quite often this is probably not true [Moore, 2005]. Biological epistasis might not necessarily be traced using statistical interactions. The relationship between these two is not completely understood as yet [Cordell, 2002]. There have been some studies pertaining to this, but these have been more widely investigated in smaller and simpler organisms like yeast [Tyler et al., 2009], where conversion of the different epistatic pathways are more understandable.

Biologically, epistasis is plausible because, quite often, this can be imagined to be an alternative pathway to buffer against sudden mutations [Moore and Williams, 2005]. This gives the species a certain amount of genetic stability and, as explained later, the phenomenon also makes the underlying genetic design more compact and versatile. Concepts like canalization and stabilizing selection, introduced in evolutionary and developmental biology, help in explaining how epistasis fits into the wider genetic architecture. Canalization was defined by Waddington [Waddington, 1942] as a concept where genetic buffering leads to some stability in complex developmental procedures in biological systems. In evolutionary biology [Gibson and Wagner, 2000], canalization is a procedure that stabilizes phenotypes against the forces of natural selection. Thus, even when there might be a few mutations over several generations, the response of the complex biological pathways would remain the same, and thus the phenotypes would not show substantial changes. This is enabled by the presence of several dormant biological pathways, which are activated when rare mutations or other adverse environmental effects occur. Theoretical models suggest [Bryant and Meffert, 1993, Bryant et al., 1986, Carson and Templeton, 1984, Goodnight, 1987, 1988, Tachida and Cockerham, 1989, Wade, 1992] that, in the presence of epistasis, population bottlenecks and subdivision expose hidden networks which favour selection and hence allow rapid adaptation to environmental changes. This ensures stability of the phenotypic characteristics of a species and overall success in the natural selection process. These inter-dependencies can be studied more easily in lower or simpler organisms like yeast [Segre et al., 2005].

Moore [2005], Moore and Williams [2005] explain the difference between genetical, biological and statistical epistasis, as reproduced in Figure 2.2. Genetical

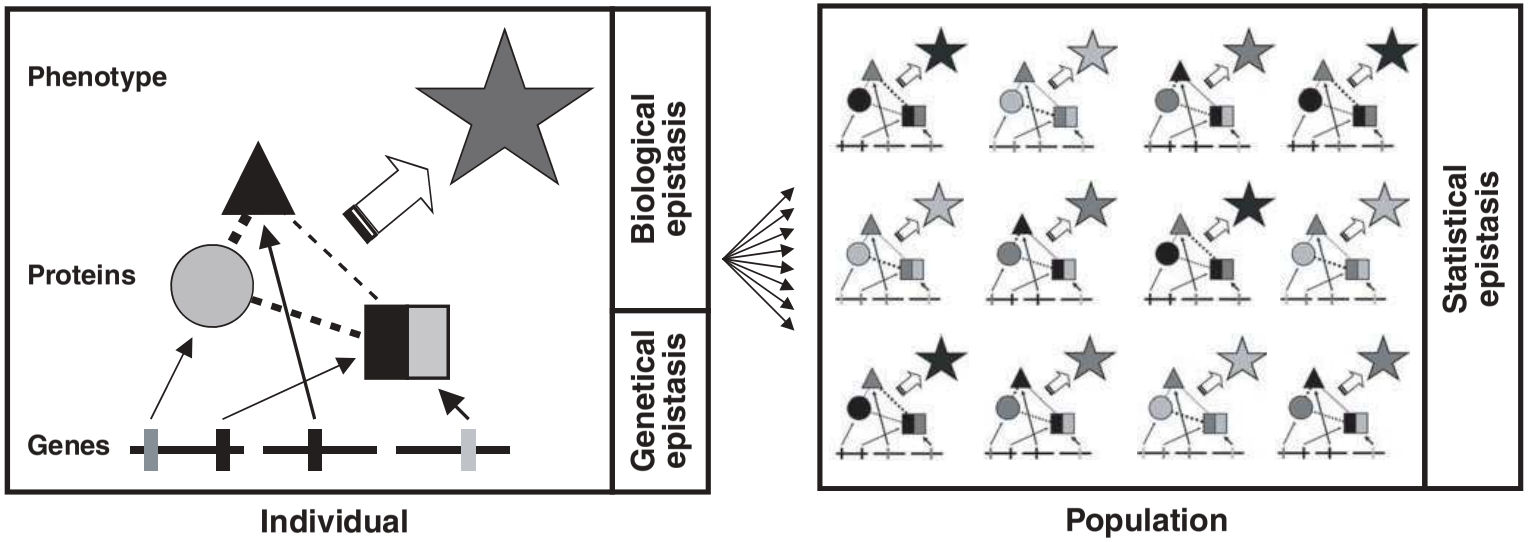


Figure 1 Genetical, biological and statistical epistasis. Genetical epistasis can be thought of as the interaction among DNA sequence variations (vertical bars) that give rise to a particular phenotype in an individual. Genetic information affects phenotype through a hierarchy of proteins (circle, square, triangle) that are involved in biological processes ranging from transcription to physiological homeostasis. The physical interactions (dashed lines) among proteins and other biomolecules and their impact on phenotype (star) constitute biological epistasis. There is a very close relationship between genetical and biological epistasis, with each occurring at the level of the individual. Differences in genetical and biological epistasis among individuals in a population give rise to statistical epistasis. It is entirely possible for genetical and biological epistasis to occur in the absence of statistical epistasis. This can happen when the DNA sequence variations and biomolecules are the same for every individual sampled from a population. Thus, genetic and biological variation is crucial for the statistical detection of epistasis. But does evidence of statistical epistasis necessarily imply genetical or biological epistasis?

Figure 2.2: Reproduced from [Moore, 2005]

and biological epistasis (jointly called physiological epistasis) occur at the level of an individual. Physiological epistasis is a genotypic phenomenon, completely independent of population parameters. Statistical epistasis, on the other hand, is a population phenomenon, being dependent on the allele frequencies at the given SNPs. Further, deviations from statistical epistasis are associated with only the interaction genetic variance component, whereas physiological epistasis also contributes to the additive or dominance (main effects) genetic variance components [Cheverud and Routman, 1995, Crow and Kimura, 1970]. If, however, the variation between these two kinds of epistasis is very small, or absent, in a given population, statistical epistasis cannot be observed. Hence it is essential that all signals of statistical epistasis are replicated and then the potential functional pathways are validated by laboratory based methods.

The merger of Mendelian and Darwinian theories created a new dawn for evolutionary biology. Modern genetics is perhaps experiencing the beginning of a new age, with the merger of biotechnology and population genetics. Producing massive datasets is now substantially cheaper than ever before, and this is only set to get cheaper, resulting in ever bigger datasets and slowly reaching a stage with genome wide sequenced datasets. So far, this has only been possible for much smaller organisms. Oliveri and Davidson's [Oliveri and Davidson, 2004] work on sea urchins, providing a gene network model for embryonic specifications provides some insight into the process of canalization, and hence epistasis. Segre et al. [2005]'s study on yeast is perhaps a start to a new age, in terms of systems-level analysis of epistasis. It was shown that modifications in genes of one functional group affected genes in another. The perturbation of respiratory genes was shown to aggravate glycolysis. These results challenge existing

notions that epistasis usually pervades in genes of similar functional groups. Similar studies in multiple model organisms would provide a much better platform for future system level analysis in humans. It is perceivable, at this point, that more fine scale scans of the human genome, combined with detailed information from transcription networks, other biochemical pathways, and physiological systems, alongside environmental information, could all be described by complex mathematical functions to describe the pathways affecting disease phenotypes.

However, experiments have shown that the effects of epistasis might not be straight forward to identify in humans [Barker, 1979, Cheverud and Routman, 1995, Crow, 1987, 1979, Hedrick et al., 1978, Simmons and Crow, 1977]. Traditional techniques, such as family designs, have proved mostly unsuccessful in trying to discern complex biological pathways involving epistasis [Falconer and Mackay, 1996]. Despite suggestions of complex two-locus theoretical models [Hastings, 1982, 1985, Karlin et al., 1970, 1975, Lewontin and Kojima, 1960], and the promise of an explanation of phenotypic variance, the understanding of the relationship between epistasis and phenotypes remains mostly incomplete.

It has been pointed out that epistasis could, indeed, account for a substantial proportion of unexplained phenotypic variance [Carlborg and Haley, 2004, Williams et al., 2004]. There has been plenty of evidence in recent years to demonstrate the existence of two-locus or multi-locus epistasis in different species [Albrechtsen et al., 2007, Coffey et al., 2004, Zee et al., 2002]. A proper understanding of these mechanisms seems vital in the next phase of GWAS. Not only should this help in explaining missing heritability, but it should take us ever closer to understanding the biological pathways of underlying complex diseases, and hence speed up the process of the translation of these findings into clinical

practice.

With increasing sample sizes, recent studies have uncovered some of the first significant gene-gene interaction signals for human diseases [Hughes et al., 2012, Strange et al., 2010]. In the case of psoriasis, evidence points to an interaction between variants mapping to HLA-C and *ERAP1*. The former plays an important role in MHC class 1 peptide processing and *ERAP1* variants only affected psoriasis susceptibility in the presence of the HLA-C risk allele.

This chapter introduces two popular tests of statistical interaction between a pair of SNPs: the additive-additive and the 'general' test. The former assumes independence of the alleles within a SNP, whereas the latter does not. This leads to differences in statistical power between the two tests under different models of epistasis. Following this, I discuss existing strategies to identify gene-gene interactions. However, most existing strategies either have sub-optimal statistical power, or lack computational efficiency to be implemented on a genome wide level, and in particular for exhaustive scans. We have then discussed the need for interaction scans on the scale of the whole genome and the challenges in achieving this.

2.2 Challenges and strategies for gene-gene interaction testing

One major challenge in scanning for epistatic effects genome-wide is the sheer amount of computational power required for all pair-wise combinations of SNPs. Even if we were not to scan for higher order interactions, we may still have to

perform almost squared amounts more than a regular main effects scan. Thus, for 100 SNPs, we would have to perform 9900 regressions and for a million SNPs we would have to perform close to 10^{12} regressions. Thus, a whole genome scan for only two locus interactions would be a significant computational task. Thus, alternative approaches appear necessary and have been proposed by several papers including Wu et al. [2010], Yang et al. [1999], Zhang et al. [2010a,b]. A systematic comparison of some more realistic approaches has been carried out in Cordell [2009].

A second challenge is correction for multiple testing - in a two locus scan of one million SNPs we perform 10^{12} regressions, for which we need to control the type 1 error rate. One approach is to make use of a Bonferroni correction, which would correspond to a genome-wide significance threshold of $p \leq 5 \times 10^{-14}$. The Bonferroni correction is extremely conservative, since it assumes all tests to be independent which is unlikely to be true here because: (i) genotypes at some SNPs will be correlated because of LD; and (ii) each SNP is involved in multiple interaction tests with other SNPs. Alternatively, permutations can be used to determine the empirical distribution of test statistics [Chapman and Clayton, 2007], although this is computationally infeasible for gene-gene interaction studies.

2.2.1 Generalised Linear models (GLM)

Testing for interaction involving two or more variants can be implemented in a variety of methods. Regression models tend to be a widely used tool for this and have advantages in terms of being able to specify the interaction model in various forms. The nature of the response function is different depending on whether the

phenotype is continuous or binary (as discussed in Chapter 1).

2.2.1.1 Additive-additive interaction test

In the case of an additive-additive interaction test, the following GLM could be used to model two locus interactions, equation (2.1).

$$g(y) = \beta_0 + \beta_{A1}x_{A1} + \beta_{A2}x_{A2} + \beta_{A1A2}x_{A1}x_{A2} \tag{2.1}$$

where, $g(y)$ is the linear predictor for the disease phenotype, y, x_{A1} and x_{A2} are the coded additive genotypes of the two SNPs as defined in 1.5.1.4, and the βs are the corresponding regression coefficients. In the absence of any interaction effects, equation (2.1) would be reduced to a main effects model, equation (2.2).

$$g(y) = \beta_0 + \beta_{A1}x_{A1} + \beta_{A2}x_{A2} \tag{2.2}$$

The log-likelihood of the main effects model (2.2) can be compared against that of the interaction model (2.1), resulting in a one degree of freedom chi-squared test of the additive-additive interaction component.

2.2.1.2 General interaction test

The general interaction test allows for deviations from additivity both within and between variants. Thus for a two-locus test, there would be four interaction terms in the saturated model.

$$\begin{aligned} g(y) &= \beta_0 + \beta_{A_1} x_{A_1} + \beta_{D_1} x_{D_1} + \beta_{A_2} x_{A_2} + \beta_{D_2} x_{D_2} + \beta_{A_1A_2} x_{A_1} x_{A_2} \\ &\quad + \beta_{A_1D_2} x_{A_1} x_{D_2} + \beta_{D_1A_2} x_{D_1} x_{A_2} + \beta_{D_1D_2} x_{D_1} x_{D_2} \end{aligned} \tag{2.3}$$

In this expression, x_{A1} and x_{A2} correspond to the additive main effects of the SNPs as defined in section 2.2.1.1, and x_{D_1} and x_{D_2} correspond to dominance main effects. Here, when the genotype is homozygous $x_{D_i} = -0.5$, and when the genotype is heterozygous $x_{D_i} = 0.5$. The parameters β_{A_i} and β_{D_i} are the coefficients of the additive and dominance main effects respectively, and $\beta_{A_1A_2}$, $\beta_{A_1D_2}$, $\beta_{D_1A_2}$ and $\beta_{D_1D_2}$ are the coefficients corresponding to the different interaction effects ("additive × additive", "dominance × additive", "dominance × additive", "dominance × dominance") respectively. In the absence of any interaction effects, equation(2.3) would be reduced to a main effects model. The saturated model would then be tested against a model with only main effects and of the form (2.4).

$$g(y) = \beta_0 + \beta_{A_1} x_{A_1} + \beta_{D_1} x_{D_1} + \beta_{A_2} x_{A_2} + \beta_{D_2} x_{D_2} \tag{2.4}$$

The log-likelihood of the mean effects model (2.4) can be compared to that of the interaction model (2.3), resulting in a four degree of freedom chi-squared test of interaction, encompassing additive and dominance effects.

2.2.1.3 Case only study

A powerful way to conduct interaction studies is a "case only" approach [Piegorsch et al., 1994, Weinberg and Umbach, 2000, Yang et al., 1999]. This method works

on the principle that the interaction term in a GLM would correspond to the relationship between the phenotype and relevant explanatory variables within the population represented by cases. Thus, a case only test of interaction is conducted by testing a null hypothesis of no association between genotypes or alleles at the two loci (for a two-locus interaction). This test can be conducted by applying a chi-squared test of association to the genotypes (for a four degrees of freedom test, equivalent to the general interaction test) or alleles (for a one degree of freedom test, equivalent to the additive-additive interaction test). A logistic or multinomial regression model could also be used in a GLM setting.

A key assumption of this method is that of no correlation between the genotypes in the population. Thus, if the genotype variables are close to each other (and thus have strong LD), or are correlated for any other reason, a case only interaction test is not appropriate. One could conduct a correlation of genotypes in the general population and then select pairs of SNPs where a case-only or case-control approach is more suitable. However, [Mukherjee et al., 2008] points out that such a strategy is likely to give biased results.

Zhao et al. [2006] suggest a method where the difference between the inter-locus allelic association between cases and controls is used to test for interaction. This was initially suggested by [Hoh and Ott, 2003]. Zhao et al. [2006] showed this method is more powerful than a four degrees of freedom interaction test in a logistic regression. It is likely that the power difference could be explained by the difference in the degrees of freedom of the tests. Mukherjee and Chatterjee [2008], Mukherjee et al. [2008] proposed an **empirical Bayes procedure** which uses a weighted average of the case-control and case-only statistics to test for interaction. This test benefits from the higher statistical power of a case-only

test of interaction. Further, it takes into account information from the controls enabling estimation of marginal effects.

2.2.1.4 Testing for association allowing for interaction

Some researchers have focussed on studying trait associations after accounting for interaction between variants or with environmental factors. This test has been hypothesised as a joint test in Kraft et al. [2007]. However, if no statistically significant interaction effects exist then this strategy will be less powerful than a traditional single variant test of association. Kraft et al. [2007] demonstrated this strategy is optimal for testing single variant associations compared to many realistic theoretical models. However, a case-only analysis can be more powerful for interaction testing [Piegorsch et al., 1994, Weinberg and Umbach, 2000, Yang et al., 1999]. Chapman and Clayton [2007] suggested a "hybrid" strategy which combined a case-control main effects model with a case-only interaction model. This strategy assumes the existence of a factor which would interact with the locus of interest. In the absence of such a factor, an average over all likely interactions is used. Ferreira [2010], Ferreira and Marchini [2011] suggests a bayesian approach to conduct such an analysis implemented in the BIA software (`www.stats.ox.ac.uk/~marchini/software/gwas/bia.html`). An alternative to averaging over all models was suggested by Chapman and Clayton [2007], where the joint test was conducted on a set of potentially interacting loci and the maximum of the computed values was tested using a permutation method. However, given the computational burden, such methods are likely to need large computational resources (and probably infeasible) on the scale of the whole-genome.

2.2.1.5 Two stage two locus methods

Marchini et al. [2005] describe methods to identify epistatic effects on a genome wide scale. They further show the feasibility of such studies for up to a few hundred thousand SNPs at a time, and demonstrate that, when interaction effects truly exist, such studies are more powerful, even after correcting for multiple testing, than traditional approaches that look for main effects only.

However, Marchini et al. [2005] only looked into a few of the many possible epistatic effects, as pointed out by Evans et al. [2006]. All of the models considered had a strong marginal effect, thus ruling out the possibility of looking for interacting loci with very small marginal effects. In such situations, single locus scans would not be very successful and an exhaustive search for all possible interactions would be required. Evans et al. [2006] compare the performance of three different strategies for GWAS, using a wide range of epistatic models from Hallgrmsdttir and Yuster [2008]. The three approaches compared are:

- **Main effects only scan**. Ignores interactions, reducing computational burden. However, in the presence of interactions without marginal effects, this approach turns out to be the least powerful.

- **Exhaustive scan**. An exhaustive scan will consider all two locus interactions and can be computationally expensive for a modern GWAS. Further, given the large number of tests that will have to be performed, multiple testing corrections can reduce statistical power. This strategy is however the most powerful method in the presence of interactions.

- **Two stage scan**. This method scans for the main effects at stage one and then checks for interactions between the significant main effects at stage

two. Thus it misses out on interactions where the main effects are too weak to be detected. This method is not as powerful as the exhaustive scan, but is substantially more efficient computationally. It has a further advantage of conducting far fewer tests, and thus the multiple comparison correction has less impact on the statistical power.

Depending on the underlying epistatic model, the two-stage scan could be almost as powerful as the exhaustive scan. However, in the absence of strong marginal effects, the exhaustive two-locus scan will be most powerful.

2.2.1.6 Higher-order interactions

It is straight forward to extend the GLM to allow for higher-order interaction terms. However, an exhaustive search would involve searching through all possible higher order interactions. This would increase the number of tests exponentially and also increase the computational time. Since even a two-way exhaustive interaction scan poses substantial computational burden, identifying higher order interactions would be impractical in a genome wide setting. Hence, two stage strategies have been suggested by multiple researchers [Hoh et al., 2000, Marchini et al., 2005, Millstein et al., 2009]. Here all loci which pass a single locus significance threshold are passed on to a second stage where higher order interactions are scanned. However, if a locus does not have a significant main effect, it would be dropped from interaction analysis, and would be missed in the second stage.

Besides the two-stage methods, some machine learning strategies are available which reduce the dimensionality of the problem without restricting loci with non-significant marginal effects. Algorithms like ReliefF and random forests will be discussed in section 2.2.2. Existing biological insights could also be used to reduce

the search space for interactions. Bochdanovits et al. [2008] used information from co-adapting mammalian loci to look for interacting SNPs in humans. Emily et al. [2009] used their existing knowledge of biological networks to reduce the number of interactions by a factor of 10^7 in analysing genotype data from WTCCC [Wellcome Trust Case Control Consortium, 2007].

2.2.1.7 Other approaches

Yang et al. [2008] suggested a partitioning method using chi-squared values from the approach of Zhao et al. [2006]. In the absence of marginal effects this method was more powerful than a GLM. Chanda et al. [2007], Dong et al. [2007], Kang et al. [2008], Moore et al. [2006] suggest entropy and information theory based methods are equivalent to GLM based methods given their conditional probability statements, but are not necessarily better performing [Zwick, 2004].

Ljungberg et al. [2004] introduces a global optimisation algorithm DIRECT which is able to test for two or three quantitative trait loci (QTL). Authors have suggested this is more efficient than traditional exhaustive search routines. Struchalin et al. [2010] proposes a variance heterogeneity analysis to identify potentially interacting loci. The algorithm has been implemented in an R package called 'VariABEL' [Struchalin et al., 2012]. An application of this method can be found in Par et al. [2010].

2.2.2 Data mining and related methods

Traditional regression modelling has limitations in exploring non-linear patterns or higher order interactions, owing to sparseness in limited sample sizes [McKinney et al., 2006, Moore and Williams, 2002, Ritchie et al., 2001]. Machine learning

or data mining methods often perform better than traditional linear models in such situations. Rather than scanning through all possible interactions, data mining approaches can step through possible models in a computationally efficient way. In many cases, the two approaches can be equivalent and cross-validation is commonly used in both methods to avoid over-fitting problems.

Data mining methods can struggle with incomplete data or imbalanced case-control designs [Velez et al., 2007]. Further, in the presence of collinearity (correlated predictor variables - for example, LD between SNPs) many data mining methods can overestimate the amount of variance explained in the phenotype. Traditional regression methods have been able to address these issues in many ways, one of them being penalised regression [Copas, 1983, Hastie et al., 2005]. Applications of such methods in genetics can be found in Lee and Silvapulle [1988] and [Le Cessie and Van Houwelingen, 1992], where penalised logistic regression was used. Efron et al. [2004] applied least-angle regression to identify gene-gene interactions for a case-control design.

2.2.2.1 Recursive partitioning methods

Recursive partitioning approaches have been demonstrated to work as an alternative to traditional regression methods in deciphering disease linked marginal or interaction effects [Culverhouse et al., 2004, Nelson et al., 2001, Zhang et al., 2000]. They are a hybrid of classification and regression trees [Breiman, 1993]. These methods map genotypes to the disease phenotypes using graphical tools, for example, an inverse tree like structure. Each node of such a tree would represent a genotype, and further sub-branches would report genotypes at other SNPs which maximise the split between the proportion of cases and controls. Each

branch of a tree is split, based on how well the predictor variable is able to maximise a difference metric (gini-coefficient for example) of the cases and controls [Bastone et al., 2005, Breiman, 1993]. The path created by each branch represents a set of genotypes which best explains the differences between cases and controls at the "mother" node. The recursive splitting is stopped when no further gains in the difference metric is observed, when all individuals in a node belong to the same disease status (all cases or all controls), or when a threshold of the difference metric is achieved. In the interest of having a parsimonious model, it is sensible to prune the tree at this stage. This can be achieved by penalising every extra branch split, or thresholding the number of splits desired.

With the introduction of an "ensemble" of trees, substantial improvements can be noticed. Random forest methods have been designed for this purpose [Breiman, 2001], and have been applied with success in some genetic studies [Bureau et al., 2005, Lunetta et al., 2004]. The main advantage of random forests is the ability to generate a hierarchical structure of SNPs, categorised by the amount of impact they have. This helps understand multi-way relationships better than more traditional non-data mining methods.

Random forests can be implemented as a fast algorithm with built-in parallel computational capabilities. In weighing the relative importance of SNPs, random sets of predictors are allocated to each branch, and many such branches (trees) can be computed in parallel. The final outcome of the analysis will almost certainly depend on the size of allocations, extent of parallel processing and the nature of the statistical test used to classify SNPs. To examine the sensitivity of the above choices, the analysis should be run several times and the differences of each run should be evaluated carefully.

These methods do not explicitly account for interaction, but possible interactions are included within SNPs included in each branch. Thus, in essence, they search for associations after accounting for interactions. However, given that the initial partition is performed on main effects, and subsequent partitions are conditional on previous main effects, interactions without large marginal effects are likely to be missed [McKinney et al., 2009].

2.2.2.2 Multifactor Dimensionality Reduction method

Inspired by a previously suggested method in Nelson et al. [2001], Ritchie et al. [2001] suggested a multi-factor dimensionality reduction (MDR) procedure, which identifies combinations of multiple loci and environmental factors that together affect a given disease phenotype. They suggest a method to identify a subset of SNPs, combinations of which would influence the disease phenotype. The MDR method does not assume the presence of any underlying standard theoretical distributions or inheritance models, and could be used for both case-control or family design studies. The procedure was able to show reasonable power in picking up two locus interactions in a relatively small simulated dataset. Application to a real life sporadic breast cancer dataset revealed a four locus interaction effect, although this has not been replicated.

However, Ritchie et al. [2001] use a binary classification of genotypes based on the ratio of phenotype values between the cases and controls. This immediately takes away a lot of information offered by the variance of genotypes. If the "spread" of cases and controls in a subset is similar to that in the entire sample, or perhaps if the number of cases or controls is quite small for the subset, the possibility of having multiple false positive or negative results cannot be ruled out.

To try to resolve this issue, Chung et al. [2007] propose a new "odds ratio" based approach to the MDR, referred to as OR-MDR. This makes sure that, rather than a binary classification of risk, there is a sequence of multi-locus combination groups, each with an attached quantitative risk factor, from the highest to the lowest risk groups. This method also provides a confidence interval, thereby assuring the level of statistical uncertainty in the respective risk groups. However, for relatively small samples, or perhaps for some loci with low allele frequencies, there could be some empty cells, which cannot be allocated to a risk group.

Lee et al. [2007] suggest a log linear model based modification(LM-MDR) to the above algorithm to estimate frequencies of the missing cells. A saturated form of this model is a special case and is equivalent to the MDR and OR-MDR methods. The unsaturated model always has less mean square error than the saturated model. Even if there were some extra bias, a smaller variance value makes it preferable compared to the saturated model or the MDR method. When there is high-order epistasis, both of these methods have similar levels of power. However, the LM-MDR works better in all other cases. This algorithm may still have problems for estimating cell frequencies when multiple loci are considered together. It has also not been tested in really large datasets, and it remains to be seen whether it is computationally feasible to run with millions of SNPs.

2.2.2.3 ReliefF, Tuned ReliefF and evaporative cooling

Tuned ReliefF [Moore and White, 2007], and its predecessor ReliefF [Moore et al., 2006], are filtering algorithms which compute distance metrics between individuals, using genome wide genetic similarity, within and across the phenotype classes. For a given predictor variable (here a SNP), the difference in value (number of

alleles) between pairs of individuals are weighted negatively and positively, for within and across phenotype groups, respectively. These measures are used to construct an importance measure for the variable. This method can be scaled for large numbers of variables and individuals. However, ReliefF is likely to be less powerful in presence of large numbers of genetic variants which do not explain the phenotype [McKinney et al., 2009]. This has prompted the development of "evaporative cooling" [McKinney et al., 2007, 2009], which combines the strengths of ReliefF and random forest methods.

2.2.3 Bayesian model selection approaches

A range of bayesian techniques have been developed to identify associated loci, and interactions between them, which together best predict complex phenotypes [Gelman et al., 2003]. Bayesian methods differ from frequentist (non-bayesian) methods by hypothesising a prior distribution for the parameters to be estimated. In a frequentist setting, the absence of a prior equates to hypothesising a "uniform" prior, which can be a compromise on statistical power if some information is already available on the distribution of the parameter in question. The product of the prior distribution and the likelihood of the data enables the deduction of the posterior distribution. If the posterior distribution does not belong to the same family as the prior distribution (not conjugate prior) then Markov Chain Monte Carlo (MCMC) based approach can be implemented to estimate the parameters of the posterior distribution, but can be computationally expensive.

A Bayesian version of stepwise regression was proposed by Lunn et al. [2006] and implemented in the software WinBUGS [Lunn et al., 2000]. The method

was designed for main effects identification, but can be extended to look for interactions. However, owing to computational limitations, this process is only able to look at a few hundred loci in each run. This makes it less scalable for large scale interaction scans.

2.2.3.1 Bayesian epistasis association mapping

Bayesian epistasis association mapping (BEAM) is an MCMC approach designed to detect interacting and non-interacting loci associated with a phenotype [Zhang and Liu, 2007]. This method classifies all SNPs into three groups: those with no association with the phenotype, those with main effect associations, and those which are part of a model with interacting alleles. Given the prior distribution of group membership of markers, and the prior for regression parameters given group membership, a posterior distribution is generated for the parameter which are estimated using an MCMC algorithm.

BEAM is unable to handle large sample sizes (5000 and above), and the large numbers of SNPs (500k-1000k) that are commonly used in modern GWAS. Further, LD between multiple variants is not accounted for, suggesting a subset of all SNPs found in a modern genotyping array should be used such that they are independent of each other.

2.3 Chapter Summary

This chapter introduced the concept of epistasis and elaborated on its relationship with statistical gene-gene interaction. It then established the need for studying statistical interactions and the challenges in doing so. A wide range of meth-

ods are available to conduct gene-gene interactions, many of which have been discussed. These methods very often involve a trade off between computational burden and statistical power. There are several limitations in scanning for interaction signals in modern GWAS datasets, some of which have been discussed.

The following chapter discusses "Coverage and Power of gene-gene interaction tests". It compares the performance of a general genotypic interaction test against an additive-additive interaction test in a GLM framework using simulated data. It further discusses the impact of considering LD in these simulations.

Chapter 3

Coverage and Power of gene-gene interaction tests

3.1 Introduction

GWAS have been able to successfully identify hundreds of loci which are associated with diseases including Crohn's disease and type 2 diabetes. However, when the main effects of SNPs at these loci are used to explain the phenotypic variance, it only explains a small fraction of heritability. Regression models, accounting for interaction effects, might identify additional loci and increase the amount of explained phenotypic variance. However, at present there are formidable statistical and computational challenges with performing such studies. A systematic approach to deal with these issues and account for interaction effects seems essential. Analysing the power and relative efficiencies of different tests under varying models is a required first step, and is the challenge addressed in this chapter.

The power of a test is defined as the probability of detecting an effect, given

that it actually exists, at a pre-determined significance threshold. Formally, it is defined as the probability of rejecting the null hypothesis, when the alternative hypothesis is true.

Depending on the nature of the test used and the underlying model relating genotypes to phenotype, the power of an interaction test can be affected by various factors. In general, we would like to optimize the power by controlling as many of these factors as possible. For a given significance threshold, power depends on sample size, the size of the genetic effect in question, allele frequencies, and properties of the statistical test, amongst other factors. In present day GWAS, given the huge sample sizes and carefully developed quality control methods, we have been able to achieve very good levels of power to detect main effects for common variants. However, even with their large sample sizes, it is not clear that the same will be true for genome-wide interaction scans.

In this chapter, I carry out detailed simulations to study the impact of genetic effect size, MAF, and underlying interaction models, on the power to detect interactions between SNPs with statistical tests in a GLM framework [Bhattacharya and Morris, 2009]. In the process of a genome wide scan for the main effects, the most powerful approach is to perform additive tests, unless there is extreme dominance [Wellcome Trust Case Control Consortium, 2007]. However for two-locus interaction models, it still remains to be seen what is the most efficient approach, and whether the decision would change under varying conditions. This should eventually help in understanding the strategy which would aid power optimization to detect interactions when performing a genome wide scan.

3.2 Models of interaction

The nature in which SNPs at loci marginally affect the phenotype in question can be assumed to be of two kinds. If the effect of the heterozygous genotype is exactly between the two homozygotes, the effect is "additive". However, any deviation from this state of 'additivity' is referred to as "dominance".

I have considered a number of theoretical models of two-locus interaction: "simple additive", "simple dominant", "additive-additive", "dominance-dominance", "pure additive", "pure dominance" and "recessive-recessive". The "simple additive" and "simple dominant" models exhibit additive and dominance main effects across the two loci but no interaction. The "additive-additive" model exhibits interaction such that presence of each additional "causal allele" causes a linear increase in the phenotype. The magnification of the phenotypic effect is proportional to the total number of "causal alleles" present at both loci. The "dominance-dominance" model exhibits interaction such that presence of one or two "causal alleles" magnify the impact on the phenotype. There is no linear relationship between the number of "causal alleles" and the phenotype. The "pure-additive" and "pure-dominance" models have no marginal effects and exhibit a magnification of phenotypic effect on the additive and dominance scales respectively. The "recessive-recessive" model exhibits a dominance effect at both loci. When one of the loci contains a recessive allele, the phenotype of the other locus is hidden.

3.2.1 Data Simulation

Datasets were simulated for various combinations of MAF, proportion genetic variance (λ) explained by the pair of SNPs, and the underlying interaction models. Interaction models in Table 3.1 were adopted from Marchini et al. [2005], Evans et al. [2006] and Hallgrmsdttir and Yuster [2008]. I further assumed that the two 'causal' SNPs are not in LD with each other.

Table 3.1: Theoretical models of interaction between two loci. For a given model and combination of genotypes, the mean of the corresponding quantitative phenotypic trait are represented in the matrices (as explained in Table 3.2).

Interaction Model		Locus B		
	Locus A	BB	Bb	bb
Simple additive no epistasis (additive main effects)	AA	+2	+1	+0
	Aa	+1	+0	-1
	aa	+0	-1	-2
Simple dominant no epistasis (dominance main effects)	AA	+2	+2	+0
	Aa	+2	+2	+0
	aa	+0	+0	-2
Additive-additive epistasis	AA	+3	+1	-1
	Aa	+1	+0	-1
	aa	-1	-1	-1
Dominant-dominant epistasis	AA	+3	+3	-1
	Aa	+3	+3	-1
	aa	-1	-1	-1
Pure additive epistasis	AA	+1	+0	-1
	Aa	+0	+0	+0
	aa	-1	+0	+1
Pure dominance epistasis	AA	+1	-2	+1
	Aa	-2	+4	-2
	aa	+1	-2	+1
Recessive-recessive epistasis	AA	+3	-1	-1
	Aa	-1	-1	-1
	aa	-1	-1	-1

For given allele frequencies, genotypes at two SNPs were simulated under

HWE for 1000 individuals. In order to simulate a continuous trait, the population trait mean for each combination of genotypes was specified (Table 3.1). Underlying interaction models of different kinds (Table 3.1) can each be represented by Table 3.2.

Table 3.2: Mean phenotypic responses for combination of genotypes

Locus 1	Locus 2		
	BB	Bb	bb
AA	μ_{AABB}	μ_{AABb}	μ_{AAbb}
Aa	μ_{AaBB}	μ_{AaBb}	μ_{Aabb}
aa	μ_{aaBB}	μ_{aaBb}	μ_{aabb}

The following relationships were used to calculate the overall mean phenotype,

$$\mu = \sum_{i=\{AA,Aa,aa\}} \sum_{j=\{BB,Bb,bb\}} f_{A_i} f_{B_j} \mu_{ij} \tag{3.1}$$

where, f_{A_i} is the population frequency of genotype i at locus A and, f_{B_j} is the population frequency of genotype j at locus B. For a given λ and genetic variance (σ^2_G), the environmental variance (σ^2_E) could be calculated using the following relationships. Genetic variance is given by,

$$\sigma^2_G = \sum_i \sum_j (f_{A_i} f_{B_j} \times (\mu_{ij} - \mu)^2) \tag{3.2}$$

$$Total\ Variance = \sigma^2_G + \sigma^2_E \tag{3.3}$$

If λ is the proportion of the total variance that is genetic, it follows that,

$$\lambda = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_E} \tag{3.4}$$

$$\Rightarrow \sigma_E^2 = \sigma_G^2 \frac{(1-\lambda)}{\lambda} \tag{3.5}$$

Now, the continuous trait was simulated for each individual, given their simulated genotypes at the pair of SNPs. For a mean, μ_{jk} (Table 3.2) and variance, σ_E^2, the phenotypic trait is simulated from a normal distribution,

$$y_i \sim \mathcal{N}(\mu_{jk}, \sigma_E^2) \tag{3.6}$$

This process was repeated $10,000$ times for each combination of interaction model, MAF (0.05 - 0.50), and λ (0.01 - 0.05).

3.2.2 Incorporating LD

To incorporate LD, a marker SNP is generated based on the genotypes at an assumed causal variant, for a given r^2 between them. To enable this, the following probabilities needed to be calculated,

$$P(M_A M_A|AA) = P(M_A|A)^2 \tag{3.7}$$

$$P(M_A m_a|AA) = 2 \times P(M_A|A) \times P(m_a|A) \tag{3.8}$$

$$P(m_a m_a|AA) = P(m_a|A)^2 \tag{3.9}$$

where, for example, $P(M_A|A)$ is the probability of having allele M_A at the marker SNP, given the allele at the causal variant is A.

$$P(M_A|A) = \frac{P(M_A \cap A)}{P(A)}, \tag{3.10}$$

$$P(M_A|a) = \frac{P(M_A) - P(M_A|A)P(A)}{1 - P(A)} \tag{3.11}$$

$$P(m_a|A) = 1 - P(M_A|A) \tag{3.12}$$

$$P(m_a|a) = 1 - P(M_A|a) \tag{3.13}$$

$P(M_A|A)$ is related to r^2 by the following,

$$r^2 = \frac{\{P(M_A A) - P(M_A)P(A)\}^2}{P(M_A)P(A)(1 - P(M_A))(1 - P(A))} \tag{3.14}$$

If we assume that the MAF at the causal variant and SNP are the same, and we assume that alleles M_A and A are positively correlated with each other, it follows that:

$$r = \frac{\{P(M_A A) - P(M_A)P(A)\}}{\sqrt{\{P(M_A)P(A)(1 - P(M_A))(1 - P(A))\}}} \tag{3.15}$$

Using $P(M_A) = P(A)$,

$$r = \frac{\{P(M_A|A) - P(A)\}}{(1 - P(A))} \tag{3.16}$$

and hence that,

$$P(M_A|A) = P(A) + r \times (1 - P(A)) \tag{3.17}$$

To incorporate LD, we can thus simulate genotypes at the marker SNP, conditional on the simulated genotypes at the causal variant. For an r^2 value of 0,

the marker SNP and causal variant are independent, and we would expect power equal to the type-I error rate. Conversely, for an r^2 value of 1, the causal variant and marker SNP are completely correlated, and we would expect power equal to that achieved for the causal variant itself.

3.2.3 Tests for interaction

An additive and a general interaction test were performed on the causal variants and marker SNPs of each of the simulated datasets (tests are described in section 2.2.1.1 and section 2.2.1.2). The p-values corresponding to the interaction regression coefficients were recorded and considered significant for values less than the traditional threshold of 0.05 (which is used here purely to compare tests).

Power was calculated using sets of $10,000$ p-values for the additive and general interaction tests.

$$power = \frac{Number\ of\ p-values < \alpha}{Number\ of\ replicates} \tag{3.18}$$

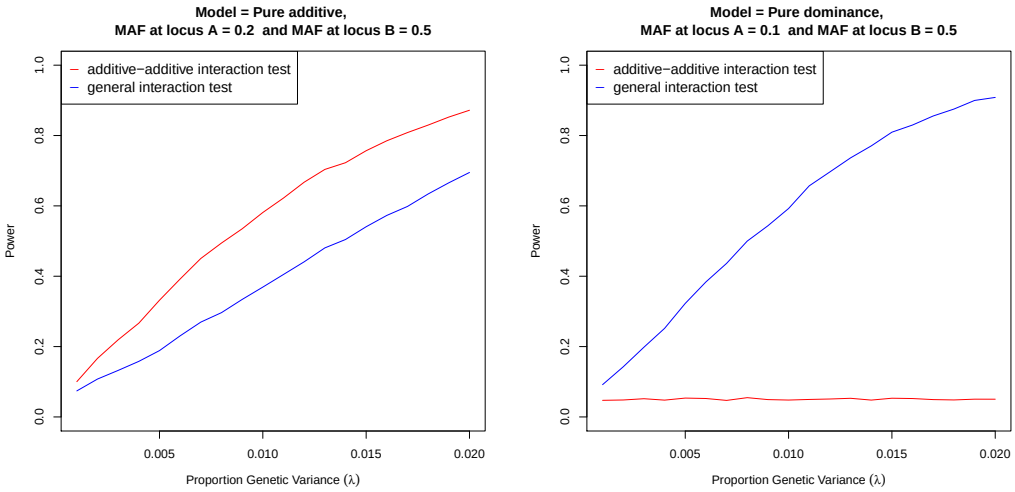
where, α is the significance threshold ($= 0.05$ here).

3.3 Results

Figure 3.1(a) plots power of an additive-additive and general interaction test of the causal variants against λ, when the MAF at one locus is 0.2, MAF at other locus is 0.5 and the underlying epistasis model is "pure additive"(Table 3.1). In this case, the additive-additive interaction test is uniformly more powerful than a general interaction test, which is not surprising given that this model of

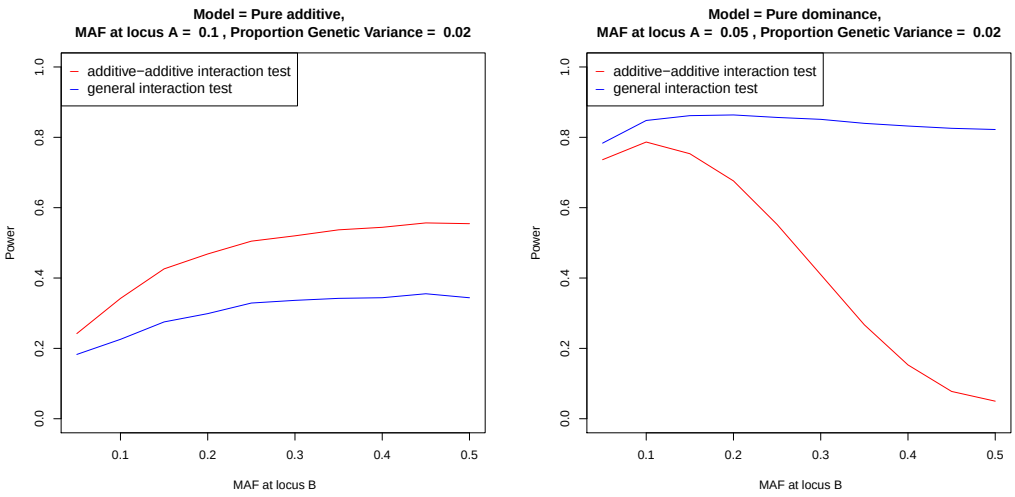
epistasis does not incorporate dominance. Figure 3.1(b) plots the power of an additive-additive and general interaction test against λ, when the MAF at one locus is 0.1, MAF at the other locus is 0.5 and the underlying model is "pure dominance" (Table 3.1). In this case, the general interaction test is uniformly more powerful than an additive-additive interaction test, suggesting additional degrees of freedom are required here to allow for dominance. Figure 3.1(c) plots the power of an additive-additive and general interaction test against the MAF of one locus, when the MAF at the other locus is fixed at 0.1, λ is fixed at 2%, and the underlying model is "pure additive" (Table 3.1). Here, the additive-additive test seems more powerful across the range of λ. Figure 3.1(d) plots the power of an additive-additive and general interaction test against the MAF of one locus, when MAF at the other is fixed at 0.05, λ is fixed at 2% and the underlying model is "pure dominance" (Table 3.1). In this case, the general interaction test is almost always more powerful than an additive-additive interaction test unless the MAF at both loci are very close to 0. In this scenario, the rare homozygous genotypes are sparse, and dominance effects are difficult to estimate.

In general, there was an appreciation of power with an increase of MAF or λ. As expected, the additive-additive interaction test was more powerful when the interaction in the underlying model did not include a substantial deviation from additive-additive effects. The power of the general interaction test far exceeded that of the additive-additive interaction test when the underlying model incorporated significant deviations from additivity. However, the "pure-interaction" models considered in Figure 3.1 do not have marginal effects, and it is not clear how likely they are to occur in reality. Figure 3.2 considers more "realistic" interaction models that incorporate some marginal effects.



(a) Power against λ, for fixed MAF

(b) Power against λ, for fixed MAF

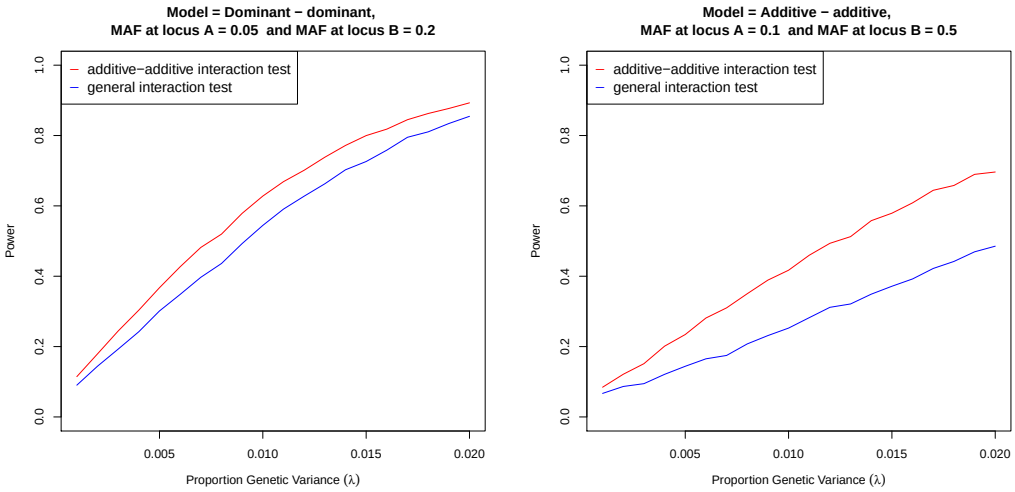


(c) Power against MAF, for fixed λ

(d) Power against MAF, for fixed λ

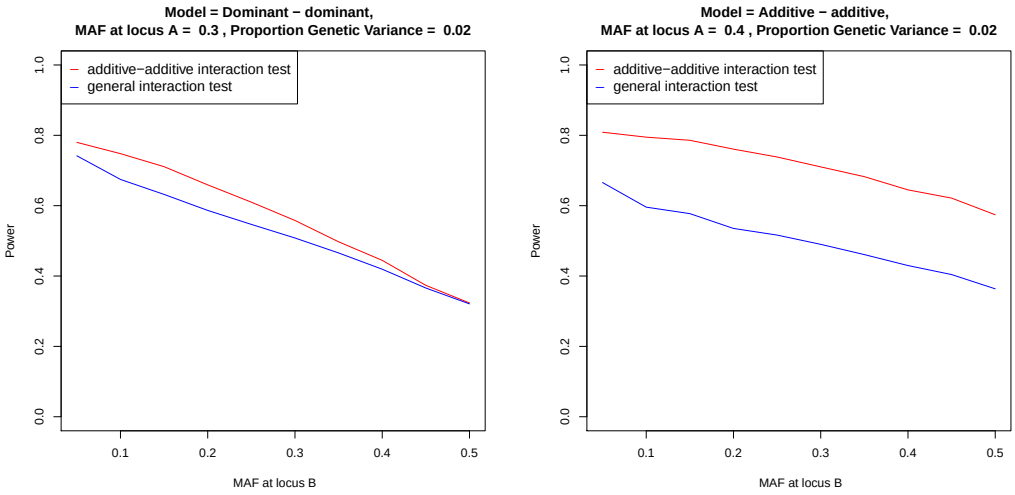
Figure 3.1: Power of additive-additive and general interaction tests against MAF and λ for "pure additive" and "pure dominance" theoretical interaction models.

Figure 3.2(a) plots power of an additive-additive and general interaction test against λ, when the MAF at locus is 0.05, MAF at other locus is 0.2 and the underlying model is "dominant-dominant" (Table 3.1). Here the additive-additive



(a) Power against λ, for fixed MAF

(b) Power against λ, for fixed MAF



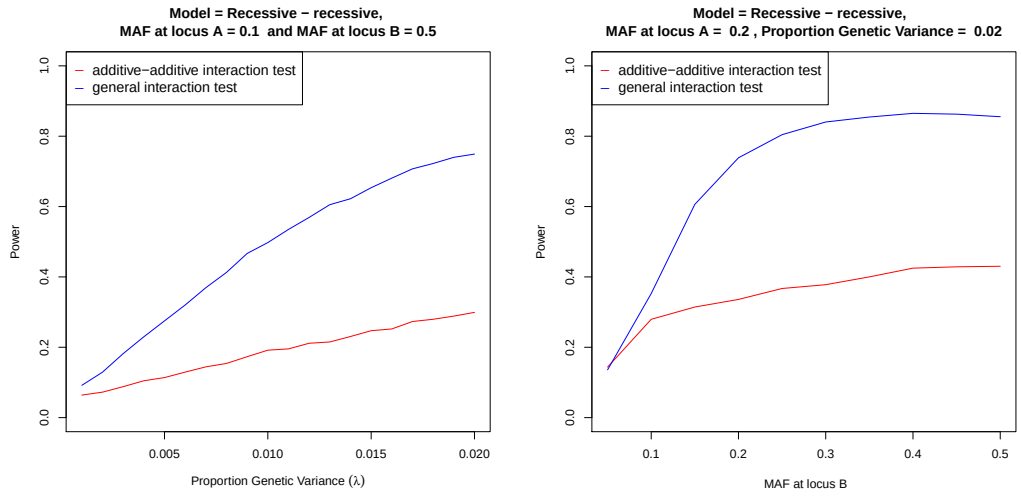
(c) Power against MAF, for fixed λ

(d) Power against MAF, for fixed λ

Figure 3.2: Power of additive-additive and general interaction tests against MAF and λ for "dominant-dominant" and "additive-additive" theoretical interaction models.

interaction test is marginally more powerful than the general interaction test. Figure 3.2(b) plots power of an additive-additive and general interaction test against λ, when the MAF at one locus is 0.1, MAF at other locus is 0.5 and the

underlying model is "additive-additive" (Table 3.1). Here the additive-additive interaction test is consistently more powerful than the general interaction test. Figure 3.2(c) plots the power of an additive-additive and general interaction test against the MAF at one locus, when MAF at other locus is fixed at 0.3, the λ is fixed at 2% and the underlying model is "dominant-dominant" (Table 3.1). In this case, the additive-additive interaction test is almost always more powerful than a general interaction test, unless the MAF were close to 0.5. In such a case, the general interaction test is slightly more powerful than an additive-additive interaction test. Figure 3.2(d) plots the power of an additive-additive and general interaction test against the MAF at one locus, when MAF at other locus is fixed at 0.4, the λ is fixed at 2% and the underlying model is "additive-additive" (Table 3.1). In this case, the additive-additive interaction test is consistently more powerful than a general interaction test.



(a) Power against λ, for fixed MAF

(b) Power against MAF, for fixed λ

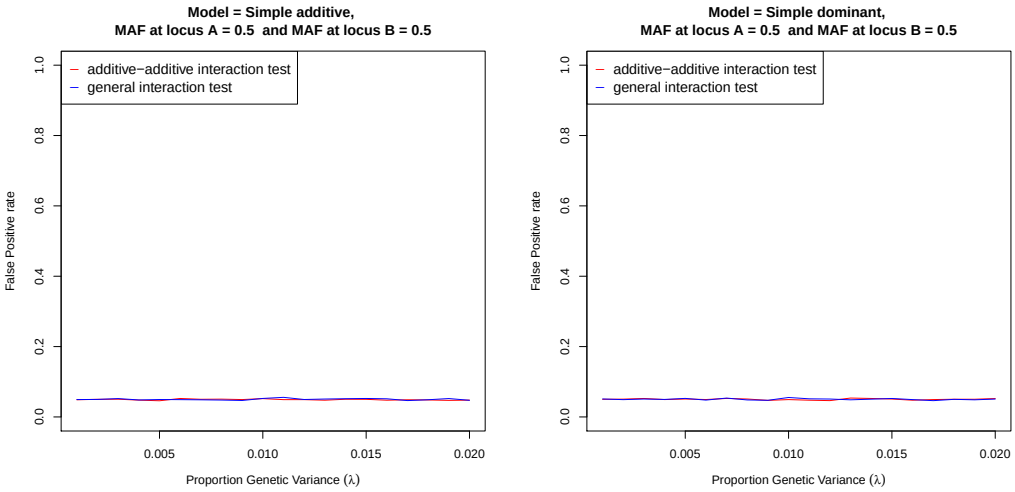
Figure 3.3: Power of additive-additive and general interaction tests against MAF and λ for a theoretical "recessive-recessive" interaction model.

Figure 3.3(a) plots power of an additive-additive and general interaction test against λ, when the MAF at one locus is 0.1, MAF at the other locus is 0.5 and the underlying model is "recessive-recessive" (Table 3.1). Here the general interaction test is substantially more powerful than the additive-additive interaction test. Figure 3.3(b) plots the power of an additive-additive and general interaction test against the MAF at one locus, when MAF at the other locus is fixed at 0.2, the λ is fixed at 2% and the underlying model is "recessive-recessive" (Table 3.1). In this case, the additive-additive interaction test is almost always more powerful than a general interaction test, unless the MAF were lesser than 10%.

Figures 3.4(a) and 3.4(b) plot the false positive rates of an additive-additive and general interaction tests against λ when the MAF at both loci is 0.3 and the underlying models are "simple additive" or "simple dominant" respectively (Table 3.1). The additive-additive interaction test and general interaction test have the same power and are equal to around 5%. Figures 3.4(c) and 3.4(d) plot the false positive rare of an additive-additive and general interaction test against the MAF at both loci, when the λ is fixed at 2% and the underlying models are "simple additive" and "simple dominant" respectively (Table 3.1). The additive-additive interaction test and general interaction test have false positive rates similar to the chosen α of the statistical tests (5%).

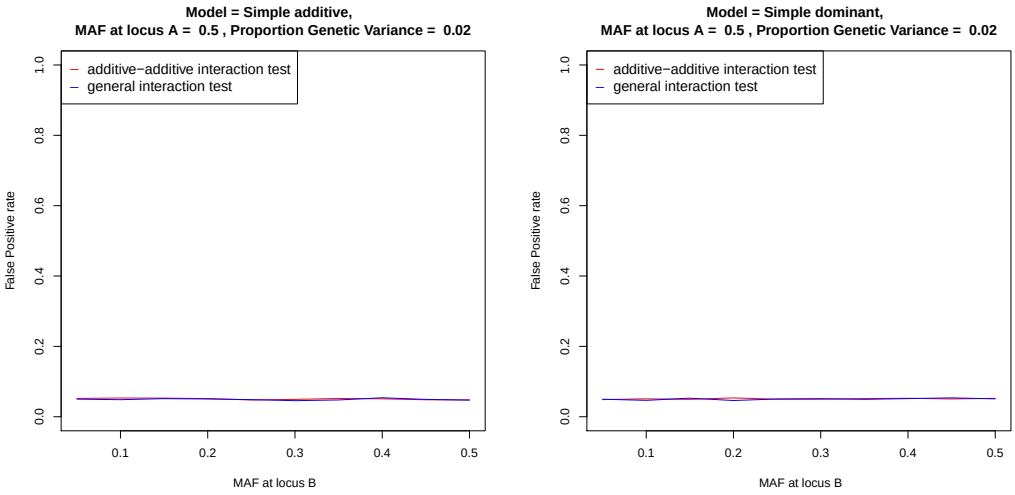
Next, I considered the impact of LD between causal variants and tested marker SNPs. I focussed on cases where either one of the two interaction tests had reached a statistical power greater than 90% and in particular cases where any of the tests was significantly more powerful than the other. I have considered specific combinations of MAF and λ for five theoretical models of epistasis.

Figure 3.5(a) plots the power of an additive-additive interaction test and a



(a) FPR against λ, for fixed MAF

(b) FPR against λ, for fixed MAF

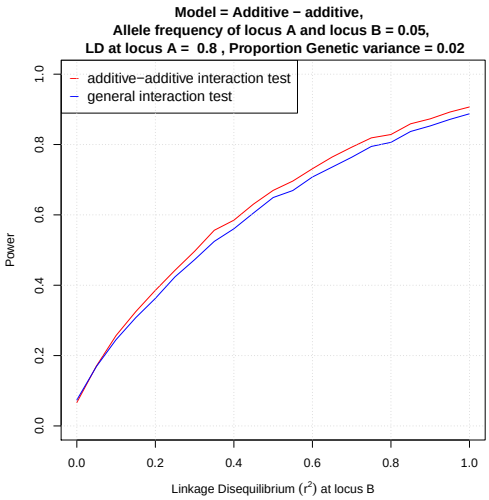


(c) FPR against MAF, for fixed λ

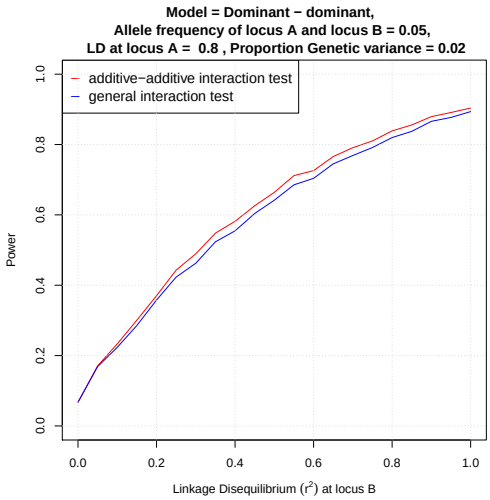
(d) FPR against MAF, for fixed λ

Figure 3.4: False positive rate (FPR) of additive-additive and general interaction tests against MAF and λ for theoretical "simple additive" and "simple dominant" interaction models. These models have no epistasis and only have additive main effects and dominance main effects respectively.

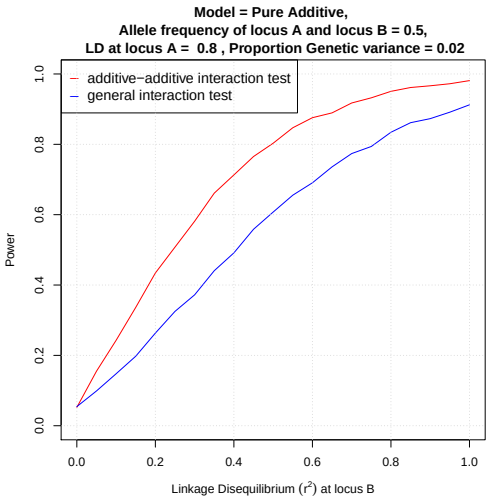
general interaction test against LD values (r^2) of one of the loci, while fixing the r^2 at the other loci (at 0.8). The MAF considered is 0.05 at both loci, $\lambda = 2\%$



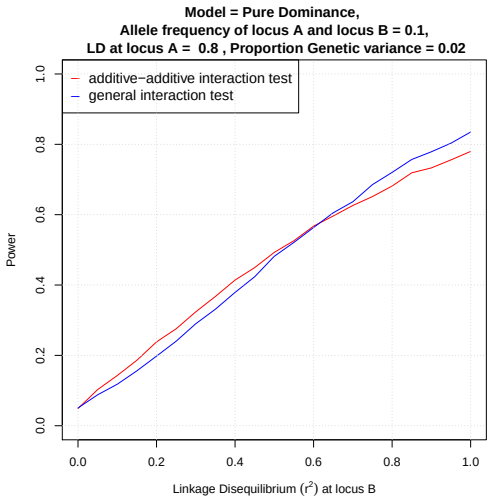
(a) Additive-additive model



(b) Dominant-dominant model



(c) Pure Additive model



(d) Pure Dominance model

Figure 3.5: **Impact of LD**: Power against LD (r^2), for specific cases of MAF and λ where either of the tests have achieved a reasonably large statistical power. We have particularly focussed on cases where the general interaction test appeared to be significantly more powerful than the additive-additive interaction test.

and an underlying "additive-additive" model. It shows that, unless the r^2 is very low (≤ 0.1) the additive-additive interaction test performs better than the general

interaction test.

Figure 3.5(b) plots the power of an additive-additive interaction test and a general interaction test against LD values (r^2) of one of the loci, while fixing the r^2 at the other loci (at 0.8). The MAF considered is 0.05 at both loci, $\lambda = 2\%$ and an underlying "dominant-dominant" model. Similar to the "additive-additive" model, unless the r^2 is very low (≤ 0.1) the additive-additive interaction test is more powerful than the general interaction test.

Figure 3.5(c) plots the power of an additive-additive interaction test and a general interaction test against LD values (r^2) of one of the loci, while fixing the r^2 at the other loci (at 0.8). The MAF considered is 0.5 at both loci, $\lambda = 2\%$ and an underlying "pure additive" model. The additive-additive interaction model is uniformly more powerful than the general interaction test here.

Figure 3.5(d) plots the power of an additive-additive interaction test and a general interaction test against LD values (r^2) of one of the loci, while fixing the r^2 at the other loci (at 0.8). The MAF considered is 0.1 at both loci, $\lambda = 2\%$ and an underlying "pure dominance" model. It shows that the general interaction test is more powerful than the additive-additive interaction test only when the r^2 values are more than around 0.7 (in this case), otherwise the additive-additive interaction test is more powerful. Thus, if the causal SNPs are in strong LD with the causal variants, and the underlying model deviated from additive-additive interaction, the general interaction test would be more powerful.

Figure 3.6 plots the power of an additive-additive interaction test and a general interaction test against LD values (r^2) of one of the loci, while fixing the r^2 at the other loci (at 0.8). The MAF considered is 0.2 at both loci, $\lambda = 2\%$ and an underlying "recessive-recessive" model. Here, the general interaction test is

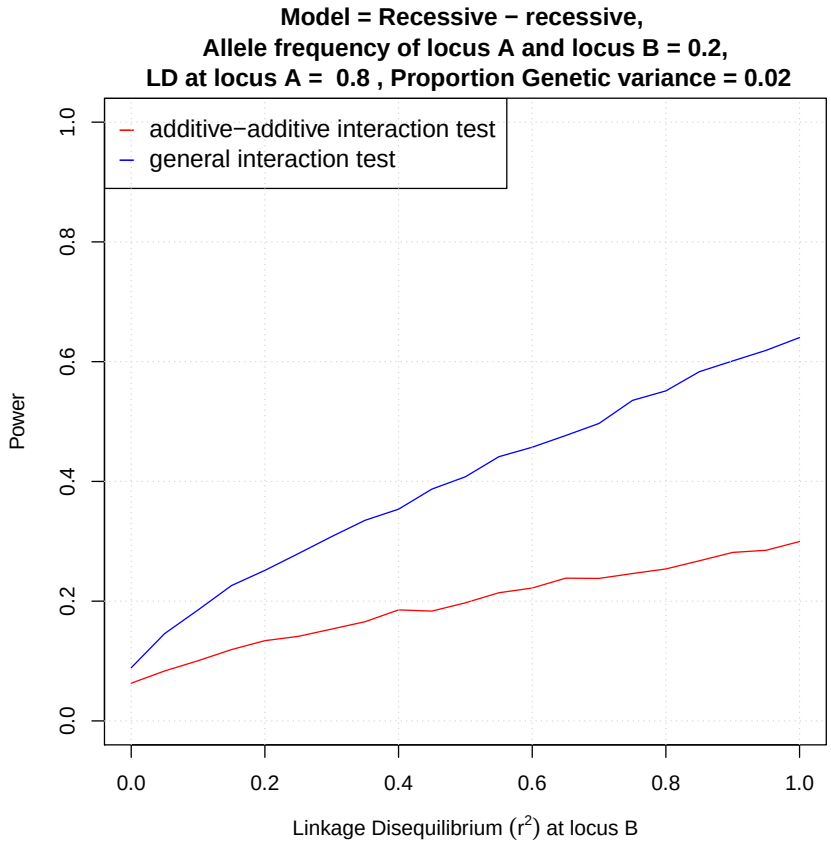


Figure 3.6: **Impact of LD in a "recessive-recessive" model**: Power against LD (r^2), for MAF of 0.02 at both loci and λ of 2%. Here, the general interaction is significantly more powerful.

uniformly more powerful than the additive-additive interaction test for a wide range of r^2 values.

I inspected several thousand different plots in total, taking into account all combinations of the 7 models, 10 different MAF at the two loci and 10 different values of λ. On top of this, we also considered 10 different values of LD(measured by $r2$) for 5 different theoretical models. Irrespective of the underlying λ or MAF, as the r^2 between the causal variant and marker SNP gets lower, the power to detect deviations from additivity is reduced. In general, for more "realistic"

interaction models i.e. those incorporating some degree of additivity within and across loci, the additive-additive interaction test was more powerful than the general interaction test. For situations where the general interaction test was more powerful, the differences in power between the tests were minimal and the difference reduced as LD decreases. We may thus conclude that, even when there is substantial deviation from additive-additive interaction, as LD decreases between causal variants and markers SNPs, the general interaction test becomes less powerful than the additive-additive interaction test.

3.4 Discussion

An additive-additive interaction model is a trade off between fitting a smaller number of parameters against fitting the more detailed model in the general interaction test. From the simulations, we notice that, in the case of a "additive-additive", "pure additive" or "dominant-dominant" interaction model, the additive-additive interaction test was more powerful than the general interaction test for almost all realistic conditions (Figure 3.1(a), 3.1(c), 3.2(a), 3.2(b), 3.2(c) & 3.2(d)). In other words, as long as there were some marginal effects, and the deviation from additive-additive interaction was not substantial, the additive-additive interaction test was more powerful than the general interaction test. The additive-additive model also appears to be more effective when the MAF is quite low, because of sparsity of the minor alleles and hence difficulties in estimating dominance effects.

However, in the case of "pure dominance" and "recessive-recessive" models, a general interaction test is substantially more powerful than an additive-additive interaction test (Figure 3.1(b), 3.1(d), 3.3(a) & 3.3(b)). The difference was more

pronounced for lower values of λ and higher r^2 values. The difference in power between the dominance and additive-additive interaction tests were more pronounced for certain combinations of MAF values. The strength of the general interaction test seems to disappear if LD is incorporated into the simulations, at least for the "pure dominance" model . With a decrease in LD between the causal variants and the genotyped marker, the additive-additive interaction test becomes more powerful than the general interaction test because dominance effects appear more additive, a phenomenon which has been well established for marginal effects [Spencer et al., 2009]. The "recessive-recessive" model is the only model where the additive-additive interaction test is less powerful than the general interaction test even with the introduction of LD. However, the former test is more powerful across a wider range of theoretical models.

The additive-additive interaction test picks up very little or no signal in extreme interaction models. However, it is not clear how likely these models are to exist in reality. Furthermore, as LD between the causal variant and marker SNP weakens, the interaction effect appears to look more additive-additive in nature.

These simulations were performed assuming a quantitative disease trait, however for large sample sizes we should find similar principles hold for case-control studies as well. These simulations used a sample of 1000 individuals, however modern GWAS use much larger samples and hence we are unlikely to find a reduction in power for even a case-control design.

3.5 Chapter Summary

My simulations suggest that the additive-additive interaction test is generally more powerful than the general interaction test over a range of realistic models. The next chapter develops a computationally efficient approach to perform additive-additive interaction tests and applies it to a genome wide scan for seven common complex disorders from the WTCCC [Wellcome Trust Case Control Consortium, 2007], and obesity. The ideal approach is to conduct an exhaustive two-locus scan, but this will pose a serious computational challenge with large numbers of SNPs. In order to address such issues, I propose to apply a novel two stage algorithm called "IntRapid".

Chapter 4

Genome wide scan for gene-gene interactions

4.1 Introduction

There has been no universal agreement on the best way to test for gene-gene interactions on a genome wide scale [Cordell, 2009]. As described in Chapter 2, a wide variety of methods have been proposed, including the use of dimension reduction methods [Ritchie et al., 2001], and MCMC based approaches [Yi et al., 2005]. Each has their own short comings, and we are perhaps far from developing a convincing approach to conduct genome wide scans for statistical interactions on a routine basis. One of the biggest challenges is that of a computational nature, particularly for the large sample sizes required to detect the modest interaction effects for complex traits and the large numbers of SNPs considered in GWAS. In this chapter, I develop a computationally efficient approach to test for additive-additive interactions on the scale of the whole genome [Bhattacharya et al., 2011].

4.2 Rapid interaction testing

In response to the severe computational burden of fitting regression models for all possible pairs of SNPs in a genome-wide scan, a faster and cruder method to remove some of these combinations is essential. The PLINK software includes a "fast-epistasis" algorithm which scans for gene-gene interactions in a computationally efficient manner [Purcell et al., 2007]. However, it has a number of shortcomings which need addressing.

- The test is **quite crude** in the sense, it does not take into account the **underlying LD** between a pair of SNPs and thereby on its own is not reliable for an interaction test.
- The PLINK website states a traditional GLM based approach and the "fast-epistasis" method generates "similar results". However, as I have demonstrated later, this is hardly the case. The "fast-epistasis" method generates very **many false-positive** results and results in a serious **compromise on statistical power**.
- The PLINK recommends using a 'reasonably liberal' threshold to screen for pairs of SNPs before applying a formal GLM. However, **no formalised threshold** has been suggested for the first stage and as such the impact on statistical power and computational demand is not thoroughly discussed.
- "Fast-epistasis" **only** works on **case-control designs**.
- **Covariates** cannot be adjusted for.

- PLINK reads entire datasets into memory and for large datasets this is **memory heavy** and very often unworkable.
- If the **volume** of desired output is **large** PLINK would struggle or **fail** owing to the way it has been written.
- It is not easy to run a PLINK epistasis scan in **parallel**. There is **no seamless way** to split a very large dataset and despatch them to remote nodes.
- Thanks to **many inefficiencies** in the code it is not as efficient as it could be.

I have tried to address all of the above limitations in my algorithm, Intrapid which is a two-stage approach. Stage 1 uses the rapid, but crude, "fast-epistasis" approach to identify pairs of SNPs for more detailed investigation in stage 2 by means of an additive model in a GLM framework (as described in section 4.2.2). I have performed simulations to formalise a threshold for stage 1 of the analysis to optimise statistical power, which also saves vast amounts of computational time. It can work with binary or quantitative traits and covariates can be accounted for quite easily. There is the option of not loading the entire dataset in memory and this allows the software to work on the dataset even when it is larger than the memory available. Similarly, is the output is very large it can write it as and when is ready, rather than save everything for the end. Intrapid has been written to access parts of a very large dataset very quickly, split it (if necessary) and then work with it in a remote node to minimise network and memory load. Profiling the code helped identify some key inefficiencies in the existing algorithm. I have

used some mapping methods to speed up the computation. This has increased the efficiency by around ten fold (more details in 4.3).

I thus implement a two step approach, where stage 1 is used to prune down the total number of interaction pathways which need to be investigated in a computationally efficient manner. Stage 2 implements a traditional GLM using pairs of SNPs selected in stage 1 as its explanatory variables. The method does not require a lot of memory and can easily be run in parallel.

4.2.1 Stage 1

The PLINK "fast epistasis" method is applicable to case-control studies. For a quantitative trait, I use the median of the phenotype to dichotomize individuals as "pseudo-cases" and "pseudo-controls". However, if there are covariates to be accounted for, I fit a GLM with the covariates and extract the residuals to use it as a "pseudo-phenotype".

Specifically, pseudo case control status is given by,

$$w_i = \begin{cases} 0 & \text{if } y_i \leq y_{med}, \\ 1 & \text{if } y_i > y_{med} \end{cases}$$

where, y_{med} is the median of the quantitative trait and y_i is the phenotype of the i^{th} individual.

For a sample of n individuals, and a given pair of SNPs, let us assume that the first has three genotypes (AA, Aa and aa) and the second has three genotypes (BB, Bb and bb). If, n_{ijkl} were the number of individuals with ij genotype at the first SNP and kl genotype at the second SNP, then Table 4.1 gives the contingency table for the pair of SNPs in question.

Table 4.1: Contingency table for a pair of SNPs

SNP 2	SNP 1			
	AA	Aa	aa	Total
BB	n_{AABB}	n_{AaBB}	n_{aaBB}	$n_{..BB}$
Bb	n_{AABb}	n_{AaBb}	n_{aaBb}	$n_{..Bb}$
bb	n_{AAbb}	n_{Aabb}	n_{aabb}	$n_{..bb}$
Total	$n_{AA..}$	$n_{Aa..2}$	$n_{aa..}$	n

The genotype frequency table, 4.1, may then be transformed into an allele frequency table (4.2),

Table 4.2: Contingency table for a pair of alleles

Locus 2	Locus 1	
	A	a
B	w	x
b	y	z

where,

$$\begin{aligned} w &= 4n_{AABB} + 2n_{AaBB} + 2n_{AABb} + n_{AaBb}, \\ x &= 4n_{aaBB} + 2n_{AaBB} + 2n_{aaBb} + n_{AABb}, \\ y &= 4n_{AAbb} + 2n_{Aabb} + 2n_{AABb} + n_{AaBb}, \\ z &= 4n_{aabb} + 2n_{Aabb} + 2n_{aaBb} + n_{AaBb} \end{aligned}$$

The odds ratio (OR) for a 2×2 table of this kind is then calculated separately for the cases and controls.

$$OR = \frac{w \times z}{x \times y} \tag{4.1}$$

and the corresponding variance could be estimated by,

$$Var(log(OR)) = \left(\frac{1}{w} + \frac{1}{x} + \frac{1}{y} + \frac{1}{z}\right) \tag{4.2}$$

However, [Ueki and Cordell, 2012] points out the above estimate (equation 4.2) is strictly speaking incorrect and results in a more conservative test. The true asymptotic variance is in fact slightly smaller. For the sake of simplicity I have continued to use the variance estimate in equation 4.2, particularly because the differences are very small.

If OR for the cases is OR_{cases} and that for the controls is $OR_{controls}$, the statistic used for testing differences between the ORs is,

$$Z = \frac{log(OR_{cases}) - log(OR_{controls})}{\sqrt{Var(log(OR_{cases})) + Var(log(OR_{controls}))}} \tag{4.3}$$

Under a model of no interaction between the SNPs, we should expect the same distribution of alleles at the two SNPs with respect to phenotype, and hence no difference in the OR between cases and controls. The Z-statistic thus provides a test of interaction between the two SNPs.

4.2.2 Stage 2

Pairs of SNPs are chosen from stage 1 using a p-value threshold (α_{RAPID}) and carried forward to more detailed testing in stage 2. The choice of α_{RAPID} determines the number of interactions performed in stage 2. Choosing a stringent threshold minimises computational burden, but may miss 'true' interactions not identified in the rapid testing stage. A GLM assuming an underlying additive model was applied to these pairs of SNPs (equation 2.1) as described in section 2.2.1.1. A maximum likelihood ratio test was performed to compute the p-value

corresponding to the interaction term.

4.3 Software development

One objective was to develop software which could be easily distributed to researchers and used with standard desktop PCs. Many research groups do not have access to large scale computing facilities, meaning that implementation of an exhaustive interaction scan is infeasible. A massive reduction in computing time and temporary memory requirements would enable researchers to conduct such analysis on their regular laptops or desktops in a reasonable amount of time. I designed and programmed the 'IntRapid' software to implement the computationally efficient two-stage approach described in section 4.2. It has managed to address a lot of the computational issues present in PLINK.

C++ was chosen as the language to write the program. Standard regression routines were used from the GNU scientific libraries. A program was created which could convert standard format files into that needed for the software. The converter also allows some basic quality control filters, including removal of SNPs with low MAF. In the case of quantitative traits, the converter further allows adjustment of the phenotype for covariates, the residuals from which are used in stage 1 analysis. Creating the contingency table for stage 1 (as described in section 4.2.1) appeared to be particularly time consuming when I profiled the code. In order to speed up the process, I created empty matrices and the genotypic data (number of minor alleles in each location) were used as the matrix indices to increment the cell counts. This made a dramatic difference to the speed of computation.

The regression scan was designed to be exhaustive in nature. However, given the large size of possible datasets, I created a map index of each large file to allow splitting, extraction and loading (map and split feature) of small sections of the dataset for analysis, and then store the output files for each section. This feature is not available in PLINK and makes a radical difference to the efficiency of data extraction from very large files. The 'map and split feature' is uses a triangular grid to decide on dividing computational tasks and later accumulating outputs. The software is able to make the processes memory efficient and irrespective of the size of data provided the program should work seamlessly. When parallel computing resources are available, 'IntRapid' is able to run the entire routine in parallel with minimal human interaction. Once the analysis is complete, the files are merged using a simple routine, which results in the final output.

The process of breaking up the analysis is done carefully and with the objective of optimizing computation time. While loading too much data in one go would slow the process down, loading too little would mean the connection to read data would be opened several times and thereby reduce the overall progress of the analysis. No real effort was made to alter the default buffer mechanism of the system. By assuming a RAM capability of about 2GB, the number of divisions of the dataset would be noted. Two of all the possible combinations of these chunks would be taken at a time and an index of jobs would then be formed. This process is then fired off in a grid format. This optimizes the time of analysis given the hardware of the computer. Some of the parameters could be defined by the user, depending on the hardware systems of the computer in use. In absence of an auto-grid engine a separate BASH or R script is required to fire the jobs using PBS. The software is publicly available and can be freely downloaded from

http://www.well.ox.ac.uk/INTRAPID/.

4.4 Simulations

A detailed simulation study was carried out to compare the power of "IntRapid" and a traditional exhaustive GLM for testing interaction between SNPs in a quantitative trait GWAS. Two interacting loci were simulated for 5000 individuals in each dataset. The p-values corresponding to the additive-additive interaction effect of the GLM were compared to a nominal threshold of $p < 0.05$ to evaluate significance. I generated 1000 replicates of data for a given combination of model parameters; epistatic model, causal allele frequency (CAF) and interaction effect size (ϵ).

I considered various theoretical epistatic models, "pure additive", "additive-additive", "dominant-dominant" and "recessive-recessive", for the simulations (as described in Chapter 3). Each model was parametrized in terms of CAF at the two causal SNPs, denoted by q_1 and q_2, and the interaction component (ϵ). For a given model, these components were used to obtain the population trait values, denoted by μ_{jk} (Table 4.3) for each genotype combination jk.

The genotypes (G_{i1} and G_{i2}) were generated based on the respective allele frequency distributions and assuming HWE. For each population trait mean, μ_{jk} (Table 4.3) and the assumed common environmental variance ($\sigma^2_E = 1$), corresponding trait values were generated from a normal distribution. For each dataset, stage 1 of the "IntRapid" method was applied and for various levels of cut-off p-values (α_{RAPID}), pairs of SNPs were selected for more detailed analysis in stage 2. When $\alpha_{RAPID} = 1$, the "IntRapid" method is equivalent to an ex-

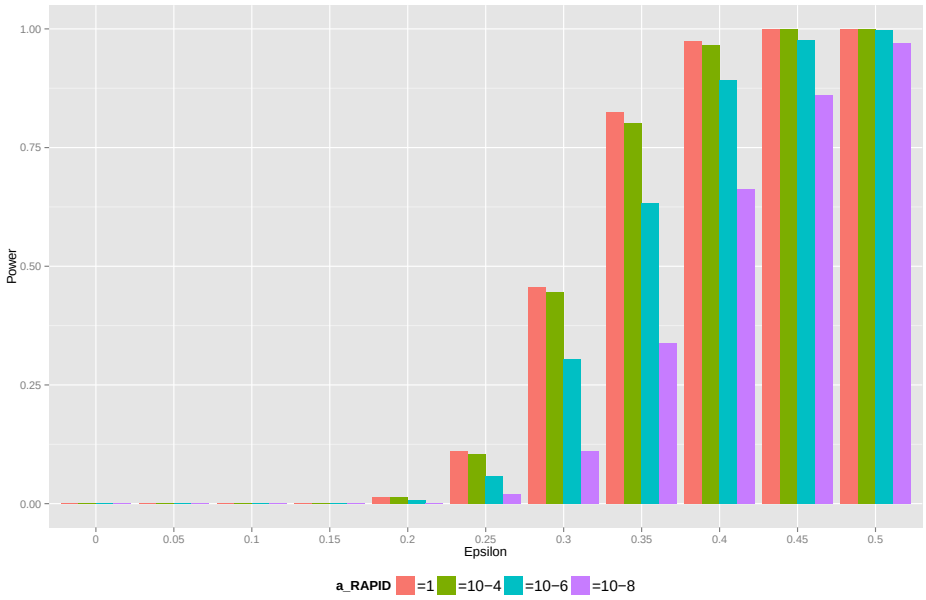
Table 4.3: Population mean trait values of theoretical epistasis models

Interaction Models:		SNP 2		
	SNP 1	MM	Mm	mm
Pure additive	MM	$\mu_{00} = \epsilon$	$\mu_{01} = 0$	$\mu_{02} = -\epsilon$
	Mm	$\mu_{10} = 0$	$\mu_{11} = 0$	$\mu_{12} = 0$
	mm	$\mu_{20} = -\epsilon$	$\mu_{21} = 0$	$\mu_{22} = \epsilon$
Additive-additive	MM	$\mu_{00} = \epsilon - 1$	$\mu_{01} = -0.5$	$\mu_{02} = -\epsilon$
	Mm	$\mu_{10} = -0.5$	$\mu_{11} = 0$	$\mu_{12} = 0.5$
	mm	$\mu_{20} = -\epsilon$	$\mu_{21} = 0.5$	$\mu_{22} = 1 + \epsilon$
Dominant-dominant	MM	$\mu_{00} = 0$	$\mu_{01} = 0$	$\mu_{02} = 0$
	Mm	$\mu_{10} = 0$	$\mu_{11} = \epsilon$	$\mu_{12} = \epsilon$
	mm	$\mu_{20} = 0$	$\mu_{21} = \epsilon$	$\mu_{22} = \epsilon$
Recessive-recessive	MM	$\mu_{00} = 0$	$\mu_{01} = 0$	$\mu_{02} = 0$
	Mm	$\mu_{10} = 0$	$\mu_{11} = 0$	$\mu_{12} = 0$
	mm	$\mu_{20} = 0$	$\mu_{21} = 0$	$\mu_{22} = \epsilon$

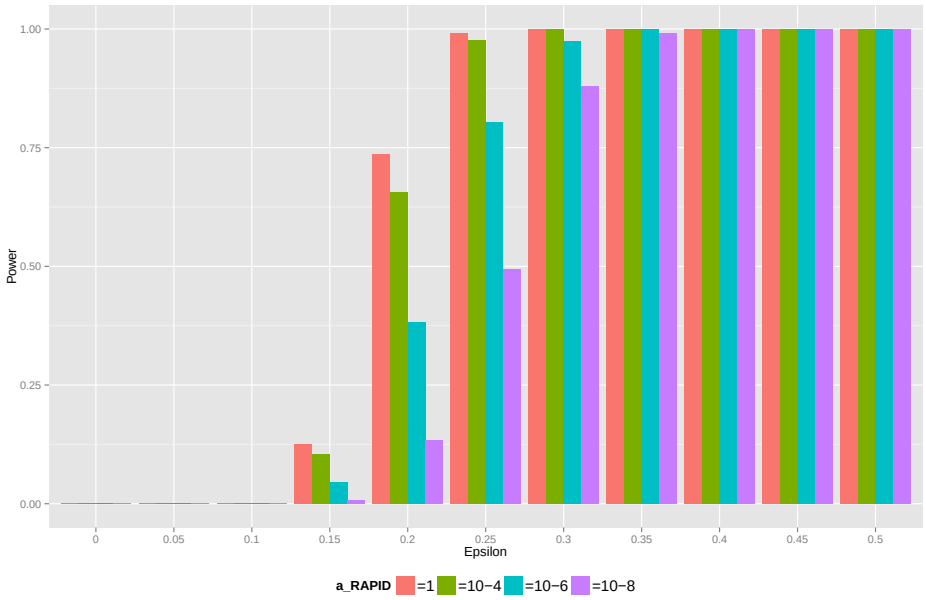
haustive linear regression search since all pairs of SNPs are carried forward for detailed analysis in the GLM. The number of pairs of SNPs which were selected for stage 2 and thereby the computation time could thus be controlled by changing α_{RAPID}. The p-values corresponding to the interaction term in the GLM, performed in stage 2, were then compared against a nominal threshold of 10^{-10}. Power was calculated as the proportion of SNP pairs which are significant at this threshold.

Figure 4.1(a) and 4.1(b) show that, for an "additive-additive" model, a maximum cut-off p-value of $\alpha_{RAPID} = 10^{-4}$ in stage 1 results in very little or no loss in power, compared to an exhaustive GLM. The results were very similar for a "dominant-dominant" model (Figure 4.2(a) and 4.2(b)), "pure additive" or a "recessive-recessive" model.

However, when the interaction effect is quite low, the resulting power would be quite low irrespective of the method used. However, for all moderate sized

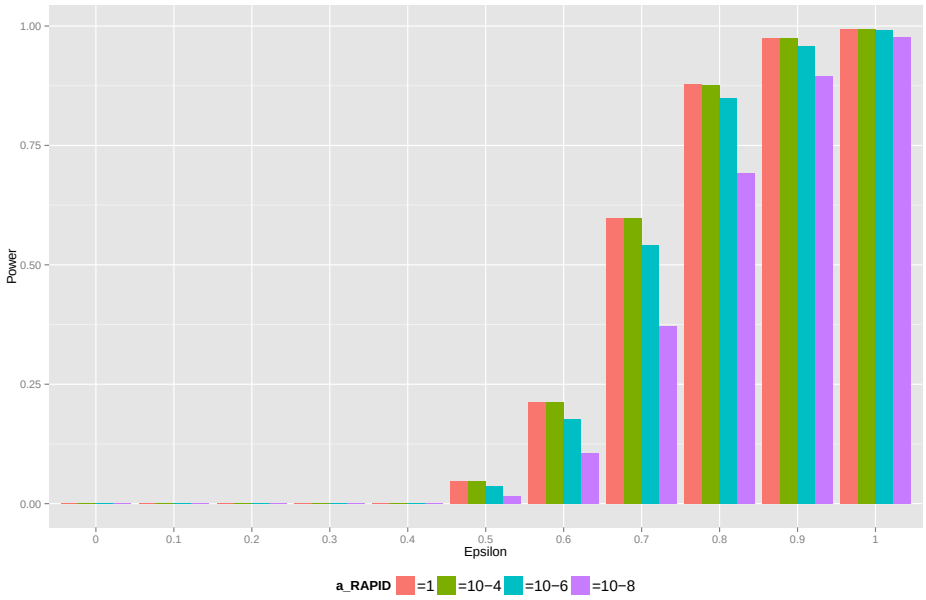


(a) CAF of 50% and 10% at the two loci respectively

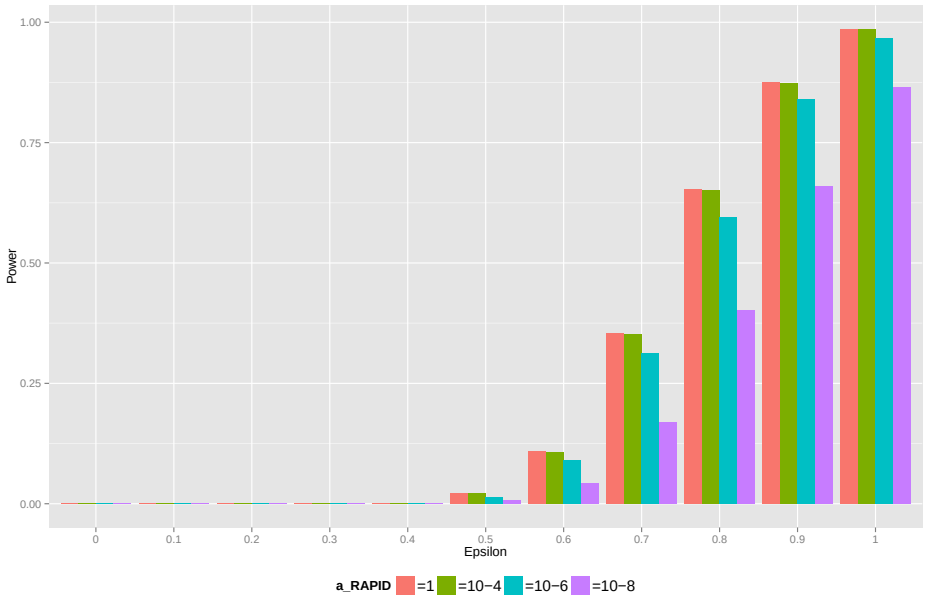


(b) CAF of 50% and 50% at the two loci respectively

Figure 4.1: **Additive-additive model**: Power levels for various cut off p-values in stage 1 of "IntRapid". Power at a nominal significance threshold of the interaction p-value $p_{GLM} < 10^{-10}$, of the rapid interaction testing strategy for an additive-additive two-locus model of association, with causal allele frequency (CAF) of 50% at one SNP and 10% or 50% at the other. Power is presented as a function of the additive-additive interaction component for a range of thresholds, α_{RAPID}, for carrying forward pairs of SNPs from the rapid testing stage to the GLM analysis framework.



(a) CAF of 50% and 10% at the two loci respectively



(b) CAF of 50% and 50% at the two loci respectively

Figure 4.2: **Dominant-dominant model**: Power levels for various cut off p-values in stage 1 of "IntRapid". Power at a nominal significance threshold of the interaction p-value $p_{GLM} < 10^{-10}$, of the rapid interaction testing strategy for an additive-additive two-locus model of association, with causal allele frequency (CAF) of 50% at one SNP and 10% or 50% at the other. Power is presented as a function of the additive-additive interaction component for a range of thresholds, α_{RAPID}, for carrying forward pairs of SNPs from the rapid testing stage to the GLM analysis framework.

interaction effects, a cut-off p-value of $\alpha_{RAPID} = 10^{-4}$ in stage 1 would make the results almost equivalent to the exhaustive GLM scans, irrespective of the underlying theoretical epistatic model.

4.5 Application to GWAS data from the WTCCC

The WTCCC was set up to take advantage of the improvement in genotyping technology to help understand the genetic architecture of seven common complex diseases. Genotypic data were collected from more than $500,000$ sites of common genetic variation (using the Affymetrix 500k chip) in approximately 2000 cases of each disease and 3000 shared controls from the 1958 British Birth Cohort and National Blood Service (Table 4.4).

For the 2000 T2D cases, body mass index (BMI) was also available to us (through collaboration with Prof Mark McCarthy). BMI is measured as the ratio of weight (in kilograms) and the square of height (in metres). Excess BMI, for example $BMI \geq 30$, can be indicative of **obesity** which is a medical condition where excessive deposition of body fat leads to increased health problems and the likelihood of a reduced life expectancy. It is a growing problem worldwide and perhaps particularly so in developed countries. The causes of obesity are typically a combination of sedentary lifestyle, excessive intake of high calorie food and genetic susceptibility. Obesity is often seen to be associated with other common diseases such as type 2 diabetes and hypertension.

Table 4.4: Disease cohorts, their abbreviations and calculated heritability in the WTCCC study

Disease Cohorts with abbreviations	**Heritability** in %
Bipolar disorder (BD)	80 - 90 [Craddock et al., 2005]
Coronary heart disease (CHD)	34 - 53 [Katzmarzyk et al., 2000]
Hypertension (HT)	31 - 68 [Ji et al., 2011]
Crohn's disease (IBD)	50 - 60 [Sofaer, 1993]
Rheumatoid arthritis (RA)	53 - 65 [MacGregor et al., 2000]
Type 1 diabetes (T1D)	72 - 88 [Hyttinen et al., 2003]
Type 2 diabetes (T2D)	26 [Poulsen et al., 1999]
Control Cohorts	
1958 Birth Cohort (58C)	-
UK Blood Service (UKBS)	-

4.5.1 Quality Control

In order to minimize the probability of finding false positives, I had to implement strict quality control (QC) measures. I employed the same set of quality control filters as published in the methods section of the WTCCC paper [Wellcome Trust Case Control Consortium, 2007]. This ensured the minimization of false positive signals in the final results. Specific details of the QC process are as follows:

- High missing data rate per individual indicates low DNA quality. Thus, individuals with missing data rates of $\geq 3\%$ were removed.
- Excessive heterozygosity suggests DNA contamination. Samples with heterozygosity rates of $\geq 30\%$ or $\leq 23\%$ were removed.
- Individuals with discrepancies between WTCCC information and external identifying criteria, such as genotypes from another experiment, blood types or incorrect disease status were removed.
- Individuals with non-European ancestry, as detected by means of multi-dimensionality scaling, were removed.
- Duplicate or apparent related samples were removed.
- SNPs with a study wide missing rate of $\geq 5\%$ were removed. The threshold was $\geq 1\%$ when the MAF was $\leq 5\%$.
- SNPs with deviation of more than 1% from HWE were removed.
- SNPs with MAF of less than 5% were removed from further analysis to avoid potential problems with sparse two-locus genotype counts.

Table 4.6 reports the sample size and number of SNPs used for each of the eight interaction scans. After filtering for individuals with large fractions of missing SNPs, missing phenotypes and many other quality control measures, the remaining sample size was short of 2000 for all diseases. Around $350k$ SNPs were left for most diseases after all standard quality filtering processes. Table 4.6 further reports the break up of the sample size into cases and controls and males and females.

4.5.2 Implementation of interaction scans using IntRapid

I have applied the "IntRapid" software to the seven diseases and BMI.

Table 4.5: Time required to complete analysis on the Type 2 diabetes and Obesity (as measured by BMI) samples depending on threshold p-value at stage 1 of the "IntRapid" method.

T2D					
α_{RAPID} Threshold	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Time(secs)	2.06×10^7	3.54×10^6	1.88×10^6	1.72×10^6	1.70×10^6
# regressions	7.66×10^9	7.45×10^8	7.18×10^7	6.89×10^6	6.60×10^5
BMI					
α_{RAPID} Threshold	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Time(secs)	1.21×10^7	2.44×10^6	1.52×10^6	1.43×10^6	1.42×10^6
# regressions	4.31×10^9	4.14×10^8	4.05×10^7	3.88×10^6	3.69×10^5

Using a significance threshold of 10^{-4} for α_{RAPID} would reduce computation

time substantially (as detailed in Table 4.5). This holds true because of the small fraction of SNP pairs which are carried over to Stage 2 of the process. Since only a small fraction of the initial 75 billion regression models is actually computed, there is approximately a 1000 times reduction in computation time.

4.5.3 Results

4.5.3.1 Disease phenotypes

Table 4.6: Sample sizes and number of SNPs used in the seven diseases

Diseases	Sample Size	Cases	Controls	Males	Females	Number of SNPs
BD	4806	1868	2938	2143	2663	357739
CHD	4864	1926	2938	2963	1901	357649
HT	4890	1952	2938	2126	2764	357640
IBD	4686	1748	2938	2221	2465	357898
RA	4798	1860	2938	1916	2882	357611
T1D	4901	1963	2938	2444	2457	357748
T2D	4862	1924	2938	2564	2298	357775

Informed by previous simulations (section 4.4), I imposed a stage 1 significance threshold of $\alpha_{RAPID} \leq 10^{-4}$ for carrying forward pairs of SNPs in the more detailed GLM framework. No pairs of SNPs met a Bonferroni correction for multiple testing of all pairs of SNPs passing QC filters (approximately $P < 7{\times}10^{13}$ depending on the disease). Table 4.7 presents lead SNPs for each disease at each pair of loci demonstrating strong evidence for interaction ($P < 10^{-10}$) in

the second-stage GLM, together with p-values from the first stage rapid test. Also presented are P-values for each SNP from a marginal single-locus test of association from a GLM incorporating only an additive effect. The majority of SNPs show no evidence of marginal association with disease, and hence would not have been discovered through single-locus GWAS analysis under the assumption of an additive effect in the log-odds ratio (hence were not previously reported by the WTCCC).

The strongest signal of interaction in BD appeared between rs10510819 on Chromosome 3 and rs4299909 on Chromosome 7, mapping near $AQP1$ (Table 4.7). Neither of these loci have been previously implicated in BD.

I selected all cases where the stage 2 p-value was lesser than or equal to $10-6$ and plotted the -log10 p-values from Stage 1 (Figure 4.3) and Stage 2 (Figure 4.4) across the genome. As can be noticed, the Stage 1 -log10 p-value result in several false-positive signals. The red line corresponds to the approximate Bonferroni corrected threshold of 10^{-13} and the blue line corresponds to a nominal threshold of 10^{-10} that I have used in this thesis.

When I performed a 'smoothScatter' plot using package 'graphics' in R, the levels of agreement between -log10 p-values appeared to be quite strong for all stage-2 p-values which are lesser than 10^{-8} (Figure 4.5(a)). A plot of all -log10 p-values show that the amount of linear agreement continues to be strong (Figure 4.5(b)), however several outliers can now be noticed. The Pearson's product moment correlation are around 0.52 and 0.89 in the two cases respectively.

I also considered the upper tail of a qqplot to illustrate the genome wide distribution of stage 2 -log10 p-values (Figure 4.6). It appears our test may have been a little conservative, but there are no signals which substantially deviate

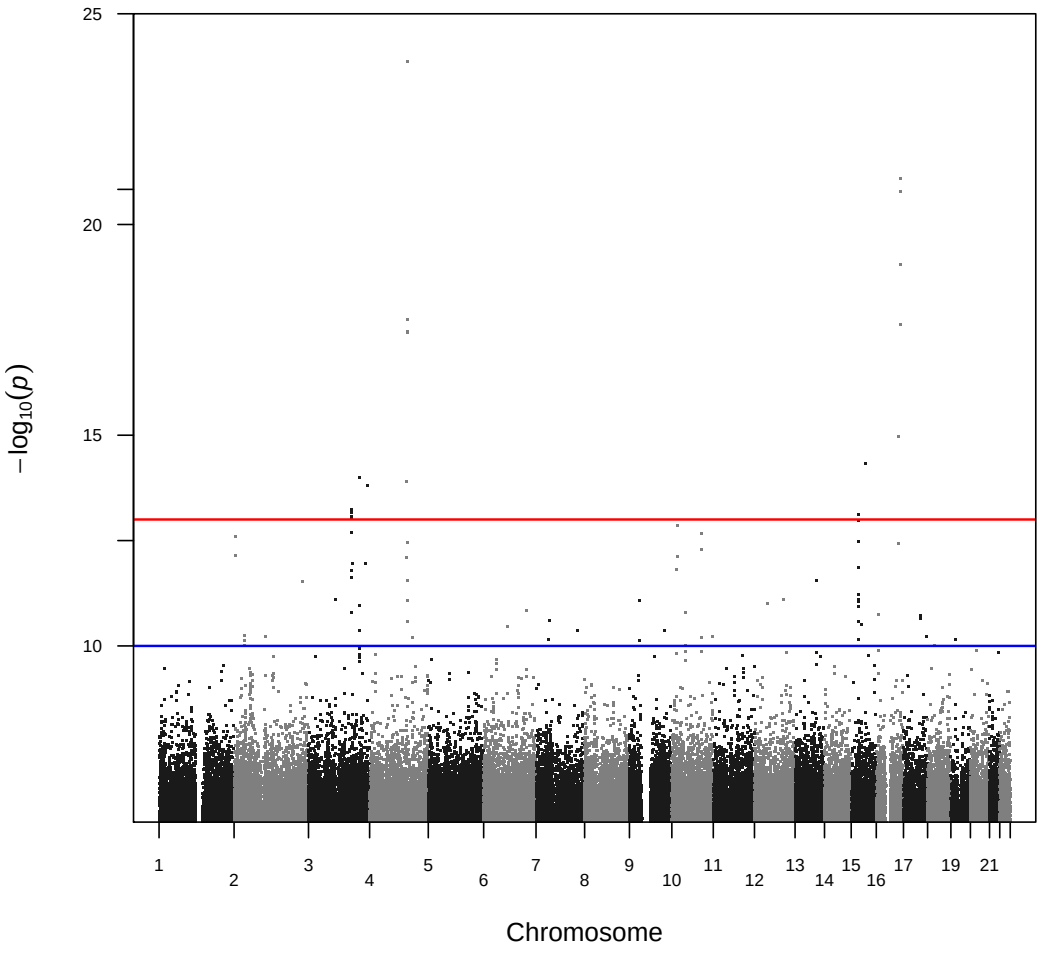


Figure 4.3: Genome wide plot of interaction p-values from Stage 1 of IntRapid in BD.

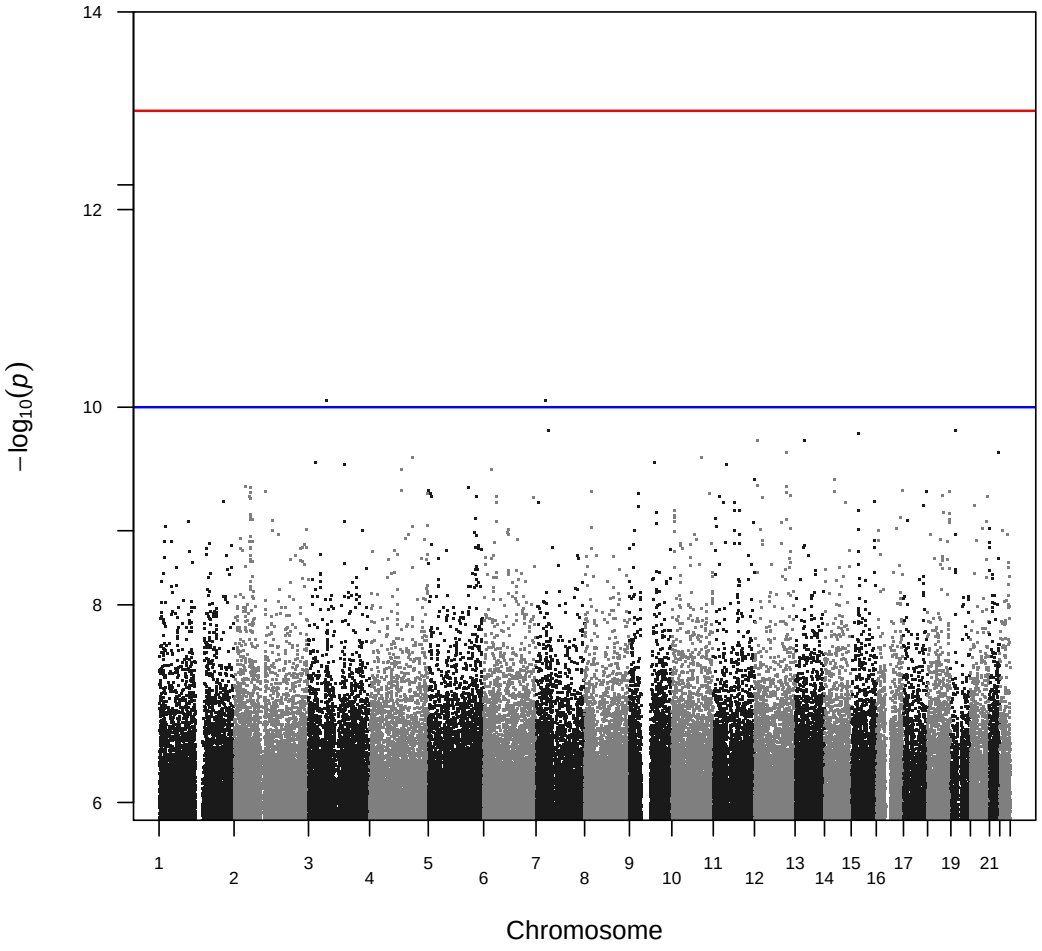
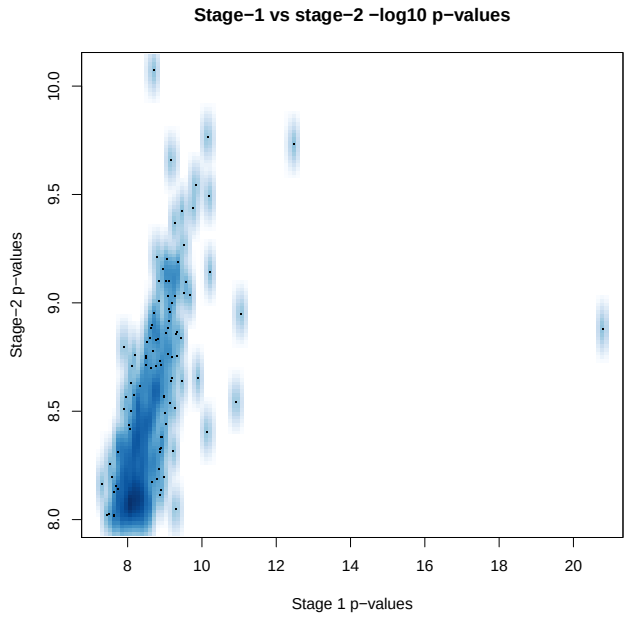
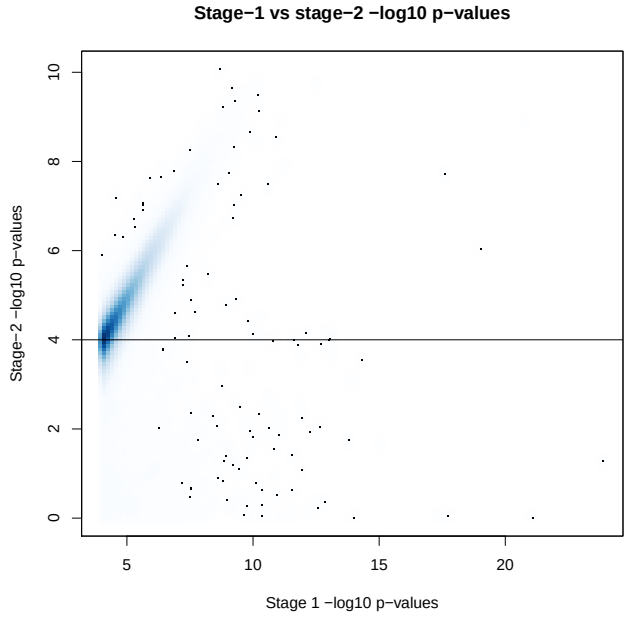


Figure 4.4: Genome wide plot of interaction p-values from a GLM in BD.



(a) Plot of -log10 p-values in cases where p-values at Stage-2 are lesser than 10^{-8}



(b) Plot for all -log10 p-values at Stage-1 and Stage-2 of Intrapid

Figure 4.5: Plot shows the -log10 p-values in Stage-1 and Stage-2 mostly agree at the smaller end of the spectrum. The level of agreement is much lower in a plot for all -log10 p-values. The horizontal line in plot (b) helps identify cases where the stage-1 p-values were large but the stage-2 p-values were lower than 10^{-4}.

from the diagonal.

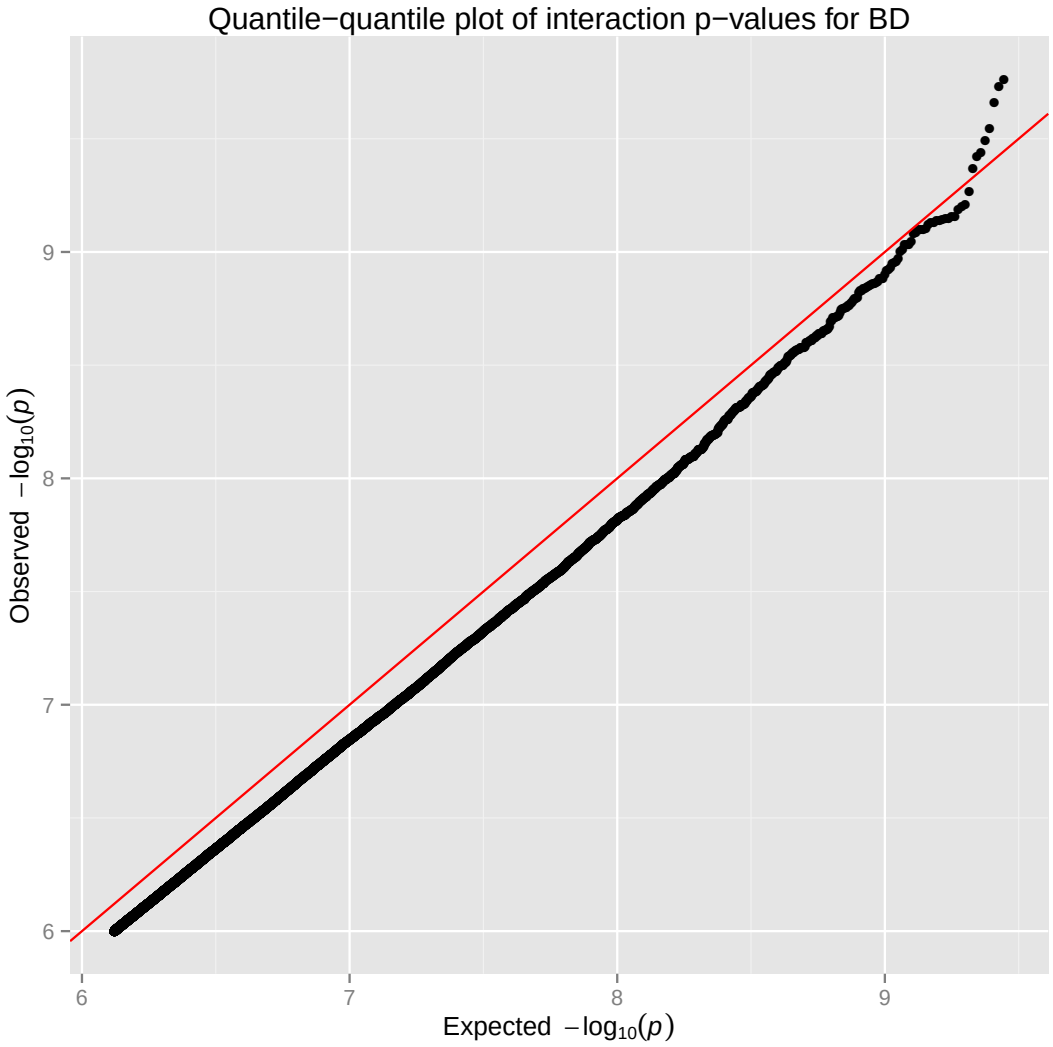


Figure 4.6: QQ plot of -log10 p-values from stage 2 of applying IntRapid to the WTCCC dataset of BD

The strongest signal identified for CHD was between rs11170276 in chromosome 12 and rs273763 in chromosome 18. Once again, no previous association with CHD has been reported for these loci. Two interaction effects were identified for HT that satisfied the nominal threshold of 10^{-10}. The first one was between

rs9521017 of chromosome 13 and rs17292844 of chromosome 20. The former lies in a neighbourhood of the $MY016$ gene and the latter is close to the $C20orf133$ gene. The second interaction for HT was between rs12492072 in chromosome 3 and rs34965037 in chromosome 7. The latter maps near the $PEX1$ gene. To my knowledge these loci have not been implicated for HT previously.

Four interaction effects were identified for IBD. These included interactions with the genes PTPRD in chromosome 9 and $TACC2$ in chromosome 10. All four of the signals in RA mapped to the MHC region of chromosome 6. The closest genes to the lead SNPs were $SEEK1$, $HCG27$ and $NOTCH4$. The SNP mapping to $NOTCH4$ also demonstrated genome-wide significant evidence of marginal association with RA. Nine interaction effects were identified for T1D. Five of these mapped the MHC region of chromosome 6. The closest genes were $LOC346147$, $ZNF192$ and $GABBR1$.

The strongest signal of interaction with T2D was detected between a pair of proximal SNPs in the ZFAT gene. The SNPs are only weakly correlated with each other ($r^2 = 0.062$ and $D' = 1.000$ in CEU 1000 Genomes Project pilot data), and this interaction could represent a haplotype effect where the risk alleles have a synergistic effect when *in cis*, for example. The $ZFAT$ gene encodes a protein that likely functions as a transcriptional regulator involved in apoptosis, but has not been previously implicated in T2D or other metabolic traits. The second strongest signal of interaction with T2D was detected between a pair of SNPs in the OBSC gene and flanking the CLU gene, respectively. These genes have not been previously implicated in T2D. However, both genes are involved in carbohydrate and lipid metabolic process and in apoptosis, and thus might be reasonable candidates for interaction within related functional pathways.

Table 4.7: Gene-gene interaction signals obtained after applying IntRapid on all seven WTCCC diseases. Pairs of SNPs where the p-value corresponding to the interaction term in the GLM satisfies a minimum threshold of 10^{-10}

	SNP 1					SNP 2					Interaction	Marginal p-value	
Disease	ID	Chr	Position	MAF	Genes	ID	Chr	Position	MAF	Genes	P-value	SNP 1	SNP 2
BD	rs10510819	3	59691572	0.11		rs4299909	7	30969120	0.06	*AQP1*	8.47E-11	7.39E-01	2.53E-03
CHD	rs11170276	12	53179673	0.20	*KRT77, KRT1*	rs273763	18	23197631	0.43		6.88E-11	1.56E-01	9.90E-01
HT	rs9521017	13	109419159	0.36	*MY016*	rs17292844	20	15119797	0.31	*C20orf133*	2.36E-11	7.41E-01	1.19E-01
HT	rs12492072	3	188884221	0.06		rs34965037	7	92117993	0.24	*PEX1*	1.81E-12	8.89E-01	4.45E-01
IBD	rs10809018	9	10035658	0.24	*PTPRD*	rs16907121	9	23341848	0.11		2.33E-11	5.29E-01	4.82E-02
IBD	rs6554056	4	53721466	0.29		rs2609850	22	34851377	0.28		4.85E-11	5.18E-01	1.10E-01
IBD	rs4693426	4	83093938	0.10		rs3898061	7	152737701	0.15		8.70E-11	5.37E-01	5.94E-01
IBD	rs10887096	10	123926310	0.11	*TACC2*	rs7318474	13	29129431	0.37		1.33E-11	6.97E-01	1.29E-02
RA	rs4959053	6	31099577	0.07	*SEEK1*	rs9295961	6	31167498	0.06	*HCG27*	2.10E-13	1.81E-23	5.72E-01
RA	rs206015	6	32182759	0.15	*NOTCH4*	rs6936863	6	32684029	0.26		2.25E-16	3.93E-12	4.95E-01
RA	rs188245	6	32955976	0.46		rs4713604	6	32979770	0.37		1.60E-11	2.23E-02	1.62E-01
RA	rs1419442	7	126734171	0.35	*GRM8*	rs10740429	10	53916564	0.20	*PRKG1*	9.51E-11	3.44E-01	6.31E-02
T1D	rs11677151	2	12759532	0.08		rs949882	6	109586062	0.34	*RP11*	8.85E-11	9.87E-01	2.10E-01
T1D	rs7765920	6	26514771	0.21		rs12529458	6	27361681	0.44	*LOC346147*	9.02E-11	1.46E-01	2.53E-03
T1D	rs1871695	6	28109960	0.06	*ZNF192*	rs362525	6	29543646	0.16	*GABBR1*	1.54E-11	2.07E-06	2.20E-01
T1D	rs362525	6	29543646	0.16		rs3129012	6	29988642	0.20		8.25E-11	2.20E-01	9.07E-04
T1D	rs29228	6	29623739	0.23		rs9267673	6	31883679	0.10		3.52E-11	2.67E-04	1.68E-06
T1D	rs3117016	6	33095516	0.41		rs1003979	6	33114171	0.45		2.04E-12	7.95E-05	6.42E-06
T1D	rs6683663	1	57356476	0.14	*C8A*	rs6818162	4	68738811	0.38	*TMPRSS110*	6.60E-11	5.50E-01	9.73E-01
T1D	rs1475398	1	65983257	0.27	*LEPR*	rs6015978	20	37516318	0.38	*PPP1R16B*	6.84E-11	5.73E-01	2.19E-01
T1D	rs4857717	3	177328322	0.33		rs7158575	14	100215383	0.14		5.11E-11	6.13E-01	2.08E-01
T2D	rs10916293	1	228431930	0.33	*OBSCN*	rs9314349	8	27474202	0.38		2.82E-11	5.66E-01	8.99E-01
T2D	rs6421008	8	135581827	0.05	*ZFAT1*	rs7827545	8	135566567	0.32	*ZFAT1*	1.57E-11	5.07E-01	1.83E-03

4.5.3.2 Quantitative trait - BMI

The quantitative trait was dichotomized at the median (=30.94) of the BMI values for stage 1 of the analysis. In stage 2 of the analysis, the phenotype is log transformed (to meet assumptions of normality) and then adjusted for possible confounding factors of gender and age. Using a rapid interaction testing threshold of $\alpha_{FAST} < 10^{-4}$, 7.14 million pairs of SNPs in total were carried forward for consideration in the more computationally intensive GLM framework. This analysis took 127 hours of computing time on a dedicated processor, compared to an expected 4,826 hours to test for the interaction between all pairs of SNPs in the GLM framework.

After Bonferroni correction, the significance threshold was calculated to be 7.1×10^{-13}. There were no signals which satisfied this conservative threshold. In general, the signals of interaction were much weaker for BMI than for the binary disease traits, reflecting smaller sample size.

Table 4.8 presents lead SNPs at each pair of loci demonstrating strong evidence for interaction ($P < 10^{-10}$) in the second-stage GLM, together with P-values from the first-stage rapid test. P-values for each SNP from a marginal single-locus test of association from a GLM incorporating only an additive main effect are also presented. None of these SNPs shows evidence of marginal association with BMI (adjusted for age and sex), and thus none would have been discovered through single-locus GWAS analysis. The loci corresponding to strongest signals of pair-wise interaction have not been previously implicated in obesity or other metabolic traits.

Table 4.8: In obesity as measured by BMI, pairs of SNPs where the p-value corresponding to the interaction term in the GLM satisfies a minimum threshold of 10^{-10}

rs ID	Chr	Position	Gene	MAF	rs ID	Chr	Position	Gene	MAF	Stage1	Stage2	SNP1	SNP2
rs12655480	5	178212877	$ZNF354B$	0.35	rs2423646	20	12212116	$TMC5$	0.40	1.34×10^{-6}	2.50×10^{-11}	6.89×10^{-1}	2.89×10^{-1}
rs17684830	3	60910010	$FHIT$	0.17	rs521446	9	133647868	$VAV2$	0.47	3.52×10^{-8}	5.48×10^{-11}	7.35×10^{-1}	4.70×10^{-1}
rs10067788	5	151313204	$GLRA1$	0.07	rs8059797	16	19311395		0.37	3.18×10^{-6}	8.11×10^{-11}	6.17×10^{-1}	1.57×10^{-1}

4.6 Summary

One of the key challenges addressed in this chapter has been that of a computational nature. Computing several billions of interaction models, without compromising on statistical power in a feasible amount of time, is a significant challenge. A two stage algorithm, as described in this chapter, was able to address this issue to a huge extent. However, an efficient implementation of the "IntRapid" algorithm, so that it can be run in parallel and only consume limited amounts of memory was another part of the challenge.

Simulations demonstrated that for a stage 1 cut off p-value of 10^{-4} there was limited or no loss in statistical power, under assumptions of an additive-additive interaction model. Stage 2 of this process had to fit only a small fraction of the initial billions of interaction models. The software "IntRapid" that was developed was able to run the entire process in parallel and thus enabled me to obtain the final interaction signals for a given disease in just a few hours.

The interaction signals reported satisfied a nominal significance level of 10^{-10}. This is larger than the Bonferroni multiple testing correction adjusted p-value of around 10^{-13}. However, contrary to Bonferroni correction assumption of independence between multiple tests, interaction signals are likely to be correlated with each other given the underlying LD. One interesting aspect of these signals is their lack of main effects. When the marginal p-values are inspected closely, only 4 of the 52 SNPs involved in interaction signals would have been detected by single locus scans. If non-exhaustive strategies were adopted, as suggested by [Evans et al., 2006], only 6 of the 26 interaction signals reported would have been identified. This provides further evidence for the need for an exhaustive scan of

interaction scans.

A closer look at the neighbourhood of these identified SNPs show the existence of previously identified genes in the interaction effects. For T1D and RA in particular, several interaction effects were identified in the "gene rich" HLA region of Chromosome 6.

We have been able to identify several potentially important interaction effects for multiple common complex disorders in this chapter. However, given the limited coverage of the 500K Affymetrix chips, used by the WTCCC, the 'causal' SNP may not have been included in the analysis so far. However, we could take advantage of the 1000 Genomes Project data and impute several more million SNPs and thereby increase the likelihood of including the 'causal' variant itself. In the next chapter, I extend the rapid interaction approach to allow for imputed genotype data. I impute the WTCCC GWAS up to reference panels from the 1000 Genomes Project across loci identified from the directly typed data alone, to allow for fine-mapping of interaction effects.

Chapter 5

Interaction between Imputed signals

5.1 Introduction

In chapter 4, potentially interesting gene-gene interaction signals were identified for multiple common complex disorders. These signals did not meet the Bonferroni multiple test corrected p-value (10^{-13}), but were lower than a nominal threshold of 10^{-10}. One strategy to improve power and refine localisation of causal variants is the imputation of 'interesting' loci up to high density reference panel, followed by a detailed interaction scan in these regions [Marchini and Howie, 2010].

Most modern genotyping chips include up to a million variants, which cover most of the common genetic variation in the genome. Imputation of SNPs based on publicly available detailed phased reference panels, like the HapMap project [The International HapMap Consortium, 2003, 2005] or 1000 Genomes Project

[1000 Genomes Project Consortium, 2010, 2012], help predict genotypes at locations not directly typed on existing chips. This "fills in" the variants which are not typed directly on GWAS chips and can help with fine-scale mapping. In this chapter, I make use of imputation to fine map loci identified from my analysis of GWAS for seven diseases from the WTCCC. Imputation provides a probability distribution of the genotype calls at each untyped SNP. This is typically converted to an "expected" genotype which reflects the uncertainty in the evaluated genotype. To implement an interaction scan on imputed data, the "IntRapid" method and the "IntRapid" were modified, as described in this chapter.

5.2 Imputation

Most genotyping chips used for GWAS include around 500,000 to 1 million SNPs. There are publicly available reference panels like the HapMap or 1000 Genomes Project which contain a much denser set of SNPs for several global populations and different ethnic groups. Each population has a unique correlation structure (LD) between the SNPs. Genotypic imputation takes advantage of this LD structure and reference panels to predict genotypes which were not initially typed in the sample. Subject to imputation quality, this increases coverage and power of the study. For a fixed reference panel, the typed SNPs affect the quality of imputation and the success of the study. Imputation is also useful in the meta-analyses of GWAS conducted using different genotyping chips. This allows for the inclusion of the reference panel SNPs in all studies and thereby boosts statistical power [Barrett et al., 2008, Zeggini et al., 2008].

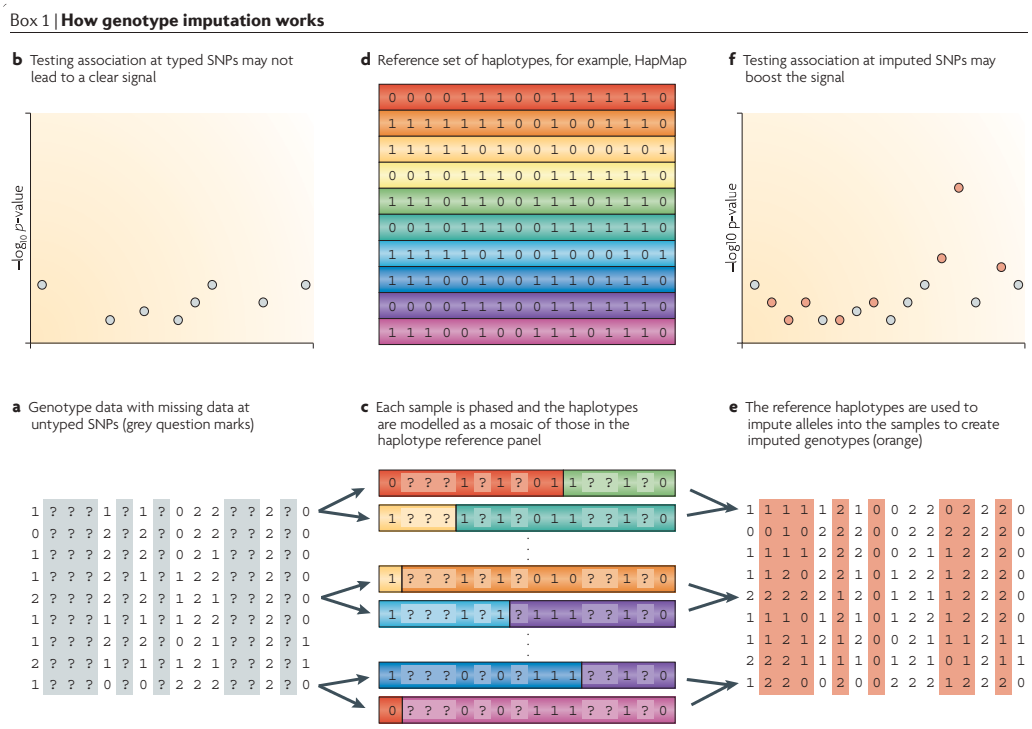
There are a wide range of methods available for imputation, including IM-

PUTE [Howie et al., 2009, Marchini et al., 2007], MACH [Li and Abecasis, 2006], fastPHASE/BIMBAM [Scheet and Stephens, 2006, Servin and Stephens, 2007], PLINK [Purcell et al., 2007], TUNA [Nicolae, 2006], WHAP [Zaitlen et al., 2007] and BEAGLE [Browning, 2006]. Simulations suggest that IMPUTE and MACH minimise error rates over a range of scenarios. In this study, I have made use of IMPUTE for imputation, which I describe in detail below.

5.2.1 Impute

In this thesis, I have used IMPUTE (version 2) to impute missing genotypes Howie et al. [2009]. An outline of imputation, as presented in the IMPUTE paper, has been reproduced in Figure 5.1. Individuals from the same population are expected to have short haplotypes similar to each other owing to identity by descent (IBD). The local structure of the IBD can be estimated by a genealogical tree, where recombination would produce micro level differences in haplotypic frequency. Imputation algorithms exploit the information from these shared haplotypes of the study sample and the reference panel in use. Methodologically, genotype imputation shares common ground with phasing algorithms and tagging SNP methods.

The algorithm used in IMPUTEv2 is slightly different from that in IMPUTEv1. Two sets of SNPs are considered initially. The first set consists of SNPs which have been typed in both the reference panel and the study samples. The second set of SNPs is only present in the reference panel but not in the study sample. IMPUTEv1 is used to estimate haplotypes in the first set. Alleles are then imputed for the second set conditional on estimated haplotypes from the first set.



In samples of unrelated individuals, the haplotypes of the individuals over short stretches of sequence will be related to each other by being identical by descent (IBD). The local pattern of IBD can be described by an (unobserved) genealogical tree, which will differ at different loci throughout the genome owing to recombination. Imputation methods attempt to identify sharing between the underlying haplotypes of the study individuals and the haplotypes in the reference set and use this sharing to impute the missing alleles in study individuals. For this reason, there are strong connections between the models and methods used to infer haplotype phase and those used to perform genotype imputation[22,37], as well as strong connections to tagging SNP-based approaches[19,21,38] and methods used in linkage studies[39,40].

The figure above illustrates imputation for a sample of unrelated individuals. The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part **a**). Testing for association at just these SNPs may not lead to a significant association (part **b**). Imputation attempts to predict these missing genotypes. Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part **c**). These haplotypes are compared to the dense haplotypes in the reference panel (part **d**). Strand alignment between data sets must be done before this comparison takes place (Supplementary information S1 (box)). The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part **e**). In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data. Testing these imputed SNPs can lead to more significant associations (part **f**) and a more detailed view of associated regions.

Figure 5.1: An outline of imputation reproduced from Marchini and Howie [2010]

Table 5.1: Expected genotype calculation from imputed data

Genotype	Additive coding	Probability
AA	0	p_{AA}
Aa	1	p_{Aa}
aa	2	p_{aa}

Owing to the haploid nature of this imputation, rather than the diploid form in IMPUTEv1, it is substantially faster. Phase uncertainty is taken into account by multiple Markov Chain Monte Carlo (MCMC) iterations.

5.2.2 Probability distribution of imputed genotypes

The process of imputation usually begins with "prephasing" of study samples, usually with SHAPEIT for IMPUTEv2. The haplotypes in the study sample are then compared to the more dense haplotypes in the reference panel. IMPUTEv2 generates a posterior distribution of genotypes for each individual at each SNP. This probability distribution can then be used directly in association tests (as implemented by SNPTEST, for example). Alternatively, we can calculate the "expected genotype" under an additive model using the following equation,

$$E(genotype) = p_{Aa} + 2p_{aa} \tag{5.1}$$

where the probabilities p_{Aa} and p_{aa} are defined in Table 5.1.

5.3 IntRapid for imputed signals

I have adapted both stages of the IntRapid software to allow for imputed genotype data.

5.3.1 Stage 1

I implemented a strategy analogous to that used in stage 1 of IntRapid (section 4.2). Consider a pair of SNPs with genotypes denoted AA, Aa and aa at the first, and BB, Bb and bb at the second. After imputation, I recover the distribution of genotypes, denoted $p(AA_i)$, $p(Aa_i)$ and $p(aa_i)$ at the first SNP, and $p(BB_i)$, $p(Bb_i)$ and $p(bb_i)$ at the second SNP, for the i^{th} individual. Across individuals, I can represent this distribution in Table 5.2.

Table 5.2: Contingency table for a pair of imputed SNPs

SNP 2	SNP 1			
	AA	Aa	aa	Total
BB	$\sum_i p(AA_i)p(BB_i)$	$\sum_i p(Aa_i)p(BB_i)$	$\sum_i p(aa_i)p(BB_i)$	$\sum_i p(..BB_i)$
Bb	$\sum_i p(AA_i)p(Bb_i)$	$\sum_i p(Aa_i)p(Bb_i)$	$\sum_i p(aa_i)p(Bb_i)$	$\sum_i p(..Bb_i)$
bb	$\sum_i p(AA_i)p(bb_i)$	$\sum_i p(Aa_i)p(bb_i)$	$\sum_i p(aa_i)p(bb_i)$	$\sum_i p(..bb_i)$
Total	$\sum_i p(AA_i..)$	$\sum_i p(Aa_i..)$	$\sum_i p(aa_i..)$	n

The genotype frequency table, 5.2, may then be transformed into an allelic probability table 5.3, as per the following equation array.

Table 5.3: Contingency table with probability of a pair of alleles

Locus 2	Locus 1	
	A	a
B	w	x
b	y	z

where,

$$
\begin{aligned}
w &= 4\sum_i p(AA_i)p(BB_i) + 2p\sum_i p(Aa_i)p(BB_i) + 2\sum_i p(AA_i)p(Bb_i) + \sum_i p(Aa_i)p(Bb_i),\\
x &= 4\sum_i p(aa_i)p(BB_i) + 2p\sum_i p(Aa_i)p(BB_i) + 2\sum_i p(aa_i)p(Bb_i) + \sum_i p(AA_i)p(Bb_i),\\
y &= 4\sum_i p(AA_i)p(bb_i) + 2p\sum_i p(Aa_i)p(bb_i) + 2\sum_i p(AA_i)p(Bb_i) + \sum_i p(Aa_i)p(Bb_i),\\
z &= 4\sum_i p(aa_i)p(bb_i) + 2p\sum_i p(Aa_i)p(bb_i) + 2\sum_i p(aa_i)p(Bb_i) + \sum_i p(Aa_i)p(Bb_i)
\end{aligned}
$$

The OR for a 2×2 table of this kind is then calculated separately for the cases and controls. The ORs are then computed and tested as described in chapter 4 (section 4.2).

5.3.2 Stage 2

In order to implement the GLM stage of IntRapid (stage 2) on imputed signals I computed the expected genotypes in each location. The expected genotype for the i^{th} individual at the first SNP is given by the following equation,

$$E(G_{Ai}) = p(Aa_i) + 2p(aa_i) \tag{5.2}$$

Once I had the expected genotypes, I could implement stage 2 of the IntRapid algorithm as described in chapter 4 (section 4.2).

5.4 Application to diseases from WTCCC

5.4.1 Imputation of signal neighbourhoods

The interaction signals found after genome wide interaction scans on all seven WTCCC diseases appear in Table 4.7. SNPs appearing in all nominally significant interaction pathways were short listed for imputation and fine-mapping. A 1Mb neighbourhood region was imputed using the European phase 1 (released March 2012) integrated set of the 1000 Genomes Project. In the case of RA and T1D, there appeared to be several significant interaction effects mapping to the HLA. Because of the extensive LD in this region, I considered a 4Mb flanking region for imputation for these two diseases. Table 5.4 gives a list of regions that were imputed. The physical locations appear in Mb (build 37) and the nearest known genes are presented in the last column.

5.4.2 Quality control and imputation

In order to strengthen the signal I decided to impute 1Mb neighbourhoods of existing signals. I used IMPUTEv2 to undertake this imputation [Howie et al., 2009]. I followed 'best practices' as recommended by the authors of IMPUTEv2. This included the following steps:

Pre-imputation filtering: I removed SNPs of low quality, as defined by 'best practices', before running initial interaction scans. As mentioned earlier,

Table 5.4: These SNPs were the top hits in their respective 1Mb neighbourhoods after using IntRapid to conduct genome wide interaction scans for all WTCCC1 diseases. An immediate 1Mb region was imputed around these signals using the 1000 genomes reference panel. The only exceptions were the HLA regions in RA and T1D, where a 4Mb region was imputed.

Disease	ID	Chr	Region(Mb)	Nearest known genes
BD	rs10510819	3	59.19 - 60.19	*LOC100421672, NPCDR1*
BD	rs4299909	7	30.47 - 31.47	*AQP1, GHRHR*
CHD	rs11170276	12	52.68 - 53.68	*KRT76, KRT3, KRT4*
CHD	rs273763	18	22.7 - 23.7	*CTD-2006O16.2*
HT	rs12492072	3	188.38 - 189.38	*TPRG1, TPRG1-AS2*
HT	rs34965037	7	91.62 - 92.62	*PEX1*
HT	rs9521017	13	108.92 - 109.92	*MY016*
HT	rs17292844	20	14.62 - 15.62	*C20orf133*
IBD	rs4693426	4	82.59 - 83.59	*U6, RP11-268F19.1, BIN2P1*
IBD	rs6554056	4	53.22 - 54.22	*RASL11B, SCFD2*
IBD	rs3898061	7	152.24 - 153.24	*ACTR3B, AF104455.1*
IBD	rs10809018	9	9.54 - 10.54	*PTPRD*
IBD	rs16907121	9	22.84 - 23.84	*SUMO2P2, ELAVL2*
IBD	rs10887096	10	123.43 - 124.43	*TACC2*
IBD	rs7318474	13	28.63 - 29.63	*FLT1, EIF4A1P7*
IBD	rs2609850	22	34.35 - 35.35	*RP1-101G11.2*
RA	rs206015	6	30.00 - 34.00	*NOTCH4, SEEK1, HCG27*
RA	rs1419442	7	126.23 - 127.23	*GRM8*
RA	rs10740429	10	53.42 - 54.42	*PRKG1*
T1D	rs1475398	1	65.48 - 66.48	*LEPR*
T1D	rs6683663	1	56.86 - 57.86	*C8A*
T1D	rs11677151	2	12.26 - 13.26	*AC096559.2, TRIB2*
T1D	rs4857717	3	176.83 - 177.83	*SNORA18, RP11-114M1.2*
T1D	rs6818162	4	68.24 - 69.24	*TMPRSS110*
T1D	rs949882	6	109.09 - 110.09	*RP11*
T1D	rs12529458	6	30.00 - 34.00	*LOC346147, ZNF192, GABBR1*
T1D	rs7158575	14	99.72 - 100.72	*EML1, CYP46A1*
T1D	rs6015978	20	37.02 - 38.02	*PPP1R16B*
T2D	rs10916293	1	227.93 - 228.93	*OBSCN*
T2D	rs7827545	8	135.07 - 136.07	*ZFAT1*
T2D	rs9314349	8	26.97 - 27.97	

standard WTCCC1 recommended quality filters were used. A failure to administer this step would have affected imputation quality and possibly given rise to false positive signals.

Match Genomic positions: With updates in the human genome reference sequence, the genomic positions change every few years. Thus it is important to update the genomic positions in the study sample when a new reference panel is used. The different builds of data are usually identified by UNSC references (b36, b37 etc) or UCSC references (denoted by h18, h19 etc). A program called 'liftOver' is available for this purpose. In my study I used PLINK to match the IDs of each SNP in build 36 to get their corresponding positions in build 37.

Strand alignment: It is essential for the strands in the study sample to be aligned with the strands in the reference panel. Typically reference panels are aligned to the '+' strand and in this case I ensured the data was similarly aligned. A failure to implement this step leads to allelic discrepancy in the two datasets. The major and minor allele may also be reversed for some variants. This affects imputation quality and thus should be taken into account when running imputation. I used PLINK to perform strand alignments on the basis of available strand files.

Reference Panel: One of the fundamental principles of genotypic imputation necessitates the use of a reference panel and sample from the same population. For a fixed sample, the choice of reference panel could be made manually or a feature in IMPUTEv2 allows automatic selection of a reference panel population. In my study, I used individuals with European

descent as released (March 2012) in the phase 1 integrated set of the 1000 genomes project [1000 Genomes Project Consortium, 2010].

Post-imputation filters: Once the imputed data are obtained, the quality of imputation needs to be checked. The SNP information is contained in a file with suffix '$imputed_{info}$'. A low quality imputation is associated with a low imputation information score. A 'low quality' SNP could inflate false positive rates. To be conservative, we are better off excluding such SNPs. All loci with an imputation quality score of less than 40% were excluded.

Low frequency alleles and imputation: Variants with MAF less than 5% were excluded from the directly types genotype data. Imputed loci with MAF less than 1% were excluded.

5.4.3 Results

I report pairs of SNPs which have the most significant interaction effect in the neighbourhood of previously reported nominally significant interaction signals (Table 5.5). Of the 19 loci considered here, 17 pairs of SNPs had interaction p-value smaller than that reported from non-imputed SNPs. The stronger signal suggests proximity of the 'causal' variant which would be expected to have stronger effect on disease. It is worth noticing that a Bonferroni level significance level was not achieved for most pairs of SNPs. For RA and T1D, there were multiple significant interactions in the HLA region of Chromosome 6. I have reported the two most significant results from these regions for each disease. Given the high LD in the HLA region, there are several significant results in close proximity to each other. In any case, there appears to be strong evidence to suggest the

existence of several disease relevant interaction effects for RA and T1D in the HLA region.

For all reported interacting SNPs, I have presented the imputation quality score under the 'Info' column of the table. As can be noticed, in most cases the scores are greater than 0.9. Among reported results, the smallest score corresponded to a SNP in the HLA region for the T1D study (0.54). However, several SNPs around this marker had very high quality scores and thus I was not particularly doubtful of this result.

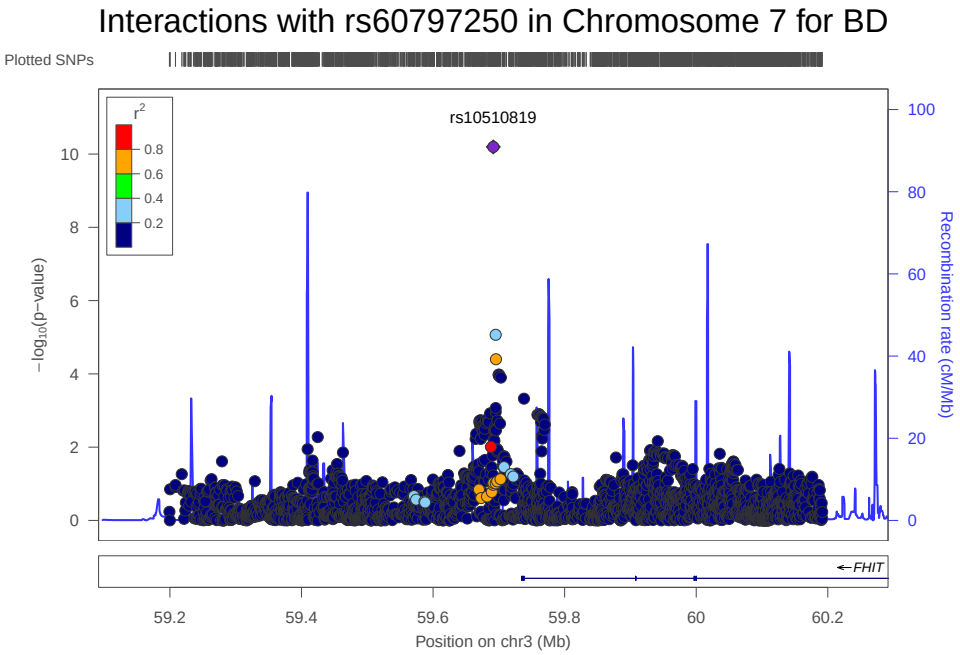
Figure 5.2(a) plots the interaction p-values of all SNPs in a 600Kb neighbourhood of rs10510819, when interaction with rs60797250 is tested. The signal corresponding to rs10510819 is surrounded by very weak signals and its LD with its immediate neighbours are quite low. There might be potential quality issues with this SNP but my investigations did not result in any convincing reason to believe that might be the case. A permutation test for this interaction result confirmed our findings. The joint minor allele count for the two loci were not small enough to raise concerns around biased results.

Figure 5.2(b) plots the interaction p-values of all SNPs in a 600Kb neighbourhood of rs60797250, when interaction with rs10510819 is tested. These plots suggest that I have been able to map causal variants to two narrow recombination intervals, mapping near $FHIT$ and $AQP1$. $FHIT$ is associated with digestive tract cancers [Al-Temaimi et al., 2013] and $AQP1$ is associated with imbalance in ocular fluid movement [Zou et al., 2013].

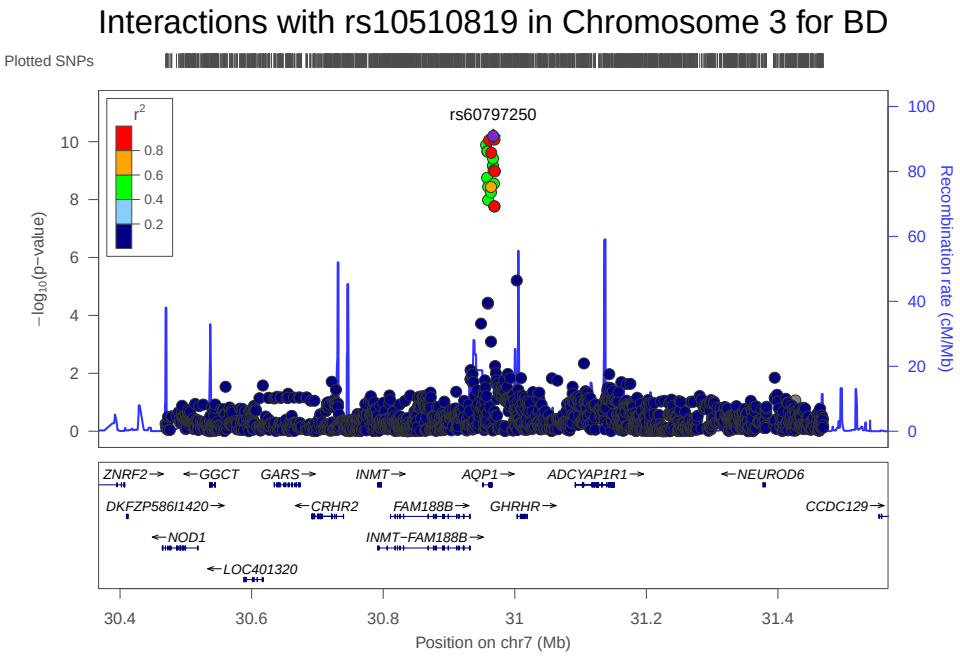
Figure 5.3(a) plots the interaction p-values of all SNPs in a 600Kb neighbourhood of rs273753, when interaction with rs12312528 is tested. Figure 5.3(b) plots the interaction p-values of all SNPs in a 600Kb neighbourhood of rs12312528,

Table 5.5: Gene-gene interaction signals obtained after applying IntRapid to imputed neighbourhoods of SNPs previously identified as part of two loci interactions. The SNPs corresponding to the strongest signals in each neighbourhood are reported and thus not all SNPs that appear in the table are imputed. 'Imp.' stands for 'Imputation'

	SNP 1						SNP 2							Marginal p-value	
Disease	ID	Chr	Position	MAF	Info	Imp	ID	Chr	Position	MAF	Info	Imp	P-value	SNP 1	SNP 2
BD	rs10510819	3	59691572	0.14	1.00	No	rs60797250	7	30967345	0.05	0.84	Yes	6.37×10^{-11}	7.32×10^{-1}	9.20×10^{-3}
CHD	rs273753	18	23190208	0.48	0.99	Yes	rs12312528	12	53184845	0.21	0.79	Yes	3.63×10^{-11}	5.35×10^{-1}	1.99×10^{-1}
HT	rs12492072	3	188884221	0.07	1.00	No	rs7791858	7	92136754	0.23	0.96	Yes	1.55×10^{-12}	8.88×10^{-1}	3.73×10^{-1}
HT	rs17292844	20	15119797	0.31	1.00	No	rs2391692	13	109425758	0.36	0.98	Yes	7.38×10^{-12}	1.19×10^{-1}	7.08×10^{-1}
IBD	rs3898061	7	152737701	0.15	1.00	No	rs2868336	4	83098078	0.11	0.96	Yes	7.23×10^{-11}	5.98×10^{-1}	6.54×10^{-1}
IBD	rs10811889	9	23325709	0.10	0.88	Yes	rs10809018	9	10035658	0.24	1.00	No	1.89×10^{-11}	1.50×10^{-1}	5.31×10^{-1}
IBD	rs7318474	13	29129431	0.36	1.00	No	rs10887096	10	123926310	0.12	1.00	No	1.33×10^{-11}	1.53×10^{-2}	6.99×10^{-1}
IBD	rs2609850	22	34851377	0.28	1.00	No	rs6844680	4	53724815	0.29	0.98	Yes	3.01×10^{-11}	1.07×10^{-1}	5.31×10^{-1}
RA	rs1419442	7	126734171	0.35	1.00	No	rs10733902	10	53917370	0.20	0.93	Yes	2.84×10^{-11}	3.43×10^{-1}	7.88×10^{-2}
RA	rs9274778	6	32638694	0.42	0.84	Yes	rs9275212	6	32658327	0.35	0.98	Yes	2.95×10^{-48}	3.00×10^{-4}	1.17×10^{-52}
RA	rs9469220	6	32658310	0.50	1.00	No	rs28891371	6	32638713	0.14	0.68	Yes	2.95×10^{-48}	1.00×10^{-4}	3.20×10^{-3}
T1D	rs6802551	3	177322821	0.33	0.96	Yes	rs7158575	14	100215383	0.14	1.00	No	2.57×10^{-11}	7.19×10^{-1}	2.05×10^{-1}
T1D	rs11677151	2	12759532	0.08	1.00	No	rs949882	6	109586062	0.35	1.00	No	8.85×10^{-11}	9.86×10^{-1}	2.08×10^{-1}
T1D	rs1171280	1	65989677	0.27	0.99	Yes	rs6015978	20	37516318	0.38	1.00	No	3.61×10^{-11}	5.12×10^{-1}	2.21×10^{-1}
T1D	rs6683663	1	57356476	0.14	1.00	No	rs2033930	4	68736979	0.39	0.92	Yes	3.48×10^{-11}	5.46×10^{-1}	9.36×10^{-1}
T1D	rs2076534	6	32362416	0.17	0.92	Yes	rs9273206	6	32613573	0.24	0.54	Yes	7.63×10^{-129}	4.04×10^{-34}	2.01×10^{-2}
T1D	rs1071630	6	32609126	0.32	0.90	Yes	rs9275524	6	32675109	0.42	0.99	Yes	4.27×10^{-115}	1.70×10^{-157}	7.22×10^{-2}
T2D	rs1771456	1	228430697	0.33	0.98	Yes	rs9314349	8	27474202	0.38	1.00	No	1.61×10^{-11}	5.68×10^{-1}	8.99×10^{-1}
T2D	8-135566575	8	135566575	0.05	0.96	Yes	rs7827545	8	135566567	0.32	1.00	No	8.21×10^{-13}	5.05×10^{-1}	2.30×10^{-3}

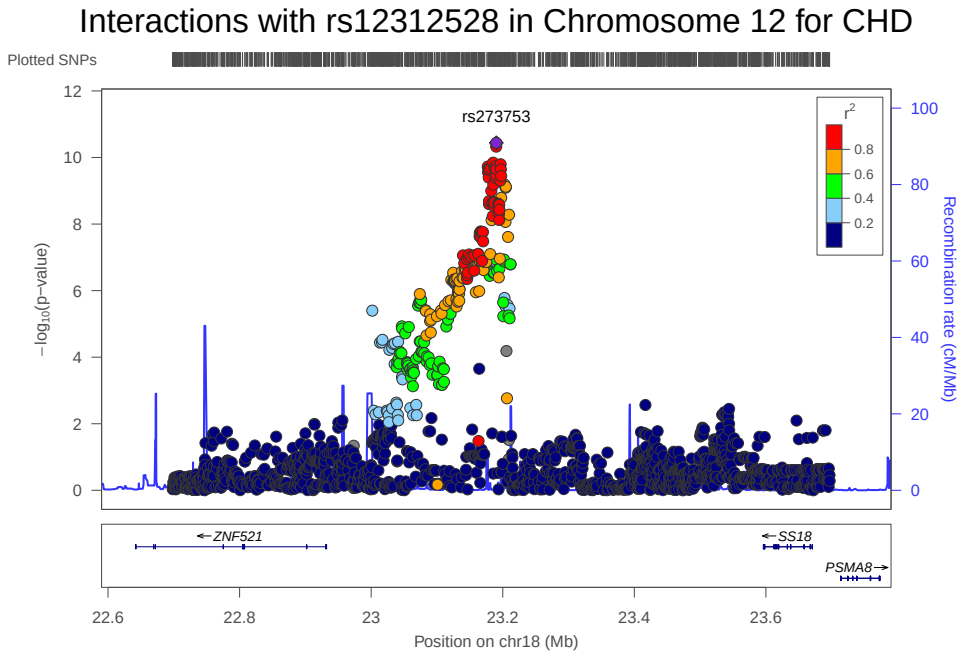


(a) Plot of interaction p-values in a 600Kb neighbourhood of rs10510819

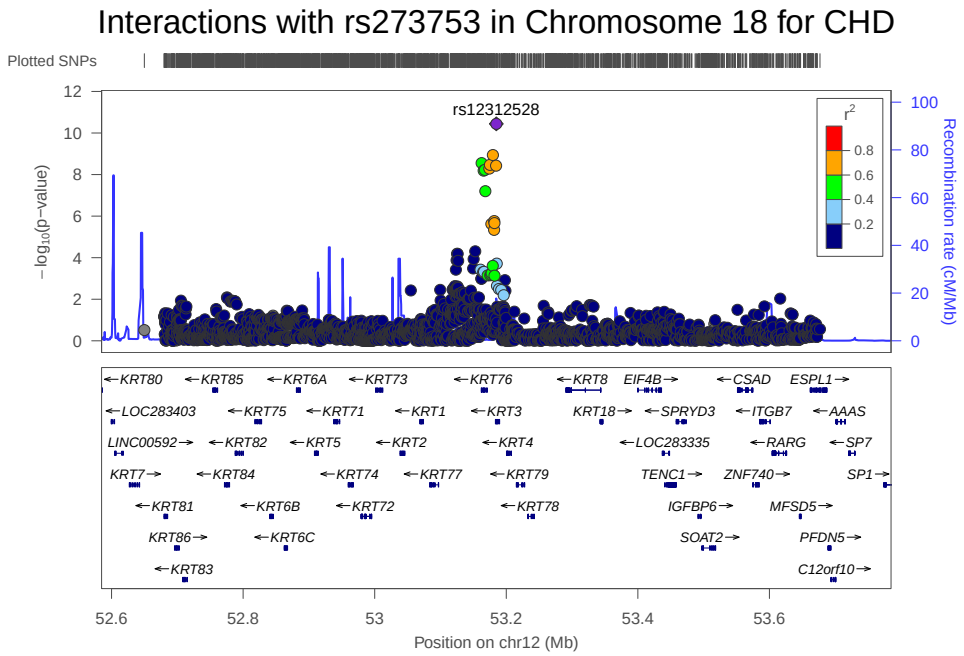


(b) Plot of interaction p-values in a 600Kb neighbourhood of rs60797250

Figure 5.2: Signal plots for interaction signals found in BD, using LocusZoom



(a) Plot of interaction p-values in a 600Kb neighbourhood of rs273753



(b) Plot of interaction p-values in a 600Kb neighbourhood of rs12312528

Figure 5.3: Signal plots for interaction signals found in CHD, using LocusZoom

when interaction with rs273753 is tested. These plots suggest that I have been able to map causal variants to two narrow recombination intervals, mapping near $ZNF521$ and $KRT76$ or $KRT3$. I was not able to find any functions of these genes which could be of direct relevance to CHD.

5.4.4 Permutation tests

Permutation tests are a form of statistical significance tests which applies re-sampling to create an empirical distribution of the test statistic [Fisher, 1951]. The location of the original p-value in this distribution is used to calculate an exact p-value. While the test is computationally demanding, it is often considered to be more reliable when the parametric distribution of the sample statistic is unknown.

The risk of finding false positive signals is particularly an issue with genetic interaction studies [Coffey et al., 2004]. In order to confirm the findings, I conducted permutation tests (with 1000 replicates) on all interaction models with a stage 2 p-value (in IntRapid) of less than 10^{-8}. I randomly sampled from the case-control status to create a new response to which a saturated GLM model with two SNPs was fitted. This process was repeated 1000 times and the proportion of p-values smaller than the original p-value was noted as the permutation p-value. In all cases I was able to find a permutation p-value of less than 0.05.

Ideally, I would have liked to conduct this test across the genome to generate an approximate genome wide threshold. However, this would have lead to scanning around 6.44×10^{13} interaction models. Even with the computational advantages offered by IntRapid this would have been a formidable computational

challenge.

5.5 Discussion

The 500k Affymetrix chip used for the WTCCC studies has limited coverage. Thus it may not have been able to include the 'causal' SNP involved in interactions. It is thus likely that any identified SNPs in interaction signals are in LD with the 'causal' variant. I could increase the chance of including these 'causal' variants by running interaction scans on imputed data. However, implementing imputation on a genome wide scale carries a very significant computational cost. To reduce computational burden, I have imputed regions of interacting SNPs identified in the previous Chapter (4), and have implemented an interaction scan on these imputed signals.

The primary motivation in looking for interaction signals in the imputed neighbourhoods of already identified signals was for fine-mapping. The imputation was carried out over a 1Mb neighbourhood (on each side) of the identified SNPs, except for the HLA region in Chromosome 6 where a 4Mb region (on each side) was imputed because the high density of signals found there and extended LD. Imputed data had to be dealt with differently in the IntRapid framework. Rather than analysing direct genotypes, I now had a probability distribution of the genotypes. I thus calculated the expected genotypes to populate association cross-tables, under an assumption of independence between a pair of loci. This process was computationally more demanding, but given the small number of scans, the overall computational burden was not large. The strongest interaction signal from each pair of imputed regions was reported.

When I compared the interaction signals obtained from imputed data with the ones initially identified, from the directly typed genotypes alone, there was a noticeable improvement in the strength of the signals in most loci. Of the 38 regions considered, an imputed SNP constituted the strongest local signal at 21. All 19 pairs of interaction signals included imputed SNPs leading to a more significant interaction p-value. Some of the interaction signals now satisfied even the conservative Bonferroni corrected threshold for genome-wide significance.

In the non-HLA interaction signals, the imputed SNPs which were part of a now stronger signal moved by up to a few Kb around the original non-imputed signal. In CHD, one of the interacting SNPs, rs11170276 on chromosome 12 was within the molecular regions of gene KRT1 and KRT77. However, the stronger imputed signal rs12312528 was not part of genes KRT76 and KRT3. In the case of HT, the stronger imputed signal for rs34965037 on chromosome 7 (position 92117993) was 18.8Kb downstream at rs7791858. Both locations map to the PEX1 gene. The stronger imputed signals in the HLA region had, in many cases, moved by several hundreds of Kb.

5.6 Chapter Summary

In the present chapter, I imputed neighbourhoods of these signals and conducted interaction modelling between these imputed SNPs. This was undertaken with the objective of finding stronger signals in the vicinity of existing interaction signals (as identified in chapter 4). I identified stronger signals of interaction for all seven diseases considered.

While these efforts help identify strong interaction signals on the autosomes,

the sex chromosomes have so far been ignored. The distribution of genotypes in males and females on the X chromosome is different. This is taken into account in the next chapter (6) where I have explored a fixed effect meta-analysis model to study interaction between and with SNPs on the X-Chromosome.

Chapter 6

X Chromosome interaction

6.1 Introduction

There has been limited research on association methods involving the X Chromosome [Clayton, 2009]. Possible reasons for this lack of attention include the statistical challenges in performing single locus association analysis in the X-chromosome. The X-chromosome is often overlooked in many GWAS, even when many common diseases show significant differences in disease occurrence between males and females [Ober et al., 2008]. This lack of attention could potentially mean a loss of several interesting signals given the X-chromosome accounts for around 5% and 2.5% of the female and male genomes respectively. Clayton [2009] discusses some issues with respect to efficient analysis of samples with both genders, in particular accounting for X chromosome inactivation when half of all cells in females have one copy of the X-Chromosome active and the other half of cells have the second copy active [Chow et al., 2005].

Under this model, males are coded as homozygous for the allele they carry.

Further, as described in Clayton [2008], a one degree of freedom statistic has been shown to be by far the most powerful test for single locus associations under various theoretical considerations [Hickey and Bahlo, 2011].

Instances of association with variants on the X-chromosome have been reported for some diseases. Chu et al. [2013] reports genes in the sex chromosomes which are significantly associated with Grave's disease; a disease which has long been considered to have gender specific effects. These signals were replicated in independent studies and further confirmed by re-sequencing the relevant sections of the X-chromosome. Loley et al. [2011] further reports two genetic regions in the X-chromosome which are significantly associated with Crohn's disease.

6.2 Methods

Association testing of variants on the X-chromosome must take into account the difference in the number of copies carried by males and females [Hickey and Bahlo, 2011]. A standard automsomal association test like the chi-squared test of association or a Cochran-Armitage trend test [Armitage, 1955, Cochran, 1954] does not appropriately account for this difference. Around 1.9% of the X-chromosome is covered by two small regions called the *pseudo-autosomal* regions (PAR), where the loci are inherited like autosomal loci. The X-chromosome specific methods should not include these regions.

Given the differences in the number of allelic copies between males and females, males have two distinct genotypes 'A-' and 'a-' and females have three distinct genotypes 'AA', 'Aa' and 'aa'. Thus, a traditional association test in males would be similar to an allelic test and that in females would be subject to

specification of a genetic model. However, an allelic test assumes HWE, which may not be valid [Clayton, 2009]. These issues could be addressed by application of a traditional GLM [Prentice and Pyke, 1979]. In the case of a significant difference in allele frequency between males and females, gender needs to be allowed for as a covariate. This would essentially stratify the dataset and thereby reduce statistical power considerably. In particular, in the absence of significant frequency differences such stratification and hence loss of power is undesirable.

The X-inactivation phenomenon [Chow et al., 2005] allows us to assume that males are equivalent to homozygous females. Thus, males are coded as 0 and 2, and females are coded as 0, 1 and 2. However, genotype distributions are different between genders and, in general, differences in sex ratios in cases and controls could act as a confounding factor. We propose an approach to perform the analysis by testing for association in males and females separately, and subsequently meta-analyse the effect estimates. This approach allows for differences in the variance of allelic effect estimates, but assumes the allelic effect to be the same in males and females. A similar testing strategy will be required when testing interactions between autosomal SNPs and those in the X-chromosome. In this chapter, I extend "IntRapid" to allow for interaction with the X-chromosome. We will further apply X-chromosome based interaction scans to all diseases included in the WTCCC scan.

6.3 Stage 1 of IntRapid for X - Chromosomes

The first stage of the analysis is somewhat analogous to the 'IntRapid' methods described in Chapter 4 for autosomes. However, for the X-chromosome, I perform

analyses separately in males and females.

6.3.1 Within X-Chromosome : Females

For a sample of n_sf female individuals, disease status 's' (Affected or Unaffected) and a given pair of SNPs, let us assume the first marker has three genotypes (AA, Aa and aa) and the second SNP has three genotypes (BB, Bb & bb). If n_{sfkl} were the number of individuals with k genotype at the first SNP and l genotype at the second SNP, the sample can be represented in a contingency table (Table 6.1).

Table 6.1: Contingency table for a pair of SNPs in the X-Chromosome of females.

SNP 1	SNP 2			
	BB	Bb	bb	Total
AA	n_{sfAABB}	n_{sfAABb}	n_{sfAAbb}	$n_{sfAA..}$
Aa	n_{sfAaBB}	n_{sfAaBb}	n_{sfAabb}	$n_{sfAa..}$
aa	n_{sfaaBB}	n_{sfaaBb}	n_{sfaabb}	$n_{sfaa..}$
Total	$n_{sf..BB}$	$n_{sf..Bb}$	$n_{sf..bb}$	n_sf

The genotype frequency table, 6.1, may then be transformed into an allelic frequency table 6.2.

Table 6.2: Allelic contingency table for a pair of SNPs in the X-Chromosome for females.

SNP 1	SNP 2		
	B	b	Total
A	n_{sfAB}	n_{sfAb}	$n_{sfA.}$
a	n_{sfaB}	n_{sfab}	$n_{sfa.}$
Total	$n_{sf.B}$	$n_{sf.b}$	n_{sf}

where,

$$
\begin{aligned}
n_{sfAB} &= 4n_{sfAABB} + 2n_{sfAaBB} + 2n_{sfAABb} + n_{sfAaBb}, \\
n_{sfaB} &= 4n_{sfaaBB} + 2n_{sfAaBB} + 2n_{sfaaBb} + n_{sfAABb}, \\
n_{sfAb} &= 4n_{sfAAbb} + 2n_{sfAabb} + 2n_{sfAABb} + n_{sfAaBb}, \\
n_{sfBB} &= 4n_{sfaabb} + 2n_{sfAabb} + 2n_{sfaaBb} + n_{sfAaBb}
\end{aligned}
$$

6.3.2 Within X-Chromosome : Males

As mentioned earlier, the alleles in the X-Chromosome in males could be assumed to act like homozygous genotypes in females because of inactivation.

If n_{smkl} were the number of male individuals with k genotype at the first SNP and l genotype at the second SNP, the sample can be represented in a contingency table (Table 6.3).

The genotype frequency table, 6.3, can be converted into a 2×2 allelic table 6.4.

Table 6.3: Contingency table for a pair of SNPs in the X Chromosome of males.

SNP 1	SNP 2		
	BB	bb	Total
AA	n_{smAABB}	n_{smaaBB}	$n_{smAA..}$
aa	n_{smAAbb}	n_{smaabb}	$n_{smaa..}$
Total	$n_{sm..BB}$	$n_{sm..bb}$	n_{sm}

Table 6.4: Allelic contingency table for a pair of SNPs in males.

SNP 1	SNP 2		
	B	b	Total
A	n_{smAB}	n_{smAb}	$n_{smA.}$
a	n_{smaB}	n_{smab}	$n_{sma.}$
Total	$n_{sm.B}$	$n_{sm.b}$	n_{sm}

where,

$$\begin{aligned} n_{smAB} &= 2n_{smAABB}, \\ n_{smaB} &= 2n_{smaaBB}, \\ n_{smAb} &= 2n_{smAAbb}, \\ n_{smab} &= 2n_{smaabb} \end{aligned}$$

6.3.3 Autosomal SNP and X-Chromosome : Females

Testing for interaction effects in stage 1 of IntRapid between autosomal SNPs and X-Chromosome SNPs in females is exactly the same as for an autosomal and autosomal SNP, as described in Chapter 4 (Section 4.2.1).

6.3.4 Autosomal SNP and X-Chromosome : Males

Testing for interaction effects in stage 1 of IntRapid between autosomal SNPs and X-Chromosome SNPs in males is quite different from the within X-Chromosome testing as described in section 6.3.1 or 6.3.2.

For a sample of $n_s m$ male individuals, disease status 's' (Affected or Unaffected) and a given pair of SNPs, let us assume the first marker has three genotypes (AA, Aa and aa) and the second SNP has three genotypes (BB, Bb & bb). If n_{smkl} were the number of individuals with k genotype at the first SNP and l genotype at the second SNP, the sample can be represented in a contingency table (Table 6.5).

Table 6.5: Contingency table for an autosomal SNP and an X-Chromosome SNP in males.

Autosomal SNP	X-Chromosome SNP		
	BB	bb	Total
AA	n_{smAABB}	n_{smAAbb}	$n_{smAA..}$
Aa	n_{smAaBB}	n_{smAabb}	$n_{smAa..}$
aa	n_{smaaBB}	n_{smaabb}	$n_{smaa..}$
Total	$n_{sm..BB}$	$n_{sm..bb}$	$n_s m$

The genotype frequency table, 6.5, may then be transformed into an allelic frequency table 6.6.

Table 6.6: Allelic contingency table for an autosomal SNP and an X-Chromosome SNP for males.

Autosomal SNP	X-Chromosome SNP		
	B	b	Total
A	n_{smAB}	n_{smAb}	$n_{smA.}$
a	n_{smaB}	n_{smab}	$n_{sma.}$
Total	$n_{sm.B}$	$n_{sm.b}$	n_{sm}

where,

$$\begin{aligned} n_{smAB} &= 4n_{smAABB} + 2n_{smAaBB}, \\ n_{smaB} &= 4n_{smaaBB} + 2n_{smAaBB}, \\ n_{smAb} &= 4n_{smAAbb} + 2n_{smAabb}, \\ n_{smBB} &= 4n_{smaabb} + 2n_{smAabb} \end{aligned}$$

6.3.5 Odds ratios

Let n_{sjkl} be defined as the number of individuals with disease status s (Affected (D) or Unaffected (U)), gender j (Males (m) or Females (f), and alleles k and l at the pair of SNPs ('a/b' for minor allele and 'A/B' for major allele).

From the definition of Odds Ratio (OR) in Chapter 4, OR (ψ_{ij}) for disease status s and gender j can be defined as,

$$\psi_{sj} = \frac{n_{sjAB}.n_{sjab}}{n_{sjaB}.n_{sjAb}} \tag{6.1}$$

or in this case,

$$\begin{aligned}
\psi_{Dm} &= \frac{n_{DmAB}.n_{Dmab}}{n_{DmAb}.n_{DmaB}}, \\
\psi_{Um} &= \frac{n_{UmAB}.n_{Umab}}{n_{UmAb}.n_{UmaB}}, \\
\psi_{Df} &= \frac{n_{DfAB}.n_{Dfab}}{n_{DfAb}.n_{DfaB}}, \\
\psi_{Uf} &= \frac{n_{UfAB}.n_{Ufab}}{n_{UfAb}.n_{UfaB}},
\end{aligned}$$

The corresponding variance of the logarithm of OR is,

$$Var(log\psi_{sj}) = \left(\frac{1}{n_{sjAB}} + \frac{1}{n_{sjAb}} + \frac{1}{n_{sjaB}} + \frac{1}{n_{sjab}}\right) \tag{6.2}$$

6.3.6 Meta analysis

The above estimates could be used to meta analyse sex specific effects for a given binary disease trait. For individuals with disease status i and gender j a weighting variable ω_{ij} is defined as the inverse of the corresponding variance of log OR.

$$ln(\psi_i) = \frac{\sum_{j\in\{m,f\}} ln(\psi_{ij}).\omega_{ij}}{\sum_{j\in\{m,f\}} \omega_{ij}}$$

$$\omega_{ij} = [Var(ln(\psi_{ij}))]^{-1}$$

Thus, for the cases I calculate two weighing factors ω_{Dm} and ω_{Df} corresponding to the males and females respectively. Similarly, for controls I have ω_{Um} and ω_{Uf} respectively.

$$\omega_{Dm} = [Var(ln(\psi_{Dm}))]^{-1}$$
$$\omega_{Df} = [Var(ln(\psi_{Df}))]^{-1}$$
$$\omega_{Um} = [Var(ln(\psi_{Um}))]^{-1}$$
$$\omega_{Uf} = [Var(ln(\psi_{Uf}))]^{-1}$$

We then calculate weighted logarithm ORs, $ln(\psi_i)$ for the affected and unaffected individuals.

$$ln(\psi_D) = \frac{ln(\psi_{Dm}).\omega_{Dm} + ln(\psi_{Df}).\omega_{Df}}{\omega_{Dm} + \omega_{Df}}$$
$$ln(\psi_U) = \frac{ln(\psi_{Um}).\omega_{Um} + ln(\psi_{Uf}).\omega_{Uf}}{\omega_{Um} + \omega_{Uf}}$$

The next step is to calculate the variances for the logarithm of these ORs. This is done by taking the inverse of the sum of the previously calculated weighing factors (ω_{ij}).

$$Var(ln(\psi_i)) = \left(\sum_{j \in \{m,f\}} (\omega_{ij}) \right)^{-1} \tag{6.3}$$

Thus for the cases and controls the variances are as follows,

$$\begin{aligned} Var(ln(\psi_D)) &= [\omega_{Dm} + \omega_{Df}]^{-1} \\ Var(ln(\psi_U)) &= [\omega_{Um} + \omega_{Uf}]^{-1} \end{aligned}$$

We test the difference between the meta-analysed logarithms of the two odds ratios using the following statistic. This is analogous to the statistic used in Chapter 4. It follows a χ^2 distribution with 1 degree of freedom.

$$\chi_{cal} = \frac{(ln(\psi_D) - ln(\psi_U))^2}{Var(ln(\psi_D)) + Var(ln(\psi_U))} \tag{6.4}$$

6.4 Stage 2 - Generalised linear model

A GLM is used to model the interaction between two SNPs in males and females separately. From each GLM model the additive-additive interaction effects (regression coefficient corresponding to the interaction term - β_{sj12}) are obtained and combined via a fixed-effects meta-analysis (details in section 6.5).

6.5 Meta-analysis of fixed effects

Under a fixed effects model (FEM) we assume the phenotypic distribution of males and females is similar. At least the first two moments of the distributions need to be similar for the model to hold. However, for a number of diseases this is hardly the case and thus admittedly this approach is an over simplification at this level.

Let disease status be represented by i (Affected or Unaffected) and gender be represented by j (Males or Females). Then we could define,

$$\omega_{ij12} = (Var(\hat{\beta}_{ij12}))^{-1} \tag{6.5}$$

We may then define a pooled regression coefficient, $\beta_{.12}$, for males and females corresponding to the interaction term, $G_{i1}G_{i2}$,

$$\beta_{i.12} = \frac{\sum_{j\in\{m,f\}} \beta_{ij12}.\omega_{ij12}}{\sum_{j\in\{m,f\}} \omega_{ij12}} \tag{6.6}$$

and the corresponding variance is,

$$Var(\beta_{i.12}) = \left(\sum_{j\in\{m,f\}} \omega_{ij12} \right)^{-1} \tag{6.7}$$

We may then test the meta-analysed estimate of regression coefficient of interaction using the following statistic,

$$\chi^2_{\beta_{i.12}} = \frac{\beta^2_{i.12}}{\sum_{j\in\{m,f\}} Var(\beta_{ij12})} \tag{6.8}$$

6.6 Implementation in IntRapid software

The above described method was implemented in IntRapid. The software would detect a X-chromosome SNP and automatically apply the appropriate method for interaction testing. When one set of SNPs come from autosomal chromosomes and the other from the X-chromosome the software still works seamlessly. Testing for interactions between the X-Chromosome and all other chromosomes was computationally efficient and took around 2 hours in one computing core for each disease. The interactions within the X-chromosome were significantly faster and were completed within approximately 10 minutes per disease.

6.7 Application to WTCCC

We tested for interaction for all seven WTCCC diseases: (i) both SNPs on the X-chromosome; and (ii) one autosomal SNP and one X-chromosome SNP. As for the autosomes, I used a stage 1 threshold of 10^{-4} for carrying forward SNPs to stage 2. After ruling out interaction signals which are likely to be false positives, no interaction effects were identified with the X chromosome for any disease in stage 2, that met the nominal significance threshold of 10^{-10}.

The strongest interaction signal observed for BD included rs11649589 in chromosome 16 and rs5942091 on the X-chromosome. The interaction p-value was 3.42×10^{-10} which is just larger of the nominal threshold of 10^{-10} used in this the-

sis so far. A total of 14 pairs of SNPs where identified for BD where the interaction p-value was less than 10^{-8}. The strongest interaction signal for CHD included rs636512 on chromosome 5 and rs2269815 on chromosome X. The strongest interaction signal found for HT included rs2877260 on chromosome 7 and rs5940207 on the X-chromosome. In IBD, rs16837064 on chromosome 4 and rs411319 on the X-chromosome displayed the lowest interaction p-value. The strongest interaction signal for RA included rs10519906 on chromosome 5 and rs5950249 on the X-chromosome. In T1D the strongest interaction signal was found for rs6569826 on chromosome 6 and rs2143488 on the X-chromosome. Finally, in T2D the strongest interaction was found between rs1452220 on chromosome 3 and rs6633576 on the X-chromosome. None of the pairs of SNPs in this section had an interaction p-value lower than 10^{-10}.

6.8 Chapter Summary

Limited research has been performed on association with the X chromosome. Interaction studies in the X-chromosome are rarer still. A confounding effect of differential gender ratio and overestimation of the risk in females owing to X-inactivation have proved formidable statistical challenges so far. Some solutions do exist and this chapter builds on those to provide a comprehensive method to scan for interactions in the sex chromosomes.

I was unable to find any statistically significant results at a nominal threshold of 10^{-10} in the interaction analysis with the X chromosome. However, given the datasets used are under powered for such interaction scans the lack of 'significant' results is not a surprise. The software demonstrates the ability of IntRapid to

implement an effective X-Chromosome wide interaction scan and as such, further studies should be done on larger sample sizes and meta-analysis settings.

Chapter 7

Discussion

GWAS have come a long way in helping us understand the genetic aetiology of complex human diseases. Linkage based association studies have proven very successful in identifying the genetic basis of Mendelian disorders or single gene disorders. However, the common complex human disorder and traits have posed a significantly more difficult challenge. More than 1200 loci have been identified as holding significant association with various complex diseases [Zuk et al., 2012]. However, a large proportion of the 'heritability' remains unexplained [Manolio et al., 2009]. While there is some debate about the nature in which 'heritability' is computed, there is clear agreement on the need to conduct interaction scans on a genome wide level [Carlborg and Haley, 2004, Cordell, 2002] to investigate the role of epistasis in complex human traits.

Epistasis is defined as a phenomenon when one gene silences or enhances the action of another. A statistical interaction between two genes may not always capture genetic epistasis. However, the identification of statistically significant interaction effects should hold clues towards uncovering epistatic pathways. Given that limited work has been undertaken on comparison of methods to scan for epistasis, there is limited agreement on best practices [Carlborg and Haley, 2004].

However, there is evidence to suggest an exhaustive genome wide study is the most powerful method to detect gene-gene interactions [Evans et al., 2006]. For main effect studies, an assumption of additive effects of alleles results in a one degree of freedom test which is typically more powerful than a two degree of freedom general genotypic test. Hence, I carried out simulation studies to compare the statistical power of an additive-additive interaction test against a general interaction test. The results highlighted that unless the model of epistasis incorporates extreme deviation from additivity, the additive-additive interaction test is most powerful across a wide range of theoretical gene-gene interaction models. To assess the effects of LD on the relative performance of the tests, I selected extreme models of epistasis where the general interaction test is most powerful and repeated my simulations assuming that nearby markers are typed, as opposed to the causal variants themselves. Even for the most extreme models, the general interaction test became less powerful than the additive-additive interaction test as the extent of LD decreases. It is not clear in general which models of epistasis are most likely to exist in reality. However, extreme dominance epistasis with large effects is unlikely to exist as it would have already been detected otherwise. Therefore, for a genome wide scan of two locus interaction effects, an additive-additive interaction test is likely to give more power over a range of realistic models than a general interaction test.

A formidable challenge in implementing genome wide exhaustive scans is that of a computational nature. Traditional linear models would require several thousands of computing hours to completely scan the genome with up to 1 million SNPs. This can quickly become very expensive and not necessary feasible for research groups without sufficient parallel computing facilities. To enable such

computations with reasonably basic computing facilities, I developed an algorithm to search for interaction models using a two stage approach. This was inspired by the PLINK "fast-epistasis" algorithm which I extended to quantitative traits. My contribution includes formalising the method by conducting simulations to highlight the appropriate threshold in stage 1 for carrying forward SNPs to more detailed examination in stage 2. The method has been implemented in optimised efficient low level code and enabled user-friendly parallel computing. I applied the software, "IntRapid", to seven WTCCC diseases and obesity as measured by BMI. After Bonferroni correction, I was unable to find any genome wide, significant interaction effects. However, given such corrections assume independence between tests (which is not true in this case), it results in a very conservative p-value threshold. Thus I report all pairs of SNPs which are able to satisfy a nominal threshold of 10^{-10}. I was able to identify several interesting interaction signals at this threshold. The most interesting part of these results was the lack of main effects for any of the SNPs implicated. Thus, I would not have been able to uncover these loci in traditional single variant GWAS analysis. Further, I have also found interactions between SNPs already known to the scientific community. This warrants further follow up studies and may well prove very useful for downstream functional analyses.

In order to extend the coverage of the genome wide interaction scans, I would have ideally have liked to repeat the analysis for genome-wide imputed data for each disease. However, this would have posed a very substantial computing challenge given the quadratic growth in the number of interaction models which need fitting. Instead, I imputed 1Mb neighbourhoods of the signals identified with directly typed data, except for the HLA region in chromosome 6 where a

4Mb neighbourhood was imputed. IMPUTEv2 was used for the process and all recommended "good practices" [Howie et al., 2009] were followed. This reduced the overall computing burden substantially. Obviously, the IntRapid algorithm needed to be modified to handle imputed data, which was further implemented within IntRapid. I was able to identify stronger interaction signals between the imputed signals. This suggested the identification of a signal closer to the "causal" variant which may not have been present in the Affymetrix chip used by the WTCCC.

The X-chromosome forms around 5% and 2.5% of the entire genome in females and males respectively, and many significant associations with a range of complex traits have been reported in the X-Chromosome. However, owing to a single copy of the X-chromosome in males, as opposed to two in females, there are statistical concerns about combining data from both sexes. The X-inactivation phenomenon allows us to assume the second copy of the X-chromosome in females to be "inactive". This allows us to simply double the allele count in males to implement an interaction on the sex chromosome. In order to test for interaction on the X chromosome, I estimated the interaction effects for males and females separately, and meta-analyse them using a fixed effects model. I carried out an interaction scan on two sets of pairs of SNPs: one set within the X-chromosome and one set between the autosomal SNPs and the X-Chromosome. After adjusting for multiple testing, I was unable to find any pair of SNPs which satisfied a nominal significance level of 10^{-10}.

7.1 Limitations and possible solutions

The datasets used in these studies were fundamentally designed for single SNP studies and are not well powered for interaction studies. Further, given the large number of statistical tests in exhaustive interaction scans, there is a huge burden of multiple testing correction. Bonferroni correction has been used in this thesis to correct for multiple testing, because this implicitly assumes independence between tests. Such an assumption is unlikely to be true and the correction is thus likely to be extremely conservative. All these factors make the present study design very under powered. One possible solution to this problem involves using larger sample sizes. Given the availability of multiple datasets for the same disease, a meta-analysis based approach should also be explored.

This thesis has primarily sought to develop a computationally efficient way of exhaustively scanning interaction signals on a genome wide scale. For the WTCCC datasets, the discussed methods are scalable and could be implemented quite easily. However, with larger sample sizes and, in particular a larger set of SNPs (using imputation up to the 1000 Genome Project reference panel for example), the number of tests can be exponentially larger. Furthermore, multi-locus interactions will introduce added complexity and computational burden. These challenges do not have an easy solution. Some machine learning methods (as introduced in Chapter 2) have shown promise, but it is not clear how powerful these methods are on a genome wide scale and how computationally efficient some of them are for multi-way interaction scans. These areas naturally need scientific research and should benefit from novel method developments.

The WTCCC datasets used in this thesis are limited to around half a mil-

lion SNPs. The coverage of these datasets can be improved by imputing SNPs up to reference panels from the 1000 Genomes Project [1000 Genomes Project Consortium, 2010, 2012] on a genome wide level, generating more than 30 million variants. However, interaction scans on such a large number of two-locus or multi-locus models would pose a large computational burden. Hence, I only focused on the neighbourhood of already identified SNPs which participates in significant gene-gene interactions. In some way, there is a compromise here to enable quick computation. A slightly more thorough approach would be to have a low threshold for implicated SNPs and impute their neighbourhoods for interaction scans. Ideally interaction scans on the entire dataset of imputed SNPs and hence better coverage would be preferred. Modern GWAS use more than a million markers and with the availability of next generation sequencing data, coverage is likely to improve in interaction scans of the future.

My analysis so far has only included common variants or SNPs with a MAF greater than 5%. As the coverage of the dataset improves, rare variants should be included in future studies. The procedures developed so far for common variants are likely to fail in such cases. One approach could involve 'collapsing' all coding variants in a gene (or some functional unit) to implement a gene-gene interaction rather than a SNP-SNP interaction scan as I have been doing so far for example, as developed by Morris and Zeggini [2010].

7.2 Conclusion

This thesis has developed novel methods to address the computational challenges of conducting exhaustive interaction scans on a genome wide level. The method

has been extended to work on imputed data and interaction within/with the X-chromosome. Software developed to implement these novel methods are freely available for use and should make genome wide interaction studies a straight forward process. Application of the software on seven WTCCC diseases and obesity (as measured by BMI) has identified many interesting interaction effects. A replication study for these signals is desirable. Further, implementation of these methods on multiple independent datasets and then meta-analysing the effects should help improve statistical power and enable the implication of smaller interaction effects.

References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010. doi: 10.1038/nature09534. 1, 11, 104, 113, 142

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012. doi: 10.1038/nature11632. 1, 11, 26, 104, 142

Alan Agresti. *An introduction to categorical data analysis*, volume 423. Wiley-Interscience, 2007. 21

Rabeah Abbas Al-Temaimi, Sindhu Jacob, Waleed Al-Ali, Diana Ann Thomas, and Fahd Al-Mulla. Reduced fhit expression is associated with mismatch repair deficient and high cpg island methylator phenotype colorectal cancer. *J Histochem Cytochem*, 61(9):627–638, Sep 2013. doi: 10.1369/0022155413497367. URL http://dx.doi.org/10.1369/0022155413497367. 114

Anders Albrechtsen, Sofie Castella, Gitte Andersen, Torben Hansen, Oluf Pedersen, and Rasmus Nielsen. A Bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176(2):1197–1208, June 2007. doi: 10.1534/genetics.107.071696. 2, 38

Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955. URL http://www.jstor.org/stable/10.2307/3001775. 123

JSF Barker. Inter-locus interactions: a review of experimental evidence. *Theoretical population biology*, 16(3):323–346, 1979. URL http://www.sciencedirect.com/science/article/pii/0040580979900212. 38

Jeffrey C Barrett, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M. Michael Barmada, Alain Bitton, Themistocles Dassopoulos, Lisa Wu Datta, Todd Green, Anne M Griffiths, Emily O Kistner, Michael T Murtha, Miguel D Regueiro, Jerome I Rotter, L. Philip Schumm, A. Hillary Steinhart, Stephan R Targan, Ramnik J Xavier, N. I. D. D. K. IBD Genetics Consortium, Ccile Libioulle, Cynthia Sandor, Mark Lathrop, Jacques Belaiche, Olivier Dewit, Ivo Gut, Simon Heath, Debby Laukens, Myriam Mni, Paul Rutgeerts, Andr Van Gossum, Diana Zelenika, Denis Franchimont, Jean-Pierre Hugot, Martine de Vos, Severine Vermeire, Edouard Louis, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Lon R Cardon, Carl A Anderson, Hazel Drummond, Elaine Nimmo, Tariq Ahmad, Natalie J Prescott, Clive M Onnie, Sheila A Fisher, Jonathan Marchini, Jilur Ghori, Suzannah Bumpstead, Rhian Gwilliam, Mark Tremelling, Panos Deloukas, John Mansfield, Derek Jewell, Jack Satsangi, Christopher G Mathew, Miles Parkes, Michel Georges, and Mark J Daly. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet*, 40(8):955–962, August 2008. doi: 10.1038/ng.175. 1, 6, 28, 104

Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Ccile Julier, Grant Morahan, Jrn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, Stephen S Rich, and The Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*, May 2009. 1, 28

L Bastone, M Reilly, DJ Rader, and AS Foulkes. Mdr and prp: a comparison of methods for high-order genotype-phenotype associations. *Human heredity*, 58(2):82–92, 2005. URL http://www.karger.com/Article/FullText/83029. 49

K. Bhattacharya and A. P. Morris. European mathematical genetics meeting, munich, germany, 14th-15th may 2009. *Annals of Human Genetics*, 73(6): 668, 2009. ISSN 1469-1809. doi: 10.1111/j.1469-1809.2009.00547.x. URL http://dx.doi.org/10.1111/j.1469-1809.2009.00547.x. 56

Kanishka Bhattacharya, Mark I. McCarthy, and Andrew P. Morris. Rapid testing of gene-gene interactions in genome-wide association studies of binary and quantitative phenotypes. *Genet Epidemiol*, 35(8):800–808, Dec 2011. doi: 10.1002/gepi.20629. URL http://dx.doi.org/10.1002/gepi.20629. 75

Zoltán Bochdanovits, David Sondervan, Sophie Perillous, Toos Van Beijsterveldt, Dorret Boomsma, and Peter Heutink. Genome-wide prediction of functional gene-gene interactions inferred from patterns of genetic differentiation in mice and men. *PloS one*, 3(2):e1593, 2008. 47

David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33 Suppl:228–237, March 2003. doi: 10.1038/ng1090. 14

Leo Breiman. *Classification and regression trees*. CRC press, 1993. 48, 49

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. URL http://link.springer.com/article/10.1023/A:1010933404324. 49

Sharon R Browning. Multilocus association mapping using variable-length markov chains. *The American Journal of Human Genetics*, 78(6):903–913, 2006. 105

E H Bryant and L M Meffert. The effect of serial founder-flush cycles on quantitative genetic variation in the housefly. *Heredity*, 70(2):122–129, 1993. doi: 10.1038/hdy.1993.20. 35

Edwin H Bryant, Steven A McCommas, and Lisa M Combs. The effect of an experimental bottleneck upon quantitative genetic variation in the housefly. *Genetics*, 114(4):1191–1211, 1986. 35

Alexandre Bureau, Josée Dupuis, Kathleen Falls, Kathryn L Lunetta, Brooke Hayward, Tim P Keith, and Paul Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genetic epidemiology*, 28(2):171–182, 2005. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20041/full. 49

L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–99, February 2001. doi: 10.1038/35052543. 13

Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604, 2003. URL http://www.sciencedirect.com/science/article/pii/S0140673603125202. 18

Orjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*, 5(8):618–625, August 2004. doi: 10.1038/nrg1407. 3, 38, 137

Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1):106–120, January 2004. doi: 10.1086/381000. 14

Hampton L Carson and Alan R Templeton. Genetic revolutions in relation to speciation phenomena: the founding of new populations. *Annual Review of Ecology and Systematics*, 15(1):97–131, 1984. URL http://www.jstor.org/stable/2096944. 35

Pritam Chanda, Aidong Zhang, Daniel Brazeau, Lara Sucheston, Jo L Freudenheim, Christine Ambrosone, and Murali Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *The American Journal of Human Genetics*, 81(5):939–963, 2007. 47

Juliet Chapman and David Clayton. Detecting association using epistatic information. *Genetic epidemiology*, 31(8):894–909, 2007. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20250/full. 40, 44

J. M. Cheverud and E. J. Routman. Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–1461, March 1995. 37, 38

Jennifer C. Chow, Ziny Yen, Sonia M. Ziesche, and Carolyn J. Brown. Silencing of the mammalian x chromosome. *Annu Rev Genomics Hum Genet*, 6:69–92, 2005. doi: 10.1146/annurev.genom.6.080604.162350. 122, 124

Kaare Christensen and Jeffrey C Murray. What genome-wide association studies can do for medicine. *N Engl J Med*, 356(11):1094–1097, March 2007. doi: 10.1056/NEJMp068126. 5

Xun Chu, Min Shen, Fang Xie, Xiao-Jing Miao, Wei-Hua Shou, Lin Liu, Peng-Peng Yang, Ya-Nan Bai, Kai-Yue Zhang, Lin Yang, et al. An x chromosome-wide association analysis identifies variants in gpr174 as a risk factor for graves disease. *Journal of medical genetics*, 50(7):479–485, 2013. URL http://jmg.bmj.com/content/50/7/479.short. 123

Yujin Chung, Seung Yeoun Lee, Robert C Elston, and Taesung Park. Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics*, 23(1):71–76, 2007. URL http://bioinformatics.oxfordjournals.org/content/23/1/71.short. 51

Elizabeth T Cirulli and David B Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010. URL http://www.nature.com/nrg/journal/v11/n6/abs/nrg2779.html. 29

David Clayton. Testing for association on the x chromosome. *Biostatis-*

tics, 9(4):593–600, 2008. URL http://biostatistics.oxfordjournals.org/content/9/4/593.short. 123

David G Clayton. Sex chromosomes and genetic association studies. *Genome Med*, 1(11):110, 2009. doi: 10.1186/gm110. 122, 124

David G. Clayton, Neil M. Walker, Deborah J. Smyth, Rebecca Pask, Jason D. Cooper, Lisa M. Maier, Luc J. Smink, Alex C. Lam, Nigel R. Ovington, Helen E. Stevens, Sarah Nutland, Joanna M M. Howson, Malek Faham, Martin Moorhead, Hywel B. Jones, Matthew Falkowski, Paul Hardenbol, Thomas D. Willis, and John A. Todd. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37(11):1243–1246, Nov 2005. doi: 10.1038/ng1653. URL http://dx.doi.org/10.1038/ng1653. 16

William G Cochran. Some methods for strengthening the common χ 2 tests. *Biometrics*, 10(4):417–451, 1954. URL http://www.jstor.org/stable/10.2307/3001616. 123

Christopher S Coffey, Patricia R Hebert, Marylyn D Ritchie, Harlan M Krumholz, J. Michael Gaziano, Paul M Ridker, Nancy J Brown, Douglas E Vaughan, and Jason H Moore. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics*, 5:49, April 2004. doi: 10.1186/1471-2105-5-49. 2, 38, 118

Gregory M Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*,

12(9):628–640, 2011. URL http://www.nature.com/nrg/journal/v12/n9/abs/nrg3046.html. 29

Jason D Cooper, Deborah J Smyth, Adam M Smiles, Vincent Plagnol, Neil M Walker, James E Allen, Kate Downes, Jeffrey C Barrett, Barry C Healy, Josyf C Mychaleckyj, James H Warram, and John A Todd. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet*, 40(12):1399–1401, December 2008. doi: 10.1038/ng.249. 1

John B Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 311–354, 1983. URL http://www.jstor.org/stable/10.2307/2345402. 48

Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 11(20):2463–2468, October 2002. 3, 34, 137

Heather J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404, June 2009. doi: 10.1038/nrg2579. 40, 75

Nick Craddock, MC Odonovan, and MJ Owen. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *Journal of Medical Genetics*, 42(3): 193–204, 2005. URL http://jmg.bmj.com/content/42/3/193.short. 88

James F Crow. Population genetics history: a personal view. *Annual review of genetics*, 21(1):1–22, 1987. URL http://www.annualreviews.org/doi/pdf/10.1146/annurev.ge.21.120187.000245. 38

James F Crow and Motoo Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, 1970. URL http://www.cabdirect.org/abstracts/19710105376.html?freeview=true. 37

JF Crow. Minor viability mutants in drosophila. *Genetics*, 92(1 Pt 1 Suppl): s165, 1979. URL http://www.ncbi.nlm.nih.gov/pubmed/488699. 38

Robert Culverhouse, Tsvika Klein, and William Shannon. Detecting epistatic interactions contributing to quantitative traits. *Genetic epidemiology*, 27 (2):141–152, 2004. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20006/full. 48

M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–232, Oct 2001. doi: 10.1038/ng1001-229. URL http://dx.doi.org/10.1038/ng1001-229. 14

B Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999. URL http://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.1999.00997.x/full. 19

Changzheng Dong, Xun Chu, Ying Wang, Yi Wang, Li Jin, Tieliu Shi, Wei Huang, and Yixue Li. Exploration of gene–gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16(2):229–235, 2007. URL http://www.nature.com/ejhg/journal/vaop/ncurrent/full/5201921a.html. 47

Chuanhui Dong, Shuang Wang, Wei-Dong Li, Ding Li, Hongyu Zhao, and R. Arlen Price. Interacting genetic loci on chromosomes 20 and 10 influence

extreme human obesity. *Am J Hum Genet*, 72(1):115–124, Jan 2003. doi: 10.1086/345648. URL http://dx.doi.org/10.1086/345648. 2

Chuanhui Dong, Wei-Dong Li, Ding Li, and R. Arlen Price. Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21. *Eur J Hum Genet*, 13(1):102–108, Jan 2005. doi: 10.1038/sj.ejhg.5201292. URL http://dx.doi.org/10.1038/sj.ejhg.5201292. 2

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. URL http://projecteuclid.org/euclid.aos/1083178935. 48

Mathieu Emily, Thomas Mailund, Jotun Hein, Leif Schauser, and Mikkel Heide Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240, 2009. URL http://www.nature.com/ejhg/journal/vaop/ncurrent/full/ejhg200915a.html. 47

Evangelos Evangelou and John PA Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013. URL http://www.nature.com/nrg/journal/v14/n6/abs/nrg3472.html. 27

David M Evans, Jonathan Marchini, Andrew P Morris, and Lon R Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, Sep 2006. doi: 10.1371/journal.pgen.0020157. URL http://dx.doi.org/10.1371/journal.pgen.0020157. 45, 58, 101, 138

Douglas S. Falconer and Trudy F. C. Mackay. *Introduction to Quantitative Genetics (4th Edition)*. Prentice Hall, February 1996. ISBN 0582243025. URL http://www.worldcat.org/isbn/0582243025. 38

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003. URL http://www.genetics.org/content/164/4/1567.short. 19

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7(4):574–578, 2007. doi: 10.1111/j.1471-8286.2007.01758.x/full. 19

T. Ferreira. *Statistical Methods for Modelling Epistasis in Genetic Association Studies*. University of Oxford, 2010. URL http://books.google.co.uk/books?id=igErnwEACAAJ. 44

Teresa Ferreira and Jonathan Marchini. Modeling interactions with known risk loci-a bayesian model averaging approach. *Ann Hum Genet*, 75(1):1–9, Jan 2011. doi: 10.1111/j.1469-1809.2010.00618.x. URL http://dx.doi.org/10.1111/j.1469-1809.2010.00618.x. 44

R.A. Fisher. *The Design of Experiments*. Hafner, 1951. URL http://books.google.co.uk/books?id=ejxBAAAAIAAJ. 118

Ronald Aylmer Fisher. On the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*,

52(2):399–433, 1918. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+correlation+between+relatives+on+the+supposition+of+mendelian+inheritance. 34

Jonathan Flint and Trudy F C Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res*, 19(5):723–733, May 2009. doi: 10.1101/gr.086660.108. URL http://dx.doi.org/10.1101/gr.086660.108. 2

Boris Freidlin, Gang Zheng, Zhaohai Li, and Joseph L Gastwirth. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human heredity*, 53(3):146–152, 2009. URL http://www.karger.com/Article/Abstract/64976. 23

Stacey B. Gabriel, Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly, and David Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576): 2225–2229, Jun 2002. doi: 10.1126/science.1069424. URL http://dx.doi.org/10.1126/science.1069424. 14

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2003. 52

G. Gibson and G. Wagner. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays*, 22(4):372–380, Apr 2000. doi: 3.0.CO;2-J. URL http://dx.doi.org/3.0.CO;2-J. 35

C J Goodnight. On the effect of founder events on epistatic genetic variance. *Evolution*, 41(1):80, 1987. URL http://www.jstor.org/stable/2408974?origin=crossref. 35

C J Goodnight. Epistasis and the effect of founder events on the additive genetic variance. *Evolution*, 42(3):441–454, 1988. URL http://www.jstor.org/stable/2409030?origin=crossref. 35

IP Gorlov, OY Gorlova, SR Sunyaev, MR Spitz, and CI Amos. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*, 82:100–112, 2008. doi: 10.1016/j.ajhg.2007.09.006. 17

Trisha Greenhalgh. The human genome project. *J R Soc Med*, 98(12):545, Dec 2005. doi: 10.1258/jrsm.98.12.545. URL http://dx.doi.org/10.1258/jrsm.98.12.545. 4

Mickael Guedj, Gregory Nuel, and Bernard Prum. A note on allelic tests in case-control association studies. *Annals of human genetics*, 72(3):407–409, 2008. 22

J. R. Gulcher, A. Kong, and K. Stefansson. The role of linkage studies for common diseases. *Curr Opin Genet Dev*, 11(3):264–267, Jun 2001. 12

Ingileif B Hallgrmsdttir and Debbie S Yuster. A complete classification of epistatic two-locus models. *BMC Genet*, 9:17, 2008. doi: 10.1186/1471-2156-9-17. URL http://dx.doi.org/10.1186/1471-2156-9-17. 45, 58

Eran Halperin and Dietrich A Stephan. Snp imputation in association studies.

Nat Biotechnol, 27(4):349–351, Apr 2009. doi: 10.1038/nbt0409-349. URL http://dx.doi.org/10.1038/nbt0409-349. 13

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue): D514–D517, Jan 2005. doi: 10.1093/nar/gki033. URL http://dx.doi.org/10.1093/nar/gki033. 5

Buhm Han and Eleazar Eskin. Interpreting meta-analyses of genome-wide association studies. *PLoS genetics*, 8(3):e1002555, 2012. 28

Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics, 4th edition*. Number ISBN-10: 0-87893-308-5. Sinauer and Associates, Sunderland, MA, 2007. 7, 9

Daniel L Hartl, Andrew G Clark, et al. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997. 8

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005. URL http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf. 48

Alan Hastings. Unexpected behavior in two locus genetic systems: an analysis of marginal underdominance at a stable equilibrium. *Genetics*, 102(1):129–138, 1982. URL http://www.genetics.org/content/102/1/129.short. 38

Alan Hastings. Multilocus population genetics with weak epistasis. i. equilibrium

properties of two-locus two-allele models. *Genetics*, 109(4):799–812, 1985. URL http://www.genetics.org/content/109/4/799.short. 38

Nancy L Heard-Costa, M. Carola Zillikens, Keri L Monda, Asa Johansson, Tamara B Harris, Mao Fu, Talin Haritunians, Mary F Feitosa, Thor Aspelund, Gudny Eiriksdottir, Melissa Garcia, Lenore J Launer, Albert V Smith, Braxton D Mitchell, Patrick F McArdle, Alan R Shuldiner, Suzette J Bielinski, Eric Boerwinkle, Fred Brancati, Ellen W Demerath, James S Pankow, Alice M Arnold, Yii-Der Ida Chen, Nicole L Glazer, Barbara McKnight, Bruce M Psaty, Jerome I Rotter, Najaf Amin, Harry Campbell, Ulf Gyllensten, Cristian Pattaro, Peter P Pramstaller, Igor Rudan, Maksim Struchalin, Veronique Vitart, Xiaoyi Gao, Aldi Kraja, Michael A Province, Qunyuan Zhang, Larry D Atwood, Jose Dupuis, Joel N Hirschhorn, Cashell E Jaquish, Christopher J O'Donnell, Ramachandran S Vasan, Charles C White, Yurii S Aulchenko, Karol Estrada, Albert Hofman, Fernando Rivadeneira, Andr G Uitterlinden, Jacqueline C M Witteman, Ben A Oostra, Robert C Kaplan, Vilmundur Gudnason, Jeffrey R O'Connell, Ingrid B Borecki, Cornelia M van Duijn, L. Adrienne Cupples, Caroline S Fox, and Kari E North. Nrxn3 is a novel locus for waist circumference: a genome-wide association study from the charge consortium. *PLoS Genet*, 5(6):e1000539, Jun 2009. doi: 10.1371/journal.pgen.1000539. URL http://dx.doi.org/10.1371/journal.pgen.1000539. 3

P Hedrick, S Jain, and L Holden. Multilocus systems in evolution [drosophila, maize]. *Evolutionary biology*, 11, 1978. 38

Peter F Hickey and Melanie Bahlo. X chromosome association testing in genome

wide association studies. *Genetic epidemiology*, 35(7):664–670, 2011. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20616/full. 123

J Hoh, A Wille, R Zee, S Cheng, R Reynolds, K Lindpaintner, and J Ott. Selecting snps in two-stage analysis of disease association data: a model-free approach. *Annals of human genetics*, 64(5):413–417, 2000. URL http://onlinelibrary.wiley.com/doi/10.1046/j.1469-1809.2000.6450413.x/full. 46

Josephine Hoh and Jurg Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4(9):701–709, 2003. URL http://www.nature.com/nrg/journal/v4/n9/abs/nrg1155.html. 43

Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, Jun 2009. doi: 10.1371/journal.pgen.1000529. URL http://dx.doi.org/10.1371/journal.pgen.1000529. 13, 26, 105, 110, 140

Travis Hughes, Adam Adler, Jennifer A Kelly, Kenneth M Kaufman, Adrienne H Williams, Carl D Langefeld, Elizabeth E Brown, Graciela S Alarcón, Robert P Kimberly, Jeffrey C Edberg, et al. Evidence for gene–gene epistatic interactions among susceptibility loci for systemic lupus erythematosus. *Arthritis & Rheumatism*, 64(2):485–492, 2012. 39

Michael Hutchinson, Cleanthe Spanaki, Sergey Lebedev, and Andreas Plaitakis. Genetic basis of common diseases: The general theory of mendelian recessive genetics. *Medical Hypotheses*, 65(2):

282 – 286, 2005. ISSN 0306-9877. doi: DOI:10.1016/j.mehy.2005.02.034. URL http://www.sciencedirect.com/science/article/B6WN2-4G1GD2K-1/2/775e2f35aebe73dc5707e81a98a527d3. 4

Valma Hyttinen, Jaakko Kaprio, Leena Kinnunen, Markku Koskenvuo, and Jaakko Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes*, 52(4): 1052–1055, 2003. URL http://diabetes.diabetesjournals.org/content/52/4/1052.short. 88

Lin-Dan Ji, Li-Na Zhang, and Jin Xu. Genome-wide association studies of hypertension: Achievements, difficulties and strategies. *World*, 1(1):10–14, 2011. 88

Guolian Kang, Weihua Yue, Jifeng Zhang, Yuehua Cui, Yijun Zuo, and Dai Zhang. An entropy-based approach for testing genetic epistasis underlying complex diseases. *Journal of theoretical biology*, 250(2):362–374, 2008. URL http://www.sciencedirect.com/science/article/pii/S002251930700478X. 47

Samuel Karlin, Marcus W Feldman, et al. Linkage and selection: two locus symmetric viability model. *Theoretical population biology*, 1(1):39–71, 1970. URL http://www.cabdirect.org/abstracts/19731608793.html. 38

Samuel Karlin et al. General two-locus selection models: some objectives, results and interpretations. *Theoretical Population Biology*, 7(3):364, 1975. 38

PT Katzmarzyk, L Perusse, T Rice, J Gagnon, JS Skinner, JH Wilmore, AS Leon, DC Rao, and C Bouchard. Familial resemblance for coronary heart disease

risk: the heritage family study. *Ethnicity & disease*, 10(2):138, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10892820. 88

Xiayi Ke and Lon R. Cardon. Efficient selective screening of haplotype tag snps. *Bioinformatics*, 19(2):287–288, Jan 2003. 14

Peter Kraft, Y-C Yen, Daniel O Stram, John Morrison, and W James Gauderman. Exploiting gene-environment interaction to detect genetic associations. *Human heredity*, 63(2):111–119, 2007. URL http://www.karger.com/Article/Fulltext/99183. 44

Saskia Le Cessie and JC Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992. URL http://www.jstor.org/stable/10.2307/2347628. 48

AH Lee and MJ Silvapulle. Ridge estimation in logistic regression. *Communications in Statistics-Simulation and Computation*, 17(4):1231–1257, 1988. URL http://www.tandfonline.com/doi/abs/10.1080/03610918808812723. 48

Seung Yeoun Lee, Yujin Chung, Robert C Elston, Youngchul Kim, and Taesung Park. Log-linear model-based multifactor dimensionality reduction method to detect gene–gene interactions. *Bioinformatics*, 23(19):2589–2595, 2007. URL http://bioinformatics.oxfordjournals.org/content/23/19/2589.short. 51

RC Lewontin and Ken-ichi Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, pages 458–472, 1960. URL http://www.jstor.org/stable/10.2307/2405995. 38

Yun Li and Gonçalo R Abecasis. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet S*, 79(3):2290, 2006. 105

Cecilia M Lindgren, Iris M Heid, Joshua C Randall, Claudia Lamina, Valgerdur Steinthorsdottir, Lu Qi, Elizabeth K Speliotes, Gudmar Thorleifsson, Cristen J Willer, Blanca M Herrera, Anne U Jackson, Noha Lim, Paul Scheet, Nicole Soranzo, Najaf Amin, Yurii S Aulchenko, John C Chambers, Alexander Drong, Jian'an Luan, Helen N Lyon, Fernando Rivadeneira, Serena Sanna, Nicholas J Timpson, M. Carola Zillikens, Jing Hua Zhao, Peter Almgren, Stefania Bandinelli, Amanda J Bennett, Richard N Bergman, Lori L Bonnycastle, Suzannah J Bumpstead, Stephen J Chanock, Lynn Cherkas, Peter Chines, Lachlan Coin, Cyrus Cooper, Gabriel Crawford, Angela Doering, Anna Dominiczak, Alex S F Doney, Shah Ebrahim, Paul Elliott, Michael R Erdos, Karol Estrada, Luigi Ferrucci, Guido Fischer, Nita G Forouhi, Christian Gieger, Harald Grallert, Christopher J Groves, Scott Grundy, Candace Guiducci, David Hadley, Anders Hamsten, Aki S Havulinna, Albert Hofman, Rolf Holle, John W Holloway, Thomas Illig, Bo Isomaa, Leonie C Jacobs, Karen Jameson, Pekka Jousilahti, Fredrik Karpe, Johanna Kuusisto, Jaana Laitinen, G. Mark Lathrop, Debbie A Lawlor, Massimo Mangino, Wendy L McArdle, Thomas Meitinger, Mario A Morken, Andrew P Morris, Patricia Munroe, Narisu Narisu, Anna Nordstrm, Peter Nordstrm, Ben A Oostra, Colin N A Palmer, Felicity Payne, John F Peden, Inga Prokopenko, Frida Renstrm, Aimo Ruokonen, Veikko Salomaa, Manjinder S Sandhu, Laura J Scott, Angelo Scuteri, Kaisa Silander, Kijoung Song, Xin Yuan, Heather M Stringham, Amy J Swift, Tiinamaija Tuomi, Manuela Uda, Peter Vollenweider, Gerard Waeber, Chris

Wallace, G. Bragi Walters, Michael N Weedon, Wellcome Trust Case Control Consortium, Jacqueline C M Witteman, Cuilin Zhang, Weihua Zhang, Mark J Caulfield, Francis S Collins, George Davey Smith, Ian N M Day, Paul W Franks, Andrew T Hattersley, Frank B Hu, Marjo-Riitta Jarvelin, Augustine Kong, Jaspal S Kooner, Markku Laakso, Edward Lakatta, Vincent Mooser, Andrew D Morris, Leena Peltonen, Nilesh J Samani, Timothy D Spector, David P Strachan, Toshiko Tanaka, Jaakko Tuomilehto, Andr G Uitterlinden, Cornelia M van Duijn, Nicholas J Wareham, Hugh Watkins, Procardis Consortia, Dawn M Waterworth, Michael Boehnke, Panos Deloukas, Leif Groop, David J Hunter, Unnur Thorsteinsdottir, David Schlessinger, H-Erich Wichmann, Timothy M Frayling, Gonalo R Abecasis, Joel N Hirschhorn, Ruth J F Loos, Kari Stefansson, Karen L Mohlke, Ins Barroso, Mark I McCarthy, and Giant Consortium. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet*, 5(6):e1000508, Jun 2009. 3

K. Ljungberg, S. Holmgren, and O. Carlborg. Simultaneous search for multiple qtl using the global optimization algorithm direct. *Bioinformatics*, 20(12): 1887–1895, Aug 2004. doi: 10.1093/bioinformatics/bth175. URL http://dx.doi.org/10.1093/bioinformatics/bth175. 47

Kirk E Lohmueller, Celeste L Pearce, Malcolm Pike, Eric S Lander, and Joel N Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33 (2):177–182, Feb 2003. doi: 10.1038/ng1071. URL http://dx.doi.org/10.1038/ng1071. 1

Christina Loley, Andreas Ziegler, and Inke R. Knig. Association tests for x-

chromosomal markers–a comparison of different test statistics. *Hum Hered*, 71 (1):23–36, 2011. doi: 10.1159/000323768. URL http://dx.doi.org/10.1159/000323768. 123

A. D. Long and C. H. Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res*, 9(8):720–731, Aug 1999. 12

Kathryn L Lunetta, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):32, 2004. 49

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000. URL http://link.springer.com/article/10.1023/A:1008929526011. 52

David J Lunn, John C Whittaker, and Nicky Best. A bayesian toolkit for genetic association studies. *Genetic epidemiology*, 30(3):231–247, 2006. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20140/full. 52

Alexander J MacGregor, Harold Snieder, Alan S Rigby, Markku Koskenvuo, Jaakko Kaprio, Kimmo Aho, and Alan J Silman. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis & Rheumatism*, 43(1):30–37, 2000. 88

Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*, 363(2):166–176, Jul 2010. doi: 10.1056/NEJMra0905980. URL http://dx.doi.org/10.1056/NEJMra0905980. 2

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009. doi: 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494. 2, 29, 137

Jonathan Marchini and Bryan Howie. Comparing algorithms for genotype imputation. *Am J Hum Genet*, 83(4):535–9; author reply 539–40, Oct 2008. doi: 10.1016/j.ajhg.2008.09.007. URL http://dx.doi.org/10.1016/j.ajhg.2008.09.007. 16

Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, Jul 2010. doi: 10.1038/nrg2796. URL http://dx.doi.org/10.1038/nrg2796. x, 103, 106

Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nat Genet*, 36(5):512–517, May 2004. doi: 10.1038/ng1337. URL http://dx.doi.org/10.1038/ng1337. 19

Jonathan Marchini, Peter Donnelly, and Lon R Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):

413–417, Apr 2005. doi: 10.1038/ng1537. URL http://dx.doi.org/10.1038/ng1537. 45, 46, 58

Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, Jul 2007. doi: 10.1038/ng2088. URL http://dx.doi.org/10.1038/ng2088. 105

Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006. URL http://link.springer.com/article/10.2165/00822942-200605020-00002. 47

Brett A McKinney, David M Reif, Bill C White, JE Crowe, and Jason H Moore. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics*, 23(16):2113–2120, 2007. URL http://bioinformatics.oxfordjournals.org/content/23/16/2113.short. 52

Brett A McKinney, James E Crowe Jr, Jingyu Guo, and Dehua Tian. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, 5(3):e1000432, 2009. 50, 52

Zhaoling Meng, Dmitri V. Zaykin, Chun-Fang Xu, Michael Wagner, and Margaret G. Ehm. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet*, 73(1):115–130, Jul 2003. doi: 10.1086/376561. URL http://dx.doi.org/10.1086/376561. 14

Joshua Millstein, David V Conti, Frank D Gilliland, and W James Gauderman. A testing framework for identifying susceptibility genes in the presence of epistasis. *American journal of human genetics*, 84(2):301, 2009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2678003/. 46

Jason H Moore. A global view of epistasis. *Nat Genet*, 37(1):13–14, Jan 2005. doi: 10.1038/ng0105-13. URL http://dx.doi.org/10.1038/ng0105-13. ix, 34, 35, 36

Jason H Moore and Bill C White. Tuning relieff for genome-wide genetic analysis. In *Evolutionary computation, machine learning and data mining in bioinformatics*, pages 166–175. Springer, 2007. URL http://link.springer.com/chapter/10.1007/978-3-540-71783-6_16. 51

Jason H Moore and Scott M Williams. New strategies for identifying gene-gene interactions in hypertension. *Annals of medicine*, 34(2):88–95, 2002. URL http://informahealthcare.com/doi/abs/10.1080/07853890252953473. 47

Jason H Moore and Scott M Williams. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*, 27(6):637–646, Jun 2005. doi: 10.1002/bies.20236. URL http://dx.doi.org/10.1002/bies.20236. 35

Jason H Moore, Joshua C Gilbert, Chia-Ti Tsai, Fu-Tien Chiang, Todd Holden, Nate Barney, and Bill C White. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology*, 241

(2):252–261, 2006. URL http://www.sciencedirect.com/science/article/pii/S0022519305005217. 47, 51

Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*, 34(2): 188–193, Feb 2010. doi: 10.1002/gepi.20450. URL http://dx.doi.org/10.1002/gepi.20450. 142

Andrew P. Morris, Benjamin F. Voight, Tanya M. Teslovich, Teresa Ferreira, Ayellet V. Segr, Valgerdur Steinthorsdottir, Rona J. Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, Inga Prokopenko, Hyun Min Kang, Christian Dina, Tonu Esko, Ross M. Fraser, Stavroula Kanoni, Ashish Kumar, Vasiliki Lagou, Claudia Langenberg, Jian'an Luan, Cecilia M. Lindgren, Martina Mller-Nurasyid, Sonali Pechlivanis, N William Rayner, Laura J. Scott, Steven Wiltshire, Loic Yengo, Leena Kinnunen, Elizabeth J. Rossin, Soumya Raychaudhuri, Andrew D. Johnson, Antigone S. Dimas, Ruth J F. Loos, Sailaja Vedantam, Han Chen, Jose C. Florez, Caroline Fox, Ching-Ti Liu, Denis Rybin, David J. Couper, Wen Hong L. Kao, Man Li, Marilyn C. Cornelis, Peter Kraft, Qi Sun, Rob M. van Dam, Heather M. Stringham, Peter S. Chines, Krista Fischer, Pierre Fontanillas, Oddgeir L. Holmen, Sarah E. Hunt, Anne U. Jackson, Augustine Kong, Robert Lawrence, Julia Meyer, John R B. Perry, Carl G P. Platou, Simon Potter, Emil Rehnberg, Neil Robertson, Suthesh Sivapalaratnam, Alena Stankov, Kathleen Stirrups, Gudmar Thorleifsson, Emmi Tikkanen, Andrew R. Wood, Peter Almgren, Mustafa Atalay, Rafn Benediktsson, Lori L. Bonnycastle, Nol Burtt, Jason Carey, Guillaume Charpentier, Andrew T. Crenshaw, Alex S F. Doney, Mozhgan Dorkhan, Sarah Ed-

kins, Valur Emilsson, Elodie Eury, Tom Forsen, Karl Gertow, Bruna Gigante, George B. Grant, Christopher J. Groves, Candace Guiducci, Christian Herder, Astradur B. Hreidarsson, Jennie Hui, Alan James, Anna Jonsson, Wolfgang Rathmann, Norman Klopp, Jasmina Kravic, Kaarel Krjutkov, Cordelia Langford, Karin Leander, Eero Lindholm, Stphane Lobbens, Satu Mnnist, Ghazala Mirza, Thomas W. Mhleisen, Bill Musk, Melissa Parkin, Loukianos Rallidis, Jouko Saramies, Bengt Sennblad, Sonia Shah, Gunnar Sigursson, Angela Silveira, Gerald Steinbach, Barbara Thorand, Joseph Trakalo, Fabrizio Veglia, Roman Wennauer, Wendy Winckler, Delilah Zabaneh, Harry Campbell, Cornelia van Duijn, Andre G. Uitterlinden, Albert Hofman, Eric Sijbrands, Goncalo R. Abecasis, Katharine R. Owen, Eleftheria Zeggini, Mieke D. Trip, Nita G. Forouhi, Ann-Christine Syvnen, Johan G. Eriksson, Leena Peltonen, Markus M. Nthen, Beverley Balkau, Colin N A. Palmer, Valeriya Lyssenko, Tiinamaija Tuomi, Bo Isomaa, David J. Hunter, Lu Qi, Wellcome Trust Case Control Consortium , Meta-Analyses of Glucose , Insulin related traits Consortium (MAGIC) Investigators, Genetic Investigation of A. Nthropometric Traits (G. I. A. N. T) Consortium , Asian Genetic Epidemiology Network-Type 2 Diabetes (A. G. E. N-T2D) Consortium , South Asian Type 2 Diabetes (S. A. T2D) Consortium , Alan R. Shuldiner, Michael Roden, Ines Barroso, Tom Wilsgaard, John Beilby, Kees Hovingh, Jackie F. Price, James F. Wilson, Rainer Rauramaa, Timo A. Lakka, Lars Lind, George Dedoussis, Inger Njlstad, Nancy L. Pedersen, Kay-Tee Khaw, Nicholas J. Wareham, Sirkka M. Keinanen-Kiukaanniemi, Timo E. Saaristo, Eeva Korpi-Hyvlti, Juha Saltevo, Markku Laakso, Johanna Kuusisto, Andres Metspalu, Francis S. Collins, Karen L. Mohlke, Richard N. Bergman, Jaakko Tuomilehto, Bernhard O. Boehm, Chris-

tian Gieger, Kristian Hveem, Stephane Cauchi, Philippe Froguel, Damiano Baldassarre, Elena Tremoli, Steve E. Humphries, Danish Saleheen, John Danesh, Erik Ingelsson, Samuli Ripatti, Veikko Salomaa, Raimund Erbel, Karl-Heinz Jckel, Susanne Moebus, Annette Peters, Thomas Illig, Ulf de Faire, Anders Hamsten, Andrew D. Morris, Peter J. Donnelly, Timothy M. Frayling, Andrew T. Hattersley, Eric Boerwinkle, Olle Melander, Sekar Kathiresan, Peter M. Nilsson, Panos Deloukas, Unnur Thorsteinsdottir, Leif C. Groop, Kari Stefansson, Frank Hu, James S. Pankow, Jose Dupuis, James B. Meigs, David Altshuler, Michael Boehnke, Mark I. McCarthy, D. I. Abetes Genetics Replication , and Meta analysis (D. I. A. G. R. A. M) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, 44(9):981–990, Sep 2012. 28

V Moskvina, N Craddock, P Holmans, MJ Owen, and MC O'Donovan. Effects of differential genotyping error rate on the type i error probability of case-control studies. *Hum Hered*, 61:55–64, 2006. doi: 10.1159/000092553. 17

Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3): 685–694, 2008. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2007.00953.x/full. 43

Bhramar Mukherjee, Jaeil Ahn, Stephen B Gruber, Gad Rennert, Victor Moreno, and Nilanjan Chatterjee. Tests for gene-environment interaction from case-control data: a novel study of type i error, power and designs. *Genetic epi-*

demiology, 32(7):615–626, 2008. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20337/full. 43

Benjamin M Neale, Jesen Fagerness, Robyn Reynolds, Lucia Sobrin, Margaret Parker, Soumya Raychaudhuri, Perciliz L Tan, Edwin C Oh, Joanna E Merriam, Eric Souied, et al. Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (lipc). *Proceedings of the National Academy of Sciences*, 107(16):7395–7400, 2010. URL http://www.pnas.org/content/107/16/7395.short. 28

MR Nelson, SLR Kardia, RE Ferrell, and CF Sing. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, 11(3):458–470, 2001. URL http://genome.cshlp.org/content/11/3/458.short. 48, 50

Dan L Nicolae. Testing untyped alleles (tuna)applications to genome-wide association studies. *Genetic epidemiology*, 30(8):718–727, 2006. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20182/full. 105

Carole Ober, Dagan A Loisel, and Yoav Gilad. Sex-specific genetic architecture of human disease. *Nature Reviews Genetics*, 9(12):911–922, 2008. URL http://www.nature.com/nrg/journal/v9/n12/abs/nrg2415.html. 122

Paola Oliveri and Eric H Davidson. Gene regulatory network controlling embryonic specification in the sea urchin. *Curr Opin Genet Dev*, 14(4):351–360, Aug 2004. doi: 10.1016/j.gde.2004.06.004. URL http://dx.doi.org/10.1016/j.gde.2004.06.004. 37

Guillaume Par, Nancy R. Cook, Paul M. Ridker, and Daniel I. Chasman. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genet*, 6(6): e1000981, Jun 2010. doi: 10.1371/journal.pgen.1000981. URL http://dx.doi.org/10.1371/journal.pgen.1000981. 47

N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, Nov 2001. doi: 10.1126/science.1065573. URL http://dx.doi.org/10.1126/science.1065573. 14

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006. 19

Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32(4):381–385, 2008. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20303/full. 27

Patrick C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11):855–867, Nov 2008. doi: 10.1038/nrg2452. URL http://dx.doi.org/10.1038/nrg2452. 34

Walter W Piegorsch, Clarice R Weinberg, and Jack A Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibil-

ity in population-based case-control studies. *Statistics in medicine*, 13(2): 153–162, 1994. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.4780130206/full. 42, 44

M. Platzer. The human genome and its upcoming dynamics. *Genome Dyn*, 2:1–16, 2006. doi: 10.1159/000095083. URL http://dx.doi.org/10.1159/000095083. 4

P Poulsen, K Ohm Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, 1999. URL http://link.springer.com/article/10.1007/s001250051131. 88

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):pp. 403–411, 1979. ISSN 00063444. URL http://www.jstor.org/stable/2335158. 124

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006. URL http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html. 19

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. URL http://www.genetics.org/content/155/2/945.short. 19

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. de Bakker, Mark J.

Daly, and Pak C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81 (3):559–575, September 2007. ISSN 0002-9297. doi: 10.1086/519795. URL http://dx.doi.org/10.1086/519795. 76, 105

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996. 12

M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1):138–147, Jul 2001. doi: 10.1086/321276. URL http://dx.doi.org/10.1086/321276. 47, 50, 75

Peter D Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, pages 1253–1261, 1997. URL http://www.jstor.org/stable/10.2307/2533494. 23

Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006. 105

N. J. Schork, L. R. Cardon, and X. Xu. The future of genetic epidemiology. *Trends Genet*, 14(7):266–272, Jul 1998. 12

Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines,

Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Gonalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, Jun 2007. doi: 10.1126/science.1142382. URL http://dx.doi.org/10.1126/science.1142382. 6, 28

D Segre, A DeLuna, G M Church, and R Kishony. Modular epistasis in yeast metabolism. *Nature Genetics*, 37(1):77–83, 2005. URL http://dx.doi.org/10.1038/ng1489. 35, 37

Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114, Jul 2007. doi: 10.1371/journal.pgen.0030114. URL http://dx.doi.org/10.1371/journal.pgen.0030114. 105

Michael J Simmons and James F Crow. Mutations affecting fitness in drosophila populations. *Annual review of genetics*, 11(1):49–78, 1977. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.ge.11.120177.000405. 38

J Sofaer. Crohn's disease: the genetic contribution. *Gut*, 34(7):869–871, 1993. 88

Chris CA Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini. Design-

ing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477, 2009. 26, 73

Steven Stone, Victor Abkevich, Deanna L Russell, Robyn Riley, Kirsten Timms, Thanh Tran, Deborah Trem, David Frank, Srikanth Jammulapati, Chris D Neff, Diana Iliev, Richard Gress, Gongping He, Georges C Frech, Ted D Adams, Mark H Skolnick, Jerry S Lanchbury, Alexander Gutin, Steven C Hunt, and Donna Shattuck. Tbc1d1 is a candidate for a severe obesity gene and evidence for a gene/gene interaction in obesity predisposition. *Hum Mol Genet*, 15(18): 2709–2720, Sep 2006. doi: 10.1093/hmg/ddl204. URL http://dx.doi.org/10.1093/hmg/ddl204. 2

Daniel O Stram, Celeste Leigh Pearce, Phillip Bretsky, Matthew Freedman, Joel N Hirschhorn, David Altshuler, Laurence N Kolonel, Brian E Henderson, and Duncan C Thomas. Modeling and e-m estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, 55(4):179–190, 2003. doi: 10.1159/000073202. URL http://dx.doi.org/10.1159/000073202. 14

Amy Strange, Francesca Capon, CC Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Céline Bellenguez, JG Bergboer, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. *Nature genetics*, 42(11):985, 2010. URL http://europepmc.org/abstract/med/20953190. 2, 39

Maksim V. Struchalin, Abbas Dehghan, Jacqueline Cm Witteman, Cornelia van Duijn, and Yurii S. Aulchenko. Variance heterogeneity analysis for detection of

potentially interacting genetic loci: method and its limitations. *BMC Genet*, 11:92, 2010. doi: 10.1186/1471-2156-11-92. URL http://dx.doi.org/10.1186/1471-2156-11-92. 47

Maksim V. Struchalin, Najaf Amin, Paul H C. Eilers, Cornelia M. van Duijn, and Yurii S. Aulchenko. An r package "variabel" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. *BMC Genet*, 13:4, 2012. doi: 10.1186/1471-2156-13-4. URL http://dx.doi.org/10.1186/1471-2156-13-4. 47

Ioannis M Stylianou, Ron Korstanje, Renhau Li, Susan Sheehan, Beverly Paigen, and Gary A Churchill. Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. *Mamm Genome*, 17(1):22–36, Jan 2006. doi: 10.1007/s00335-005-0091-2. URL http://dx.doi.org/10.1007/s00335-005-0091-2. 2

Meredith Tabangin, Jessica Woo, and Lisa Martin. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, 3 (Suppl 7):S41, 2009. ISSN 1753-6561. doi: 10.1186/1753-6561-3-S7-S41. URL http://www.biomedcentral.com/1753-6561/3/S7/S41. 17

H Tachida and C C Cockerham. A building block model for quantitative genetics. *Genetics*, 121(4):839–844, 1989. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203667&tool=pmcentrez&rendertype=abstract. 35

Yik Y Teo, Andrew E Fry, Taane G Clark, E. S. Tai, and Mark Seielstad. On the usage of hwe for identifying genotyping errors. *Ann Hum Genet*, 71(Pt

5):701–3; author reply 704, Sep 2007. doi: 10.1111/j.1469-1809.2007.00356.x. URL http://dx.doi.org/10.1111/j.1469-1809.2007.00356.x. 16

The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003. 1, 11, 13, 14, 103

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005. 1, 11, 103

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007. doi: 10.1038/nature06258. URL http://dx.doi.org/10.1038/nature06258. 1, 11, 26

I. P. Tu and A. S. Whittemore. Power of association and linkage tests when the disease alleles are unobserved. *Am J Hum Genet*, 64(2):641–649, Feb 1999. doi: 10.1086/302253. URL http://dx.doi.org/10.1086/302253. 12

Stephen Turner, Loren L Armstrong, Yuki Bradford, Christopher S Carlson, Dana C Crawford, Andrew T Crenshaw, Mariza Andrade, Kimberly F Doheny, Jonathan L Haines, Geoffrey Hayes, et al. Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, pages 1–19, 2011. URL http://onlinelibrary.wiley.com/doi/10.1002/0471142905.hg0119s68/full. 15, 17

Anna L Tyler, Folkert W Asselbergs, Scott M Williams, and Jason H Moore. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, 31(2):220–227, Feb 2009. doi: 10.1002/bies.200800022. URL http://dx.doi.org/10.1002/bies.200800022. 34

Masao Ueki and Heather J. Cordell. Improved statistics for genome-wide interaction analysis. *PLoS Genet*, 8(4):e1002625, 2012. doi: 10.1371/journal.pgen.1002625. URL http://dx.doi.org/10.1371/journal.pgen.1002625. 80

Digna R Velez, Bill C White, Alison A Motsinger, William S Bush, Marylyn D Ritchie, Scott M Williams, and Jason H Moore. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic epidemiology*, 31(4):306–315, 2007. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20211/full. 48

C H Waddington. Canalisation of development and the inheritance of acquired characters. *Nature*, 150:563 – 565, 1942. 35

M J Wade. Sewall wright: gene interaction and the shifting balance theory. *Oxford Surveys in Evolutionary Biology*, 8:35–62, 1992. URL http://books.google.com/books?hl=en&lr=&id=ekB-RNrI1RoC&oi=fnd&pg=PA35&dq=Sewall+wright:+gene+interaction+and+the+shifting+balance+theory&ots=DQR8w0cQPE&sig=rl4xwWxvXj8aANCNpKZQ1cxibz0. 35

Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012. URL http://www.nature.com/nbt/journal/v30/n11/abs/nbt.2422.html. 28

Craig H Warden, Nengjun Yi, and Janis Fisler. Epistasis among genes is a universal phenomenon in obesity: evidence from rodent models. *Nutrition*, 20 (1):74–77, Jan 2004. 2

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids. a structure for deoxyribose nucleic acid. 1953. *Rev Invest Clin*, 55(2):108–109, 2003. 4

Michael E Weale. Quality control for genome-wide association studies. In *Genetic Variation*, pages 341–372. Springer, 2010. URL http://link.springer.com/protocol/10.1007/978-1-60327-367-1_19. 15

Mike E. Weale, Chantal Depondt, Stuart J. Macdonald, Alice Smith, Poh San Lai, Simon D. Shorvon, Nicholas W. Wood, and David B. Goldstein. Selection and evaluation of tagging snps in the neuronal-sodium-channel gene scn1a: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet*, 73 (3):551–565, Sep 2003. doi: 10.1086/378098. URL http://dx.doi.org/10.1086/378098. 14

CR Weinberg and David M Umbach. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *American Journal of Epidemiology*, 152(3):197–203, 2000. URL http://aje.oxfordjournals.org/content/152/3/197.short. 42, 44

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447 (7145):661–678, 2007. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature05911.html. 1, 6, 16, 17, 18, 31, 47, 56, 74, 89

Wellcome Trust Case Control Consortium and Australo-Anglo-American Spondylitis Consortium (TASC). Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. *Nat Genet*, 39(11):1329–1337, November 2007. doi: 10.1038/ng.2007.17. 1

Bateson William and Mendel Gregor. *Mendel's principles of heredity*. Cambridge [Eng.]University Press, 1909. URL http://www.biodiversitylibrary.org/item/16926. http://www.biodiversitylibrary.org/bibliography/1057. 34

Scott M Williams, Marylyn D Ritchie, John A Phillips, Elliot Dawson, Melissa Prince, Elvira Dzhura, Alecia Willis, Amma Semenya, Marshall Summar, Bill C White, Jonathan H Addy, John Kpodonu, Lee-Jun Wong, Robin A Felder, Pedro A Jose, and Jason H Moore. Multilocus analysis of hypertension: a hierarchical approach. *Hum Hered*, 57(1):28–38, 2004. doi: 10.1159/000077387. URL http://dx.doi.org/10.1159/000077387. 3, 38

Sewall Wright. Genic and organismic selection. *Evolution*, 34(5):825–843, 1980. URL http://www.jstor.org/stable/2407990. 33

Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–285, 2010. URL http://onlinelibrary.wiley.com/doi/10.1002/gepi.20459/abstract. 40

Quanhe Yang, Muin J Khoury, Fengzhu Sun, and W Dana Flanders. Case-only design to measure gene-gene interaction. *Epidemiology*, pages 167–170, 1999. URL http://www.jstor.org/stable/10.2307/3703092. 40, 42, 44

Yaoming Yang, Anne-Marie Houle, Julien Letendre, and Andrea Richter. Ret gly691ser mutation is associated with primary vesicoureteral reflux in the french-canadian population from quebec. *Human mutation*, 29(5):695–702, 2008. URL http://onlinelibrary.wiley.com/doi/10.1002/humu.20705/full. 47

Nengjun Yi, Brian S. Yandell, Gary A. Churchill, David B. Allison, Eugene J. Eisen, and Daniel Pomp. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344, Jul 2005. doi: 10.1534/genetics.104.040386. URL http://dx.doi.org/10.1534/genetics.104.040386. 75

Jingkai Yu, Svetlana Pacifico, Guozhen Liu, and Russell L Finley. Droid: the drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9:461, 2008. doi: 10.1186/1471-2164-9-461. URL http://dx.doi.org/10.1186/1471-2164-9-461. 2

Noah Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin. Leveraging the hapmap correlation structure in association studies. *The American Journal of Human Genetics*, 80(4):683–691, 2007. 105

R. Y L Zee, J. Hoh, S. Cheng, R. Reynolds, M. A. Grow, A. Silbergleit, K. Walker, L. Steiner, G. Zangenberg, A. Fernandez-Ortiz, C. Macaya, E. Pintor, A. Fernandez-Cruz, J. Ott, and K. Lindpainter. Multi-locus interactions predict risk for post-ptca restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics J*, 2(3):197–201, 2002. doi: 10.1038/sj.tpj.6500101. URL http://dx.doi.org/10.1038/sj.tpj.6500101. 2, 38

Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul I W de Bakker, Gonalo R Abecasis, Peter Almgren, Gitte Andersen, Kristin Ardlie, Kristina Bengtsson Bostrm, Richard N Bergman, Lori L Bonnycastle, Knut Borch-Johnsen, Nol P Burtt, Hong Chen,

Peter S Chines, Mark J Daly, Parimal Deodhar, Chia-Jen Ding, Alex S F Doney, William L Duren, Katherine S Elliott, Michael R Erdos, Timothy M Frayling, Rachel M Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A Hitman, Thomas E Hughes, Bo Isomaa, Anne U Jackson, Torben Jrgensen, Augustine Kong, Kari Kubalanza, Finny G Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M Lindgren, Valeriya Lyssenko, Amanda F Marvelle, Christa Meisinger, Kristian Midthjell, Karen L Mohlke, Mario A Morken, Andrew D Morris, Narisu Narisu, Peter Nilsson, Katharine R Owen, Colin N A Palmer, Felicity Payne, John R B Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W Rayner, Matthew Rees, Jeffrey J Roix, Anelli Sandbaek, Beverley Shields, Marketa Sjgren, Valgerdur Steinthorsdottir, Heather M Stringham, Amy J Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M Watanabe, Michael N Weedon, Cristen J Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J Wareham, Ins Barroso, Andrew T Hattersley, Francis S Collins, Leif Groop, Mark I McCarthy, Michael Boehnke, and David Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–645, May 2008. doi: 10.1038/ng.120. URL http://dx.doi.org/10.1038/ng.120. 1, 104

Heping Zhang, George Bonney, et al. Use of classification trees for association

studies. *Genetic epidemiology*, 19(4):323–332, 2000. 48

Xiang Zhang, Shunping Huang, Fei Zou, and Wei Wang. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26 (12):i217–i227, 2010a. URL http://bioinformatics.oxfordjournals.org/content/26/12/i217.short. 40

Xiang Zhang, Feng Pan, Yuying Xie, Fei Zou, and Wei Wang. Coe: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *Journal of Computational Biology*, 17(3):401–415, 2010b. URL http://online.liebertpub.com/doi/abs/10.1089/cmb.2009.0155. 40

Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007. URL http://www.nature.com/ng/journal/vaop/ncurrent/full/ng2110.html. 53

Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006. 43, 47

Li-Bo Zou, Shuai Shi, Run-Ju Zhang, Ting-Ting Wang, Ya-Jing Tan, Dan Zhang, Xiao-Yang Fei, Guo-Lian Ding, Qian Gao, Cheng Chen, Xiao-Ling Hu, He-Feng Huang, and Jian-Zhong Sheng. Aquaporin-1 plays a crucial role in estrogen-induced tubulogenesis of vascular endothelial cells. *J Clin Endocrinol Metab*, 98(4):E672–E682, Apr 2013. doi: 10.1210/jc.2012-4081. URL http://dx.doi.org/10.1210/jc.2012-4081. 114

Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: genetic interactions create phantom heritability. *Pro-*

ceedings of the National Academy of Sciences, 109(4):1193–1198, 2012. URL http://www.pnas.org/content/109/4/1193.short. 137

Martin Zwick. An overview of reconstructability analysis. *Kybernetes*, 33 (5/6):877–905, 2004. URL http://www.emeraldinsight.com/journals.htm?articleid=876119&show=abstract. 47