Discipline of Mathematical Sciences

Queensland University of Technology

Bayesian Methodology for Genetics of Complex Diseases

Carla Chia-Ming Chen

Bachelor of Science, James Cook University

Master of Applied Sciences, James Cook University

A thesis submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy in the Faculty of Science and Technology,

Queensland University of Technology.

Principal supervisor: Professor Kerrie Mengersen

Associate supervisor: Dr Jonathan Keith

2010

Keywords

Bayesian, statistics, genetics, phenotype analysis, complex diseases, complex etiology, model comparison, latent class analysis, grade of membership, fuzzy clustering, item response theory, migraine, twin study, heritability, genome-wide linkage analysis, deviance information criteria, model averaging, MCMC, genome-wide association studies, epistasis, logistic regression, stochastic search algorithm, case-control studies, Type I diabetes, single nucleotide polymorphism, gene expression programming, logic tree, logicFS, Monte Carlo logic regression, genetic programming for association study, random forest, GENICA

Abstract

Genetic research of complex diseases is a challenging, but exciting, area of research. The early development of the research was limited, however, until the completion of the Human Genome and HapMap projects, along with the reduction in the cost of genotyping, which paves the way for understanding the genetic composition of complex diseases.

In this thesis, we focus on the statistical methods for two aspects of genetic research: phenotype definition for diseases with complex etiology and methods for identifying potentially associated Single Nucleotide Polymorphisms (SNPs) and SNP-SNP interactions.

With regard to phenotype definition for diseases with complex etiology, we firstly investigated the effects of different statistical phenotyping approaches on the subsequent analysis. In light of the findings, and the difficulties in validating the estimated phenotype, we proposed two different methods for reconciling phenotypes of different models using Bayesian model averaging as a coherent mechanism for accounting for model uncertainty.

In the second part of the thesis, the focus is turnd to the methods for identifying associated SNPs and SNP interactions. We review the use of Bayesian logistic regression with variable selection for SNP identification and extended the model for detecting the interaction effects for population based case-control studies. In this part of study, we also develop a machine learning algorithm to cope with the large scale data analysis, namely modified Logic Regression with Genetic Program (MLR-GEP), which is then compared with the Bayesian model, Random Forests and other variants of logic regression.

Acknowledgements

Rome was not built in a day, nor was it built by one man

The completion of this PhD is the result of the help of many people. My heartfelt thanks go to my principal supervisor, Prof Kerrie Mengersen, for giving me this opportunity to work on such an exciting and cutting edge topic three years ago. Kerrie is an amazing supervisor with a wealth of knowledge and patience to teach. She supports the active participation in a wide variety of academic activities, conferences, workshops and visiting scholars around the world. I am grateful for her passion, encouragement, guidance and empathy. A short acknowledgement is insufficient to express my gratitude.

Also a big thanks to Jonathan Keith, my day-to-day manager, who is currently a senior lecture with the University of Melbourne. Jon's abundance of knowledge, endless patience and availability had a significant impact on the completion of this work. Thanks also to my colleagues at the Queensland Institute of Medical research, Nick Martin, Dale Nyholt, David Duffy, Peter Visscher and others for their provision of the data and their contributions to concepts and journal articles editing for the first part of this thesis.

I would also like to thank Prof Katja Ickstadt and Holger Schwender for their collaborations on Chapter 6 and Chapter 8, and their generous support of my trip to Dortmund. Also I have great appreciation for Prof Christian Robert of Université Paris-Dauphine and Prof Peter Donnelly of Oxford University for the eye opening experience they provided. How can I not be thankful to the funding sources, the ARC linkage, NHMRC, School of Mathematical sciences QUT and Faculty of Science and Technology QUT, whom made this PhD possible and also their support for the astonishing overseas research trip.

I am grateful to the past and present colleagues and friends of the Brag group, especially those located in RM415, for their companionship, friendship, support and encouragement. You guys are awesome!!

Most important of all, my beloved husband, Glen and Him who we belief. Glen is a pillar of my life whom left his nearly-perfect job in Townsville and sold his recreation toys to support my dream. His selfless attitudes and endless love have made this possible. I owe him greatly and am eternality grateful.

Contents

List of Figures

List of Tables

Statement of Original Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: _____

Date: _____

List of Publications arising from this Thesis

Chapter 3: Chen, C. C.-M., Mengersen, K. L., Keith, J. M., Martin, N. G., and Nyholt, D. R. (2009). Linkage and heritability analysis of migraine symptom groupings: a comparison of three different clustering methods on twin data. *Human Genetics* 125, 591 - 604.

Chapter 4: Chen, C. C.-M., Keith, J. M., Nyholt, D. R., Martin, N. G., and Mengersen, K. L. (2009). Bayesian Latent Trait Modeling of Migraine Symptom Data. *Human Genetics* 126, 277 - 288.

Chapter 5: Carla C-M Chen, Keith, J., and Kerrie Mengersen (2010). From Phenotype to Genotype: reconciling approaches through Bayesian model averaging. *Journal of the Royal Statistical Society C* (Submitted: March 2010).

Chapter 6: Chen, C. C.-M., Mengersen, K., and Keith, J. M. (2010). Bayesian Method for Genome-Wide Association Studies: Review and Illustration (In Prep).

Chapter 7: Macrossan, P., Chen, C. C.-M., and Mengersen, K. L. (2010). Using gene expression programming with modified logic regression for the investigation of SNP interactions in large dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Submitted: February 2010).

Chapter 8: Chen, C. C.-M., Macrossan, P., Schwender, H., Nunkesser, R., Keith, J., and Mengersen, K. (2010). Methods for identifying SNP interactions: A Review on variation of Logic regression, Random Forest and Bayesian Logistic regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Submitted: March 2010).

1

Introduction

Since that pivotal moment in history about 145 years ago when Bohemian monk Gregor Mendel published the results of his pea breeding experiment, scientists have endeavoured to build a foundational understanding of hereditary genetics. Despite limitations in technology during the late 60's, the genetic dissection of plants and model organisms was successfully pursued [127]. The genetic study of human traits however, did not gain much ground until advancements in molecular and computational technologies during the 80's. Many of the successes which have occurred are due to the results of genome-wide linkage analysis and position cloning [230]. Linkage analysis is a method for identifying regions of the genome with higher-than-expected shared alleles among affected individuals within a family. This method has recorded tremendous successes in mapping genes in various diseases/disorders such as Duchenne muscular dystrophy, cystic fibrosis and

Huntington's disease. However, the successes are restricted largely to Mendelian disorders [230].

Most of the common disorders do not follow the Mendelian pattern of inheritance and are believed to have "complex" genetic make-up, therefore, in contrast to Mendelian disorders, these traits are often referred to as complex traits. A more formal definition of a complex trait is given in p370. [275], that is

> A trait that appears to have a genetic component but with no simple Mendelian pattern of single-gene inheritance; multiple genes, poly genes, environmental factors, age effects, and their interaction may be involved.

Although genome-wide linkage analysis has been carried out for many complex diseases, including Crohn's disease [126], migraine [212] and schizophrenia [261], the success is limited given that the mapped genes usually explain only a small fraction of the heritability. Furthermore, the lack of replication of the linkage results has suggested that linkage analysis is not suited for mapping complex diseases. [119] identified various other factors contributing to the lack of success, including the low heritability of most complex traits, the inability of the standard set of microsatellite markers to extract complete information about inheritance, imprecise definition of phenotypes and inadequately powered study designs. Most importantly, linkage analysis is less powerful in identifying large number of loci, each with moderate to little effect. Therefore, for a better understanding of the genetic architecture of complex traits, linkage analysis may no longer be a preferable option.

A practical, less expensive (compared with sequencing) approach, which still retains the scale of the genome-wide approach for gene mapping, is the genome-wide association study (GWAs). A GWAs is designed to identify associations between potential causal loci from hundreds of thousands of single nucleotide polymorphisms (SNPs) and traits. Since the completion of the Human Genome project [271, 272] and HapMap [269, 270], along with the reduction in the cost of genotyping, GWAs have become more prevalent. During the past five years, more than 300 replicated associations have been reported for 70 common phenotypes [66]. As more SNPs are included in the commercially available gene chips, more and even larger scale GWAs will emerge, e.g. WTCCC 2 project (`https://www.wtccc.org.uk/ccc2/`).

Without a doubt, the increasing number of larger scale GWAs increases understanding of the genetic dis-

section of complex traits. However, the virtual avalanche of data generated from GWAs has raised another array of challenges. These include the development and application of sound statistical methods for data analysis, the need for higher-level computational resources and quality interpretation of the findings. With all these in mind, the overall objective of this thesis is to

> *develop sound statistical methods to enhance understanding of the genetic architecture of complex traits.*

Given the constraints imposed by the time frame of PhD candidature, the areas focused upon in this thesis are confined to 1) the definition of phenotype and 2) methods for identifying epistatic effects, that is, gene-gene interaction effects.

Although phenotype definition prior to conducting genome-wide analysis may seem to be trivial and often ignored, without thorough consideration to defining the phenotype, the subsequent gene mapping results may not be meaningful. When phenotypes can be clearly asserted using biomarkers, the definition of phenotype is less relevant; however, for phenotypes that cannot be asserted using biomarkers, and also having complex clinical etiology, this becomes a very important issue. Examples of the latter conjuncture include various psychological disorders such as Alzheimer's disease, Parkinson's diseases and various types of headache. When carrying out genetic research for these types of disorders, ascertaining the phenotype often relies on the clinical diagnostic procedure, which is based on the fulfillment of symptom criteria. Various authors have argued that this method of ascertainment is not ideal for genetic research due to heterogeneity among affected and non-affected groups [287, 224].

An alternative method for deriving the phenotype is to use statistical approaches. Statistical methods for clustering and classification problems have been well developed. Given the true phenotype is not observable, a latent type of clustering approach may be more suited for this type of problem.

Various latent type clustering methods have been used for deriving phenotypes, such as latent class analysis (LCA), grade of membership (GoM) and item response theory (IRT). Given there are many different choices of clustering methods and the "true" phenotype is unobservable, it is uncertain if applying different clustering methods will affect the subsequent genetic analysis. Therefore, the first sub-objective in the first

part of this thesis is to

- investigate the effect of different statistical methods of phenotyping on the subsequent genetic analysis

This is addressed in Chapters 3 and 4 of this thesis, and twin migraine data are used for illustration.

Migraine is a common, painful and debilitating disorder with heritability ranging from 34 to 57%. The diagnosis of migraine is difficult, due to the absence of a clear biomarker, hence the diagnosis of the disorder depends on matching self-reported symptoms against criteria suggested by the International Headache Society [115]. Although a variety of independent genetic research has been carried out using this phenotype standard, under this phenotype definition, no common gene has been replicated across studies. Due to the overlap in symptoms between the subtypes of migraine, scientists have suspected that the two subtypes of migraine, migraine with aura and migraine without aura are actually not separate entities [211, 10, 164, 287]. [212] pioneered the use of LCA for phenotyping migraine, identifying potential linkage to chromosome 5q21 and replicating previous reported loci. Besides LCA, there are other clustering methods that can be used for deriving phenotypes. In Chapter 3, we compare the phenotypes derived from LCA, GoM and fuzzy clustering and the results of the subsequent linkage analysis. The phenotypes derived from LCA and fuzzy clustering are largely similar; therefore, the loci identified by the linkage analysis are also similar. In contrast, the results of GoM are very different from the other two approaches. This work has been published in *Human Genetics* [47] and presented as a poster presentation at the Indo-Australasia Biotechnology Conference, Brisbane 2007 and at GeneMapper, Brisbane 2007.

Using the same dataset, in Chapter 4, we focus on two different types of latent methods, LCA and IRT. Unlike LCA, IRT estimates the latent value without postulating clustering structure, but by direct association with the symptoms responses, which also takes into account symptom prevalence. A notable difference between Chapters 4 and 3 is that models of the former Chapter are proposed and compared in a Bayesian context. Furthermore, the use of the MCMC algorithm for parameter estimation provides credible intervals for each of the model parameters, which accounts for the uncertainty resulting from parameter estimation. Even though LCA and IRT have different underlying algorithms, phenotypes derived from these models are highly correlated, so the results of the subsequent analysis are in general agreement. This work has also

been published in *Human Genetics* [45] and was presented as a poster at the Biometrics conference, Coffs Harbour 2007, BioInfoSummer, Canberra 2007 and ISBA conference, Hamilton Island 2008.

Given that the phenotype derived from each model cannot be easily validated, and even though using statistical model comparison criteria gives insight as to how well the model fits the data, it does not provide full support to the phenotype estimated by a single model. Moreover, we note that the disagreement about the phenotype estimated from different models is mainly for individuals with the phenotype that is at the borderline of being a case or control. Therefore, methods for consolidating phenotypes estimated from different models may potentially be more advantageous than relying on a single model. This motivates the work of Chapter 5, the objective of which is to

- develop statistical methods for the integration of estimated phenotypes obtained from multiple models.

In this chapter, we propose two new methods to overcome the problems associated with defining phenotype classes and use Bayesian model averaging [142, 121] as a coherent mechanism for accounting for model uncertainty [121]. The idea of model averaging is to average the posterior distributions of different models, where the models are weighted according to model probability. The methods we propose here allow for the integration of estimated phenotypes obtained from multiple models both within and across phenotype classification approaches. The two models used for illustration in this chapter are latent class analysis (LCA) and grade of membership (GOM) and the proposed method for integration is similar to the"$\mathscr{M}$-open perspective" discussed in [24] and [121]. Moreover, the focus of the methods is not on the state parameters, but on the latent parameters. The methods are demonstrated using a real dataset on migraine and a simulated dataset obtained from the Genetic Analysis Workshop 14 [105]. This work is submitted to *Journal of the Royal Statistical Society C*. Thus, Chapters 3, 4 and 5 form the first part of this thesis.

In the second part of thesis, the focus is turned to statistical methods for identifying the epistasis effects in large-scale SNP data generated from GWAs. Epistasis is generally defined as the interaction between different genes that is suspected to be an important factor for the expression of a complex trait. Although there are different definitions of epistasis in the literature [50, 214], the definition of epistasis in this thesis remains more general; that is, the risk of having a phenotype can increase or decrease as a result of the

combination of two or more genes. The interaction among genes can be either additive or multiplicative.

Chapter 6 is a study of a Bayesian regression model with variable selection to identify the potentially causal loci. Because the number of variables in GWAs is excessively larger than the sample size, when considering all loci simultaneously, the results are often unreliable if there is no consideration of dimension reduction. Some excellent methods for dimension reduction have been developed within a Bayesian context, including variable selection and shrinkage.

The method used for variable selection in Chapter 6 is more aligned with [95], who introduced the use of a latent indicator for the identification of subsets of variables. Similar methods have been implemented for smaller datasets [90] and QTL studies [49, 297]. In contrast to these studies, the focus here is on application to larger-scale SNP data. The model is validated using simulated Rheumatoid arthritis data obtained from the Genetic Analysis Workshop 15, and tested on two real datasets. This work has been presented to the students and colleagues of Fakultät Statistik, Technische Universität Dortmund, Germany, and presented as a poster presentation at 17th International Conference on Intelligent Systems for Molecular Biology, Stockholm. This work is currently being revised for submission to an international refereed journal, such as *Computational Statistics and Data Analysis*.

Although the results of Chapter 6 are promising, the major drawback of the model is computational inefficiency, especially given the scale of GWAs. Therefore, in Chapter 7, we explore the use of the machine learning algorithm for identifying epistasis effects. The model proposed in Chapter 7 is an extension of Logic regression [235]. Logic regression is a hybrid approach that has a tree like structure comprised of Boolean expressions, such as AND and OR, and model fitting element. Various approaches have sprouted from the original logic regression [153, 245]. Even though these new variants improve the ability of LR in detecting interaction effects, limitations exist with respect to the number of factors which can be analysed at once within the written code, which currently stands at a maximum of 1000 SNP's. With these issues in mind, we propose an alternative method, which also builds on the framework of logic trees but has the advantage of the genetic expression programming algorithm. The proposed algorithm has shown promising ability in analyzing at least up to 30,000 SNPs within a reasonable time. This work has been submitted to *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

In Chapter 7, we consider different variations of logic trees for identifying interaction effects in association studies. Often these methods are introduced independently and it is uncertain how they differ from each other. In this Chapter, we also include the random forest (RF) for comparison. RF is also a tree-like algorithm, but has a very different morphological structure compared with the logic tree. Therefore, the purpose of Chapter 8 is to address the differences within the variations of logic regression as well as between the tree-like algorithms. Since it is also not clear how the tree-like algorithm compares with model based approaches, we also include the model proposed in Chapter 5 for comparison. This work has been submitted to *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

This thesis is written in fulfillment of the requirement for thesis by publication, such that Chapters 3 to 7 are comprised of journal articles. Therefore, each chapter contains some materials which may partially overlap with the content of Chapter 2. Moreover, the same migraine data has been used throughout Chapters 3 to 5 of this thesis and hence is repeatedly described in these chapters for the purpose of publications. Furthermore, each chapter has a self-contained bibliography, although for completeness these are merged into a comprehensive bibliography at the end of the thesis.

In summary, the overall objective of this thesis is to develop statistical methods for enhancing the understanding of genetic composition of complex diseases. To achieve this object, we explore the following aspects:

- phenotype definition
  - investigate the effect of different statistical methods of phenotyping on the subsequent analysis. (Chapters 3 and 4)
  - develop methods for reconciling the phenotypes estimated from different methods. (Chapter 5)
- methods for identifying the associated SNPs or SNP interactions
  - explore the potential of using Bayesian logistic model with variable selection (Chapter 6)
  - explore the potential of machine learning algorithms to improve the speed of computation (Chap-

ter 7)

– documenting the strengths and weaknesses of machine learning and model based approaches. (Chapter 8)

2

Literature Review

2.1 Introduction

The literature review is organized as follows. This chapter starts with the basics of human genetics (Section 2.2), then in Section 2.3, I review common practices for defining phenotypes for traits with complex etiology and then review various statistical methods for clustering. Because the next three chapters of this thesis implement genome-wide linkage analysis, we also include an overview of the underlying algorithms. The subsequent section includes a brief overview of the transition from genome-wide linkage analysis to genome-wide association studies. The last section of this chapter contains a discussion of the statistical methods which are commonly used for identifying important genes and gene-gene interaction effects. Be-

cause methodology for identifying epistasis effects is a new and popular area of research with more and more methods emerging in the literature, the methods reviewed here are confined to those that are commonly discussed and implemented.

2.2 Human Genetics

The aim of this section is to provide a short summary on human genetics. Because the contents of this section can be commonly found in genetic text books, the materials of this section are summarized from four main sources, biology by [147], introduction to human genetics by [106], the statistical methods in genetic epidemiology by [275] and genetic analysis of complex diseases by [108].

An understanding of subjects in molecular genetics, such as gene networks, can be useful in conjunction with the methods proposed in this thesis. The review of this topic is beyond the scope of this thesis, however, [148] provides an overview for gene networks.

2.2.1 Gene, Chromosome and DNA

The human body is made up of cells and the materials produced by these cells. The genetic material can be found in every cell of the human body, where a large proportion of information is contained in the chromosome, which is located in the nucleus of the cell, and a small part of genetic material is located in various mitochondria. The nucleus of the human somatic cells contains 22 pairs of autosomes and a pair of sex chromosomes, a total of 46 chromosomes. An individual inherits half of the chromosomes from the father and the other half from the mother. Because the autosomal chromosomes are arranged in pairs, they are called homologous chromosome pairs. The chemical structure of the chromosomes is deoxyribonucleic acid (DNA), which comprises the gene and encodes information for synthesizing both protein and RNA.

DNA is composed of three elements, a sugar, a phosphate and a base. There are four possible bases in DNA, which are pyrimidines adenine (A), guanine (G), purines cytosine (C) and thymine (T). A DNA sequence is often described as an ordered list of bases, each denoted by the letter of its name, e.g ATCCGA. Because a

single strand of DNA is unstable, it has a double-helical structure where two strands of DNA are arranged in anti-parallel orientation, and a hydrogen bond linking a base with its complimentary base, i.e A-T and G-C. The length of the sequence is different from chromosome to chromosome, and there are approximately 3×10^9 base pairs in the human genome. Even though there are a large number of base pairs in the human genome, a large proportion of the genetic sequence is actually similar between unrelated individuals. Genetic polymorphism is the term used to describe the difference in two genetic sequences between two individuals.

There are various types of genetic polymorphism, such as a single nucleotide polymorphism, a short tandem repeat and an insertion-deletion. A single nucleotide polymorphism (SNP) is when a base pair in a sequence is replaced by another base pair. A short tandem repeat (STR) is when a sequence of bases is repeated a different number of times between two individuals. An insertion-deletion polymorphism is when a base or a sequence of base is inserted or deleted from the original sequence. Therefore, the length of the chromosome can also differ among individuals.

The physical location of a stretch of DNA on a chromosome is called a genetic locus. At any particular locus, there can be different forms of the gene, which is called an allele. Because the autosomes are arranged in pairs, an individual also has a pair of alleles at the same locus, one from each chromosome. The combination of the two alleles is called the genotype of the individual at that genetic locus. For a biallelic gene with possible alleles a and A, there are three possible genotypes, AA, Aa and aa. Homozygosity is when two alleles are identical (i.e. AA and aa), and heterozygosity is when when the alleles are different (i.e. Aa).

2.2.2 Meiosis

The biological foundation for linkage analysis is meiosis, which is a process of producing gametes (i.e. sperm and egg cells) in sexual organisms. Human reproduction starts with the production of gametes, with the gamete of each parent uniting during the process of fertilization to forms zygote. A zygote is then developed into a human by the process of cell division. Unlike the somatic cells in the human body, which have 46 chromosomes, the gametes only have 23 chromosomes. Therefore, the human somatic cells are diploid and the gametes are haploid.

Meiosis consists of two stages: meiosis I and meiosis II. In meiosis I, each chromosome in a cell replicates to form two sets of duplicated homologous chromosomes. During meiosis I, physical contact between chromatids (arms of chromosomes) may occur which results in the formation of chiasmata. Chiasmata are physical manifestations of crossing over or recombination, which is the exchange of the DNA fragment in the adjacent homologous chromosome region. Chiasma occurs at least once per chromosome pair and the frequency of the recombination is not uniform through the whole genome. For instance, some areas of some chromosomes have higher rates of recombination (hot spots) and others have fewer recombination (cold spots).

After crossing over, cell division occurs to form two unique diploid cells that are different from the parent cells. This concludes the first stage of meiosis. In the second stage of meiosis, the cell division occurs again and genetic material is transmitted independently without recombination. At the end of meiosis II, two diploid cells become four haploid cells.

2.2.3 Genetic maps

Genetic markers are the loci where the locations on the chromosome are well established and are polymorphic among individuals in a population. The length of genetic markers varies, it can be a short DNA sequence or a long one, such as microsatellite markers. These markers are essential for developing dense genetic maps, which are important for finding out the locations of disease loci.

There are two type of maps, physical maps and genetic maps. Physical maps quantify the distance between genetic markers by counting the number of base pairs in between, whereas genetic maps arrange genetic markers by specifying the number of recombinations occurring between markers. Although both maps are essential for mapping disease loci, there are substantial variations in the estimates of an identical region from physical and genetics maps. Table 1.5 of [108] shows the discrepancy in the estimated length for different chromosomes using physical and genetic maps.

The measurement of distance in the physical map is often described in the thousands of kilobases (kbp, 1 kbp= 1000 base pair), whereas the unit for the genetic map is a centimorgan (cM). When two loci are one

Morgan (1 Morgan=100cM) apart, the expected number of recombinations between these two loci is one per meiosis. According to the genetic map, the length of chromosomes is different between male and female. The overall length of the autosomal chromosomes for males is 28.5 Morgans, and for females is 43 Morgans .

The availability of the genetic maps allow scientists to link a loci of unknown location to a genetic marker where the location on the chromosome is known. Suppose there are two loci on the same chromosome, with possible allele A, a at first loci and B, b at the second loci. Let the genotype of the father's chromosome is AB and ab at the mother's chromosome. There are four possible combinations from the meiosis: AB, Ab, aB and ab. If a gamete receives aB or Ab during meiosis, the loci is said to be recombinant. Conversely, if a gamete received AB or ab during the meiosis, it is said to be non-recombinant even if recombination occurred. Therefore if an odd number of recombination occur during meiosis, two loci are said to be recombinant. If an even number of recombinations occur during meiosis, two loci are said to be non-recombinant. The recombination fraction (θ) is the probability that two loci become recombinant during meiosis given the distance between two loci. The simplest probabilistic model for estimating the recombination fraction is the Haldane's map function. Let x denote the distance between two loci which is measured in cM, the recombination fraction is then

$$\theta(x) = 0.5(1 - exp(-0.02x)). \tag{2.1}$$

Two loci are linked if $\theta \leqslant 0.5$. Conversely, two loci are unlinked if $\theta \approx 0.5$. For loci on different chromosomes, the recombination fraction is always 0.5. This model, however, is oversimplified for discovering human disease genes. This is because the generation time in humans is relatively longer and the multi-generational pedigrees with a segregated disease or trait is rare. Moreover, the mating scheme can not be systematically designed and there are other ethical issues. Therefore, the linkage analysis in humans requires different assumptions, and hence more complicated statistical models are necessary. Section 2.3.2 reviews statistical models commonly used for linkage analysis in humans.

2.2.4 Epistasis

Epistasis is an interaction between two different genes or loci. The idea of epistasis has been around for more than 100 years. It was initially used by William Bateson to describe the distortions of mendelian segregation ratios that were due to one gene masking the effect of another [19]. The early definition of epistasis is similar to the concept of dominance, that is, a variant of one gene can prevent the variant at another loci from manifesting its effect. An example of this type of epistasis is the coat color of pigs [39]. Two loci, *KIT* and *MC1R*, are known to jointly influence the coat color of pigs. If the dominant allele (*I*) is present at the *KIT* loci, it masks the effect of *MC1R* loci and all pigs have white coats. When the recessive genotype (*ii*) is present in the *KIT* loci, the color of the pigs will depend on the variants at the *MC1R*. Pigs with the dominant allele (*E*) at the MC1R will have a brown color coat and pigs with the recessive genotype (*ee*) will have a black color coat. This definition of epistasis is similar to the concept often used by biologists and molecular geneticist when investigating the interaction between proteins [50].

Another definition of epistasis was suggested by Fisher in 1918 [86] as a deviation from the additive combination of different loci to their effect on a phenotype. Unlike Bateson's definition, Fisher's definition of epistasis is closer to the statistical definition of interaction, which departs from a specific linear model describing the relationship between predictive factors [50]. This view of epistasis is often adopted by population geneticists.

Another definition of epistasis relates to the molecular interactions present in proteins, such as if proteins operate within the same pathway, or consist of proteins which directly interact with one another. The discrepancy in the term 'epitasis' has resulted in the separation of three definitive categories by [214], which are functional epistasis, compositional epistasis and statistical epistasis. Functional epistasis describes the protein interaction and the latter two type of epistasis are equivalent to Bateson's and Fisher's definition of epistasis, respectively.

Besides the dominance interaction, for a biological interpretation, gene and gene can interact in other ways to influence the phenotype. [39] suggested two other types of statistical epistasis in QTL, co-adaptive and dominance-by-dominance epistasis, which may be interesting to biologists. These two types of epistasis

belong to the statistical epistasis defined by Fisher. The co-adaptive epistasis is when the homozygous genotype appear in two loci (e.g the genotype of two loci are *aa* and *bb* or *AA* and *BB*), which increases the level of the phenotypic trait. This type of epistasis is found to affect the hatch-weight of chickens. Dominance-by-dominance epistasis is when double heterozygous alleles in two loci resulted in a deviation of the phenotypic trait from the expected. For instance, a negative dominance-by-dominance epistasis is when the heterozygous genotype is at two loci, the phenotype is lower then expected . This type of epistasis is found in the maternal performance for offspring survival in mice.

[50] and [214] provide more thorough definitions and interpretation on epistasis. [39] and [108] review the importance of epistasis in genetic research of complex traits.

2.3 Phenotype Definition for Diseases with Complex Etiology

Before carrying out genetic research on any diseases/disorders, an essential step is to define the targeted disease/disorder. When a disease can be identified using a pathological test(s) for assurance, the procedure becomes straight-forward. However, it is common for a disease to have no objective markers or for practioners to be uncertain about the causes of a disease, for example various psychological disorders (e.g. Schizophrenia, obsessive compulsive disorders or depression), migraine and Alzheimer's disease. The most common method for identifying these diseases/disorders is to rely on medically recognised criteria. For example, migraine is a common and painful disorder, the diagnosis of which depends on classifying the self-reported headache characteristics using International Headache Society (IHS) published criteria [115]. These criteria were developed to standardise headache definition. The most common subtypes of migraines are migraine with aura (MA) and migraine without aura (MO). Tables A.1 and A.2 list symptoms for each subtype.

Various published studies have used this scheme for the identification of migraine phenotype and focused on either the MO or MA group [28, 257, 37, 286, 41, 136]. Certainly these criteria have refined the diagnosis of migraine and consequently have improved epidemiological research of the disorder. However, some scientists are questioning the homogeneity of the subgroups and the validity of using these for genetic

analysis. Although [238], [236] and [237] argued that MA and MO are distinct entities due to insignificant co-occurrence of MO and MA in population and twin pairs, other authors contradict this finding [160, 211, 287]. A study by [160] found that 42% of individuals who reported having migraine with aura often have migraine without aura. Furthermore, the Italian Headache Centre reported that 45% of families have members with both MA or MO [196]. [10] and [287] point out that IHS criteria may oversimplify the complex variability among sufferers and argue that there is overlap in the symptoms of the two subtypes of migraine. Furthermore, no gene that potentially differentiates these two subtypes has been successful replicated across studies [28, 257, 37, 286, 136, 41].

There are currently two main types of methods for identifying the phenotypic structure of the collective symptoms, one based on the use of statistical methods to obtain more homogenous groups and the other based on treating individual symptoms as separate phenotypic traits, i.e trait component analysis [TCA, 10]. In the following section, I review various methods that have been implemented for the identification of phenotypes. Theses methods are not limited to only migraine, but are relevant to a larger scope of genetic research.

2.3.1 Statistical Methods

Hierarchical Clustering Two common approaches to hierarchical clustering are agglomerative and divisive. The agglomerative hierarchical approach starts with each individual in a separate cluster and then merges two clusters at each step until there is only one cluster remaining or a stopping threshold is reached. In contrast, divisive clustering starts with all individuals in one cluster, then splits clusters at each step until the number of clusters is equivalent to the number of individuals or, again, a stopping threshold is reached.

Both of these approaches are often based on a measure of dissimilarity between individuals. The dissimilarity coefficient is the distance between two individuals. The most commonly used dissimilarity measures are Euclidean distance

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + ... + (x_{ip} - x_{jp})^2}, \tag{2.2}$$

its square, $d^2(i, j)$ or the Manhattan distance

$$d(i, j) = |x_{i1} - x_{j1}| + \ldots + |x_{ip} - x_{jp}| \tag{2.3}$$

where x_{ip} and x_{jp} are the observations for individual i and j for the pth factor respectively. After obtaining the dissimilarity coefficient matrices, splitting or merging is chosen to optimise some criteria, i.e. single linkage, complete linkage or maximum likelihood.

Hierarchical clustering may not be the best way to discover interesting groupings and is considered by some as more a visualization tool [281]. The disadvantage of using hierarchical methods is that they can never repair what was done in the previous steps of merging or splitting [143]. Several factors can result in different dendrogram clustering structure, such as different criteria used in optimization or changes in the data [281, 113]. In addition to the above problems, hierarchical clustering enforces hierarchical structure even if there is no such structure in the data [113]. Consequently, any inference based on hierarchical clustering should be treated with caution.

Partition Clustering (Relocating Clustering) In contrast to hierarchical approaches, partitioning methods often specify the number of groups (k) in advance. The K-means method [173] is the most commonly used partitioning method and is intended for quantitative variables.

The aim of the K-means method is to minimise the average dissimilarity measure between each observation and the mean within each cluster. In general, the steps involved for the K-means cluster algorithm are:

1. Partition the data into K initial sets at random or using some heuristic.

2. Compute the centroid (or seed points) for each current cluster $\{m_1, m_2, ..., m_K\}$.

3. Assign individuals to the closest cluster then update the centroids.

4. Repeat Step 2 and 3 until the assignment no longer changes.

Although the K-mean cluster is popular due to the speed of convergence, it is not guaranteed to give the

global optimum.

The clustering methods described above are frequently used for clustering individuals into affected and not-affected clusters. Therefore, phenotypic values derived from these methods are dichotomous.

Mixture models The concept of utilising a model-based approach for clustering was first introduced by [16]. In comparison to hierarchical and partitioning clustering approaches, the model-based approach gives better performance [87], yields the optimum number of groups within the data according to some criteria, and has the ability to handle outliers [87]. Moreover, the model incorporates a measurement of classification uncertainty which can be easily estimated using the expectation-maximization (EM) or MCMC algorithms.

In model-based clustering, data are assumed to be generated from a mixture of clusters or components, each represented by a probability distribution. Given observations $x_1, \ldots, x_n$, where n is the number of individuals, $f_k(x_i|\theta_k)$ is the density function of observation x_i belonging to component k given θ_k, where θ_k is the corresponding vector of parameters for that component, k.

The likelihood is then

$$L(\theta_1, \theta_2, .., \theta_k) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k) \qquad \pi_k \geq 0 \text{ and } \sum_k \pi_k = 1 \tag{2.4}$$

where π_k is the weight of each component k.

The early solution for the mixture of multivariate normal by [247] has some limitations, i.e. constant covariance matrices for different clusters, restriction to Gaussian distribution and inability to model noise. [16] suggested reparameterization to overcome these problems. For the first two limitations, they proposed a general framework for geometric cross-cluster constraints by parameterizing the covariance matrix, Σ_k, of the multivariate normal distribution through eigenvector decomposition in the form

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{2.5}$$

where D_k is the orthogonal matrix of eigenvectors, which determine the orientation of the clusters. A_k is a diagonal matrix whose elements are proportional to the eigenvalue and λ_k is scalar. A_k and λ_k specify the contour of the clusters, the former is the shape and the latter is the size of clusters. [16] then extented the mixture model to incorporate the poisson noise.

[88] has developed a software package, **MCLUST** for applying the Gaussian mixture model with the EM algorithm. **MCLUST** was written in FORTRAN and interfaced to the S-Plus and R software packages, which can be downloaded from the developer's website.

The mixture model for multivariate discrete data is also known as Latent class analysis (LCA), which is discussed in the following section.

Latent Class Analysis Latent class analysis is a multivariate technique which can be applied to clustering, regression and factor analysis. The classes are *latent* because they are not directly observed, but are identified based on a function of a set of observed variables. LCA was developed in the 1950s for dichotomous variables [161]. However, the potential and wide practical application of LCA only became evident after the introduction of more general latent class analysis and a simpler method of obtaining maximum likelihood estimates of the parameters in the 1970s [103]. LCA [103] is capable of dealing with both dichotomous and polytomous variables and more than one latent variable could be included in the model.

During the same period, the connection between LCA and clustering analysis was first introduced. However, the structure of latent class clustering was not developed until the late 1990s. Latent class clustering analysis has been used in a wide spectrum of epidemiology studies such as the studies of attention-deficit/hyperactivity disorder (ADHD) [282], migraine [211, 212], depressive syndromes [144], Alzheimer's disease [172, 193, 199] and investigating the nosologic structure of psychotic illness [145].

Suppose there are n individuals, J observed (manifest) variables and each variable has L_j levels of response, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$ and $l = 1, 2, \ldots, L_j$; Let y_{ijl} denote a binary response pattern of the ith individual to variable j with level l, and Y_i is a J by L_j matrix of subject i's response pattern. Assuming there are K latent classes within the latent variable, let λ_{kjl} denote the class conditional probability that an

observation in class k produces the lth outcome on the jth variable. Therefore, for each j, $\sum_l \lambda_{kjl} = 1$. Assuming local independence, the probability of an individual i in class k having a particular set of response patterns is

$$f(Y_i|\lambda_k) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{ijl}}$$

where p_k denotes prior probability of belonging to latent class k, $\mathbf{p} = (p_1, \ldots, p_K)$. Let Λ be a matrix containing all members of λ_{kjl}, the joint distribution for all J variables under the latent class model is

$$Pr(Y_i|\Lambda, \mathbf{p}) = \sum_{k=1}^{K} p_k \prod_{j}^{J} \prod_{l}^{L_j} (\lambda_{kjl})^{y_{ijl}}$$

LCA analysis can be carried out using the poLCA [167] package of R2.4.1. The parameters are estimated using the expectation-maximization (EM) algorithm [60]. The details of the EM algorithm for LCA are given by [167]. Unlike other models described in this report, the conditional probability of being kth class membership, given Y_i is estimated using Bayes's formula:

$$Pr(k|Y_i) = \frac{p_k f(Y_i|\hat{\lambda}_k)}{\sum_r p_r f(Y_i|\hat{\lambda}_r)}.$$

where $\hat{\lambda}_k$ is an estimated of outcome probability conditioning on class k.

Grade of Membership Grade of membership (GoM) is another popular statistical method which also fits into the latent class framework. GoM was first developed by Max Woodbury in the 1970s for medical classification and it has been widely used in the analysis of survey data in various disciplines ranging from determining the subtype of medical conditions such as mania [42], depression [266], and Alzheimer's disease [85], to identifying the genetic component in inheritable illness [178], and in social studies [80, 81].

GoM has a very similar algorithm to latent class clustering analysis, and the PhD thesis of [80] gives a detailed overview of the similarity and dissimilarity of these two models. The most fundamental difference between latent class analysis and GoM is that the latter model gives partial membership rather than full membership.

Let $g_i = (g_{i1}, g_{i2}, \ldots, g_{iK})$ be the latent vector of grade membership score for individual i belonging to component k and $\sum_{k=1}^{K} g_{ik} = 1$. Unlike LCA, the membership score of an individual is estimated directly from data. Let λ_{kjl} denote the probability of a positive response to level l of variable j for a complete membership of component k, $\lambda_{kjl} = Pr(x_{ijl} = 1|g_{ik} = 1)$ where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$. The parameter λ_{kjl} has to be greater than or equal to zero while for each j, the sum of λ_{kjl} across all levels is equal to one. Let y_{ijl} be a binary indicator variable for the response of individual i to level l of question j. The joint likelihood of GoM is

$$Pr(Y|\lambda, g) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\sum_{k} g_{ik}\lambda_{kjl})^{y_{ijl}}. \tag{2.6}$$

Equation 2.6 is maximized through iteratively optimizing with respect to one set of parameters while keeping the other set of parameters constant. This iterative procedure is referred to as the missing information principle. Details of the parameter estimation procedure are on page 68 of [179].

Missing values are important and yet common in genetic research. The missing values are the result of various causes. They can be generated by a random mechanism which is independent of the membership score, g_{ik}. In GoM, this type of missing data can be treated as unobserved and independent observations. In this case, y_{ijl} for the missing observation is set to be 0 for $l = 1, .., L_j$. Consequently, missing data are dropped in the calculation of the likelihood value. Another more complicated cause of missing data is a non-random process, such that certain items have a higher rate of missing data in a specific latent class. One way to deal with this problem is to increase the dimension of the data by adding an extra category called "missing" for each variable in the model. In this study, the missing data are assumed to be random and independent from the membership score; therefore, we applied the first strategy to handle missing values.

FANNY GoM Unlike the maximum likelihood approach, FANNY forms clusters based on a dissimilarity matrix. Here, the dissimilarity matrix is calculated using a contingency table. Considering two objects, i and j and the contingency table of i and j for variable p given in Table 2.1,

Table 2.1: The contingency table of object i and j.

$i \setminus j$	1	0
1	a	b
0	c	d

The dissimilarity between i and j is

$$d(i, j) = \frac{b + c}{a + b + c + d}.$$

Let ν denote the cluster ($\nu = 1, 2, \ldots, K$) and let $u_{i\nu}$ be the membership of object i to cluster ν. The objective of FANNY is to iteratively minimize the following criterion:

$$\sum_{\nu=1}^{K} \frac{\sum_{i,j=1}^{n} u_{i\nu}^2 u_{j\nu}^2 d(i, j)}{2 \sum_{i=1}^{n} u_{j\nu}^2}. \tag{2.7}$$

At each iteration, membership, $u_{i\nu}$ has to be greater or equal to zero for all $i = 1, \ldots, n$ and membership for i has to sum to 1 among all clusters.

Although this method has an uncomplicated algorithm, it is not commonly used in genetic phenotyping. [110] implemented this approach for subtyping schizophrenia and [141] used this approach for phenotyping anxiety disorder prior to linkage analysis.

Traditionally, the phenotype definition relies on either grouping patients based on the criteria proposed by medical associations or by frequentist statistical methods. To date, there is limited literature on applying Bayesian statistical models for phenotype definition prior to genetic analysis. Therefore, this research is different from others in developing Bayesian statistical methods for phenotype definition.

Item Response Theory Item response theory (IRT), also known as the latent trait analysis. It is a class of popular statistical methods that are commonly used for modeling psychological and educational survey

responses. The model assumes an underlying continuous latent value which has direct influence on the response to survey questions. This underlying continuous latent value is unobservable, which represents the ability of an individual in school tests or a propensity score for an individual have a diseases given the items, which are measurable.

IRT is a collective term for many different models which can be categorized based on the type of dependent variable. For instance, the partial credit models, the graded response models and the sequential scale models are designed for polytomous data and the Rasch model, the two-parameter logistic model and the three-parameter logistic model are mainly for dichotomous data. Because medical symptom data are often dichotomous, the latter models are more frequently implemented in genetic research. The examples of using IRT for phenotype definition include [73] explore the genetic and environmental influence on the timing of pubertal change with the two-parameter logistic model and the same method was also used by [74] and [290] for the analysis of multiple symptom genetic data. Therefore, in this review, we focus on the Rasch, two- and three- parameters models.

The IRT models entail three assumptions, which are unidimensionality, conditional independence and monotonicity. Unidimensionality refers to the existence of a one-dimensional, unobservable quantity associated with each respondent in the sample which describes the individual's propensity score to the items (symptoms). Conditional independence means that given an individual's propensity score (or called latent trait), the item responses are independent. This assumption is also associated with the propensity (latent trait value) and it states that individuals with high propensity (latent trait) are more likely to endorse the items than the ones with a smaller propensity.

To formalise the IRT, let x_{ij} denote the response of individual $i, i \in \{1, \ldots, N\}$ to item j. Let θ_i be the propensity (latent trait) value and $P_j(\theta_i)$ be the probability that x_{ij} is positive given the latent value, θ_i. The probability $P(\theta_i)$ is often referenced as the item response function (IRF). The main difference between the Rasch, two- and three- parameter models is in the number of parameters incorporated in IRF: the Rasch model has one parameter, and the other two models have two and three parameters respectively. The two-parameter logistic model is now described, follow by the Rasch and the three-parameter models.

Two-parameter Logistic Model

The two-parameter logistic model [27] has a similar form to the ordinary logistic model, that is

$$\text{logit}(P_j(\theta_i)) = \alpha_j(\theta_i + \beta_j) \tag{2.8}$$

$$P_j(\theta_i) = \frac{1}{1 + \exp(-\alpha_j(\theta_i - \beta_j))} \tag{2.9}$$

where α_j is the slope of the item response function, also called the item discrimination. This is a measure of how much information an item provides about the latent value θ. The parameter β_j is the intercept of the IRF. In educational settings, β_j is an indicator of the item difficulty and in medical applications, the parameters α and β can be interpreted as the measures of the symptom prevalence. The parameter α indicates the prevalence of the symptom in the affected individuals while the product of α and β provides an insight into the symptom prevalence in the overall population. For instance, if $\alpha_j = \alpha_{j+1}$, when β_j is larger than β_{j+1}, it indicates that symptom j is more prevalent in the population than symptom $j + 1$.

The Rasch Model

The Rasch model [225], also called a one parameter logistic model, assumes that all items have the same discrimination ability, so that α_j is fixed for all j. Common values for the discrimination are $\alpha = 1$ and $\alpha = 1.7$; under this setting, IRF is similar to the cumulative density function of the normal distribution. Although this model is suited for various educational or psychological settings, it is less relevant to the setting of disease, where individual symptoms often differ in prevalence. Therefore, the Rasch model is not implemented for phenotyping.

Three-parameter model

Besides having extra parameters, the three-parameter model is designed with a different scenario in mind. In the previous models, the response function $P_j(\theta) \to 1$ as $\theta \to \infty$ and $P_j(\theta) \to 0$ as $\theta \to -\infty$. However, in the education examples, the latter assumption is not reasonable if j can be correctly 'guessed'. Therefore, [27] developed a generalization of the two-parameter model that allows the IRF to have an asymptote different from zero. The IRF of this three-parameter model is

$$P_j(\theta_i) = \gamma_j + \frac{1-\gamma_j}{1+\exp(-\alpha_j(\theta_i-\beta_j))} \tag{2.10}$$

where γ_j is the probability that an examinee correctly guessed the answer of item j. In the application of medical research, this parameter is similar to probability of an individual having the symptom j without actually having the diseases/disorder. This is a common scenario for psychological disorders, where some symptoms are prevalent in the controls.

2.3.2 Methods for Linkage analysis

Linkage analysis is a statistical method to determine the approximate location of the phenotype locus with respect to some genetic markers, where genetic markers are at known locations on the chromosomes and contain multiple alleles. Linkage analysis is based on the concept of co-segregation between the disease and marker gene [275].

Linkage analysis can be divided into parametric and nonparametric methods. In nonparametric linkage analysis, the assumptions of penetrance and allele sharing are not required. In Chapter 10 of [108], the author listed four major advantages of model-based linkage analysis. Firstly, if the assumed genetic model is correct, then the model-based approach is more powerful than any nonparametric method. Moreover, model-based linkage analysis exploits all genotype and phenotype information within a pedigree, and also provide an estimate of the recombination fraction between markers and disease alleles and a statistical test for linkage and gene locus heterogeneity.

However, parametric linkage analysis is not suitable for complex disorders, whose manifestation depends on the joint action of various genes and perhaps environmental agents. Furthermore, in order to construct parametric linkage models, variables such as the model of inheritance, the trait and marker allele frequencies, the penetrance values for each disease genotype, phenocopies and the sex-specific recombination fractions are required to be specified in advance.

These are difficult to specify for complex disorders. It is important to note that a parametric linkage test is a

test of all assumptions; the failure of linkage analysis could be due to a misspecification of model parameters rather than a lack of evidence for linkage [108].

Due to the complexity of the inheritance pattern of migraine and the lack of the knowledge about the parameters required for parametric linkage analysis, we review two nonparametric linkage analysis methods, namely affected sib-pairs and variance component linkage analysis in this report. For more in-depth knowledge of various linkage analyses, the recent book published by [108] provides a comprehensive description of both parametric and nonparametric analysis.

Affected Sib Pair Affected sib pairs (ASP) analysis is the most commonly used nonparametric linkage analysis for dichotomous traits. The most important element of ASP is the probability distribution for number of alleles shared identity by descent (IBD). Two individuals are said to share an allele IBD in a given marker is when a common ancestor in the pedigree passes one of its two alleles in this locus to both individuals.

Let $z_k(x)$ denote the probability of k alleles being shared between related pairs at marker locus x. For a random sibling pair, these values are expected to be 1/4, 1/2 and 1/4 for $k = 0, 1, 2$ respectively and for a monogenic disease. If there is linkage between a marker and the disease locus, the observed and expected distributions of allele sharing will be significantly different. This can be tested using a χ^2 test with two degrees of freedom.

Alternatively, we can compare the average IBD sharing in the sample of pairs with the expected value of 0.5. [29] found that this approach performs better than different level of expected values under a large range of genetic models.

Variance Component Variance component linkage analysis involves partitioning total variance into various components. For linkage analysis, the aim is to separate the unmeasured genetic variance from unmeasured non-genetic variance. [6] developed a mixed effect variance component approach for quantitative traits which can be used for general pedigree data.

Let X_i denote the quantitative trait value for the ith individual and let z_{ik} be the kth covariate value for subject i. A general model is:

$$X_i = \mu + g_i + G_i + \sum_{k=1}^{s} \beta_k z_{ik} + \varepsilon_i \tag{2.11}$$

where μ is the overall mean, G_i is a random polygenic effect and g_i is a fixed and unobserved genetic component where alleles A and a affect the trait as follows:

$$g_i = \begin{cases} a & \text{if individual i has unobserved genotype AA} \\ d & \text{if individual i has unobserved genotype Aa} \\ -a & \text{if individual i has unobserved genotype aa} \end{cases}$$

The term β_k is the covariate effect and ε_i is the residual for subject i. Both of these parameters are uncorrelated with the genetic factors. Since the average effects of G_i, g_i and ε_i can be included in the overall mean, the expectation of these factors are zero.

Assuming the identity-by-descent sharing of a pair of individuals i and j is observable (denote π_{ij}) then the first moment of Equation 2.11 becomes

$$E(X_i) = \mu + \sum_{k=1}^{s} \beta_k z_{ik}$$

Given the genetic variability of two individuals, i and j can be decomposed into additive and dominance components, i.e. $\sigma_g = \sigma_a + \sigma_d$. Let p and q denote the gene frequency of A and a, then

$$\sigma_a = 2pq(a - d(p - q))^2$$
$$\sigma_d = 4p^2q^2d^2.$$

When $i \neq j$, the second moment of model is

$$Cov(X_i, X_j) = \pi_{ij}\sigma_a^2 + \Delta_{ij}\sigma_d^2 + \Phi_{ij}\sigma_G^2 \tag{2.12}$$

where Δ_{ij} is the probability that i and j sharing two genes at the major locus IBD and Φ_{ij} is the coefficient of the relationship between i and j.

Often the typed markers do not have a direct effect on the phenotype, therefore [6] extend the model to include data from linked markers by considering the cosegregation of trait and marker allele. For a pair of relatives, if there is a linkage, then there is a linear regression relationship between the square difference of the pair's trait value, i.e. $(X_i - X_j)^2$ and the estimated proportion of genes IBD at marker allele. Assuming the $E(X_i^2) = E(X_j^2)$,

$$\begin{aligned} E(X_i - X_j)^2 &= E(X_i^2) + E(X_j^2) - E(X_iX_j) \\ &= 2Var(X_i) - 2Cov(X_i, X_j). \end{aligned}$$

and using the same notation as above,

$$Cov(X_i, X_j) = f(\theta, \pi_{ij})\sigma_a^2 + g(\theta, \Delta_{ij})\sigma_d^2 + \Phi_{ij}\sigma_G^2 \tag{2.13}$$

where θ is the recombination fraction and $f(\theta, \pi_{ij})$ is associated with the additive major-gene component and the value of the function depends on the kinship. Table 1 of [6] details the value of $f(\theta, \pi_{ij})$ for different degrees of kinship. The second function, $g(\theta, \Delta_{ij})$, is the dominance component which is often ignored in linkage analysis because it is much smaller than the additive component and can only be assessed in bilineal relatives [7].

[7] compared three different parameter estimation approaches: maximum likelihood estimation assuming traits have a multivariate normal distribution, quasilikelihood and regression procedures. Using simulation studies, they found the last two procedures provide unbiased estimation of additive genetic effect. In contrast,

maximum likelihood methods are less robust to error in the specification of the distribution of residual variance, and the estimates were downward biased for small samples.

The variance-component linkage analysis has been further developed in various aspects. [57] have extended the model to incorporate longitudinal family data and genetic marker information in a quasilikelihood framework. [194] extended the current method to simultaneously obtain estimates for additive effects of multiple loci on phenotype variation and additive interaction effects among loci (epistatic effect). [30] also extended the variance component linkage model to allow application of full pedigree data.

Multipoint QTL analysis

The previous section was confined to consideration of a sequence of pairwise comparisons between the trait and each of the marker loci. Multipoint linkage analysis is useful for establishing the chromosomal order of a set of linked loci and resolves the problem caused by the limited informativeness of markers.

Multipoint linkage analysis is particularly computationally demanding for computing likelihood values [4]. Traditionally, the Lander-Green algorithm [156] is used for a large number of loci and small pedigree and a peeling-based algorithm for a few loci and large pedigree [4]. For cases with a large number of loci and a large pedigree, the Lander-Green algorithms can be applied, but some sampling methods are required.

[107] applied MCMC methods to calculate Monte Carlo estimates of the likelihood. This was previously infeasible due to the large and complex pedigree. [117] described the implementation of the reversible jump MCMC sampler [104] to estimate the map position of the linked QTL, the effects of frequencies of all QTL and other model parameters, such as residual variance. Therefore, instead of searching a small region of chromosomes for evidence of linkage, a joint analysis can be performed when a large number of markers throughout the genome is available. [117] found that RJMCMC allows a more natural modeling of genetic heterogeneity due to not forcing the genetic model to be the same across all families.

[30] developed a computer package called SOLAR for linkage analyses of multivariate quantitative traits and discrete traits using a threshold model and mixed traits. This program also incorporates gene $\times$ gene

and gene × environment interactions.

MCMC methods have been applied for mapping multiple QTL for complete and incomplete genotypic data [133, 252] and various types of pedigrees [252].

2.3.3 Overview of Bayesian Model Averaging (BMA)

Bayesian model averaging (BMA) provides a coherent mechanism to account for model uncertainty [121]. The idea of BMA is to average the posterior distributions of different models, where the weight for each model depends on the posterior model probability. [175] and [223] noted the use of BMA can improve predictive performance.

Various works have been published on the methods of BMA [142, 175, 222, 223, 121]. In particular, [121] provides a thorough overview of the history and challenges of BMA and provides solutions.

BMA has been widely applied to different models, and [120] provides a summary of the methodologies that have been implemented with BMA and lists corresponding software for carrying out the analysis. Although the use of BMA in genetic research is not as common as in some other areas of science, a few published works have incorporated BMA in the analysis. For instance, [295] applied BMA for gene selection and classification of microarray data. [9] further extend the former research by incorporating iterative BMA for survival analysis. The use of BMA has also been implemented in the study of phylogenetics [215].

Let Δ denote a quantity of interest (in the area of genetic studies, Δ can be treated as a phenotypic trait of interest). Given a data set D, the posterior distribution of Δ is

$$p(\Delta|D) = \sum_{s=1}^{S} p(\Delta|M_s, D)p(M_s|D) \tag{2.14}$$

where M_s is the model s of all models considered, $s = 1, \ldots, S$. Using Bayes theorem, the probability of M_s given data set D becomes

$$p(M_s|D) = \frac{p(D|M_s)p(M_s)}{\sum_l p(D|M_l)p(M_l)} \tag{2.15}$$

where

$$p(D|M_s) = \int p(D|\theta_s, M_s)p(\theta_s|M_s)d\theta_s. \tag{2.16}$$

The former is the marginal likelihood of model M_s, where θ_s denotes the model parameters of model s and $p(D|M_s)$ is the marginal likelihood. Therefore, Equation 2.15 can be seen as providing weights for the predictions of different models in Equation 2.14. During the early introduction of BMA, it was not as well accepted as model selection due to the difficulties associated with a potentially infinite number of models ($S \rightarrow \infty$) to be included in Equation 2.14, the choice of priors on the models, and the computational difficulties in the estimation of the marginal likelihood. Although the former concern is less relevant in this thesis, various methods have been developed to overcome this problem, such as exploring the model spaces stochastically via MCMC approaches [96, 222, 99]. Moreover, [142, 83] listed various methodologies for approximating the marginal likelihood when it is intractable.

2.4 From linkage analysis to genome wide association studies

In the previous section, I reviewed various methods commonly used in genome-wide linkage analysis. Although linkage analysis has had some success in mapping genes for Mendelian diseases, such diseases are rare [119]. Various common diseases/disorders have a genetic component have been identified by familial aggregation, but they do not follow the Mendelian pattern of inheritance. Such diseases/disorders include Type I and II diabetes [273], cardiovascular diseases [273], obesity [276] and various psychological disorders [217]. These common diseases/disorders often have complex genetic architecture, thus they are often referred to as complex traits. Complex traits are presumably derived from multiple genetic and non-genetic effects, as well as the interactions among genes and between genes and the environment.

Although linkage analysis has been carried out for mapping complex traits, the success is limited. [5]

reviewed 101 whole-genome scan linkage studies in 31 different complex human diseases and found that most of the studies (~66%) do not have a significant result when using the threshold proposed by [157]. Moreover, they also noted that the findings for the same diseases are often inconsistent among studies. Although the sample size may be an important factor for the success of a linkage scan [5], [229] show the sample size required to achieve a relative high power may not be feasible. For instance, when the genetic risk ratio is less than 2, the number of families needed in order to achieve 80% power is well over 2000.

Another important disadvantage of linkage analysis is that it has low power for identifying multiple low-penetrance variants on a phenotype [229, 119]. It is long noted that the genetic component of complex traits is oligogenic (a few genes, each with moderate effects) or even polygenic (many genes, each with small effects) [224]. Considering the limitations of the linkage analysis, there is a need for an alternative method for understanding the genetic architecture of complex traits.

Since the completion of the Human Genome Project [271, 272], genetic epidemiology has entered an era of single nucleotide polymorphisms (SNP) and the realization that the human genome is organized into haplotype blocks (Linkage disequilibrium, LD) [224]. The International HapMap Consortium [269, 270, 130] has recently completed characterization of over 3.1 million human SNPs with a SNP density of approximately one per kilobase. The 3.1 million SNPs is approximately 25-35% of all the 9-10 million common SNPs (with minor allele frequency $\geq$ 0.05). The completion of characterization of the linkage disequilibrium (LD) pattern across these SNPs provides the most informative subset of 'tagging' SNPs. Subsequently, the genome-wide association study (GWAs) is made possible. The initial GWA scans had 10,000 SNPs with improvements in genotyping technology, the new Affymetrix Genome-Wide Human SNP Array 6.0 features nearly 1 million SNPs.

With the reduction in the cost and commercial availability of SNP genotyping comes large scale GWAs. The most referenced work to date is the study published in 2007 by the Wellcome Trust Case-Control consortium [WTCCC, 273]. This study contains 14,000 cases of 7 common diseases (including bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, Type I and II diabetes) with 3000 shared controls. It was the largest at its time. Since then, more and more large scale GWAs have emerged. In the past five years, in *Nature Genetics* alone, there are 76 published studies related to GWAs, the variety

of phenotypes including Parkinson's [253], lung function [226], erythrocyte phenotype [92], obesity [276], Crohn's disease [17] and hematological parameters [258]. From a statistical and computational perspective, the main challenges for GWAs is the finding of informative markers among hundreds of thousands or even million of markers with relatively small sample size (that is compared with the number of parameters). In the remaining chapter, I review methodologies that have been implemented/developed for association studies. Note that although many methods have been proposed, only a few have been tested in the genome-wide scale of study. Therefore, we enlarge the scope of the review to methods for association studies, which includes GWAs and candidate gene search.

The remainder of this chapter is in two sections: methods for detecting single marker effects and methods for detecting multilocus effects. The latter section is arranged into 1) model-based, 2) data mining or machine learning approaches and 3) two-stage approaches. Note that three review papers published in Nature and Nature Review Genetics provide valuable reviews on the statistical methodologies used in GWAs. [15] gives a comprehensive tutorial on some of the frequentist methods for population association study. [262] reviewed Bayesian methods for single-SNPs testing in GWAs. [51] gives a great overview on some of the methods used for the detection of the gene-gene effects and related computer softwares. In light of this research, the review here is concentrated on the statistical aspect of the methods.

2.5 Methods for association Studies

2.5.1 Single Marker effect

Of all the methods available to date, the most widely implemented approach is the SNP-by-SNP searching algorithm. For case-control studies, the most natural analysis of SNP genotype and case-control status is the use of a 2-by-3contingency table that contains the count of case-control status and count of genotype (e.g. AA, Aa and aa). The common choice are either Pearson's χ^2 test or Fisher's exact test. Even though the latter method is more computationally demanding, it does not depend on the χ^2 approximation. Moreover, Fisher's exact test is implemented in the R genetic package. In this aspect, [15] suggested Fisher's exact

test is better for GWAs than Pearson's χ^2 test.

Table 2.2: Example of 2×3 contingency table of case-control study

	aa	*aA*	*AA*	Sum
Case	r_0	r_1	r_2	r
Control	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

For a complex trait, the contributions of individual SNPs are thought to be additive to the disease risk, that is, the number of disease alleles correlates to the risk of having a disease, therefore simple tests on the contingency table are not as powerful as other tests [15]. Furthermore, these simple tests are only appropriate when Hardy-Weinberg equilibrium (HWE) holds [240] and do not lead to interpretable risk estimates. The Cochran-Armitage trend test [12] is an alternative testing method which is more robust, conservative and does not rely on the assumption of HWE [15].

Let *a* and *A* denote two marker alleles and suppose each person has one of three possible genotypes, *aa*, *aA* and *AA*. Table 2.2 is an example of a 2×3 contingency table for a case-control study with those marker alleles, where (r_0, r_1, r_2) and (s_0, s_1, s_2) are the number of genotypes, *aa*, *aA* and *AA* observed in cases and controls, respectively. Let i denote the number of *A* alleles in a genotype, and $i = \{0, 1, 2\} = \{aa, aA, AA\}$. Let $\phi = \frac{r}{n}$ be the proportion of cases, the Cochran-Armitage trend test statistic is then

$$T^2(x) = \frac{n^{-1} \sum_{i=0}^{2} x_i (s r_i - r s_i)^2}{n \sigma_1^2(\phi)} \tag{2.17}$$

where

$$\sigma_1^2(\phi) = \phi(1-\phi)[\sum_{i=0}^{2} x_i^2 p_i - (\sum_{i=0}^{2} x_i p_i)^2] + \phi^2(1-\phi)[\sum_{i=0}^{2} x_i^2 q_i - (\sum_{i=0}^{2} x_i q_i)^2] \tag{2.18}$$

where $x_0 = 0$, $x_1 = x$, $x_2 = 1$ and $0 \leq x \leq 1$. The value of x is required to be specified *a priori* based upon the model of interest. For instance, three possible genetic models are recessive, additive and dominant, and therefore x is often set to 0, 0.5 and 1, respectively. Variables p_i and q_i of Equation 2.18 are the probability of being a case or a control given the genotype is $i, i = \{aa, aA, AA\}$, which are often not known. [304] summarized three different estimators for these variables, and the most common choice for p_i and q_i is $\hat{p}_i = \hat{q}_i = \frac{n_i}{n}$, which has been implemented in [240] and [256].

Under the null hypothesis, the test statistic, $T^2(x)$, has an asymptotic χ^2 distribution with 1 degree of freedom.

The Cochran-Armitage trend test is the most commonly employed model in GWAs. Examples of implementation of this method include: Breast cancer [72], coronary artery disease [239], type I diabetes [284] and Parkinson's disease [253].

Departing from the conventional χ^2 approach, a more advanced method for identifying single SNP effects is by implementing logistic regression for a case-control study. Let π_i be the probability that individual i is a case,

$$\log(\frac{\pi_i}{1-\pi_i}) = \mu + \sum_{j=0}^{2} \beta_j x_{ij} \tag{2.19}$$

where μ is the population mean and x_{ij} is a binary indicator variable for genotype, taking the value of 0 or 1. The effect of the SNP is then determined by testing the null hypothesis, $\beta_0 = \beta_1 = \beta_2$ against the hypothesis that at least two β are different via the likelihood ratio test.

Logistic regression is a very common approach used for detecting single SNP effects in GWAs. Examples of such studies include acholic liver disease [277], ulcerative colitis [13], systemic lupus erythematosus [111] and some clinically relevant hematological parameters [258].

For continuous phenotypic traits, the natural choice of statistical tool is linear regression and analysis of variance (ANOVA) [15].

The multiple testing problem is the major concern for the SNP-by-SNP search algorithm. Without a proper adjustment of the power, it is likely to have false positive results (i.e. Type I error). The frequentist method for controlling the false positive is by controlling the significance level, α. The usual choice of α is 0.05, which implies that the probability for being false positive in all the tests carried out is less than 5%. Here we list three approaches for adjusting for the multiple testing problem, but various other methods have been proposed for controlling issues derived from multiple testing in association studies and the comparison of

various methods has been discussed in several recent publications [15, 227, 77, 197].

Bonferroni Correction Bonferroni correction is an often discussed example of controlling α level. If n SNPs are tested for association, the Bonferroni corrected α for each test is $\alpha' = \alpha/n$. For GWAs, the value of n can be substantially large and depending on the SNP chip used, n can vary between 500,000 to approximately 1,800,000. Thus Bonferroni correction can be overly conservative and not suitable for tightly linked SNPs [15].

Permutation Permutation testing is a simulation based resampling method, which controls the issues of multiple testing by comparing observed p-values with p-values estimated by repeated perturbation of the data and evaluating how often the observed p-value can be obtained by chance [227]. There are various method for obtaining the permuted p-values. For association studies, a sample of p-values can be obtained by keeping the individual genotype unchanged while the phenotype of individuals are replaced with randomly generated values. This method ensures the 'new' data contains the observed LD structure, but shows no association in the phenotype. Although the permutation test is robust, it is computationally intensive.

False Discovery Rate False discovery rate (FDR, [22]) is comparatively less computationally intensive, yet provides increased power over Bonferroni correction [203]. The aim of FDR is to estimate the desirable error rate to control the expected proportion of error among the rejected hypotheses. This criteria is designed to reduce the number of errors made and the probability of false rejection.

Suppose there are n hypothesis testings, $H_1, H_2, \ldots H_n$ and let $p_1, p_2, \ldots, p_n$ denote the corresponding p-values. These p-values are then arranged from the most significant to the least significant, that is $p_{(1)} \leq p_{(2)} \ldots \leq p_{(n)}$. At a preset value α, let

$$k = \arg\max_i \{p_{(i)} \leq \frac{i\alpha}{n}\} \tag{2.20}$$

where i is the order of p-value; then reject all $H_{(i)}$ where $i \leq k$.

This procedure is versatile and can be simply modified to accommodate different genetic problems. For instance, the weighting of p-values can be stratified according to prior knowledge. [234] propose to stratify the weighting of p-values based on the results of linkage analysis. In other words, p-values of loci show suggestive linkage are upweighted and conversely, the p-values of less informative regions are downweighted.

2.5.2 Multiple SNPs Effect

A SNP-by-SNP searching algorithm is optimal if SNPs are widely spaced (have little to no LD strcuture) in the data and one of the typed SNP is exactly causal. However, this is a rare event. The other disadvantage of considering only a single SNP is that it potentially neglects the joint effect of multiple SNPs, where some variants may have little marginal effect, but the effect of the variant is more obvious when it is altered or highlighted by another variant or variants. Furthermore, such interaction effects have been suspected for the expression of complex diseases. Therefore, a superior approach is the multiple SNPs test which examines the association of a phenotype with multiple SNPs simultaneously.

Statistical methodologies for detecting multiple SNP effects (both including and excluding epistasis effects) is a popular topic on which a large amount of literature has emerged in the last decade. The early methods focus on linkage analysis. However, as SNP data becomes more widely available, methods are evolving for association studies.

The most prominent paper for identifying multiple loci effects is by [181]. In their study, they simulated three plausible two-locus effects and compared three different searching strategies for identifying the interaction effect in different plausible scenarios. The first scenario is when the genetic loci have multiplicative effect, that is the odds of disease increases in a multiplicative fashion, within and between loci. For example, for two diallelic loci (denoted a and b), let the upper case of each letter be the disease allele, then having either the A or B allele increases the risk by $(1 + \theta_1)$ or $(1 + \theta_2)$ fold, where θ_1 and θ_2 are the risk increment due to disease alleles A and B. The second scenario is a statistical interaction with explicit marginal effects, that is, at least one of the disease alleles must be present at each locus for the odds to increase. Furthermore, the presence of each additional disease allele will increase the disease risk by $(1 + \theta)$ fold. The last scenario

represents the threshold model effect. Like scenario 2, at least one of the disease allele must be present at each locus for the odds to increase, however, the disease risk is thereafter constant regardless of the number of disease alleles present in the genotype combination.

The three searching strategies included in the study are 1) locus-by-locus search, 2) exhaustive pari-wise search (i.e full search) and 3) a two stage approach. Their results show that the interaction base searching algorithm is more powerful than locus-by-locus search for all three scenario. However, they also conclude it is difficult to determine a single best searching method for identifying multilocus effects given that the number of interaction loci and the form of interaction can vary from trait to trait.

Various studies have since emerged for detecting multilocus effects. Here, I review some of the popular methods currently used for association studies. Because a vast number of methods is readily available, in this chapter, our focus is on methods for case-control studies that are also capable of identifying epistasis effects. Based upon the underlying algorithms, methods are grouped into model-based, non-model based and two stage methods. Within the model-based approaches, methods are further divided into frequentist and Bayesian methods.

Model Based Approaches

Frequentist Approaches

Logistic Regression Logistic regression (LR), discussed earlier, can be simply extended for multiple SNPs by allowing extra terms in the model. To accommodate epistasis effects, interactions among the SNPs an be easily added to the model. Using similar notation as Equation 2.19, let π_i be the probability that individual i is a case, the logistic model for two way interaction becomes

$$\log(\frac{\pi_i}{1-\pi_i}) = \mu + \sum_{k}^{K}\sum_{j=0}^{2}\beta_{k_j}x_{ik_j} + \sum_{k_1}^{K}\sum_{k_2=k_1+1}^{K}\sum_{j_1=0}^{2}\sum_{j_2=0}^{2}\gamma_{k_{1_{j_1}}k_{2_{j_2}}}x_{ik_{1_{j_1}}k_{2_{j_2}}} \tag{2.21}$$

where K is the total number of SNPs, β_{k_j} is the coefficient of genotype j of SNP k and $\gamma_{k_{1_{j_1}}k_{2_{j_2}}}$ is the

coefficient of the SNP interactions at the genotype level.

As the number of SNPs becomes large, the parameter estimation becomes unmanageable, hence the power is lost. Also, some genotype combinations may have low frequency or zero responses, thus the parameter estimations can be poor.

Another important issue with this model is the correlation among SNPs due to the LD structure. When the predictors are highly correlated (collinearity), the model gives little or no information about the corresponding parameters [185]. However, the problem associated with the collinearity can be addressed by using a stepwise selection method or shrinkage. For example, Lasso regression is a well known example of the shrinkage method which is discussed in detail in the following section.

Many standard statistical packages perform automated stepwise selection. In a forward selection, the initial model contains only the population mean, that is $logit(\pi_i) = \mu$. At each step, a new SNP or SNP interaction which results in highest improvement in the model fit is selected and included in the model. This process continues until adding no more SNP or SNP interaction can significantly improve the model fit. A backward stepwise selection, as its name suggests, is a counterpart to the forward stepwise selection. Instead of starting with a noninformative model, the initial model contains all SNPs and SNP interactions. At each step, a SNP or SNP interaction which results in the least model fitting deterioration is deleted. This procedure continues and stops when the deletion of any SNP or interaction results in significant reduction in the model fit. The other type of stepwise selection which is more flexible allows both a SNP or interaction to be added or removed at each step depending on which move is more beneficial for the model fit. This is called "stepwise selection". The model improvement/deterioration is evaluated using a parsimony criteria such as Mallows' CP [177], AIC [3] and BIC [243].

Stepwise selection procedures show a promising ability to find informative SNPs and SNP interactions with fewer false positive discoveries [166]. The main drawback of this procedure is that it is not capable of handling large scale datasets. Therefore, studies which implemented this method are limited to candidate gene studies [166, 52].

Lasso Regression The least absolute shrinkage and selection operator [Lasso, 278] is a shrinkage procedure which shrinks the noninformative coefficients to nearly or equal to zero. This is achieved by minimizing the residual sum of squares with a constraint that the sum of the absolute values of the coefficients is less than a constant.

Given a dataset, $(x^i, y_i), i = 1, \ldots, N$, where $x^i = (x_{i1}, \ldots, x_{ip})$ are predictors and y_i are the responses. Let $\hat{\alpha}$ and $\hat{\beta} = (\hat{\beta_1}, \ldots, \hat{\beta_p})$ be the Lasso estimates, such that

$$\arg\min_{\hat{\alpha},\hat{\beta}} \{ \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \} \tag{2.22}$$

with the constraint that $\sum_j |\beta_j| \leq t$. Here $t \geq 0$ and is a tuning parameter that controls the degree of shrinkage. The Lasso estimates of coefficients can be efficiently computed via the LARS algorithm of [76]. [291] recently applied Lasso penalized logistic regression to case-control GWAs for the detection of SNP and interaction effects. Their concluded that the Lasso is computationally efficient and when the predictors are not correlated, the interaction effects are identifiable.

Bayesian Approaches

In the frequentist approach, the assessment of the association between genetic variants and a phenotype is based on a p-value for null hypothesis of no association. Although it is still widely used, various studies have shown limitations of p-values [263, 249, 131]. Bayesian methods provide an alternative for assessing the association that alleviates the limitations of p-values. Note that frequentist and Bayesian approaches have different interpretations of "probability". For a frequentist, the probability is a long-run expected frequency of occurrence. In contrast, Bayesians view probability as related to degree of belief in the absence of complete knowledge. Thus the frequentist approach assumes that a population mean is real, but unknown, and can only be estimated from the data. An other difference between frequentist and Bayesian methods is in the methods for parameter estimation, the former often uses maximum likelihood estimation, Newton-Ralphson or EM algorithms while the latter often uses Markov Chain Monte Carlo methods.

Another difference between Bayesian and frequentist approaches is that the former requires the specification

of a prior distribution on the unknown parameters. Let θ denote the model parameters and $p(\theta)$ be the prior probability of theta, the posterior probability is then

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \tag{2.23}$$

where D is the data and $p(D|\theta)$ is the model likelihood. The denominator of Equation 2.23 is also known as the normalizing constant. It does not depend on θ and with a fixed D, $p(D)$ is constant. Therefore the unnormalized posterior density is

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{2.24}$$

For genetic association studies, the use of a prior can be valuable and under the Bayesian formulation, this information can be easily incorporated into the model. For example, in an association study, heavier weighting can be assigned to SNPs in region that other studies previously identified.

When considering only single marker effect (as in Section 2.5.1), there is also a Bayesian version of the searching algorithm which also makes use of the contingency table. Instead of computing p-values, the association is assessed using the posterior probability of association (PPA) [262]. The calculation of PPA can be split into three parts, choosing the prior probability (δ) on H_1, computing the Bayes factor and calculating the posterior odds ratio on H_1.

The value of δ governs the number of SNPs to be selected. Typically only a minority of SNPs are expected to have association, therefore [273] suggested δ ranges between 10^{-4} and 10^{-6}. The value of δ can differ across SNPs depending on biological information. For example, different value of δ can be given to SNPs closer to the gene of interest.

A Bayes factor (BF) is the ratio of the posterior probability of two competing models: in this cases, the ratio of H_1 over H_0. A stronger value of BF indicates stronger support to H_1 over H_0.

Once the value of δ is prespecified and BF is known, the next step is to computer the posterior odds ratio on H_1, that is

$$\text{PO} = \text{BF} \times \frac{\delta}{1-\delta}$$

The posterior probability of association is then

$$\text{PPA} = \frac{PO}{1+PO}.$$

PPA is a product of BF and prior probability on H_1. Because the prior probability is often set to be very small, the value of BF needs to be large to result in a higher value of PPA. In other words, the prior probability of H_1 controls the number of SNPs associated with the phenotype.

This method has been implemented in [273] for identifying SNPs associated with seven common diseases. In their study, they reported the PPA and the traditional *p*-value. A detailed description of the Bayesian SNP-by-SNP search method is in [262].

Logistic Regression The logistic regression discussed in the frequentist approach can be easily converted to a Bayesian method by assigning prior distributions to all parameters in the model. However, like the frequentist LR, the Bayesian LR is subject to the same problems pointed out earlier, which are the excessively large number of predictors (SNPs) and collinearity across SNPs. Therefore, to overcome these problems, it is necessary to performs model/variable selection or shrinkage as discussed in the frequentist LR.

Model/Variable selection Excellent methods for the variable selection problem have been developed within a Bayesian context, including stepwise selection, stochastic search variable selection and reversible jump MCMC. [171] proposed a Bayesian version of stepwise regression which is built on the method firstly proposed by [50] to identify the relative importance of genetic variants within a candidate region.

For a case-control study, let X denote an $N \times C$ matrix where N is the total number of individuals and C is the total number of predictors, and $x_i = (x_{i1}, \ldots, x_{iC})$. Let y_i denote the phenotype of individual i, and

$$y_i \sim \text{Bernuolli}(\pi_i) \quad \text{and} \quad f(\pi_i|x_i) = \mu_i$$

where $\mu_i = \omega_i\beta$ and $\omega_i = (1, x_{i\theta_1}, \ldots, x_{i\theta_k})$. The vector ω_i is similar to the design matrix of regression models and $\theta = (\theta_1, \ldots, \theta_k)^T$ is the column indices of X that correspond to variables selected to be included in the model. Parameter β is a vector that contains coefficients of selected columns. [171] employed the generic reversible jump Markov Chain Monte Carlo [GRJMCMC, 170] to estimate model parameters. Unlike the traditional RJMCMC [104], GRJMCMC permits multiple deaths/births moves with a single proposed move. This allows chains to move freely between subspaces without getting stuck in local maxima [171]. A detailed description of the implementation of GRJMCMC is in [170].

This method is flexible for different types of phenotypic data, e.g. count data, with the ability to simultaneously impute the missing genotype and easily expand for the inclusion of covariates. Furthermore, the WinBUGs code for carrying out the analysis is available in the appendix of the paper. Unfortunately, the major drawback of this model is that it is not scalable to a large number of predictors, which is often encountered in GWAs. The authors pointed out the maximum number of predictors under the current setting is less than 200.

The same algorithm with a different MCMC method has also been implemented for a genetic association study. [90] employed the traditional reversible jump MCMC [104] for variable selection. Even so, this is still limited to a small number of predictors.

An alternative method for variable selection is to use the stochastic search variable selection (SSVS) developed by [95]. SSVS involved embedding a model in a hierarchical normal mixture model where latent variables are used to identify subsets of variables. Unlike previous methods that involve searching across trans-dimensional spaces, the dimension of models visited is constant in SSVS. This is achieved by limiting the posterior distribution of non-informative terms in a small neighbourhood of zero. This method can

be easily implemented using Gibbs samplers and provides information on the posterior probability of each prediction. SSVS was originally introduced for regression models, however it can be easily modified for a binary response. We will firstly discuss SSVS in its formulation for a continuous phenotype.

Let $\mathbf{Y}$ be a $n \times 1$ vector of quantitative phenotype, $\mathbf{X} = [X_1, \dots X_p]$ be a $n \times p$ matrix of p predictors for n individuals and σ^2 be a scalar. Consider the canonical regression set up:

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2).$$

In SSVS, variable selection is achieved by considering $\boldsymbol{\beta}$ as modeled from a mixture of two normal distributions with different variances. Let $\gamma_i, i = 1, \dots, p$ denote latent binary variables, taking a value of 0 or 1, then the mixture of normal distributions for β_i is

$$\beta_i|\gamma_i \sim (1 - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau^2) \tag{2.25}$$

where τ_i and c_i are hyperparameters that control the variance of sampling distributions of β_i. For example, when $\gamma_i = 0$, β_i is sampled from a normal distribution with mean of zero and variance of τ_i^2. Normally τ_i is set small, so when $\gamma_i = 0$, β_i is sampled from a narrow region centered at zero, i.e. $\beta_i \approx 0$. However, to avoid $\beta_i \approx 0$ when $\gamma_i = 1$, the value of c_i is often set large. [95] and [283] provide some valuable advice on choosing these two hyperparameters.

This mixture of normal distributions can be included in the model as a multivariate normal prior distribution for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \text{MVN}(0, \mathbf{D}_\gamma \mathbf{R} \mathbf{D}_\gamma) \tag{2.26}$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$, $\mathbf{R}$ is the prior correlation matrix that is usually assigned to be the identity matrix, $\mathbf{I}$, and $\mathbf{D}_\gamma = \text{diag}[a_1\tau_1, \dots, a_p\tau_p]$ with $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$.

MCMC methods, mainly the Gibbs sampler, are used to fit the model. [297] provides a detailed description of using the Gibbs sampler to generate samples from a posterior distribution. Markers with large effect are often selected in the model, therefore, markers with high posterior probability are the important markers.

This model or the extended version of SSVS have been implemented in small scale association and QTL studies. [90] illustrates the use of this model along with two other approaches: Bayesian model averaging and Bayesian variable selection with RJMCMC- for a small scale association study. The same method is also implemented in [297] for identifying multiple QTL. However, it is still not clear if SSVS is suitable for detecting epistasis effects in large scale GWAs.

[49] proposed a genotype level analysis named SNPs Interaction Model with Phase Information (SIMPIe). SIMPIe is similar to model selection using SSVS but for binary data. Although the model of SIMPIe contains interactions terms, the aim of their model is not to identify the epistasis effects but to incorporate the phase information using SNP interactions. This is accomplished by strategically coding the interaction terms. Let y_i denote the binary phenotype of subject i and π_i be the disease penetrance. Let X_m be a variable coding the genetic effect on disease and $X_m = G_m$ where G_m indicates the number of variant allele at marker m. A logistic regression for a joint main effects model with second order interactions has the form

$$\text{logit}[P(Y = 1|X_1, \ldots, X_m)] = \alpha + \sum_{m=1}^{M} \beta_m X_m + \sum_{m=1}^{M} \sum_{\ell=m+1}^{M} \beta_{m\times\ell} X_{m\times\ell}. \tag{2.27}$$

The haplotype information can be approximated by modifying the second-order interaction terms in Equation 2.27 to describe the phase between pairwise SNPs, m and ℓ. Given that the two haplotypes for individual i are h_{i1} and h_{i2} and assuming additivity, X_m is coded as

$$X_{m\times\ell} = \begin{cases} 2 & \text{if } G_m \times G_\ell = 4 \\ 1 & \text{if } G_m \times G_\ell = 2 \\ 1 & \text{if } G_m \times G_\ell = 1, \text{ and } h_{i1} \text{ and } h_{i2} \text{ is a double variant haplotype} \\ 0 & \text{if } G_m \times G_\ell = 1, \text{ and } h_{i1} \text{ and } h_{i2} \text{ is not a double variant haplotype} \\ 0 & \text{if } G_m \times G_\ell = 0 \end{cases}$$

In this model, SSVS is also implemented for variable selection. Unlike the SSVS described in [95], when $\gamma_j = 0$, regression coefficients, β_j are directly assigned value zero; when $\gamma_j = 1, \beta_j$ is drawn from $N(0, c_i^2\tau^2)$. In contrast to the original SSVS, SIMPIe adopted a fully Bayesian approach. That is, parameter c_i is not pre-specified but estimated.

Incorporating the phase information in the genotype level of analysis improves the interpretability of the results. However, this model has only been tested for small scale association studies. Therefore, it is not clear if this model is suited for GWAs. Since the interaction terms are recoded to incorporate the haplotype information, it is uncertain how the model can be extended for accounting for the epistasis effects.

Shrinkage method As in the frequentist approach, an alternative method to variable selection is a shrinkage method. However, in the Bayesian context, shrinkage is much easier to implement by using a density that sharply peaks at zero as the prior distribution for regression coefficients. The double exponential distribution (DE) and the normal exponential gamma distribution (NEG) are the most commonly used prior distributions. Both densities have peaks at zero and heavy tails. The advantage of heavy tails is that they prevent heavy shrinkage to the parameter once the predictor is included in the model.

MCMC algorithms are the typical choice for model fitting [298]. However, MCMC is computationally burdensome when the number of predictors is large. Since the posterior variance of regression coefficients is not essential, [122] used a Bayesian-inspired penalized maximum likelihood to estimate the posterior *mode* of regression coefficients and implemented the CLG algorithm [20] to speed up the convergence. This approach has demonstrated promising ability in analyzing main effects for up to 500,000 SNPs within a

relatively small period of time. It can be easily extended for quantitative traits and haplotype or interaction effects. However, for the latter examples, the authors suggests reducing the model space prior to implementing the approach.

There are some other shrinkage methods that have been applied to mapping multiple QTL. In stead of using NEG or DE, [292] assigned a normal distribution with mean of zero. Variances of the normal distribution are estimated using a hierarchical approach and the noninformative Jeffery's prior.

BEAM The Bayesian epistasis association mapping (BEAM, [302]) algorithm aims to identify both single marker and epistasis effects in a population based case-control study. Let N_d and N_u denote the number of cases and controls, assuming that L SNPs were genotyped and case genotype is represented as $D = (d_1, \ldots, d_{N_d})$ where $d_i = (d_{i1}, \ldots, d_{iL})$ is the genotype of affected individual i. Similarly, let $U = (u_1, \ldots, u_{N_u})$ be control genotypes where $u_i = (u_{i1}, \ldots, u_{iL})$ is the genotype of unaffected individual i. Markers are then divided into three groups: group 0 contains markers unlinked to the disease, group 1 contains markers independently contributing to the disease risk and group 2 contains markers that jointly influence the disease risk. Let $I = (I_1, \ldots, I_L)$ be the membership of the markers where $I_j = 0, 1$ and 2 indicates that marker j belongs to group 0, 1 and 2 respectively. Let l_0, l_1 and l_2 be the number of markers in each group and let D_0, D_1 and D_2 be case genotype markers in group 0, 1 and 2. Because case genotypes should have different distributions compared to the genotype of controls, the likelihoods of groups 1 and 2 are thus independent from group 0 and controls and the posterior probability for I is proportional to

$$P(I|D, U) \propto P(D_1|I)P(D_2|I)P(D_0, U|I)P(I) \tag{2.28}$$

where $P(D_1|I)$ is the marginal probability of case genotypes in group 1 and $P(I)$ is the prior distribution for the membership of markers. The detailed mathematical procedure for deriving the marginal distribution of each group is in [302].

BEAM also uses the MCMC algorithm to draw I from Equation 2.28. At each iteration, I has two potential moves: randomly change a marker's group membership, and randomly exchange two markers between

groups 0, 1 and 2. The acceptance of the move will depend on the Metropolis-Hastings ratio.

Besides the fully Bayesian inferential framework, BEAM also incorporates the frequentist hypothesis testing procedure by calculating a 'B' statistic to check the significance in the association between marker(s) and the disease.

BEAM has shown promising ability in analyzing data sets containing up to 100,000 SNPs. However, under the current configuration, it is not able to handle more than 500,000 SNPs [51]. Although BEAM is able to account for the LD structure of adjacent SNPs, it is still not clear if it accounts for LD structure of the non-adjacent SNPs [51].

Non-Model based approaches

Traditional model-based approaches are often criticized for their inability to deal with nonlinear models [50] and inefficiency in handling large dimensional data. Machine learning or data mining algorithms provide alternatives to the model-based approaches. Data mining or machine learning algorithms do not rely on a single pre-specified model, but step through the space of possible predictor combinations. Thus they are more flexible for identifying main and higher order interaction effects. Although [51] suggests that it is false to exclude regression models from the data mining paradigm because some data mining algorithms involve stepping through multiple regression models, we still decided to treat them as two separate sections because algorithms discussed in this section do not rely on any model assumption. Perhaps it is more sensible to call this section 'non-model based approaches'.

Various machine learning algorithms have been implemented for detecting gene-gene interactions, [187] overviews four approaches, including neural networks, cellular automata, random forests and multifactor dimensionality reduction for detecting gene-gene interactions. A recent paper by [267] provides a great overview of the machine learning methods for genome wide association studies.

The most common and popular data mining methods for identifying gene-gene interactions are Random Forests and multifactor dimensionality reduction (MDR). The former method has been implemented in

several genetic studies [36, 35, 242, 169] and it is discussed in detail in Chapter 8 of this thesis. Therefore, it is not included here. Moreover, in the same chapter, we also provide reviews of some other machine learning algorithms.

Multifactor-Dimensionality Reduction (MDR) Multifactor-Dimension Reduction (MDR) [231] is a model-free and nonparametric method which reduces the dimensionality of multilocus information to improve the identifiability of marker combinations associated with disease risk. The MDR is directly applicable to the case-control study; therefore it has been widely used for mapping genetic variants in various phenotypes, including sporadic breast cancer [231], type 2 diabetes [48], cardiovascular disease [18] and rheumatoid arthritis [138].

The algorithm of MDR starts with dividing the data equally into 10 parts, where 9 parts are used for model ranking while the remaining portion of data is used for the estimation of prediction error (i.e cross-validation). In the 9 parts, a set of *n* factors are selected. These can be either genetic variants or other co-variates. The set of *n* factors and their possible multi-factor classes are represented in *n* dimensional space. For example, at 2 diallelic loci, there are 9 possible 2-locus genotype combinations. For each combination, based upon the case-control ratio of the combination and the pre-specified threshold value, the 2-locus genotype combination is labeled as high-risk or low- risk. The collection of these multifactor classes composes the MDR model for the particular combinations of factors. Among all *n* factor combinations, the model with the least misclassification rate is the optimal *n* locus model. The prediction error is then the error of the optimal model validated using the remaining portion of the data. This procedure is repeated 10 times to avoid spurious results due to data partitioning.

The main problem of MDR, according to [51] is that it is not suited for a data set with a large number of factors (e.g GWAs). When considering higher order interactions, [50] recommends using this method when there are only a small number of genetic variants.

Two Stage approaches

So far, only a handfull of methods can potentially analyse large GWAs datasets. Most methods proposed to date are limited in their ability to cope with the computational burden required for analyzing large scale GWAs. On the other hand, with methods that are appropriated for GWAs, the results of the analysis are often less than ideal. Therefore, instead of conducting only a single analysis, scientists have suggested two-stage approaches for identifying interaction effects [118].

The first stage is to select a subset of SNPs or genetic variants from the complete data set, then model interactions among the selected markers and between the markers and trait. Because of the conceptually simplicity, there are many variations of two-stage approaches. The SNP-by-SNP searching methods are the most common for the first stage filtering [181, 82, 152]. Logistic regression can then be applied to identify the interactions.

Alternatively, [192] proposed using Random Forests as the screening procedure for identifying a smaller set of variables and using Bayesian networks to develop complex etiological models. In their study, data was reduced from 9190 variables to about 53 variables at the first stage of analysis. They found the screening strategy was able to successfully filter out SNPs unassociated with disease loci, while keeping the surrogates for risk SNPs.

Part I: Phenotype of Complex diseases

3

Linkage and heritability analysis of migraine symptom groupings: a comparison of three different clustering methods on twin data

Chapter Summary

The first objective of this thesis is to improve phenotype definitions for diseases with complex etiology. These diseases often lack clear biomarkers, which would normally provide for more exact phenotyping. To address this objective, it is important to firstly understand how different methods of phenotyping can impact the results of the subsequent analysis. Therefore, the aim of the chapter is to compare the use of the most commonly used statistical methods for phenotyping with respect to the results of the subsequent genome-wide linkage analysis and heritability estimates. In this Chapter, we focus only on the clustering type of approaches, namely latent class analysis (LCA), grade-of-membership (GoM) and fuzzy clustering methods (Fanny).

In this Chapter, migraine data is used for the illustration of different phenotyping tools, and also we present results on the genetics of migraine. Furthermore, this chapter provides better understanding of the LCA, GoM and Fanny, and we attempt to clarify some confusion associated with these methods.

Chapter Conclusion

Using migraine data as a baseline of comparisons, the main conclusion of this chapter is that different clustering methods may produce a range of results in the subsequent analyses, ranging from similar to completely different. Phenotypes obtained using LCA and fanny are highly correlated, and therefore the heritability and loci identified by the linkage analysis are in agreement. However, the phenotype of GoM is very different from the two other methods, therefore the heritability and loci identified by the linkage analysis are distinctly different. GoM is more closely related to LCA than to Fanny, because both of these models are forms of mixture model. The main difference between these two models is that the mixture of components occurs at a finer level for GoM. When comparing the models using a parsimonious measure, i.e. BIC, even though GoM has the highest likelihood, it is heavily penalised due to the model complexity, and therefore less preferable.

In this chapter, we were able to replicate some previously identified loci and estimate the heritability of migraine within the previous published range.

Authorship

Carla C.M. Chen, Kerrie L. Mengersen, Jonathan M. Keith

Discipline of Mathematical Sciences, Queensland University of Technology

Nicholas G. Martin, Dale R. Nyholt

Genetic Epidemiology Unit, Queensland Institute of Medical Research

Reference

Chen, C. C.-M., Mengersen, K. L., Keith, J. M., Martin, N. G., and Nyholt, D. R. (2009). Linkage and heritability analysis of migraine symptom groupings: a comparison of three different clustering methods on twin data. *Human Genetics* 125, 591 - 604.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for the associated publication is:
Chen CC-M, Mengersen KL, Keith JM, Martin NG, Nyholt DR (2009) Linkage and heritability analysis of migraine symptom groupings: a comparison of three different clustering methods on twin data. *Human Genetics* 125: 591 - 604

Contributor	Statement of contribution
C C.M Chen	conception and conduct the research, write the code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature & Date:	
K. L. Mengersen	conception, execution, interpretation, editing
Jonathan M. Keith	editing
N. G. Martin	design of questionnaire, interpretation, editing
D. R. Nyholt	conception, execution, interpretation, editing, design of questionnaire, data collection

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

3.1 Abstract

Migraine is a painful disorder for which the etiology remains obscure. Diagnosis is largely based on International Headache Society (IHS) criteria. However, no feature occurs in all patients who meet these criteria, and no single symptom is required for diagnosis. Consequently, this definition may not accurately reflect the phenotypic heterogeneity or genetic basis of the disorder. Such phenotypic uncertainty is typical for complex genetic disorders and has encouraged interest in multivariate statistical methods for classifying disease phenotypes.

We applied three popular statistical phenotyping methods - latent class analysis (LCA), grade of membership (GoM) and grade of membership ”fuzzy” clustering (Fanny) - to migraine symptom data, and compared heritability and genome-wide linkage results obtained using each approach. Our results demonstrate that different methodologies produce different clustering structures and non-negligible differences in subsequent analyses. We therefore urge caution in the use of any single approach and suggest that multiple phenotyping methods be used.

3.2 Introduction

The essential first step for linkage analysis or association studies is to accurately identify the phenotype. For complex diseases such as migraine, identification of the phenotype is challenging due to the lack of objective markers and uncertainty about the causes of the disease. The diagnosis of this type of disorder is often based on satisfaction of clinically accepted criteria. Although they may not be useful for diagnosis and treatment, these clinical based phenotypes may not be optimal for genetic research, in particular finding genetic loci contributing to disease inheritance (eg., [109]) and this has led to a call for the development and use of new phenotyping strategies in genetic research (e.g., [287]).

Migraine is a common, painful and debilitating disorder. Numerous researchers have shown that there is a significant genetic component to risk of this disorder [306, 202, 264, 265, 211, 212], with estimates of

heritability ranging between 34 and 57% in twin-cohort studies across six countries [202]. The diagnosis of migraine is found to be difficult due to lack of biological markers and overlap with other types of neurological disorders, such as tension type headache and brain tumour. To date, the diagnosis of migraine relies on classifying self-reported headache characteristics using International Headache Society (IHS) criteria [115, 213, 251]. These criteria were developed for standardising headache definition (e.g., [164]). The two major subtypes of migraine are migraine without aura (MO) and migraine with aura (MA); the definitions of both types are listed in Tables 3.1 and 3.2, respectively.

Table 3.1: The 1988 International Headache Society diagnostic criteria for migraine without aura (MO).

Item	Description
A	At least five attacks fulfilling B-D
B	Headache attacks lasting 4-72 hours
C	Headache has at least two of the following characteristics:
	Unilateral Locations
	Pulsating quality
	Moderate or severe intensity(inhibits or prohibits daily activities)
	Aggravation by walking stairs or similar routine physical activity
D	During headaches at lease one of the following:
	Nausea and (or) vomiting
	Photophobia and phonophobia

Table 3.2: The 1988 International Headache Society diagnostic criteria for migraine with aura (MA).

Item	Description
A	Headache fulfilling criteria B-D list in Table 3.1
B	At least five attacks fulfilling B-D
C	Aura consisting of at least one of the following but no motor sickness
	Fully reversible visual symptoms including positive features
	(*ie* flicking of lights) and (or) negative features (*ie* loss of vision)
	Fully reversible sensory symptoms including positive (*ie* pins and needles)
	and (or) negative features (*ie* numbness)
	Fully reversible dysphasic speech disturbance
D	At least two of the following:
	Homonymous visual symptoms and (or) unilateral sensory symptoms
	At least one of the aura symptom develops gradually over $\geq$5 minutes
	Each symptom lasts $\geq$5 minutes and $\leq$60 minutes.

These criteria have improved migraine diagnosis and subsequently, epidemiological research. However, none of the features occur in all patients who meet a strict definition of IHS migraine, and no single symptom is required for diagnosis. In other words, migraine is a complex of symptoms with variable symptom profiles and individuals presenting with dissimilar symptoms can equally satisfy the same diagnosis. Furthermore, although individuals may not quite satisfy IHS criteria they would nonetheless be treated as such in a clinical

setting; indeed there is an IHS classification of "probable migraine" (previously termed "migrainous disorder not fulfilling the above criteria"). The majority of genetic studies for migraine to date have concentrated on either MO or MA and found various chromosome regions associated with each (Table 3.3). Under these phenotype definitions, no common gene was replicated across studies. However, when migraine phenotypes were identified using a statistical (rather than medical) phenotyping classification via latent class analysis, [165] successfully replicated two susceptibility loci: chromosome 5q21 and 10q22-q23 [212, 10, 11, 165].

Table 3.3: Table showing the significant linkage signals which are identified in the literature for IHS criteria defined migraine with aura (MA) and migraine without aura (MO)

Phenotype	Cohort	Chromosome	Reference
MO	Icelandic	4q21	[28]
MO	Italian	14q21.2-q22.3	[257]
MA	Canadian	11q24	[37]
MA	Finnish	4q24	[286]
MA and MO	Sweden	6p12.2-p21.1	[41]
MA*	Finnish and Australian	10q22-q23	[11]

* Including three types of migraine with aura
- Pure MA, individuals fulfilling IHS criteria for migraine with aura
- Unclassified MA, a group of individuals that cannot be grouped into the IHS defined categories, but clearly suffer from aural features.
- Mixed migraine, a group of individuals that commonly have both MA and MO type of attacks.

A wide variety of statistical methods have been employed to identify clusters and classes based on symptomatic data. Classical methods such as principal component analysis (PCA) and discriminant analysis (DA) have previously been used in genetic linkage analysis. However, these approaches assume individuals belong to only one of potentially many clusters, which may neglect the phenotypic heterogeneity present in complex human diseases [140, 180]. In contrast, "fuzzy" clustering such as latent class analysis (LCA) and grade of membership (GoM) resolve the heterogeneity by assigning individuals to multiple clusters and quantified measures of the probability of belonging to each group.

Latent class analysis [186] has been widely used in subtyping complex diseases such as migraine [211, 212], attention-deficit/hyperactivity disorder (ADHD) [282] and schizophrenia [132] in the field of genetics. Another type of fuzzy clustering, Grade of Membership (GoM), has also been frequently used to obtain empirical phenotypes. This clustering method was first used for medical classification in 1978 [289] and is now commonly employed for disease subtyping. It has been employed in genetic research for diseases with complex etiology [42, 53, 85, 179].

Most recently, [140] proposed a different type of clustering method which is also called Grade of Membership (GoM). Unlike the model-based GoM proposed by [289], the method suggested by [140] is based on partitioning the data into a pre-determined number of clusters. To avoid confusion in nomenclature, the grade of membership proposed by [140] will be referred to as Fanny [143] in this thesis. This method has been used to identify loci causing anxiety disorder [141].

Although some literature has compared the mathematical and statistical differences between LCA and GoM [179, 216, 81, 79], the effects of these three common phenotyping methods, LCA, GoM and Fanny in genetic analyses such as heritability and genome-wide linkage have not been investigated. Therefore, the aim of this study is to 1) compare these three methods as they apply to common migraine symptomatic twin data, 2) benchmark their performance in genetic research and 3) investigate whether different clustering methods result in different loci being implicated in linkage analysis.

3.3 Methods and Materials

The symptomatic data were first analyzed by three different phenotyping methods: latent class analysis (LCA), grade of membership (GoM) and fuzzy clustering (Fanny) to obtain a continuous (quantitative) phenotype trait (score) for individuals. The value of phenotypic measures derived from these three models was constrained to be between 0 and 1, which was then used as a continuous trait in the genome-wide linkage analysis. LCA and GoM are both model-based approaches in which the optimum number of clusters was determined by likelihood ratio, Bayesian Information criteria (BIC) and Akaike information criteria (AIC). For Fanny, the number of clusters was set to 2, analogous to previous Fanny-based genetic studies [140, 141].

3.3.1 Phenotype Data

Migraine data were obtained from extensive semi-structured telephone interviews as part of a study designed to assess physical, psychological and social manifestations of alcoholism and related disorders [116]. The sample was unselected with regard to personal or family history of alcoholism or other psychiatric or medical

disorders [202]. The interviews were conducted during two periods of time: 1993-1995 and 1996-2000. The earlier interviews were administered to Australian twins listed with the volunteer-based Australian Twin Registry who were born between 1902 and 1964, whereas the second interviews were focused on twins born between 1964 and 1975.

Participants of both cohorts were first asked the screening question: "Do you have recurrent attacks of headaches?" If the participant screened positive, then he/she was asked a number of questions which were developed by an experienced migraine researcher based on the IHS diagnosis criteria (Table 3.4). A total of 13062 individuals from 6764 families participated in this study, with 2716 MZ twin pairs (63.6% females and 36.4% males), 3399 DZ twin pairs (34.52% female twins, 22.36% male twins and 43.13% mixed sex twins), 15 twins with unknown zygosity and 817 first degree family members, including both siblings and parents. The mean age of participants was 37.5±11.3 and ages ranged from 23 to 90 years at the time of interview.

Table 3.4: The survey questions designed based on 1988 International Headache Society diagnostic criteria.

Notation	Abbreviation	Descriptions
a	≥ 5 episode	Have at least 5 episode of headaches in your life time.
b	4-72 hr	Average headache lasts between 4 to 72 hours
c1	Unilateral	Headache often occurs at one side of head
c2	Pulsating	Headache pain can be described as throbbing, pulsating or pounding
c3a	Moderate/severe	Headache pain can be described between moderate and severe
c3b	Prohibitive	Headache pain prohibits daily activities
d1	Nausea/vomiting	Headache associated with vomiting or feeling nausea
d2a	Photophobia	Enhanced sensitivity to light
d2b	Phonophobia	Enhanced sensitivity to sounds
Aura	Aura	Have visual problems such as light shower, blurring, blind spot or double vision

Although the wording of questions was identical for both cohorts, not all questions in Table 3.4 were included for the older cohort. The questions relating to having more than 5 migraine/episodes of headache during lifetime ("≥5 episodes"), average duration of migraine/episodes between 4 and 72 hours ("4-72 hours"), and pain associated with headache described as moderate to severe ("mod/severe") were not include in the questionnaire for the older cohort. We conducted separate analyses for older, younger and two cohorts combined data.

3.3.2 Models

Latent Class Analysis (LCA) Latent class analysis is a multivariate technique which can be applied to clustering, regression and factor analysis. The classes are latent because they are not directly observed, but are identified based on a function of a set of observed variables. LCA was developed in the 1950s for dichotomous variables [161]; however, the full potential and practical application of LCA only became evident after the introduction of more general latent class analysis and a simpler method of obtaining maximum likelihood estimates of the parameters in the 1970s [101, 102]. The latter LCA is capable of dealing with both dichotomous and polytomous variables and more than one latent variable can be included in the model.

Suppose there are n individuals, J observed (manifest) variables and each variable j has L_j levels of response, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$ and $l = 1, 2, \ldots, L_j$. Let y_{ijl} denote the binary response of the ith individual to symptom j with level l and Y_i is then the vector of subject i's response to all symptom questions. Assuming there are K latent classes within the latent variable, let λ_{kjl} denote the class conditional probability that an observation in class k produces the lth outcome on the jth variable; therefore, within each j, $\sum_l \lambda_{kjl} = 1$. In this thesis, the data consist of binary responses, and thus L_j is two. Assuming local independence, the probability of a particular set of responses from an individual i in class k is:

$$f(Y_i|\lambda_k) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{ijl}} \tag{3.1}$$

Let p_k denote the weight of latent component k. Then the joint distribution for all J variables under the latent class model is

$$Pr(Y_i|\lambda, p) = \sum_{k=1}^{K} p_k \prod_{j}^{J} \prod_{l}^{L_j} (\lambda_{kjl})^{y_{ijl}}$$

The LCA analyses were carried out using the poLCA [167] package of R2.4.1 [219]. The parameters were estimated via the expectation-maximization (EM) algorithm [60]. The details of the EM algorithm for LCA are in [167]. Unlike the other models described in this thesis, the class membership probabilities are

estimated post-hoc using Bayes' formula:

$$p_{ik} = Pr(k|Y_i) = \frac{p_k f(Y_i|\hat{\lambda}_k)}{\sum_k p_k f(Y_i|\hat{\lambda}_k)}, \tag{3.2}$$

where $\hat{\lambda}_k$ is an estimate of outcome probability conditioning on class k. Because the parameters are estimated using the EM algorithm, the latent class for the observations with missing value(s) can still be estimated. This is achieved by excluding cases with missing values when calculating Equation 3.1 and the denominator of Equation 3.2 [161].

Grade of Membership (GoM) Grade of membership (GoM) also fits into the latent class framework. GoM was first developed by [289] for expressing non-stochastic heterogeneity in a population by direct latent variable estimation. This method has been further developed by various researchers and is frequently applied in medical and genetic research [80].

Let $g_i = (g_{i1}, g_{i2}, ..g_{iK})$ be the latent vector of grade membership scores for individual i having a partial membership of component k, where $g_{ik} \geq 0$ for each i and k and $\sum_{k=1}^{K} g_{ik} = 1$. The value g_{ik} can be interpreted as the intensity of membership in each component. Unlike LCA, the membership scores of individuals are estimated directly from data. Let λ_{kjl} denote the probability of positive response to level l of variable j for a complete membership of component k, $\lambda_{kjl} = Pr(x_{ijl} = 1|g_{ik} = 1)$ where $i = 1, 2, .., n$, $j = 1, 2, ..J$ and $k = 1, 2, .., K$. Within each variable j, $\lambda_{kjl} \geq 0$ and, the sum of λ_{kjl} across all levels, is equal to one. The joint likelihood of GoM is

$$Pr(Y|\lambda, g) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\sum_k g_{ik}\lambda_{kjl})^{y_{ijl}} \tag{3.3}$$

Equation 3.3 is maximized by iterative optimization with respect to one set of parameters while keeping the other set of parameters constant. This iterative procedure is referred to as the missing information principle. The details of the parameter estimation procedure are in [179].

GoM can deal with missing values in two different ways, depending on the nature of the missing values. When the missing data are generated by a random mechanism which is independent of model parameters, missing data can be treated as unobserved and independent observations. In this case, y_{ijl} for a missing observation is set to be 0 for $l = 1, \ldots, L_j$ and is consequently excluded in the computation of the likelihood. When the missing data are due to a non-random process, such that certain items have a higher rate of missing data on a specific latent class, GoM deals with this problem by increasing the dimension of the measurement spaces by adding an extra category called "missing" for each variable in the model. In this study, in light of no information to the contrary, we assume the missing value is due to random causes.

The above models were tested using the Akaike information criterion (AIC, [3]), Bayesian information criterion (BIC, [243]) and log-likelihood values for each value of K. AIC and BIC strike a balance between goodness of fit and model complexity, thus avoiding both over-fitting and under-fitting. Models with lower AIC and BIC values are preferred. Log-likelihood measures model fit but not complexity, and thus must be used cautiously to avoid over-fitting.

Phenotype Conversion In this study, the maximum number of components tested in the LCA and GoM analyses is 6 ($max(K) = 6$). The optimum number of components for LCA is determined by the Bayesian information criteria (BIC) [243] whereas the likelihood ratio test is used to determine the optimum number of components in GoM. Because both models yield only multinomial estimates, an intermediate step is added to obtain a continuous phenotypic measure. When the optimum value of K is 2, the membership score for the "affected" component (the component with more and stronger symptoms, such as p_{ik}=affected of LCA and g_{ik}=affected of GoM) is taken to be representative of the trait value. Currently, genome-wide linkage analysis is limited to either a continuous or a dichotomous trait value, and is not designed for multiple clusters. Therefore, in the past, when the optimum number of clusters in the model exceeded two, the phenotype was determined by a threshold value [211, 212, 164]. To avoid the difficulty in determining an appropriate threshold, we implemented the following method to convert multinomial values to continuous values bounded between 0 and 1.

When the optimum number of components in a model exceeds 2, we used the following equation to estimate

each individual trait value. Since this trait value aims to capture the presence of the symptom, we set l to 2:

$$\text{Phenotypic Trait}_i = \sum_{k=1}^{k=K} \frac{\sum_{j=1}^{j=J} \lambda_{kj2}}{J} \times g_{ik}$$

where g_{ik} is membership score for individual i having partial membership of cluster k and J is the total number of manifest variables.

The use of a single, continuous-valued summary of phenotype such as this is not appropriate if two or more distinct disorders were producing the observed symptoms. We note that in the analysis of the migraine data, the clusters can be ordered sequentially such that the probability of experiencing each of the ten symptoms decreases monotonically. This is highly suggestive of a single underlying determinant of severity. The justification is less clear for the GoM model, because the clusters cannot be ordered in the same way. Nevertheless, the GoM clusters can be ordered such that the endorsement probabilities decrease monotonically for eight of the ten symptoms. Moreover, as we discuss below, there is reason to believe (on the basis of information criteria) that the LCA clustering is the more appropriate data model.

Grade of Membership-Fanny Unlike the two model-based approaches described above, Fanny forms clusters based on the dissimilarity between subjects, such that where subjects resemble each other they tend to be clustered into the same group. Dissimilarity between two objects can be calculated in various ways. Due to the type of variables in the migraine dataset, the dissimilarity matrix is calculated using a contingency table. Considering two objects, i and j, and the contingency table of i and j for variable p,

Table 3.5: The contingency table of object i and j.

$i \setminus j$	1	0
1	a	b
0	c	d

the dissimilarity between i and j is estimated as

$$d(i, j) = \frac{b + c}{a + b + c + d}.$$

Let u_{ik} denote the strength of membership of object i to cluster k, $u_{ik} \geq 0, \sum_{k=1}^{K} u_{ik} = 1$. u_{ik} is analogous (but not equal) to g_{ik} and p_{ik} above. The objective of Fanny clustering is to iteratively minimize:

$$\sum_{k=1}^{K} \frac{\sum_{i,j=1}^{n} u_{ik}^2 u_{jk}^2 d(i,j)}{2 \sum_{i=1}^{n} u_{jk}^2}.$$

Unlike LCA and GoM, Fanny does not provide a measure of how many clusters best fit the data; the user must choose the value of K. We therefore followed the approach utilized in previous Fanny-based genetic studies [140, 141] and fixed the number of clusters in the model to two. Whether this is appropriate or not would depend on the underlying architecture of the trait (symptomatology) under investigation. As a result, the phenotypic value of the individual subject was simply the score, u_{i2}, for the membership of the affected group. The Fanny algorithm procedure is implemented by the Fanny function of the contributed package cluster [176] of the R [219] statistical package.

3.3.3 Genetic Data

The genotypic data are from a collection of four smaller genome-wide linkage scans performed for studies at the Queensland Institute of Medical Research (QIMR). Genotyping for four scans was undertaken at Gemini Genomics with 426 microsatellite markers, Sequana Therapeutic with 519 markers, the Center of Mammalian Genetics at the Marshfield Clinic Research Foundation with 776 markers and the University of Leiden with 435 markers. The recruitment of participants for genotyping was based on individuals involved in phenotype collection. The details of DNA collection, genotyping methods and data are provided in [305] and [54].

Graphic Representation of Relationships (GRR) [2] and RELPAIR [78, 71] were applied for the examination of the pedigree structure and identification of inconsistencies between the genotypic data and self-reported pedigree relationships. Potential pedigree misspecification, incorrect zygosity labelling of twins and potential sample mix-up were identified and investigated; the problematic individuals or families were removed from further analysis. The SIB-PAIR program by [70] was then implemented for identifying and cleaning

the Mendelian inconsistencies in the genotype data.

After combining all four scans, there were 485 markers which were typed in two or more scans. Therefore to ensure the consistency of genotypic information for these 458 markers, the duplicated markers are included separately on the genetic map for the combined scan, which is separated by a small distance of 0.001cM. The consistency of the genotypes of these 458 markers was checked using various methods described in [54]. Markers with genotypic data inconsistent between different genome scans were excluded and unlikely genotypes were identified by MERLIN [1] and omitted from further analysis. Potential map errors were identified by GENEHUNTER [155] and MENDEL [159]. Map positions were in Kosambi cM, which is estimated using locally weighted linear regression from the National Center for Biotechnology Information (NCBI) Build 34.3 physical map positions, as well as published deCODE and Marshfield genetic map positions [150]. Where the results suggested inconsistencies between genetic map distance and recombination fraction, the primer sequences for all markers in the region were BLASTed against the entire human genome sequence (http://www.ensembl.org, NCBI build 34.3). The genetic map was then revised to include the updated physical positions of all markers in the problematic regions. The revised map and the original genotype data were cleaned of unlikely genotypes using MERLIN and map errors were resolved using GENEHUNTER.

The final cleaned data contains 1770 unique markers. The main intermarker distance for all sib-pairs in the samples was 7.1cM, calculated for each sib-pair and analyzed across all sib-pairs. The combined genome scan included 4148 individuals from 919 families, which included 143 MZ and 776 DZ twin pairs.

3.3.4 Heritability

Heritability of the continuous phenotype values was estimated with the ACE model. The ACE model assumes the phenotype variation is due to the additive genetic effect (A), shared environment effect (C) and random environment effect (E). The heritability is then the proportion of the total variance which is due to the additive genetic effect. The analysis was carried out using Mx [206] which performs maximum likelihood estimation of the variance components.

3.3.5 Linkage Analysis

A non-parametric quantitative trait linkage analysis was carried out using Merlin-qtl, developed under the general framework of [149] and [288]. The membership score of the three models (g_{ik} of LCA and GoM and u_{ik} of Fanny) was treated as a quantitative trait.

3.4 Results

The results of clustering and linkage analyses performed separately for the older and younger cohorts lack the power to identify any significant loci. Moreover, the analysis of the older cohort itself is not representative of the true migraine population due to lack of three symptom responses. However, by combining two cohorts, we obtained a representative sample and power to identify disorder-related loci, hence we restrict our subsequent results to the combined data set.

Table 3.6 provides goodness of fit statistics for the choice of K in the two model based approaches, LCA and GoM. For LCA, there is little improvement in AIC or BIC as K increases above four, where there is a local minimum in BIC (Table 3.6). We therefore selected $K = 4$ as the best compromise between model fit and complexity. For GoM, both AIC and BIC indicate that the best model has $K = 2$, but even this best-scoring GoM model is substantially worse than any of the LCA models. The reason for this is that although GoM models have better fit (that is, higher log-likelihood), they achieve this at the cost of including additional parameters, namely the membership scores g_{ik}. In light of this, we based goodness of fit assessment on the log-likelihood ratios and noting that the largest reductions in log-likelihood occur as K increases to four, we again chose the four clusters GoM model.

Even though four clusters were chosen for both GoM and LCA, the characteristics of the clusters differ between these phenotyping approaches. Figure 3.1 shows the characteristics of each LCA cluster. Each bar shows the probability of having the symptom, given a full membership to cluster k. For instance, the probability of having “aura” for a member in cluster 1 is 0.90. There is a progressive reduction of endorsement

Table 3.6: The log-likelihood value, AIC and BIC values of LCA and GoM models with different numbers of clusters.

Model	Number of cluster (K)	Log-Likelihood	AIC	BIC
LCA	2	-38752.642	77549.28	77713.79
	3	-36677.701	73423.4	73677.63
	4	-36456.261	73004.52	73348.49
	5	-36401.638	72948.79	73382.48
	6	-36333.290	72806.58	73330
GoM	2	-28616.94	109561.9	305202
	3	-22696.07	123884.1	417344.6
	4	-20978.36	146612.7	527892.4
	5	-20322.00	171464.0	660564.8
	6	-18838.39	194660.8	781581.8

probability for all symptoms when cluster 2 is compared to cluster 1, when cluster 3 is compared to cluster 2 and when cluster 4 is compared to cluster 3. The only departure from this pattern is the slight increase in the probability of a positive response to the question “have you had more than 5 episodes of headaches in your life time?” when cluster 3 is compared to cluster 2. The clusters are thus in a natural order, suggesting, as mentioned earlier, that migraine phenotypes can be organised on a linear scale of severity.

This linear pattern is not apparent for the GoM clusters. It is apparent that cluster 1 has the highest endorsement probabilities for all symptoms and cluster 4 has the lowest. However, although cluster 2 has equal or higher endorsement probabilities than cluster 3 for most symptoms, this situation is reversed for the symptoms “≥5 episodes” and “moderate/severe” (Figure 3.2).

Table 3.7: The weight of each cluster under different phenotyping analysis. According to AIC and BIC, the optimum number of clusters for LCA is 4. Using the log-likelihood as selection criteria for goodness of fit, the optimum number of clusters for GoM is also 4.

Model	No. Clusters	Class 1 (Affected)	Class 2	Class 3	Class 4 (Less Affected)
LCA	4	0.136	0.206	0.103	0.554
GoM	4	0.215	0.076	0.105	0.604
Fanny	2	0.405	0.590	-	-

- Not applicable.

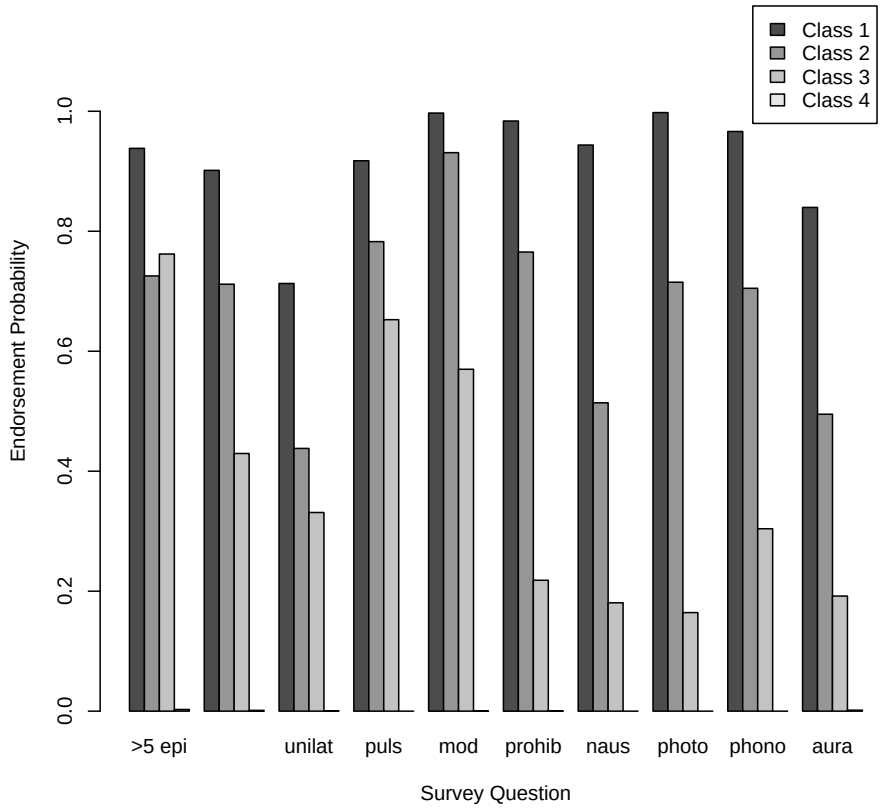

Figure 3.1: The characteristics of the four clusters under LCA K=4 model. X-axis corresponds to the items listed in Table 3.4 and the y-axis is the probability of displaying the symptom given full membership to cluster k.

The profile plot showing the characteristics of the Fanny clusters is depicted in Figure 3.3. There is a large difference in the endorsement probabilities of the two clusters, and more than 55% of individuals in cluster 2 have all symptoms listed in Table 3.4. Individuals in cluster 2 are not exempt from all symptoms; a small proportion in this cluster had the first five symptoms of Table 3.4 during their headache episode. Since there are only two clusters in this analysis, cluster 1 can be referred to as the "Affected" class and cluster 2 as the "Unaffected" class.

Of the total 13062 individuals, 14% were assigned to cluster 1, 21% to cluster 2, and 10% and 55% were in cluster 3 and 4, respectively, according to LCA (Table 3.7). In contrast to LCA, a slightly higher proportion of the population were classified into the two extreme clusters of GoM with 22% falling into cluster 1 and

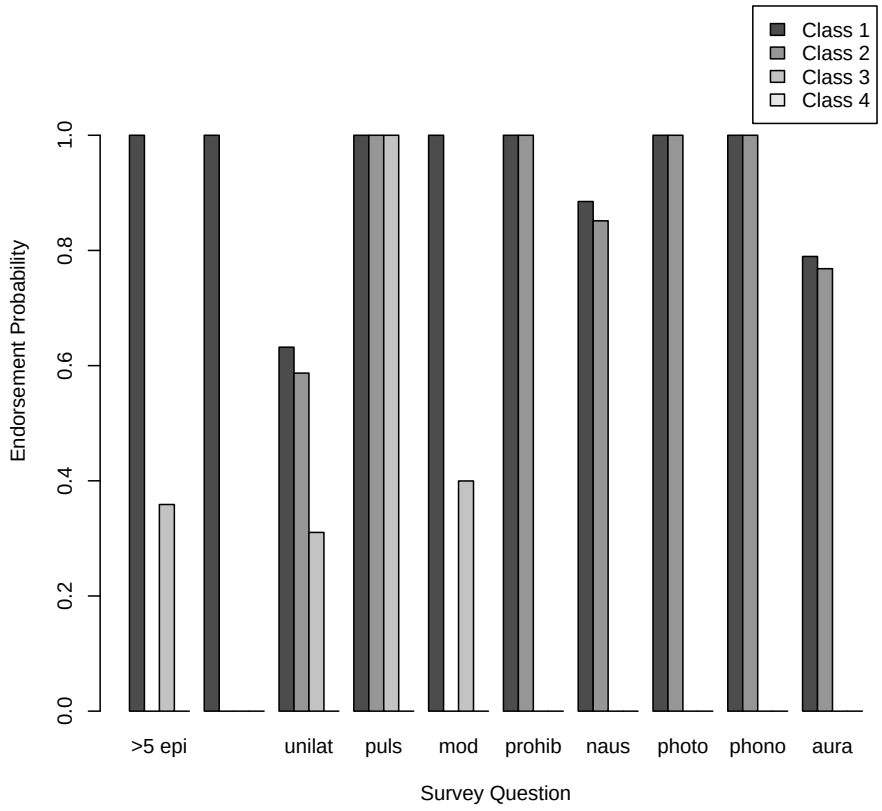

Figure 3.2: The characteristics of the four clusters under GoM K=4 model. X-axis corresponds to the items listed in Table 3.4 and the y-axis is the probability of displaying the symptom given full membership to cluster k.

60% into cluster 4. Under the Fanny clustering method, around 40% of the population are classified into cluster 1 and 60% are in cluster 2 (Table 3.7).

After phenotype conversion, all three models agreed that a large proportion of the subjects in this study have a very small probability of having had migrainous headaches (Figure 3.4). However, we observed some variations in the tail end of the histograms. According to GoM, there is an even distribution in the individuals with scores between 0 and 1, with a slightly higher proportion having scores closer to 1. This is different from the results obtained using Fanny and LCA, in which only a very small number of people had a phenotypic score between 0 and 0.4. However, unlike the tail end of the Fanny histogram which shows a slight increase in score distribution, the LCA histogram shows small peaks at 0.5 and 0.7. The maximum

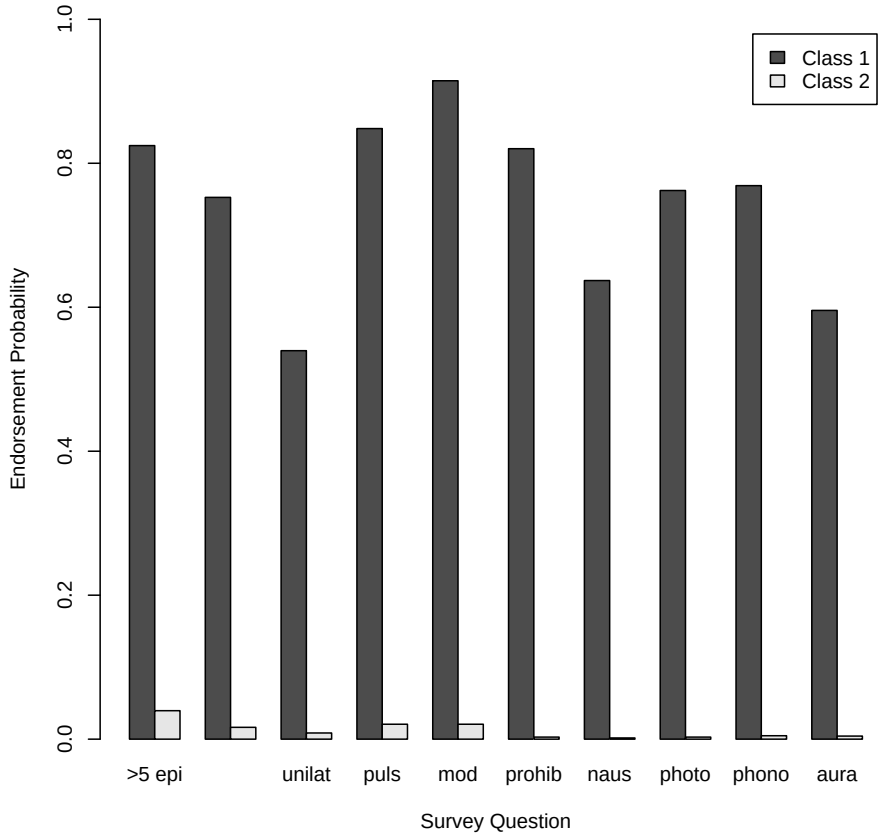

Figure 3.3: The characteristics of the four clusters under Fanny K=2 model. X-axis corresponds to the items listed in Table 3.4 and the y-axis is the proportion of individuals having the symptom given cluster k.

trait scores estimated in LCA and GoM approach 1, whereas the maximum trait score using Fanny is 0.86.

At the individual level, LCA and Fanny gave similar phenotypic estimates. Figure 3.5 contains scatter plots showing the predicted scores of individuals under the different methods. LCA and Fanny show very similar predicted scores when the score is larger than 0.4. Although Fanny tends to give higher phenotypic scores to individuals with a score lower than 0.4, generally there is a strong correlation between LCA and Fanny phenotypic scores (correlation= 0.99). In contrast, although the correlation is still high (correlation = 0.85), there is a notable discrepancy between LCA and GoM predicted scores. This is also observed in the comparison of phenotypic scores obtained using the Fanny and GoM approaches.

Table 3.8 contains the heritability estimates when using the phenotypic scores of the three models where A

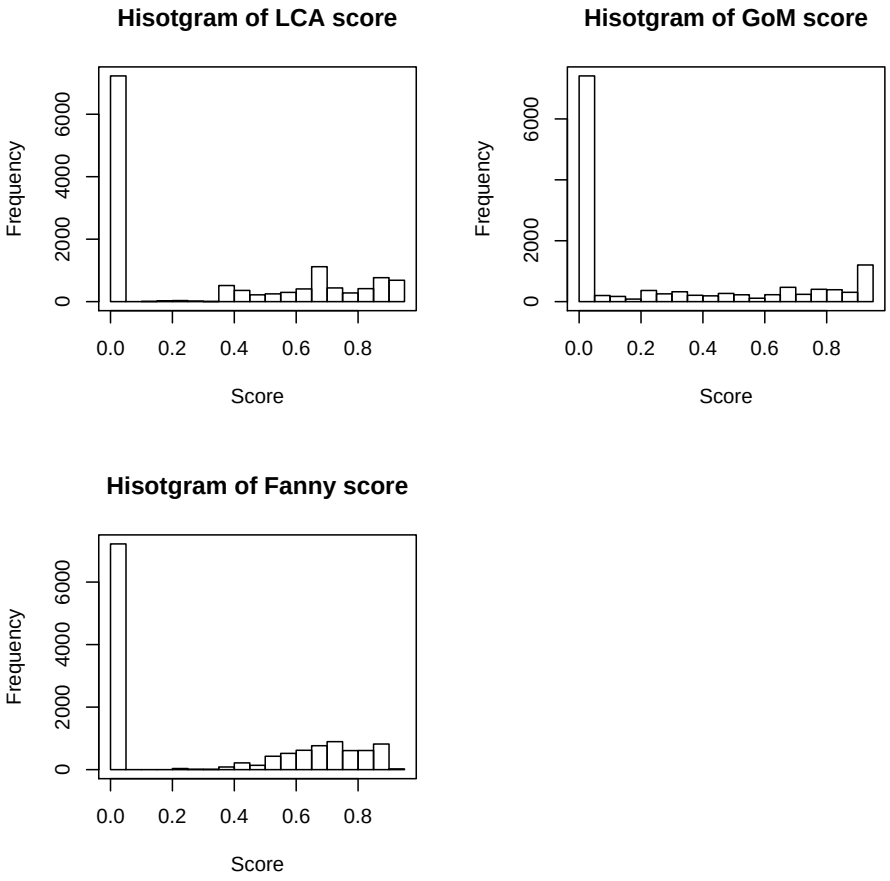

Figure 3.4: Histogram showing the distribution of the phenotypic scores estimated under LCA, GoM and Fanny. A score of 0 indicates not having migrainous headache and a score of 1 indicates having strong migrainous headache.

indicates the variation due to genetic variation, C is the variation due to the shared environmental effects and E is the effect due to non-shared environmental effects. The range of heritability is between 0.36 and 0.46. The highest heritability occurs when using the phenotype derived from the GoM model, which is 0.46 with a 95% confidence interval of 0.43 to 0.49. This indicates, if the assumptions for the ACE model hold, that 46% of total variation is due to genetic variation, none of the variation is due to shared environment effects and nearly 54% is due to the random environmental effects.

The heritability estimates obtained using LCA and Fanny phenotypes are close: respectively 37% and 36%. The variation due to shared environmental effects is consistent between these two approaches, and is in line with that obtained for the GoM approach. The non-shared environmental effects for these two approaches

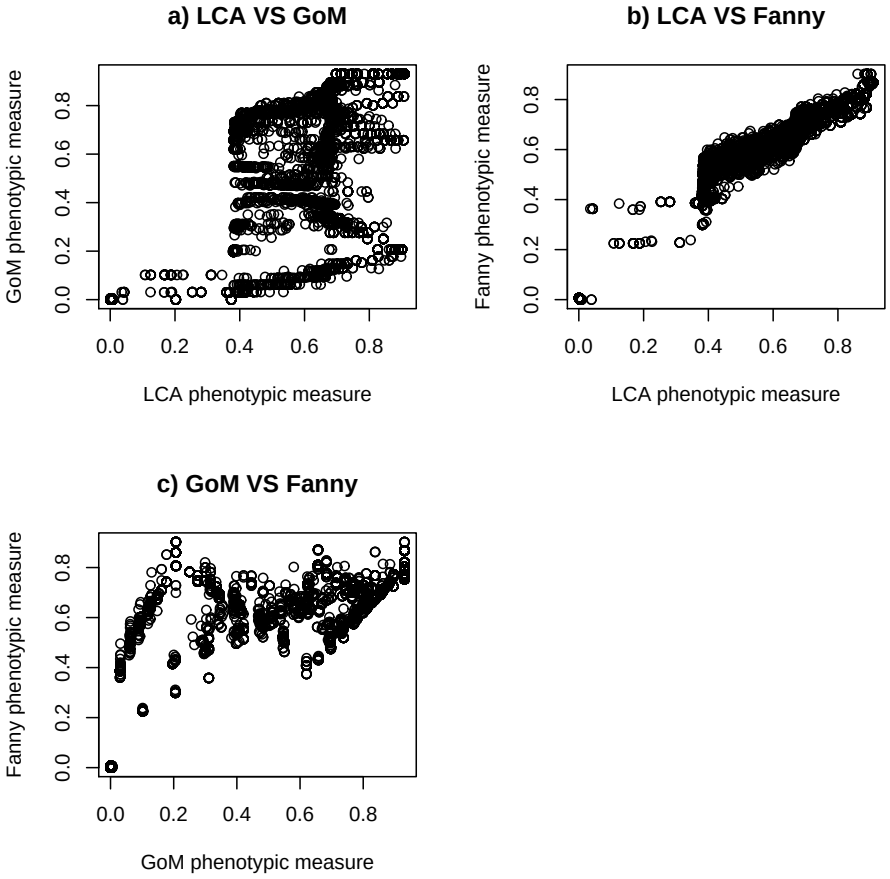

Figure 3.5: Scatter plots showing the relationship between phenotypic scores estimated under different methods. The top left plot is the estimated phenotypic score from LCA vs GOM. The top-right hand plot is the comparison in estimated trait between LCA and Fanny approaches; the bottom plot is the comparison of estimated trait between GoM and Fanny approaches.

are 63% and 64%.

Merlin-qtl multipoint LOD scores using the three different phenotypic measures were calculated at 1-cM increments; see Figure 3.6. The black solid line is the LOD score corresponding to the LCA phenotype; the red dashed line corresponds to GoM and the green dotted line corresponds to Fanny. The LOD scores based on LCA and Fanny show very similar patterns with several regions on chromosome 7 having LOD scores over 3. The highest LOD scores are on chromosome 7 at the 136 cM region (LCA LOD=3.7; Fanny LOD=4.12) followed by chromosome 7 at the 133 cM region. (LCA LOD=3.28, Fanny LOD=3.47). The third highest LOD score is also found in chromosome 7 at 127cM (LCA LOD=2.72; Fanny LCA=3.05).

Table 3.8: The migrainous headache heritability estimates from the ACE model, where A is the variability due to genetic variation and C is the variability due to environmental effect.

Model	BIC	Components	Mean	Lower CI	Upper CI
LCA	-48352.60	A	0.3710	0.3365	0.4007
		C	0.0000	0.0000	0.0000
		E	0.6290	0.5993	0.6569
GoM	-48429.35	A	0.4625	0.4308	0.4905
		C	0.0000	0.0000	0.0000
		E	0.5375	0.5095	0.5665
Fanny	-48079.38	A	0.3592	0.3266	0.3877
		C	0.0000	0.0000	0.0000
		E	0.6408	0.6104	0.6720

Although the LOD score signals are not as high as in chromosome 7, the genomewide linkage analysis shows possible evidence of linkage on chromosomes 2 and 1 in LCA and Fanny traits. Markers D28364 G, GATA194A05 and D2S1391, which are between 187 and 188 cM of chromosome 2, have a LOD score of 1.89 based on the LCA traits and 2.25 for the Fanny traits; and marker ATA73A08 (156cM) on chromosome 1 shows a small peak.

In contrast, the LOD scores based on the GoM phenotypes show a very different pattern. The highest LOD score of the GOM trait is on Chromosome 2 between 210 cM (LOD=3.10); followed by chromosome 2 at the 206 cM region (LOD=2.81). Some signals are detected on chromosome 1 and 7; marker AGAT119 M (153 cM) on chromosome 1 has a LOD score of 2.59 and marker ATA55A05 M (127cM) on chromosome 7 has a LOD score of 2.51.

3.5 Discussion

Genetic research of diseases with a complex etiology firstly requires the identification of phenotypes which capture the underlying phenotypic and genetic variance. In this study, the aim was to investigate the effects of different clustering methods on the output of genetic analyses using a previously described [212] and subsequently updated migraine dataset. We tested three commonly used statistical clustering phenotyping methods: LCA, GoM and Fanny. Of these, the first two are model-based approaches, whereas Fanny is based on partitioning of a dissimilarity matrix. Our results show that with the same symptom response

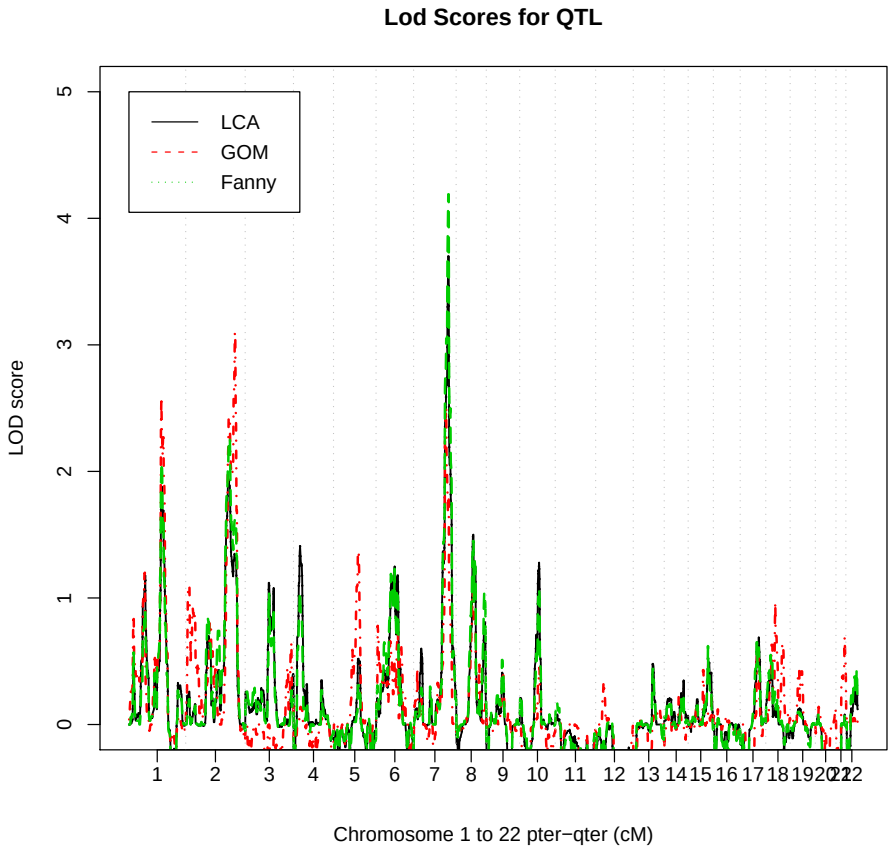

Figure 3.6: Results of MERLIN-qtl genomewide linkage analysis using traits derived from different statistical clustering methods. The solid black line is the LOD score of traits derived from LCA; red dashed line is the LOD score of trait from GoM and green dotted line is LOD score of Fanny traits. The dotted vertical lines indicate the boundaries between chromosomes.

data, different phenotype clusters are derived and as a consequence different genetic loci are implicated via linkage.

The heritability estimated here with three different migraine phenotypic traits is within the range of previously published findings [202]. [202] show that the heritability of MA and MO varies for different populations. For the Australian population, previously published results indicate the heritability varies as different phenotyping methods are applied [211]; this is supported by our findings. The ACE model fitting indicated the greater genetic contribution to migraine using GoM, followed by LCA and Fanny, which are 46%, 37% and 36%, respectively. Some of these estimates are higher than the heritability for the IHS criteria defined phenotype published in [211, 212]. We also noted that differences in heritability can occur within a model.

For instance, using the same LCA model, the heritability (h^2) of the converted continuous trait is slightly lower than the h^2 of the dichotomous trait in [211, 212]. We failed to identify the shared environmental effects for these phenotypic traits, as also occurred in [211]. [211] found that when additive genetic effects are present, the power to detect the shared environmental effects is low.

The difference between the continuous trait values derived from the LCA and GoM models is mainly due to the different clustering structure. Although the number of clusters in both models are the same, the characteristics of clusters are very different. The clusters of the GoM model differ in symptom composition but the clusters of the LCA model are different in the probability of having all ten symptoms.

The two model-based clustering methods implicated different genetic loci. However, based on the GoM phenotype, linkage was obtained to a locus near marker D2S2944 on chromosome 2 and to loci on chromosomes 1 and 7. Conversely, the two most unlike clustering methods, LCA and Fanny, not only produced linkage at the same positions but also gave the same ranking to those positions. The linkage analysis of LCA- and Fanny-based traits had highest LOD scores at Chr7q33 and Chr7q32.3 regions, respectively.

Although the markers with the highest LOD scores in the LCA and Fanny phenotype analyses are not implicated in the GoM linkage results, the genetic analysis of GoM produced linkage to other possible markers on chromosome 1 and 7. Marker AGAT119M of chromosome 7 has the fifth highest LOD score for the GoM trait, and the third highest LOD score ranking of the LCA and Fanny traits. In contrast, although linkage analysis of LCA and Fanny traits did not provide strong evidence for linkage to marker AGAT119M on chromosome 1 (LOD scores less than 2), there is still some evidence of linkage.

Although the LOD scores for some loci are less than 3, our analysis was able to replicate some previously identified regions. The small peak on chromosome 1 of LCA and Fanny traits is within 2cM of the familial hemiplegic migraine (FHM) type 2 ATP1A2 gene [59, 280]. The small peak in chromosome 2 is also within a small distance of the SCN1A FHM3 gene found by another study [62]. Another important marker is GGAT1A4, which is located on the chr 10q22.3-10q23.1 region. Our genome-wide linkage results indicated a suggestive linkage in this region. This is encouraging because the same region has been identified previously by [11, 10] and [212]. Unlike much other research, [11] adopted three different methods to

phenotype the migraine patients of the Australian and Finnish populations; this includes the less stringent form of IHS defined MA, LCA and trait-component analysis. Note the phenotypic traits derived from their LCA is calculated using a different algorithm from the one used here and [11] implement the same algorithm as the one described in [211]. We will later explain the difference between these two approaches and discuss the effects of these algorithms on linkage analysis. Previously detected loci, chr 6p12.2-p21.1 and 8q21 [212], are also detected here with suggestive linkage when the trait values are derived from LCA and Fanny.

Some previously identified loci were not detected here; this includes 4q21[28], 4q24 [286, 10, 162], 4q21-q31 [10], 5q21 [212], 8q21[11, 212], 14q21-q23 [257, 11], 15q11-q13 [10], 17q13 [10], 18q12 [28, 286, 11]. Here are some possible causes of this difference. Firstly, the common form of migraine, according to IHS criteria, is an ensemble of multiple symptoms; each symptom may be caused by specific loci and these loci contribute to susceptibility to migraine [212, 10, 162]. For the formation of common migraine, genes may need to act synergistically. One drawback of single-locus linkage analysis is that it is not able to detect epistasis effects, which commonly present in a complex disease. Therefore, the development of genome wide association studies in conjunction with statistical tools for detecting epistasis effects is more suitable for detecting the genetic architecture of migraine.

Another possible cause for not replicating previously detected loci is the variation of phenotyping methods adopted in other studies. Our results indicate that different phenotyping methods can result in different loci being identified in linkage analysis; hence it is not surprising that some previously prominent genes go undetected here. We do not advocate our findings as superior to others, or vice versa, but they do demonstrate the need to base linkage analysis on different trait values derived from various methods to ensure the validity of the conclusion. This is especially true for diseases with complex etiology.

Differences in the results of genetic analyses can occur not only between models, but also within a model. [211, 212] applied LCA to migraine survey data and identified four subgroups of migraine/severe headaches. Individuals classified into clusters 2 and 3 were treated as “affected” and given a trait value of 1 and conversely individuals in the other two clusters were given a trait value of 0. The authors then conducted a regression using MERLIN and found the highest LOD scores on chromosomes 5, 10, 8, 1 and 6. Although

the current results cannot be readily compared to those present in [212], due to differences in available phenotype data and modelling approach, we replicated their procedure and generally we found lower LOD scores but in similar positions to those identified by [212]. The main difference between the approaches used by [212] and those in the current paper is that the former employed discrete cluster membership as an ”affection” trait, whereas the current results utilized a continuous phenotypic score related to cluster membership.

To investigate further the effect of different clustering approaches on within-model effects, we separately tested the LCA and GoM models with predefined values of K. When K = 2, the results of the genetic analysis based on both the LCA and GoM are different from those obtained when K=4. Within a GoM phenotyping analysis, when K is 2, the highest LOD score is 2.29 at D1S484 on chromosome 2, which is 53 cM from the loci identified using the optimum GoM model. The within-model effect is more apparent for the LCA phenotypes, where not only the linkage position changed, but the highest LOD score reduced from 3.70 to 2.03. This demonstrates the influence of the number of clusters on the model-based clustering approaches.

The likelihood ratio test statistics and BIC used in the present analysis for model selection are common parsimony criteria but are not ideal for mixture models [182]. More sophisticated methods, such as bootstrapping [190] or reversible jump Markov chain Monte Carlo methods (RJMCMC) [228], may be more effective in searching for the optimum number of clusters in a finite sample space. The work by [23] provides a framework for using Bayes factors for component selection in mixture models.

Despite the fact that LCA and GoM are both forms of mixture models, they are quite different in practice. In GoM, the membership scores of individuals are estimated as model parameters, so the number of parameters in the model increases dramatically with the sample size. The increase in number of parameters not only slows down the computation of the model, but it also has an effect on the determination of the optimum number of components, where the criteria for model selection are based on a parsimony measure.

Another drawback of GoM, which is also shared by LCA, is in the algorithm for parameter estimation. Both of these methods are implemented using an iterative algorithm, such as EM, to find maximum likelihood

estimates. These procedures may only find the local maximum as the model becomes complex [167]. Therefore, to ensure the achievement of a global maximum, re-estimation of the model parameters with multiple starting points is recommended. As is common in such cases, it is difficult to provide guidance as to how many starting points should be used, but one rule of thumb is to repeat the optimization until each observed local maximum is attained from more than one starting point.

The large number of parameters involved in the GoM model can also result in instability of the estimation of membership score, g_{ik}. [179] has suggest various modifications to improve consistency: in particular, by assuming g_{ik} for individual i is a realization of a random vector, with a distribution function.

Although the Fanny algorithm is relatively simpler and computationally easier, there are some limitations associated with this approach. Firstly, the Fanny algorithm clusters data without taking into account any structure in the data. It is therefore essential to have two extreme response patterns in the data, ideally individuals with all symptoms, and individuals without all symptoms with heavy weights on both patterns.

Clustering using the Fanny algorithm is highly dependent on the dataset and consequently the clustering structure often changes when extra data are included in the analysis. Unless the sample is representative of the population, the phenotypic measures determined from a small sample may be biased. Another limitation of the Fanny algorithm is that as sample size and the number of questions increases, the computational requirements for the dissimilarity matrices also increase.

Of all three models, LCA is most computationally efficient, but it is not fully exempt from the effects of increasing parameter dimension. Computational time also increases rapidly with the number of latent classes (K), manifest variables (J) and levels within each manifest variable (L_j). When the number of parameters exceeds the number of samples, or one fewer than the total number of cells in the cross-classification table of manifest variables, the LCA will not be identifiable [167].

This study is based on the assumption that the migrainous population is composed of multiple subgroups. But it remains uncertain that the population that suffers from migrainous headaches is unidimensional. Therefore, models such as latent trait analysis may exhibit better performance than any clustering based statistical methods.

In this thesis, we adopted the somewhat innovative practice of converting cluster memberships to continuous phenotype scores. We regard this practice as preferable to the arbitrary imposition of a threshold, which effectively separates individuals into cases and controls. However, we urge caution in the use and interpretation of such phenotype scores. In particular, the practice assumes that the disease can be satisfactorily modeled as the result of a single, unidimensional, continuous determinant of severity. One should therefore investigate whether the clusters can be placed in a natural order of monotonically decreasing severity, as we have done here. We suggest further research into the relative merits of using continuous phenotype scores as opposed to thresholds.

In conclusion, different phenotyping methods have different properties; not knowing the true phenotypic structure of the population, phenotyping methods can therefore only provide approximations to the trait. To minimise the impact of phenotypic uncertainties, we suggest the following alternative approaches:

1. Phenotype Integration Run multiple phenotyping methods and integrate the results of different methods to produce a single phenotype. Then perform linkage analysis on this integrated phenotype.

2. Eliminate ambiguous cases Eliminate cases with phenotypes that differ for different phenotyping methods, thus limiting subsequent analysis to those individuals for which all methods produce essentially the same classification.

3. Multiple linkage analysis Run multiple linkage analysis on the phenotypic classifications derived from different models, using different clustering techniques and different numbers of classes. Then combine the results of these multiple analyses with a voting mechanism.

Such approaches may facilitate more stable estimation of genetic linkage for diseases with complex etiology. We recommend further research into the relative success of such approaches.

Bibliography

[1] Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics 30*, 97–101.

[2] Abecasis, G. R., S. S. Cherny, W. O. C. Cookson, and L. R. Cardon (2001). *GRR: graphical representation of relationship errors*, Volume 17. Oxford University Press.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

[10] Anttila, V., M. Kallela, G. Oswell, M. A. Kaunisto, D. R. Nyholt, E. Hamalainen, H. Havanka, M. Ilmavirta, J. Terwilliger, and E. Sobel (2006). Trait components provide tools to dissect the genetic susceptibility of migraine. *The American Journal of Human Genetics 79*(1), 85–99.

[11] Anttila, V., D. R. Nyholt, M. Kallela, V. Artto, S. Vepsäläinen, E. Jakkula, A. Wennerström, P. Tikka-Kleemola, M. A. Kaunisto, and E. Hämäläinen (2008). Consistently replicating locus linked to migraine on 10q22-q23. *The American Journal of Human Genetics 82*(5), 1051–1063.

[23] Berkhof, J., I. van Mechelen, and A. Gelman (2003). A bayesian approach to the selection and testing of mixture models. *Statistica Sinica 13*, 423–442.

[28] Björnsson, A., G. Gudmundsson, E. Gudfinnsson, M. Hrafnsdóttir, J. Benedikz, S. Skúladóttir, K. Kristjánsson, M. L. Frigge, A. Kong, K. Stefánsson, and J. R. Gulcher (2003). Localization of a gene for migraine without aura to chromosome 4q21. *The American Journal of Human Genetics 73*(5), 986–993.

[37] Cader, Z. M., S. Noble-Topham, D. A. Dyment, S. S. Cherny, J. D. Brown, G. P. A. Rice, and G. C. Ebers (2003). Significant linkage to migraine with aura on chromosome 11q24. *Human Molecular Genetics 12*(19), 2511–2517.

[41] Carlsson, A., L. Forsgren, P. O. Nylander, U. Hellman, K. Forsman-Semb, G. Holmgren, D. Holmberg,

and M. Holmberg (2002). Identification of a susceptibility locus for migraine with and without aura on 6p12.2-p21.1. *Neurology 59*(11), 1804–1807.

[42] Cassidy, F., C. F. Pieper, and B. J. Carroll (2001). Subtypes of mania determined by grade of membership analysis. *Neuropsychopharmacology 25*(3), 373–83.

[53] Corder, E. H. and M. A. Woodbury (1993). Genetic heterogeneity in alzheimer's disease: A grade of membership analysis. *Genetic Epidemiology 10*, 495–499.

[54] Cornes, B. K., S. E. Medland, M. A. R. Ferreira, K. I. Morley, D. L. Duffy, B. T. Heijmans, G. W. Montgomery, and N. G. Martin (2005). Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 australian twin families. *Twin Research Human Genetics 8*(6), 616–632.

[59] De Fusco, M., R. Marconi, L. Silvestri, L. Atorino, L. Rampoldi, L. Morgante, A. Ballabio, P. Aridon, and G. Casari (2003). Haploinsufficiency of atp1a2 encoding the na+/k+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. *Nature Genetics 33*(2), 192–6.

[60] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

[62] Dichgans, M., T. Freilinger, G. Eckstein, E. Babini, B. Lorenz-Depiereux, S. Biskup, M. D. Ferrari, J. Herzog, A. van den Maagdenberg, and M. Pusch (2005). Mutation in the neuronal voltage-gated sodium channel scn1a in familial hemiplegic migraine. *The Lancet 366*(9483), 371–377.

[70] Duffy, D. L. (2002). Sib-pair version 0.99. 9. *Queensland Institute of Medical Research, Brisbane, Australia*.

[71] Duren, W. L., M. P. Epstein, M. Li, and M. Boehnke (2003). Relpair: A program that infers the relationships of pairs of individuals based on marker data.

[78] Epstein, M. P., W. L. Duren, and M. Boehnke (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics 67*(5), 1219–1231.

[80] Erosheva, E. A. (2002a). *Grade of membership and latent structure models with application to disability survey data*. Ph.d., Carnegie Mellon University.

[81] Erosheva, E. A. (2002b). Partial membership models with application to disability survey data. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery*, pp. 117–133. Boca Raton, FL: Chapman and Hall/CRC.

[79] Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, rasch, and latent class models. *Psychometrika 70*(4), 619.

[85] Fillenbaum, G. G. (1998). Typology of alzheimer's disease: findings from cerad data. *Aging and Mental Health 2*(2), 105–127.

[101] Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. part ia modified latent structure approach. *American Journal of Sociology 79*(5), 1179.

[102] Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

[109] Hallmayer, J. F., A. Jablensky, P. Michie, M. Woodbury, B. Salmon, J. Combrinck, H. Wichmann, D. Rock, M. D. Ercole, S. Howell, M. Dragovic, and A. Kent (2003). Linkage analysis of candidate regions using a composite neurocognitive phenotype correlated with schizophrenia. *Molecular Psychiatry 8*(5), 511.

[115] Headache Classification Committee of the International Headache Society (1988). Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalagia 8*, 1–96.

[116] Heath, A. C., W. Howells, K. M. Kirk, P. A. Madden, K. K. Bucholz, E. C. Nelson, W. S. Slutske, D. J. Statham, and N. G. Martin (2001). Predictors of non-response to a questionnaire survey of a volunteer twin panel: findings from the australian 1989 twin cohort. *Twin Research 4*(2), 73–80.

[132] Jablensky, A. (2006). Subtyping schizophrenia: implications for genetic research. *Molecular Psychiatry 11*, 815–836.

[140] Kaabi, B. and R. C. Elston (2003). New multivariate test for linkage, with application to pleiotropy: Fuzzy haseman-elston. *Genetic Epidemiology 24*(4), 253–264.

[141] Kaabi, B., J. Gelernter, S. W. Woods, A. Goddard, G. P. Page, and R. C. Elston (2006). Genome scan for loci predisposing to anxiety disorders using a novel multivariate approach: Strong evidence for a chromosome 4 risk locus. *The American Journal of Human Genetics 78*, 543.

[143] Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. New York: Wiley.

[149] Kong, A. and N. J. Cox (1997). Allele-sharing models: Lod scores and accurate linkage tests. *The American Journal of Human Genetics 61*(5), 1179–1188.

[150] Kong, X., K. Murphy, T. Raj, C. He, P. S. White, and T. C. Matise (2004). A combined linkage-physical map of the human genome. *The American Journal of Human Genetics 75*(6), 1143–8.

[155] Kruglyak, L., M. J. Daly, M. P. Reeve-Daly, and E. S. Lander (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics 58*(6), 1347–1363.

[159] Lange, K., D. Weeks, and M. Boehnke (1988). Programs for pedigree analysis: Mendel, fisher, and dgene. *Genetic Epidemiology 5*(6), 471–2.

[161] Lazersfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. S. Stouffer (Ed.), *Measurement and Prediction*, pp. 362–412. Princeton, NJ: Princeton University Press.

[162] Lea, R. A., D. R. Nyholt, R. P. Curtain, M. Ovcaric, R. Sciascia, C. Bellis, J. MacMillan, S. Quinlan, R. A. Gibson, and L. C. McCarthy (2005). A genome-wide scan provides evidence for loci influencing a severe heritable form of common migraine. *Neurogenetics 6*(2), 67–72.

[164] Ligthart, L., D. I. Boomsma, N. G. Martin, J. H. Stubbe, and D. R. Nyholt (2006). Migraine with aura and migraine without aura are not distinct entities: Further evidence from a large dutch population study. *Twin Research and Human Genetics 9*(1), 54–63.

[165] Ligthart, L., D. R. Nyholt, J. J. Hottenga, M. A. Distel, G. Willemsen, and D. I. Boomsma (2008).

A genome-wide linkage scan provides evidence for both new and previously reported loci influencing common migraine. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.

[167] Linzer, D. A. and J. Lewis (2007). polca: Polytomous variable latent class analysis.

[176] Maechler, M. M. and M. Hubert (2008). The cluster package.

[179] Manton, K. G., M. A. Woodbury, and H. D. Tolley (1994). Statistical applications using fuzzy sets. pp. 68. Wiley.

[180] Manton, K. G., G. Xiliang, H. Hai, and M. Kovtun (2004). Fuzzy set analyses of genetic determinants of health and disability status. *Statistical Methods in Medical Research 13*(5), 395–408.

[182] Marin, J. M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics 25*, 459507.

[186] McCutcheon, A. L. (1987). *Latent Class Analysis*. Quantitative Applications in the Social Science. Newbury Park: Sage Publications.

[190] McLachlan, G. J., D. Peel, K. E. Basford, and P. Adams (1999). The emmix software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software 4*(2).

[202] Mulder, E. J., C. Van Baal, D. Gaist, M. Kallela, J. Kaprio, D. A. Svensson, D. R. Nyholt, N. G. Martin, A. J. MacGregor, and L. F. Cherkas (2003). Genetic and environmental influences on migraine: a twin study across six countries. *Twin Research 6*(5), 422–31.

[206] Neale, M. C., V. I. for Psychiatric, G. Behavioral, P. Department of, and V. Medical College of (1997). *MX: Statistical Modeling*. Department of Psychiatry, Medical College of Virginia.

[211] Nyholt, D. R., N. G. Gillespie, A. C. Heath, K. R. Merikangas, D. L. Duffy, and N. G. Matrin (2004). Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities. *Genetic Epidemiology 26*, 231–244.

[212] Nyholt, D. R., K. I. Morley, M. A. R. Ferreira, S. E. Medland, D. I. Boomsma, A. C. Heath, K. R. Merikangas, G. W. Montgomery, and N. G. Matrin (2005). Genomewide significant linkage to migrainous headache on chromosome 5q21. *American Journal of Human Genetics 77*, 500–512.

[213] Olesen, J. and T. J. Steiner (2004). The international classification of headache disorders, 2nd edn (icdh-ii). *British Medical Journal 75*(6), 808.

[216] Potthoff, R. F., K. G. Manton, and M. A. Woodbury (2000). Dirichlet generalizations of latent-class models. *Journal of Classification 17*(2), 315–353.

[219] R Development Core Team (2006). R 2.4.1 a language and development.

[228] Richardson, S. and P. J. Green (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(4), 731–792.

[243] Schwarz, G. (1979). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

[251] Silberstein, S., J. Olesen, M. G. Bousser, H. C. Diener, D. Dodick, M. First, P. Goadsby, H. Gobel, M. Lainez, and J. Lance (2005). The international classification of headache disorders, (ichd-ii)-revision of criteria for 8.2 medication-overuse headache. *Cephalalgia 25*(6), 460–465.

[257] Soragna, D., A. Vettori, G. Carraro, E. Marchioni, G. Vazza, S. Bellini, R. Tupler, F. Savoldi, and M. L. Mostacciuolo (2003). A locus for migraine without aura maps on chromosome 14q21.2-q22.3. *American Journal of Human Genetics 72*(1), 161.

[264] Svensson, D. A., B. Larsson, E. Waldenlind, and N. L. Pedersen (2003). Shared rearing environment in migraine: Results from twins reared apart and twins reared together. *Headache: The Journal of Head and Face Pain 43*(3), 235–244.

[265] Svensson, D. A., E. Waldenlind, K. Ekbom, and N. L. Pedersen (2004). Heritability of migraine as a function of definition. *The Journal of Headache and Pain 5*(3), 171.

[280] Vanmolkot, K. R., E. E. Kors, J. J. Hottenga, G. M. Terwindt, J. Haan, W. A. Hoefnagels, D. F. Black, L. A. Sandkuijl, R. R. Frants, and M. D. Ferrari (2003). Novel mutations in the na+, k+-atpase pump gene atp1a2 associated with familial hemiplegic migraine and benign familial infantile convulsions. *Annals of Neurology 54*(3), 360–6.

[282] Volk, H. E., R. J. Neuman, and R. D. Todd (2005). A systematic evaluation of adhd and comorbid

psychopathology in a population-based twin sample. *Journal of the American Academy of Child and Adolescent Psychiatry 44*(8), 768(8).

[286] Wessman, M., M. Kallela, M. A. Kunisto, P. Marttila, E. Sobel, J. Hartiala, G. Oswell, S. M. Leal, J. C. Papp, E. H m l inen, P. Broas, G. Joslyn, I. Hovatta, T. Hiekkalinna, J. Kaprio, J. u. r. Ott, R. M. Cantor, J.-A. Zwart, and M. Ilmavirta (2002). A susceptibility locus for migraine with aura, on chromosome 4q24. *American Journal of Human Genetics 70*(3), 652.

[287] Wessman, M., G. M. Terwindt, M. A. Kaunisto, A. Palotie, and R. A. Ophoff (2007). Migraine: a complex genetic disorder. *The Lancet Neurology 6*(6), 521–532.

[288] Whittemore, A. S. and J. Halpern (1994). A class of tests for linkage using affected pedigree members. *Biometrics 50*(1), 118–127.

[289] Woodbury, M. A., J. Clive, and A. Garson Jr (1978). Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and Biomedical Research 11*(3), 277–98.

[305] Zhu, G., D. M. Evans, D. L. Duffy, G. W. Montgomery, S. E. Medland, N. A. Gillespie, K. R. Ewen, M. Jewell, Y. W. Liew, and N. K. Hayward (2004). A genome scan for eye color in 502 twin families: most variation is due to a qtl on chromosome 15q. *Twin Research 7*(2), 197–210.

[306] Ziegler, D. K., Y.-M. Hur, T. J. Bouchard, R. S. Hassanein, and R. Barter (1998). Migraine in twins raised together and apart. *Headache: The Journal of Head and Face Pain 38*(6), 417–422.

4

Bayesian Latent Trait Modeling of Migraine Symptom Data

Chapter Summary

Similar to the previous chapter, the aim of this chapter is to understand how different phenotyping methods affect the results of the subsequent genetic analysis. However, in contrast to the previous chapter, models included here are two very different latent variable models.

The two models included here are latent class analysis (LCA) and item response theory (IRT). LCA is a mixture of Bernoulli distributions, and IRT, which is also known as latent trait analysis, assumes the underlying latent value measures an individual's propensity, which associates with symptom responses by fitting logistic curves. Another major difference in this chapter is that these models are proposed and compared in a Bayesian context, which allows common ground for comparing the two models.

From a statistical perspective, the main contribution of this chapter is introducing the use of Bayesian LCA and IRT for phenotyping, as well as comparing models using a recently proposed deviance information criteria that is suited for comparing latent models. Because models are proposed in a Bayesian context, it provides a common framework for model comparison.

Chapter Conclusion

Again the same migraine data as used in Chapter 3 is used here as the baseline of comparison. Even though BLCA and BIRT have a very different underlying structure, the phenotypes derived from these two models are highly correlated. Subsequently, the estimated heritability and the loci identified by the linkage analysis are nearly identical under both approaches. The estimated heritability for migraine is around 36%, which matches previous published results.

Unlike the previous chapter, even though BIRT model is structurally more complicated than BLCA, due to the use of deviance information criteria (DIC), BIRT is not heavily penalized and thus comparable to its counterpart.

Authorship

Carla C.M. Chen, Kerrie L. Mengersen, Jonathan M. Keith

Discipline of Mathematical Sciences, Queensland University of Technology

Nicholas G. Martin, Dale R. Nyholt

Genetic Epidemiology Unit, Queensland Institute of Medical Research

Reference

Chen, C. C.-M., Keith, J. M., Nyholt, D. R., Martin, N. G., and Mengersen, K. L. (2009). Bayesian Latent Trait Modeling of Migraine Symptom Data. *Human Genetics* 126, 277 - 288.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for the associated publication is:
Chen CC-M, Keith JM, Nyholt DR, Martin NG, Mengersen KL (2009) Bayesian Latent Trait Modeling of Migraine Symptom Data. *Human Genetics* 126: 277 - 288

Contributor	**Statement of contribution**
C C.M Chen	conception and conduct the research, write the code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature & Date:	
Keith JM	conception, interpretation, editing
Nyholt DR	conception, interpretation, design of questionnaire, data collection, editing
Martin NG	conception, design of questionnaire, data collection, interpretation, editing
Mengersen KL	conception, execution, editing, interpretation

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

4.1 Abstract

Definition of disease phenotype is a necessary preliminary to research into genetic causes of a complex disease. Clinical diagnosis of migraine is currently based on diagnostic criteria developed by the International Headache Society. Previously, we examined the natural clustering of these diagnostic symptoms using latent class analysis (LCA) and found that a four-class model was preferred. However, the classes can be ordered such that all symptoms progressively intensify, suggesting that a single continuous variable representing disease severity may provide a better model. Here, we compare two models: item response theory and latent class analysis (LCA), each constructed within a Bayesian context. A deviance information criterion (DIC) is used to assess model fit. We phenotyped our population sample using these models, estimated heritability and conducted genome-wide linkage analysis using Merlin-qtl. LCA with four classes was again preferred.

After transformation, phenotypic trait values derived from both models are highly correlated (correlation = 0.99) and consequently results from subsequent genetic analyses were similar. Heritability was estimated at 0.37, while multipoint linkage analysis produced genome-wide significant linkage to chromosome 7q31-q33 and suggestive linkage to chromosomes 1 and 2. We argue that such continuous measures are a powerful tool for identifying genes contributing to migraine susceptibility.

4.2 Introduction

Research into the genetics of complex diseases often involves the identification of genes associated with groups of patients that exhibit different combinations of disease symptoms or phenotypes. This analysis depends crucially on the careful classification of patients. Commonly, the clustering of patients depends on the criteria established by medical societies, such as the International Headache Society [213, 251, 115] for migraine. Without doubt, these criteria are valuable for the diagnosis of diseases, but their effectiveness for genetic research is debatable [109, 287] as discussed below.

Migraine is a hereditary disorder with estimated heritability between 34% and 57% [306, 202, 264, 265,

211, 212]. The two most common forms of migraine are migraine without aura (MO) and migraine with aura (MA), where aura typically concerns a visual disturbance. The genetic research of migraine is mainly focused on these two subgroups. To date, except for CACNA1A, ATP1A2 and SCN1A - genes that contribute to a rare mendelian form of MA, familial hemiplegic migraine (FHM), no gene has been convincingly implicated in migraine (Table 4.1). This may be due to clinical and genetic heterogeneity of the disease. The phenotype defined by IHS criteria may oversimplify the complex variability among sufferers of this complex disease [10, 287]. Furthermore, there is overlap in the symptoms of MO and MA. Clinically, the symptoms of MA are a superset of the symptoms of MO. The work of [211] and [164] provides further support for the argument that MA and MO are not separate entities. Therefore, the development of an endophenotype or an alternative phenotype may give better insight into the genetics of common migraine.

Table 4.1: The chromosome regions associated with the common forms of migraine.

Phenotype	Cohort	Chromosome	Reference
MO	Icelandic	4q21	[28]
MO	Italian	14q21.2-q22.3	[257]
MA	Canadian	11q24	[37]
MA	Finnish	4q24	[286]
MA	North American Caucasians	19q13	[136]
TCA and LCA	Finnish and Australian	10q22-10q23	[11]

There are currently two main types of method for investigating the phenotypic structure of symptom survey results, one based on the use of statistical methodologies to convert the symptoms to a unidimensional value and the other based on trait component analysis (TCA), which treats each individual symptom as a response variable for the purpose of linkage analyses. [211] pioneered the use of latent class analysis (LCA) of the phenotype for migraine. The authors applied LCA to migraine symptomatic data in an Australian twin population sample and found that the best fit to the data was obtained using a model with three symptomatic latent classes; these correspond to a mild form of recurrent non-migrainous headaches, a moderately severe form of migraine and a severe form. Moreover, the estimated heritability using LCA was found to be slightly higher than the heritability estimated using IHS criteria. [212] then applied this method for genome wide linkage analysis and identified linkage to chromosome 5q21. They also replicated previously reported susceptibility loci on chromosomes 6p12.2-p21.1 and 1q21-q23.

Since migraine is a suite of symptoms and the subphenotype analysis in [212] found that individual symp-

toms are associated with specific linkage peaks in their data, there have been several attempts to identify gene loci linked to individual symptoms [10, 11]. This method is referred to as trait component analysis (TCA). [10] applied TCA to dissect the genetic susceptibility of migraine in a Finnish cohort. They found strong evidence that various migraine symptoms are linked to chromosome 4q24, including photophobia, phonophobia, intensity, unilaterality, nausea, vomiting and attack length. They also found that pulsation is linked to chromosome 17p3 and reported some suggestive linkage of the phonophobia trait to chromosome 10q22 and the "aggravation by physical exercise" trait to chromosome 12q21, 15q14 and Xp21.

Besides LCA, other clustering methods have been applied to genetic research of diseases with complex aetiology. These include grade of membership (GoM), used to analyse schizophrenia [109], mania [42] and Alzheimer's [85, 53]; model-based clustering, used to analyse anorexia nervosa [61]; and fuzzy clustering, used to analyse anxiety disorder [141]. All these algorithms aim to identify homogenous classes/components in the data, based on specified traits of interest, and estimate the parameters associated with each class.

For some diseases composed of many individual symptoms, the data may be better modeled using a continuous representation. Indeed, in earlier analyses of multi-symptom migraine data using LCA and GoM [47, 211, 212, 164, 165], the classes could be ordered in such a way that there was a gradual reduction in all symptoms, suggesting that there is a single latent continuous trait underlying the observed pattern of symptoms. It is therefore reasonable to hypothesize that the data may be modeled using a single continuous variable representing severity of the disease instead of classes.

Item Response Theory (IRT), which is also known as latent trait analysis, is a popular statistical method for modeling psychological and educational survey responses. It assumes an underlying continuous latent value which has direct influence on the responses to items. Indeed, items are designed to capture this latent value. In this thesis, the item variables are equivalent to the symptom variables. IRT has been found to be useful in behavioural genetics and genetic epidemiology, where the phenotype is often determined by the questionnaire or interview data. This method has been used for exploring the genetic and environmental influence on the timing of pubertal change [75] and the analysis of multi-symptom genetic data [74, 290].

In this thesis, we test the hypothesis proposed above by firstly introducing IRT for analysing multi-symptom

migraine data, then comparing this non-clustering method to latent class analysis (LCA). Both models are introduced in a Bayesian framework and compared using statistical measures that take into account goodness of fit and model complexity. The models are then compared further by assessing the utility of their resulting trait measures in genetic heritability and linkage analysis.

4.3 Methods

4.3.1 Data

Phenotype data Data were collected by the Queensland Institute of Medical Research (QIMR) during the course of extensive and semi-structured telephone interview studies 1993-2000. The surveys were primarily designed to assess physical, psychological, and social manifestations of alcoholism and related disorders. The sample was unselected with regard to personal or family history of alcoholism or other psychiatric or medical disorders. The data were collected over two periods, 1993-1995 and 1996-2000. The earlier interview was administered to Australian twins listed with the volunteer-based Australian Twin Registry who were born between 1902 and 1964 while the second interview was focused on twins born between 1964 and 1971. Participants of both cohorts were first asked the screening question:"Do you have recurrent attacks of headaches?" If the participant screened positive, then he/she was asked a number of questions which were developed by an experienced migraine researcher based on the IHS diagnosis criteria (Table 4.2). Although the wording of the questions is identical for both periods, the older cohort was not asked questions related to having at least 5 episodes of headaches, the duration of headaches (4-72 hours) and the severity of the pain associated with headache ("moderate to severe").

There are 13062 individuals from 6764 families participating in this analysis, with 2716 MZ twin pairs (63.6% females and 36.4% males), 3399 DZ twin pairs (34.5% female twin pairs, 22.4% male twin pairs and 43.1% opposite sex twin pairs), 12 twins with unknown zygosity and 817 non-twin siblings. The mean age of participants was 37.5 years and ages ranged from 23 to 90 years at the time of survey. Details of the collection of the migraine data are provided by [211, 212].

Although it may be argued post-survey that it would have been more complete for all members of the cohort to be asked all symptom questions, this was considered to be an unacceptable impost by the survey designers. Possible ascertainment bias was considered and discounted since analysis showed little difference in prediction of LCA and IRT by including and excluding the "no" cohort.

Table 4.2: The survey questions based on IHS criteria.

Notation	Abbreviation	Descriptions
a	≥ 5 episode	Have at least 5 episode of headaches in your life time.
b	4-72 hr	Average headache last between 4 to 72 hours
c1	Unilateral	Headache often occurs at one side of head
c2	Pulsating	Headache pain can be described as throbbing, pulsating or pounding
c3a	Moderate/severe	Headache pain can be described between moderate and severe
c3b	Prohibitive	Headache pain prohibits daily activities
d1	Nausea/vomiting	Headache associated with vomiting or feeling nausea
d2a	Photophobia	Enhance sensitivity to light
d2b	Phonophobia	Enhance sensitivity to sounds
Aura	Aura	Have visual problems such as light shower, blurring, blind spot or double vision

Genotype data The genotypic data are composed of four smaller genome-wide linkage scans performed for other studies at the Queensland Institute of Medical Research (QIMR). Genotyping for the four studies was conducted at Gemini Genomics with 426 satellite markers, Sequana Therapeutics with 519 markers, the Center for Mammalian Genetics at Marshfield Clinic Research Foundation with 776 markers and the University of Leiden with 435 markers. The details of DNA collection, genotyping methods and data cleaning are discussed in other literature [305, 54].

Graphic Representation of Relationships (GRR) [2] and RELPAIR [78, 71] were applied for the examination of the pedigree structure and identification of inconsistencies between the genotypic data and self-reported pedigree relationships. The potential misspecification, incorrect zygosity labelling of twins and sample mix-ups were identified and corrected. A small number of cases with errors could not be corrected, so were excluded in further analysis. The SIB-PAIR version 0.99.9 program by [70] was then implemented for identifying and cleaning the Mendelian inconsistencies in the genotype data.

Markers from four sources were included separately on the genetic map for the combined scan, separated

by a small distance of 0.001cM. Markers with genotypic data inconsistent between different genome scans were excluded and unlikely genotypes were identified by MERLIN [1] and omitted from further analysis. Potential map errors were identified by GENEHUNTER [155] and MENDEL [159]. Map positions were in Kosambi cM, which is estimated using locally weighted linear regression from the National Center for Biotechnology Information (NCBI) Build 34.3 physical map positions, as well as published deCODE and Marshfield genetic map positions [150]. Where the results suggested inconsistencies between genetic map distance and recombination fraction, the primer sequences for all markers in the region were BLASTed against the entire human genome sequence (http://www.ensembl.org, NCBI build 34.3). The genetic map was then revised to include the updated physical positions of all markers in the problematic regions. The revised map and the original genotype data were cleaned of unlikely genotypes using MERLIN and map errors were resolved using GENEHUNTER. More details on the collapsing of markers is in [54]. There are a total of 1770 unique markers and the combined genome scan included 4148 individuals from 919 families (143 MZ and 776 DZ twin pairs and some parent genotype).

4.3.2 Model

Latent Class analysis Suppose that there are n individuals and J observed (manifest) item response variables ($i = 1, \ldots, n;\ j = 1, 2 \ldots, J$). Let y_{ij} denote the binary response of the ith individual to symptom question j such that $y_{ij} = 1$ when the symptom j is present in person i, else $y_{ij} = 0$. Y_i is then the vector of the ith subject's responses to all symptoms. Assume that there are K latent classes embedded in the data. Let λ_{kj} be the probability of a positive response on variable j for a person in latent class k ($k = 1, \ldots, K$). Then

$$P(Y_i|\lambda, p) = \sum_{k=1}^{K} p_k \prod_{j}^{J} (\lambda_{kj})^{y_{ij}} (1 - \lambda_{kj})^{1-y_{ij}}$$

where p_k denotes the probability that a randomly chosen individual belongs to latent class k. We used the following noninformative priors:

$$p_k \sim Dirichlet(1, ...1)_k$$

$$\lambda_{kj} \sim Beta(1, 1)$$

representing equal probability of membership to any of the k classes and equal probabilities of a 0 or 1 response for the jth variable in the kth class. The posterior probability that subject i belongs to class k is given by:

$$p_{ik} = \frac{p_k \prod_j f(Y_i|\lambda_{kj})}{\sum_l p_l \prod_j f(Y_i|\lambda_{lj})}$$

where λ_k is the expected probability of membership of the kth class and $f(Y_i|\lambda_k)$ represents the probability distribution for Y_i given this probability, that is,

$$f(Y_i|\lambda_k) = \prod_j (\lambda_{kj})^{y_{ij}} (1 - \lambda_{kj})^{1-y_{ij}}.$$

The parameter vectors p and λ are estimated by Markov Chain Monte Carlo (MCMC) simulations using WinBUGS1.4 [259]. Then the latent trait value for the ith subject is given by

$$\text{Phenotypic Trait}_i = \sum_{k=1}^{K} \frac{\sum_{j=1}^{J} \lambda_{kj}}{J} \times p_{ik}. \tag{4.1}$$

Item response Theory (IRT) As before, let y_{ij} denote the binary response of person i to variable j, $y_{ij} = \{0, 1\}, i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$. Let θ_i denote the latent trait value of subject i, $\theta_i \in \mathbb{R}$ and $P_j(\theta_i)$ be the probability of observing a response score of 1 (symptom present) given the latent trait value $P_j(\theta_i) = P_j(y_{ij} = 1|\theta_i)$, which is called the item response function (IRF). Different types of IRF constitute the

subtypes of IRT. Variations of the IRT model include the Rasch model, 2-parameter logistic model (2-PL), 3-PL model and the Birnbaum model.

In this thesis, we adopt the 2-PL model, which is commonly implemented for phenotyping. The IRF for the 2-PL model is

$$P_j(\theta_i|a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \tag{4.2}$$

where variables a_j and b_j are described in the education/psychology literature as the item discriminant parameter and item difficulty parameter. Higher values of a_j indicate that item j has higher correlation with the latent trait value. The item difficulty parameter represents the point on the latent trait scale at which the probability of having the symptom is 0.5. The likelihood is thus

$$P(Y|\theta) = \prod_{i}^{n} \prod_{j=1}^{J} (p_j(\theta_i))^{y_{ij}} (1 - p_j(\theta_i))^{1-y_{ij}}$$

As in the LCA model, noninformative priors are used for parameters θ_i, a_j and b_j:

$$\theta_i \sim N(0, 1), \theta_i \in \mathbb{R}$$

$$a_j \sim N(0, 10000)$$

$$b_j \sim N(0, 10000).$$

As for Bayesian LCA, estimation was carried out by Markov Chain Monte Carlo (MCMC) using WinBUGS1.4 [259].

For both LCA and IRT models, MCMC chains were generated with 10000 iterations. The initial 5000 iterations were considered as burn-in and every fifth case of the remaining 5000 (total of 1000 cases) was extracted to build the marginal posterior distribution of the parameters. For the LCA model, a chain was generated for each value of K, $K = 2, \ldots, 7$.

4.3.3 Model Comparison

The Deviance Information Criterion (DIC) is a popular and useful method for assessing model fit and complexity for the purpose of comparing Bayesian models. The early DIC proposed by [260] is only suitable when the competing likelihood models have a closed form. It is not ideal for comparing models with missing values or mixture models [43]. [43] suggested various alternative forms of DIC for these models and compared the performance of these criteria. Here we employ the third DIC (DIC3) of their work to determine the optimum number of classes for the Bayesian LCA and compare Bayesian LCA and IRT models. DIC3 is defined as

$$DIC = -4\mathbb{E}_{\theta}[\log f(y|\eta)|y] + 2\log \hat{f}(y) \tag{4.3}$$

where y is observed data, η is a vector of model parameters and $\hat{f}(y)$ is the posterior expectation of model parameters. Further details on the calculation of DIC for Bayesian LCA and IRT can be found in Appendix A.2.

4.3.4 Genetic analysis

Heritability of the quantitative phenotype values was estimated with the ACE model, which is well suited for twin studies.The ACE model assumes that phenotypic variation is due to additive genetic effect (A), shared environmental effect (C) and random (non-shared) environmental effect (E). The heritability is then the proportion of the total variance which is due to the additive genetic effect. The analysis was carried out using Mx [206]. Mx applies a maximum likelihood method to estimate the variances and the corresponding heritability. The goodness of fit criterion used in Mx for assessing the ACE model is the Bayesian Information Criteria (BIC) [243].

Non-parametric quantitative trait linkage analysis was carried out using Merlin-qtl. Merlin-qtl was devel-

oped under the general framework of [149] and [288]. The p_{ik} of LCA and the latent trait θ_i of IRT are treated as phenotypic traits for the genetic analysis.

4.4 Results

Bayesian LCA Table 4.3 contains the DIC values for different values of K, ($K = 2, \ldots, 7$). The DIC changes most dramatically when K changes from 2 to 3 but there is little improvement after $K = 4$. Therefore, the four class model is preferred.

Table 4.3: DIC and deviance values for $K = 2, \ldots, 7$ and Bayesian IRT model.

Model	K	DIC value	Deviance
LCA	2	60801.19	60721.39
	3	51390.08	51097.95
	4	49442.02	49062.91
	5	48531.12	47577.47
	6	47236.32	45910.07
	7	46687.76	45120.79
IRT	-	51718.36	51370.00

With K equal to 4, the deviance stabilized after 5000 iterations with an approximately normal distribution, a mean of 49062.91 and standard deviation of 126.315. The posterior marginal distributions for the majority of parameters were also approximately normal, with the exceptions of the conditional probabilities of classes 1 and 4, which are bounded by the values 0 and 1, respectively. Table 4.4 lists the posterior means and the credible intervals (analogous to frequentist confidence intervals) of all parameters of Bayesian LCA for $K = 4$.

We observed a gradual increase in the probability of each symptom across the four classes. Class 1 is composed of participants with limited symptoms (Figure 4.1). In contrast, class 4 is a collection of participants with all symptoms. Except for symptoms related to the location of the pain (unilateral, C3 of Table 4.2, 74%), more than 84% of individuals in this class have all other symptoms. Nearly all members in this class described their headache pain as moderate to severe, experienced sensitivity to light as the headache occurred and described the headache attacks as inhibiting their daily activities (c1: moderate/severe, $\lambda_{4,5} = 0.997$; d2a: photophobia, $\lambda_{4,8} = 0.996$; c3b: prohibitive, $\lambda_{4,6} = 0.983$, Table 4.4).

Table 4.4: The posterior statistics of LCA model parameters and their credible intervals.

K	1 (CI)	2 (CI)	3 (CI)	4 (CI)
P_k	0.55 (0.55 - 0.56)	0.10 (0.09 - 0.12)	0.20 (0.19 - 0.22)	0.14 (0.12 - 0.2)
$\lambda_{k,1}$	0.00 (0.00 - 0.01)	0.76 (0.73 - 0.81)	0.72 (0.69 - 0.75)	0.94 (0.92 - 1.0)
$\lambda_{k,2}$	0.00 (0.00 - 0.00)	0.43 (0.39 - 0.48)	0.70 (0.67 - 0.74)	0.90 (0.88 - 0.9)
$\lambda_{k,3}$	0.00 (0.00 - 0.00)	0.34 (0.30 - 0.37)	0.43 (0.41 - 0.46)	0.71 (0.68 - 0.7)
$\lambda_{k,4}$	0.00 (0.00 - 0.00)	0.65 (0.62 - 0.69)	0.78 (0.76 - 0.80)	0.92 (0.90 - 0.9)
$\lambda_{k,5}$	0.00 (0.00 - 0.00)	0.56 (0.50 - 0.62)	0.93 (0.90 - 0.95)	1.00 (0.99 - 1.0)
$\lambda_{k,6}$	0.00 (0.00 - 0.00)	0.20 (0.15 - 0.26)	0.76 (0.72 - 0.80)	0.98 (0.97 - 1.0)
$\lambda_{k,7}$	0.00 (0.00 - 0.00)	0.18 (0.14 - 0.22)	0.51 (0.47 - 0.54)	0.93 (0.90 - 1.0)
$\lambda_{k,8}$	0.00 (0.00 - 0.00)	0.16 (0.12 - 0.20)	0.70 (0.66 - 0.75)	1.00 (0.99 - 1.0)
$\lambda_{k,9}$	0.00 (0.00 - 0.00)	0.30 (0.26 - 0.34)	0.70 (0.66 - 0.74)	0.96 (0.94 - 1.0)
$\lambda_{k,10}$	0.00 (0.00 - 0.00)	0.19 (0.16 - 0.23)	0.48 (0.45 - 0.52)	0.84 (0.81 - 0.9)

The main difference between the two intermediate classes 2 and 3 lies in five symptoms: duration of headache, severity of pain associated with headache, ability to carry out daily activities and the physical reactions associated with headache such as nausea/vomiting, sensitivity to light and sound and visual problems (b, c3a, c3b, d1, d2a, d2b, aura of Table 4.2). Members in class 3 showed higher probability of these symptoms than members in class 2. The only item experienced by more individuals in class 2 is '>5 headaches occurring in your lifetime'. Individuals in class 2 exhibited a higher frequency of headache episodes. Class 1 is the largest class with more than 55% of the total 13062 participants. The second largest class is class 3 which contained 20% of participants followed by class 4 (14%) and class 3 (10%) (Table 4.4).

Bayesian IRT Because of the very large number of parameters sin this model, the MCMC analysis required a large amount of computational memory and a long computational time. The marginal distributions of the item discriminant parameters and item response parameters (parameters *a* and *b* of Equation 4.2) were approximately normal, with posterior means and credible intervals as listed in Table 4.5.

Figure 4.2 displays results for each symptom, using the 2-PL model. The x-axis is the latent trait value; the y-axis is the probability of having the symptom and each line represents one symptom. Given a trait value, symptoms on the right side of Figure 4.2 are less likely to be described by subjects than the symptoms on the left. For instance, nearly all subjects with latent value of 1 described the headache as moderate to severe but only 60% described the headache as unilateral (Figure 4.2). Overall, the results indicate that the symptom

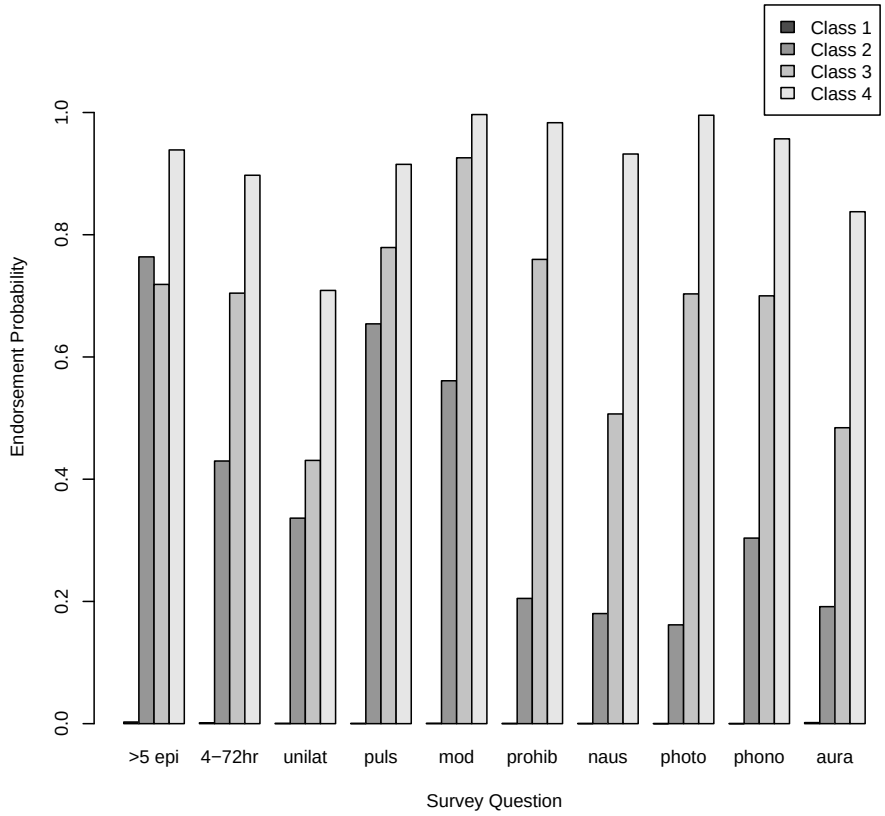

Figure 4.1: Barplot showing the symptomatic characteristics of each class under the 4 class model.

"unilateral" is the least prevalent, followed by aura and nausea/vomiting. The other symptoms have similar values of item response probability (6, Table 4.5) ranging from 0.43 to 0.65.

Table 4.5: The posterior statistics of item response probability and item discrimination parameters.

Item	**Mean**	**SD**	**2.5%**	**25%**	Median	**75%**	**97.5%**
a-a	4.074	0.12	3.844	3.992	4.073	4.154	4.311
a-b	4.245	0.135	3.983	4.152	4.244	4.336	4.518
a-c1	2.874	0.073	2.73	2.824	2.874	2.922	3.021
a-c2	4.454	0.109	4.24	4.38	4.453	4.525	4.672
a-c3a	8.368	0.361	7.688	8.117	8.36	8.598	9.095
a-c3b	6.562	0.217	6.164	6.415	6.553	6.702	7.003
a-d1	4.646	0.136	4.392	4.551	4.644	4.739	4.919
a-d2a	6.608	0.219	6.194	6.46	6.601	6.755	7.047
a-d2b	5.263	0.148	4.981	5.164	5.258	5.359	5.567
a-Aura	3.732	0.104	3.54	3.659	3.728	3.799	3.943
b-a	0.493	0.015	0.464	0.482	0.492	0.502	0.524
b-b	0.618	0.015	0.59	0.608	0.618	0.627	0.647
b-c1	0.936	0.016	0.907	0.925	0.936	0.947	0.969
b-c2	0.49	0.013	0.466	0.48	0.489	0.499	0.516
b-c3a	0.427	0.014	0.401	0.418	0.427	0.436	0.454
b-c3b	0.61	0.013	0.585	0.601	0.609	0.618	0.635
b-d1	0.781	0.014	0.756	0.771	0.781	0.79	0.807
b-d2a	0.648	0.012	0.625	0.64	0.648	0.656	0.673
b-d2b	0.623	0.013	0.6	0.615	0.623	0.632	0.648
b-Aura	0.845	0.014	0.818	0.835	0.845	0.854	0.872

A lower value of the item discrimination parameter a indicates a weaker correlation between the symptom and underlying latent trait. Of all ten symptoms, the estimated latent value correlates most strongly with the severity of pain during the headache, followed by the symptoms 'prohibitive of daily activities', photophobia and phonophobia (indicated by the posterior mean discrimination parameters of 8.368, 6.608, 6.562 and 5.263 respectively; Table 4.5). Location of pain ('unilateral') and aura correlated least strongly with the latent value.

Model Comparison The DIC estimated for the Bayesian IRT model of the migraine symptomatic data is 51718.36 (Table 4.3). This value is slightly higher than the equivalent value of 49442.02 for the best LCA model ($K = 4$). This suggests that, by this criterion, Bayesian LCA with $K = 4$ classes provides a slightly better model for these data than the Bayesian IRT model.

The models were also compared using deviance, which is -2*log-likelihood and measures the fit of a model but not its complexity. Although the difference in the deviance values between LCA with $K = 4$ and IRT

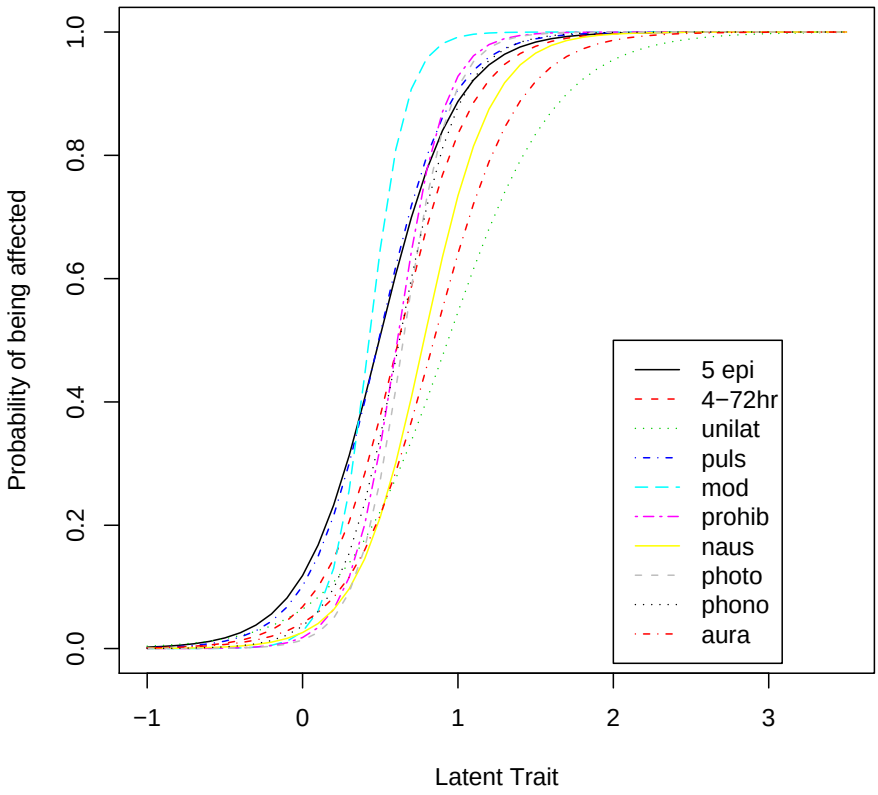

Figure 4.2: Plot showing the relationship between the latent trait and each symptom for IRT model.

is less than for DIC, lower deviance is still observed for LCA with $K = 4$ (Table 4.3). This supports the observation that LCA with $K = 4$ is a slightly better model for these data.

Figure 4.3 is a scatter plot showing the relationship between the phenotype trait values estimated using Bayesian LCA and Bayesian IRT. There is a strong correlation between phenotype values estimated with the two models (*correlation* = 0.99).

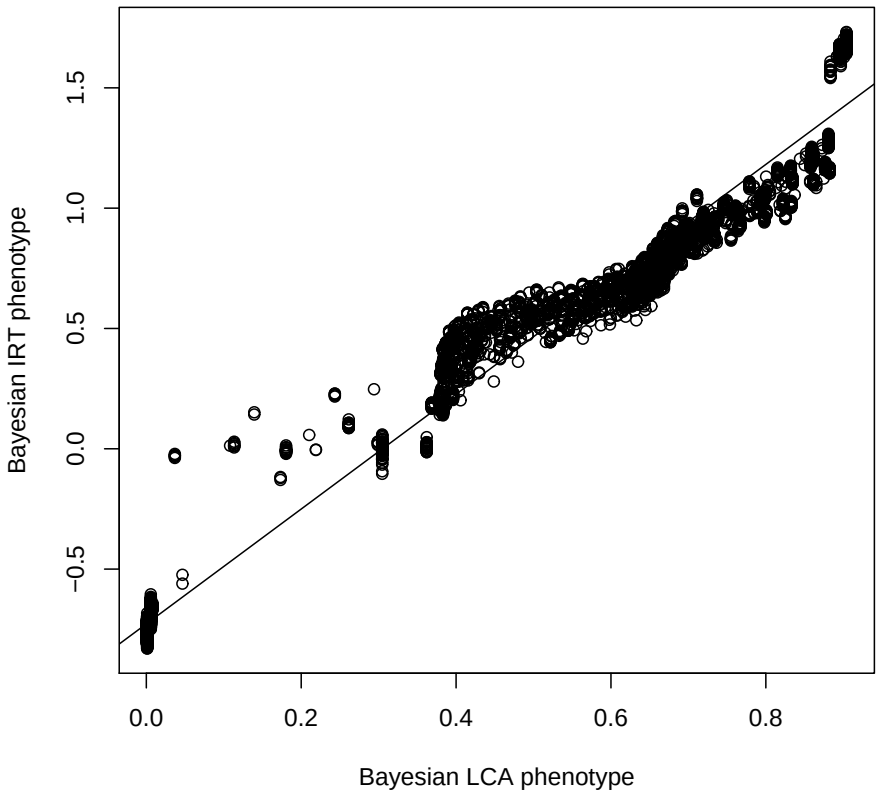

Figure 4.3: Scatter plot showing the relationship between predicted continuous phenotypic values by Bayesian LCA and Bayesian IRT model. The continuous phenotypic trait is bounded between 0 and 1, where 1 represented a severe type of common migraine and 0 indicated no evidence of common migraine. The straight line is the predicted linear relationship between these two phenotypes. The correlation between the phenotypic traits is 0.99

Genetic analysis The ACE model was fitted to the latent trait value θ of the Bayesian IRT model and the converted continuous estimate derived from Bayesian LCA (Equation 4.1), to estimate the heritability of common migrainous headache. Although the trait values derived from the Bayesian LCA model are preferable (as indicated by the smaller BIC value in Table 4.6), there is little difference in the heritability between the traits (component A of Table 4.6) due to the high correlation in the phenotypic trait values of the two models. The estimated heritability for both models is 0.37 (CI: 0.34-0.40). The non-shared environmental factor is the main contributor to the variation in the twin migraine status (62%, component *E* of Table 4.6). Interestingly, the common shared environment in twins has negligible effect on the variation

of migraine "severity" (as measures by our latent trait measures) between twin pairs.

Table 4.6: The parameters of ACE model estimated using Mx, where A is the variation due to genetic variation and C is the variability due to environmental effect. In this analysis, sex is included as a covariate.

model	BIC	Component	Mean	Lower CI	Upper CI
Bayesian LCA	-48290.53	A	0.3719	0.3413	0.4017
		C	0.0000	0.0000	0.0000
		E	0.6281	0.5983	0.6587
Bayesian IRT	-39159.34	A	0.3760	0.3475	0.4037
		C	0.0000	0.0000	0.0000
		E	0.6240	0.5963	0.6525

Figure 4.4 summarizes the results of linkage analysis using the phenotypic measures derived from Bayesian LCA with four classes using MERLIN-qtl. The black solid line of Figure 4.4 shows the LOD score of the trait derived from the posterior mean of the model parameters using Equation 4.1. Strong evidence for linkage was observed at 7q31-q33 where LOD scores are between 2.37 and 3.54. The highest LOD score (3.54) was observed for marker D7S640 on Chromosome 7, followed by a nearby marker, GATA43C11 (LOD=3.33). Besides chromosome 7, there is some suggestive evidence of linkage on chromosomes 1 and 2. The LOD scores for the area around marker ATA73A08 (153-157cM) on chromosome 1 are between 2.14 and 2.23. Marker GATA194A05 on chromosome 2 also has a LOD score above 2.0 (LOD=2.04). The next highest peak is on chromosome 8 at 86.314cM, with a LOD score of 1.85. Figure 4.5 presents similar results for trait values derived from Bayesian IRT; the black solid line shows the LOD score for the posterior mean trait. Linkage to the posterior means of Bayesian IRT indicates a maximum LOD score on chromosome 7 at 136cM. This coincides with the maximum LOD score linking to the trait estimated using the Bayesian LCA model. Similarly, the loci with the second and third highest LOD scores in the Bayesian LCA are also identified under the Bayesian IRT analysis [marker ATA73A08 (LOD=2.2) and GATA194A05 (LOD=1.99)].

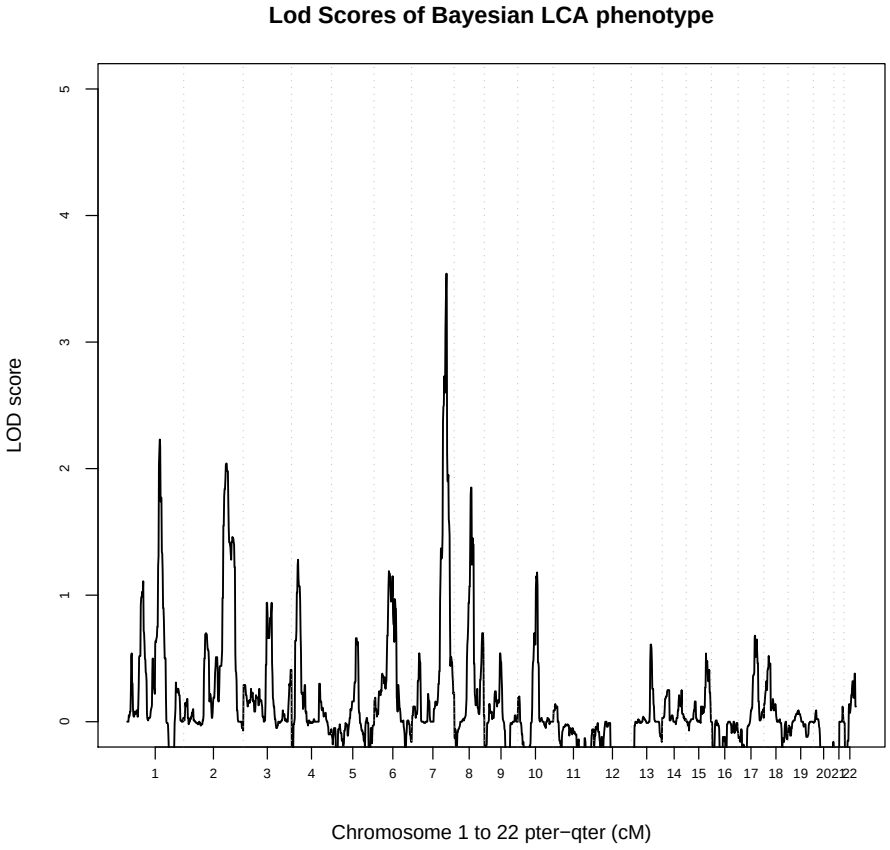

Figure 4.4: Linkage plot of phenotype derived using Bayesian LCA.

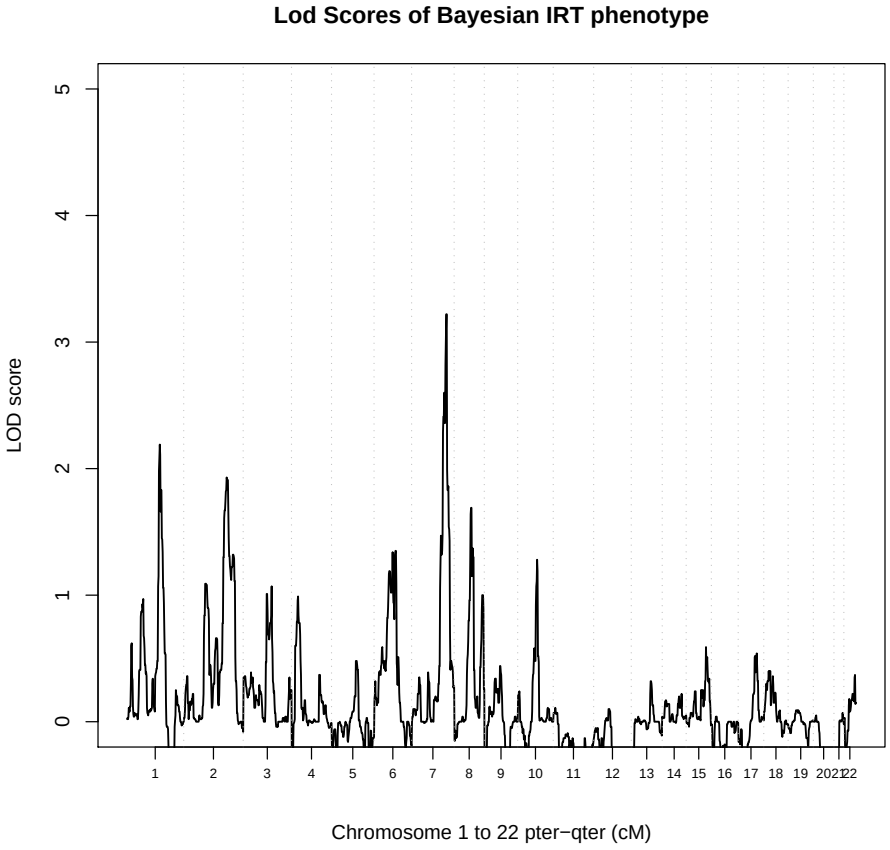

Figure 4.5: Linkage plot of phenotype derived using Bayesian IRT.

4.5 Discussion

This chapter aimed to compare two latent variable models in describing the phenotype of migraine and investigate the impact of model choice on the subsequent genetic analysis. Whereas the LCA model assumes that the subject population comprises multiple distinct subgroups or subtypes of migraine, the IRT model assumes a single continuous latent value for each subject. Both models were fitted within the Bayesian framework. Based on DIC, the LCA model with four classes provides a better fit to the data than the IRT model, but the classes could be ordered in such a way that there was a clear progression from minimum symptoms ('non-affected') in the first class through to nearly all symptoms (the most severe type of migrainous headache) in the last class. Members of the non-affected class in the Bayesian LCA model also had the lowest latent trait values under the Bayesian IRT model, compared with other classes. The two intermediate classes differ in the last five symptoms, which may be related to individual reaction during the headache episodes. An exception to the increasing progression of symptoms was the frequency of headache, which was larger in class 2 than class 3. The importance of this symptom as an indicator of the severity of migraine has been questioned by [164] in a Dutch cohort.

The characteristics of the classes identified using Bayesian LCA in this analysis are very similar to those reported by [211], but quite different from those found by [164] . The latter authors observed that except for the items related to the severity of pain and sensitivity to light and sound, the prevalence of other symptoms was much lower in their least severe class compared with the finding here. Moreover, the differences we observed here for classes 2 and 3 were not present in their cohort, with their classes 1 and 2 (corresponding to our classes 2 and 3) both composed of individuals with low physical reaction during the headaches.

A potential problem with the LCA model is that the classes identified via this method may be influenced by the composition of the population or the method of sampling–factors which have nothing to do with the aetiology of the disease. For instance, when the data are dominated by individuals with moderate migrainous headache and only a small proportion of subjects have the severe type of headache, classes derived from LCA may not represent "affected" and "non-affected" disease status. Therefore, as for all clustering approaches, the results of LCA need to be interpreted with a degree of caution and ideally with reference to clinical

criteria.

Although the IRT model fit less well with respect to the DIC and is less parsimonious than its LCA counterpart in terms of the number of parameters, it provides a valuable insight into the relationship between individual symptoms and the underlying latent value which is not directly available in the LCA model. The analysis of the Bayesian IRT revealed that the symptom 'unilateral' is less important in prediction of migraine status. This finding is supported by the Bayesian LCA with low prevalence of this symptom in all classes. This may be due to the participants' understanding of this item, or difficulty in remembering the location of the pain during the time of the survey. Surprisingly, the symptom 'aura' was reported to be the second least correlated variable to the latent trait of the Bayesian IRT model, yet this is the major symptom used in the IHS criteria in separating subjects into two subtypes, migraine with aura (MA) and migraine without aura (MO). As much as LCA and IRT are different methods, these two models complement each other and together provide a better investigation, interpretation and explanation of these data than either can provide by itself.

In our previous work [47], we found that the results of genetic analysis using traits derived from grade of membership (GoM, [289]) are very different from those obtained using traits derived from LCA and fuzzy clustering (Fanny, [143]). Based on information criteria, LCA out-performed the GoM model for these migraine symptom data. The current study demonstrates that a fourth model, IRT, produces similar results to LCA and therefore Fanny, leaving GoM as the odd method out. Further research is suggested to confirm whether the GoM model is suitable for data analyses such as those reported here.

Currently, linkage analysis is designed for either dichotomous or continuous traits and multinomial traits can only be analysed by introducing a threshold value or by conversion. As an example of the former, [212] fitted migraine symptom data using LCA with four classes, then separated the subjects into "affected" and "non-affected" based on the predicted allocation to the first two and last two classes, respectively.

Here we employed a simple conversion function to convert the multinomial trait to a continuous measure. This simple conversion included the clustering feature of LCA, as well as the uncertainty of belonging to multiple classes. Without any other manipulation, this continuous measure has a high correlation with the

latent trait of the IRT model; therefore, with some confidence, this converted value is representative of migraine "severity".

Indeed, more advanced methods could be considered for the conversion of the clustering output of LCA to a continuous phenotypic trait. Factor mixture analyses [189] which provides a general framework for combining LCA and factor analysis, is one such method.

As expected, the high correlation in the trait values of the two models resulted in minimal differences in the results from genetic analyses. Interestingly, the heritability of both traits is 0.37, which is comparable to the heritability estimated in an Australian cohort when the status is determined by the IHS criteria [$h^2 = 0.34$ [202]; $h^2 = 0.36$ [211]], despite these values being derived from substantially different data.

Analogous to the heritability results, linkage to the latent trait values from the IRT model is also nearly identical to that of the LCA continuous trait. There is strong evidence for linkage to chromosome 7q31-q33, which has not been previously identified by other studies. In addition, marker ATA73A08 and GATA194A on chromosomes 1 and 2 respectively are reported in other studies. Marker ATA73A08 is close to the familial hemiplegic migraine (FHM)-implicated ATP1A2 gene [59, 280] and GATA194A on chromosome 2 is close to the SCN1A FHM3 gene [62]. The other interesting locus identified here is on chromosome 10q22.3. Recent work by [11] applied both LCA and TCA to Australian and Finnish cohorts and successfully identified this locus linked to migraine.

Building upon our earlier work on the empirical clustering of migraine symptomatology, the results from our Bayesian latent trait modeling indicate that migraine symptom data may be modeled using a single continuous variable representing severity of the disease. The purpose of such quantitative measures is not to diagnose migraine but to provide new research tools for geneticists. For example, as in other complex diseases, the use of quantitative traits such as lipid values in hyperlipidaemia or allergy-related phenotypes in asthma provides an option for refined analysis. We therefore propose that the use of such continuous measures, which directly reflect migraine severity, provides a powerful and useful approach to identifying genes contributing to migraine susceptibility.

5

From phenotype to genotype: reconciling approaches through Bayesian model averaging

Chapter Summary

In the previous two chapters, the results show that different methods of phenotyping can result in either similar or else very different findings in the subsequent analysis. To address the first object of this thesis, the next step is to develop methods for reconciling the phenotype estimated from different models. In this chapter, we propose two new methods for achieving this goal.

The conventional approach when more than two models are used for phenotyping is to select a single model using goodness of fit measures. From the previous two chapters, we noted that even when models are comparable in likelihood; some models are less preferred due to their complexity. Given the true phenotype is unobservable, and thus validating the estimations is difficult, it is unwise to choose one model. Moreover, the choice of goodness of fit measure can be arbitrary and debatable. Therefore, instead of selecting a model, we propose a method to average models. Furthermore, one of the methods proposed here can reflect the model uncertainty in the subsequent analysis.

In addition, we propose a method to combine the model evaluation criteria by introducing an additional parameter to capture the uncertainty associated with the approximation to the marginal likelihood.

Chapter Conclusion

Using Bayesian model averaging as the foundation, we introduce two new methods for reconciling the phenotype estimated by the different models. LCA and GoM are again selected here for demonstration. Because the marginal distributions of the models are intractable, we tested two different methods of approximating the marginal likelihood within each proposed method. The methods are then validated using simulated data, and again using the migraine data.

Both methods show promising ability in integrating the phenotypes of different models by consolidating the cores of the clusters commonly identified by models, as well as reflecting model uncertainty for individuals at the borders of the clusters. We also noted that the proposal for combining the model evaluation criteria has shown promising results in overcoming the disputes associated with the weighting of the models. Therefore, the results to date indicate the value of the proposed methods.

Authorship

Carla C-M. Chen, Jonathan M. Keith, Kerrie L. Mengersen

Discipline of Mathematical Sciences, Queensland University of Technology

Reference

Chen, C. C.-M., Keith, J., and Mengersen, K. (2010). From phenotype to genotype: reconciling approaches through Bayesian model averaging. *Journal of the Royal Statistical Society C, submitted.*

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for the associated publication is:
Chen CC-M, Keith J, Mengersen K (submitted) From Phenotype to Genotype: reconciling approaches through Bayesian model averaging. *Journal of the Royal Statistical Society C*

Contributor	**Statement of contribution**
C C.M Chen	conception and conduct the research, write the code for the statistical approach, interpretation, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature & Date:	
Keith JM	conception, interpretation, editing
Mengersen KL	conception, interpretation, editing

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

5.1 Abstract

Genetic research of diseases with complex etiology is hindered by a lack of clear biomakers for phenotype ascertainment. The phenotypes for these diseases are often identified on the basis of clinically defined criteria; however such criteria may not be suitable for understanding the genetic composition of the diseases. Different statistical approaches have been proposed for phenotype definition; however the results of our previous studies have shown that differences in the phenotypes estimated by different approaches have substantial impact on the subsequent linkage analysis. Instead of obtaining results based upon a single model, we proposed two new methods, using Bayesian model averaging as the foundation, to overcome the problems associated with defining phenotype classes. Both methods reconcile phenotypes obtained from multiple models both within and across phenotype classification approaches. We illustrate the methods using latent class analysis and grade of membership, and demonstrate their application using simulated data and real data on migraine. Our methods have shown promising ability to consolidate the cores of clusters and reflect model uncertainty by increasing the fuzziness at the boundaries of clusters. Thus, in subsequent linkage analysis, loci which are strongly differentiated at the cluster cores may have stronger LOD scores under the combined model than under an individual model.

5.2 Introduction

An important goal of genetic research is to understand the composition and genetic architecture of a heritable phenotype. Springboarding from the rapid reduction in the cost of genotyping and increases in computational ability, many studies have been published on the identification of different classes or subgroups of individuals based on phenotype data. In humans alone, phenotypic classes have been identified for diverse problems ranging across food acceptance [e.g. 68], social behaviour [e.g. nicotine dependence, 26], psychological disorders [e.g schizophrenia, 109] and a wide variety of diseases [e.g. 211, 53, 129]. The results of these analyses are often then subjected to genetic analyses, typically based on linkage methods, in order to identify genes that are associated with, or can differentiate between, the phenotype classes.

For many diseases without clear biomarkers, phenotypes are identified on the basis of clinically defined criteria. While these criteria assist in clinical diagnostics, they may not be suitable for understanding the genetic architecture of the disorder [287]. Thus different statistical methods for phenotype definition have been proposed, including latent class analysis [186], grade of membership [179], item response theory [75], factor analysis [see 268], discriminant analysis [see 113] and factor mixture analysis [188]. However, different approaches can result in the identification of slightly, or sometimes substantially, different phenotype classes, which can in turn result in different linkage analysis results [47].

This problem can be addressed by either model selection or model averaging. In model selection, one chooses a single approach and, within this approach, a single model, based on a criterion such as the Likelihood Ratio (LR), Akaike Information Criterion (AIC), Bayesian information criterion (BIC), Bayes Factor (BF) or posterior predictive probabilities (PPP). However, a number of researchers have recognised that this practice ignores model uncertainty [121, 142, 67, 65, 44, 241], which can result in underestimation of the uncertainty in the quantities of interest [175]. Furthermore, the choice of criterion for model selection is often arbitrary and sometimes debatable; see, for example, the discussion on the validity of the DIC for different models by [260].

Bayesian model averaging (BMA) provides a coherent mechanism for accounting for model uncertainty [121]. The idea of BMA is to average the posterior distributions of different models, where the weight for each model depends on the posterior model probability. [175] and [223] have noted that the use of BMA can improve predictive performance. Various works have been published on the methods of BMA [142, 175, 222, 223, 121]. [121] provides a thorough overview of the history, implementation, challenges and solutions for BMA. [120] also provides a summary of BMA methodologies and lists corresponding software for carrying out the analyses.

Although the use of BMA in genetic research is not as common compared with other areas of science, some published papers have incorporated these ideas in analysis. For instance, [295] applied BMA for gene selection and classification of microarray data. [9] extended earlier research by incorporating iterative BMA for survival analysis. The use of BMA has also been implemented in the study of phylogenetics [215] and genome-wide association studies for identifying subsets of SNPs [90].

We propose here two new methods to overcome the problems associated with defining phenotype classes. Both methods allow for the integration of estimated phenotypes obtained from multiple models both within and across phenotype classification approaches. The two approaches used for illustration in this chapter are latent class analysis (LCA) and grade of membership (GOM). Both of these are commonly implemented in genetic research for deriving phenotypic traits prior to linkage or association studies, as described below. This approach to integration is similar to the "$\mathscr{M}$-open perspective" discussed in [24] and [121]. Moreover, the focus of the methods is not on prediction, but on parameter estimation. The methods are demonstrated using a real dataset on migraine and a simulated dataset obtained from the Genetic Analysis Workshop 14 [105].

5.3 Methods

Let Δ denote a quantity of interest; in the area of genetic studies, this is typically a phenotypic trait of interest. Given a data set D, the posterior distribution of Δ is

$$p(\Delta|D) = \sum_{s=1}^{S} p(\Delta|M_s, D)p(M_s|D) \tag{5.1}$$

where M_s is the model s of all models considered, $s = 1, \ldots, S$. Using Bayes theorem, the probability of M_s given data set D becomes

$$p(M_s|D) = \frac{p(D|M_s)p(M_s)}{\sum_l p(D|M_l)p(M_l)} \tag{5.2}$$

where

$$p(D|M_s) = \int p(D|\theta_s, M_s)p(\theta_s|M_s)d\theta_s \tag{5.3}$$

which is the marginal likelihood of model M_s, θ_s denotes the model parameters of model s and $p(D|M_s)$ is the marginal likelihood. In the context of this chapter, as described in Section 5.4.2, $S = 2$, M_1 is the LCA

model and M_2 is the GOM model. Various methods have been proposed for stochastic search of the model space [96, 222, 99] and alternatives have been discussed for approximating the marginal likelihood where this is intractable [99, 142, 83, 94].

Let ϕ_{is} denote the phenotype of individual i predicted by model s and let ϕ_i be the 'model averaged' phenotype for individual i, averaged over models $1, \ldots, S$. In the first method considered here (Method 1), ϕ_i is estimated as a weighted average of the posterior means of ϕ_{is} is estimated as

$$p(\phi_i|D) = \sum_{s=1}^{S} p(\phi_{is}|M_s, D)p(M_s|D) \tag{5.4}$$

which is then applied to the linkage analysis as the phenotype. In contrast, the second method (Method 2) utilises all post burn-in samples, estimates ϕ_i at each iteration and takes a weighted average of these estimates. At each iteration, the weighted average, ϕ_i, is applied to the linkage analysis as the phenotype of individual i. Let ϕ_{is}^t be the predicted phenotype of individual i by model s at the tth iteration. The posterior probability of ϕ_{is}^t is

$$p(\phi_i^t|D) = \sum_{s} p(\phi_{is}^t|M_s, D)p(M_s|D) \tag{5.5}$$

where the posterior model probability is given by Equation 5.2. However the marginal likelihood of model s becomes

$$P(D|M_s) = \int P(D|\theta_s^t, M_s)P(\theta_s^t|M_s)d\theta_s^t. \tag{5.6}$$

Given that the nature of genetic study of complex disease is hierarchical, the use of Method 2 propagates the uncertainties acquired from the first stage of model fitting into the subsequent genetic analysis.

We selected two approximations to the marginal likelihood based on the Laplace-Metropolis algorithm [163] and the BF [142]. Acknowledging the uncertainty of these approximations, we extend the algorithm further

to allow for the inclusion of Q such approximations in the analysis by introducing an extra variable c_q, $q = 1, .., Q$. The marginal likelihood is then

$$p(D|M_s) \propto \sum_q p(D|M_s, c_q)p(c_q) \tag{5.7}$$

and the posterior distribution becomes

$$p(M_s|D) \propto \frac{\sum_q p(D|M_s, c_q)p(M_s)p(c_q)}{\sum_l \sum_q p(D|M_l, c_q)p(M_l)p(c_q)} \tag{5.8}$$

The Laplace-Metropolis algorithm is based on Laplace's asymptotic approximation

$$\int e^{\log(p(D|\theta,M_s)p(\theta|M_s))} \approx (2\pi)^{\frac{d}{2}}|H*|^{\frac{1}{2}}p(D|\theta^*, M_s)p(\theta^*|M_s) \tag{5.9}$$

where d is the dimension of the parameter vector θ, θ^* is the MAP value of theta and H* is minus the inverse of the Hessian matrix which is evaluated at θ^*. Due to the difficulties in analytical estimation of θ^*, [221] suggests the use of the posterior simulation outputs to estimate the quantities required for Equation 5.9, and called it a Laplace-Metropolis algorithm. [163] provide four methods for estimating θ^*, which are simple to implement.

The BIC also uses the Taylor series expansion and the Laplace method for integrals to approximate the marginal likelihood, but is a simplified version of the approximation by [220]. The main difference between the Laplace approximation and BIC is in the error of approximation. This is discussed in details in [142]. The log marginal is approximated as the log likelihood minus a correction,

$$\log p(D|M_s) = \log p(D|\theta_s, M_s) - \frac{d}{2}\log n \tag{5.10}$$

where n is the sample size. In our examples, the first term on the right hand side is estimated using the posterior mode as θ_s.

5.4 Examples

5.4.1 Data

Data 1: Genetic Analysis Workshop 14

The first study is a simulated dataset proposed for the Genetic Analysis Workshop 14 [105]. The aim of the simulation was to reflect uncertainty difficulties and controversy associated with defining a phenotype for a hypothetical psychiatric condition, Kofendred Personality Disorder (KPD; see Table 4 of [105]). A complicated underlying genetic structure was constructed for KPD, with the involvement of four loci, denoted as D1, D2, D3 and D4. These loci interact in complex ways to produce three different phenotypes (P1, P2, P3) in which the symptoms of each sub phenotype overlap (Figure 5.1). The causal loci for each phenotype strongly overlap. The interaction of D1 and D2 results in P1; the combination of D2+D3 and D3+D4 results in P2, and the combination of D1+D4 and D2+D3 results in P3. The disease related loci are located on different parts of the genome: D1, D2, D3 and D4 are located on Chromosomes 1, 4, 5 and 9 respectively. Further details of the exact location and other genetic parameters are shown in Tables 1, 2 and 3 of [105].

The traits (symptoms) of each phenotype are also highly interchangeable. P3 has all the traits (symptoms) of P1 and P2; and P2 has nearly all the symptoms of P1. A full description of each symptom is given in Appendix A.2.2.

Four populations were generated in the original simulation study in order to test the effect of different ascertainment schemes. One of the populations is included here, namely Aipotu. The Aipotu families are selected in the analysis when at least two of the offspring have any of the true phenotypes. There are 100 replicates and each replicate contains 100 families (approximately 700 individuals). To avoid the complications associated with small sample size, 20 replicates were randomly selected to form a larger

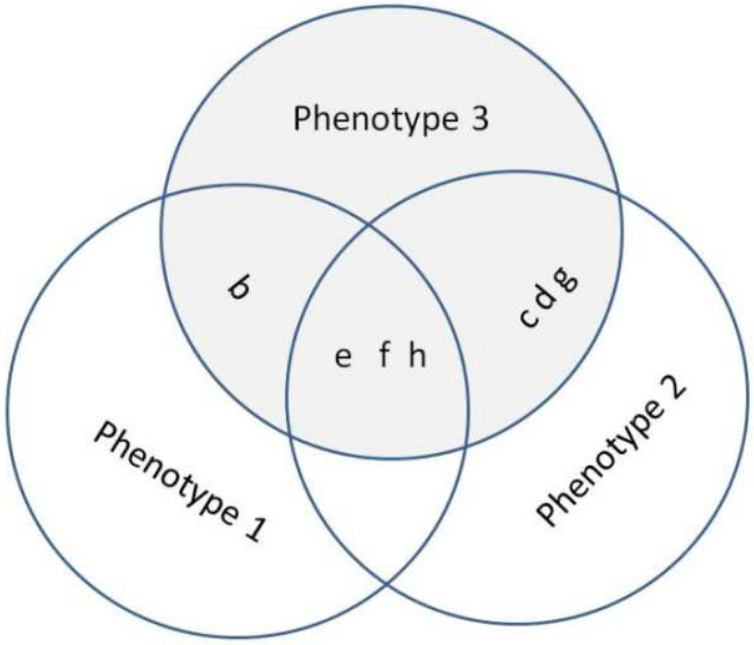

Figure 5.1: The overlapping of the traits for each of the true phenotypes. Letters b, c, d, e, f, g and h correspond to the symptoms listed in Table 4 of [105] (also in Appendix A.2.2).

dataset.

The simulation contained other interesting elements, such as single nucleotide polymorphism data and linkage equilibrium. For the purpose of reflecting the real life data (Dataset 2: migraine), only the microsatellite data are considered here. On average, the microsatellite markers are 7.5 cM apart and there are 400 markers available without missing data.

Data 2: Migraine

Migraine is a common, painful and debilitating disorder with various researchers showing a strong genetic component to the risk of this disorder [306, 202, 264, 265, 212, 211]. The diagnosis of migraine is challenging due to a lack of biomarkers and overlapping symptoms with other neurological disorders. To date, diagnosis of the disorder relies on classifying the self-reported headache characteristics using International Headache Society (IHS) criteria. According to IHS, there are two major subtypes of migraine, migraine

without aura (MO) and migraine with aura (MA); symptoms of each class are listed in Tables A.1 and A.2 in Appendix A.2.1. The early genetic research on migraine is concentrated on either MA or MO, but no genes have been convincingly replicated in follow-up studies. As a consequence, various researchers have questioned the adequacy of defining the phenotype using such criteria [10, 11, 211, 287] and have advocated instead the use of statistical methods for the identification of clusters and classes based on the symptomatic data for genetic research of this disorder [211, 212, 165].

However, our recent work has shown that when different statistical methods are used for identification of phenotype classes, the results of the subsequent linkage analysis designed to identify genes that differentiate between these classes can be surprisingly different [47].

Migraine data were obtained from an extensive semi-structured telephone interview as part of a study designed to assess physical, psychological and social manifestations of alcoholism and related disorders [116] at QIMR. The sample was unselected with regard to personal or family history of alcoholism or other psychiatric or medical disorders [202]. The interview was conducted during two periods of time, 1993-1995 and 1996-2000. The earlier interview was administered to Australian twins listed with the volunteer-based Australian Twin Registry who were born between 1902 and 1964 while the second interview was focused on twins born between 1964 and 1975.

Participants of both cohorts were first asked the screening question: “Do you have recurrent attacks of headaches?” If the participant screened positive, he/she was then asked ten questions which were developed by an experienced migraine researcher based on the IHS diagnosis criteria. A total of 13062 individuals from 6764 families participated in this study, with 2716 MZ twin pairs (63.6% females and 36.4% males), 3399 DZ twin pairs (34.52% female twins, 22.36% male twins and 43.13% mixed sex twins), 15 twins with unknown zygosity and 817 first degree family members, including both siblings and parents. Within the total of 13062 samples, 60 samples were devoid of responses, so were excluded from the analysis.

The genotypic data were obtained from four smaller genome-wide linkage studies performed at QIMR and are available for 4148 individuals from 919 families. Genotyping for the four studies was carried out at four different centers: Gemini Genomics, with 426 satellite markers; Sequana Therapeutics, with 519 markers;

the Center for Mammalian Genetics at Marshfield Clinic Research Foundation, with 776 markers; and the University of Leiden, with 435 markers. Detailed description on the DNA collection, genotyping methods and data sorting are published in [305] and [54].

Graphic Representation of Relationships (GRR) [2] and RELPAIR [78, 71] were applied for the examination of the pedigree structure and identification of inconsistencies between the genotypic data and self-reported pedigree relationships. Potential misspecification, incorrect zygosity labelling of twins and sample mix-ups were identified and corrected. A small number of cases with errors could not be corrected, so were excluded in further analysis. The SIB-PAIR version 0.99.9 program by [70] was then implemented for identifying and cleaning Mendelian inconsistencies in the genotype data.

Markers from four sources were included on the genetic map for the combined scan, separated by a small distance of 0.001cM. Markers with genotypic data inconsistent between different genome scans were excluded and unlikely genotypes were identified by MERLIN [1] and omitted from further analysis. Potential map errors were identified by GENEHUNTER [155] and MENDEL [159]. Map positions were in Kosambi cM, which is estimated using locally weighted linear regression from the National Center for Biotechnology Information (NCBI) Build 34.3 physical map positions, as well as published deCODE and Marshfield genetic map positions [150]. Where the results suggested inconsistencies between genetic map distance and recombination fraction, the primer sequences for all markers in the region were BLASTed against the entire human genome sequence (http://www.ensembl.org, NCBI build 34.3). The genetic map was then revised to include the updated physical positions of all markers in the problematic regions. The revised map and the original genotype data were cleaned of unlikely genotypes using MERLIN and map errors were resolved using GENEHUNTER. More details on the collapsing of markers is in [54]. The final genotypic data contains information on 1770 unique markers.

5.4.2 Models and Settings

As discussed in Section 5.2, in this study, we choose two common statistical methods used in genetic research for deriving phenotype classes, namely latent class analysis and grade of membership. Both of these

models are considered in a Bayesian framework.

For LCA, following [186], suppose that there are n individuals and J symptoms ($i = 1, \ldots, n$; $j = 1, \ldots, J$). Let y_{ij} denote a binary response of individual i to symptom j, such that $y_{ij} = 1$ indicates that symptom j is present in person i. Let K denote the total number of clusters. Then LCA is a mixture of Bernoulli distributions,

$$p(Y_i|\lambda, p) = \sum_{k=1}^{K} p_k f(Y_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j}^{J} (\lambda_{kj})^{y_{ij}} (1 - \lambda_{kj})^{1-y_{ij}} \tag{5.11}$$

where p_k is the weight of each component, Y_i is a vector of responses of individual i and λ_{kj} is the probability of a positive response on variable j for a subject in cluster k. Non-informative priors were adopted, namely

$$p_k \sim Dirichlet(1, ...1)_K; \qquad \lambda_{kj} \sim Beta(1, 1) \tag{5.12}$$

Introducing an auxiliary (latent) variable $z_i = \{z_{i1}, \ldots, z_{iK}\}$ as an unobservable cluster indicator for y_i, and using an MCMC approach [183], the conditional posterior distributions of p and λ are

$$\begin{aligned} p_k &\sim Dirichlet(\sum_i z_{i1} + 1, .. \sum_i z_{iK} + 1) \\ \lambda_{kj} &\sim Beta(\sum_i (z_{ik} y_{ij}) + 1, \sum_i (z_{ik} - z_{ik} y_{ij}) + 1) \end{aligned} \tag{5.13}$$

where

$$z_i \sim multinomial(\delta_{i1}, ...\delta_{iK}); \qquad \delta_{ik} = \frac{p_k \prod_j^J (\lambda_{kj})^{y_{ij}} (1 - \lambda_{kj})^{1-y_{ij}}}{\sum_l p_l \prod_j^J (\lambda_{lj})^{y_{ij}} (1 - \lambda_{lj})^{1-y_{ij}}}$$

For GoM, following [80], let g_{ik} be a latent variable of membership score, representing the probability that

individual i belongs to cluster k. Constraining the number of levels of responses in symptom j to 2, GoM is similar to a mixture of Bernoulli distributions,

$$Pr(Y|\gamma, g) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ \sum_{k} g_{ik}^{y_{ij}} \gamma_{kj}^{y_{ij}} (1 - \gamma_{kj})^{(1-y_{ij})} \right\} \tag{5.14}$$

where γ_{jk} is similar to λ_{kj} of the LCA model, and is the probability of having symptom j for an individual in cluster k. Similarly, the non-informative priors are used here,

$$g_{ik} \sim Dirichlet_i(1, ...1)_K; \qquad \gamma_{kj} \sim Beta(1, 1) \tag{5.15}$$

We introduce J categorical variables $\omega = (\omega_1, \ldots, \omega_J)$ in which each ω_j can take on K values from $\{1, \ldots, K\}$. The latent class is then defined as $\omega \in \Omega = \{1, 2, \ldots, K\}^J$.

A Gibbs sampler is again used to estimate the model parameters based on the conditional posterior distributions,

$$\begin{aligned} g_{ik} &\sim Dirichlet(\sum_{j} \omega_{ij1} + 1, .. \sum_{j} \omega_{ijK} + 1) \\ \gamma_{kj} &\sim Beta(\sum_{i} (\omega_{ijk} y_{ij}) + 1, \sum_{i} (\omega_{ijk} - \omega_{ijk} y_{ij}) + 1) \end{aligned} \tag{5.16}$$

where

$$\omega_{ij} \sim multinomial(\kappa_{ij1}, \ldots, \kappa_{ijK}); \qquad \kappa_{ijk} = \frac{\left\{ g_{ik}^{y_{ij}} \gamma_{kj}^{y_{ij}} (1 - \gamma_{kj})^{(1-y_{ij})} \right\}}{\prod_{j=1}^{J} \left\{ \sum_{l} g_{il}^{y_{ij}} \gamma_{lj}^{y_{ij}} (1 - \gamma_{lj})^{(1-y_{ij})} \right\}}$$

In light of the computational burden imposed by the large number of parameters in the GoM model, and in order to maintain comparability of the two approaches, the number of phenotype clusters was restricted to

two. The results of the pilot analysis showed that under this regime both models tended to identify clusters with extreme characteristics, that is a cluster of individuals with all symptoms and a cluster of individuals with limited to no symptoms. The cluster of individuals with all symptoms is then described as the "affected" cluster.

Depending on individual research, the quantity of interest can be either a binary variable indicating the status of a patient, i.e. affected/not affected, or a continuous variable representing the probability of an individual having the disorder, considering all symptoms. From our past experience with migraine data, the choice of representation has no effect on the outcome of the linkage analysis, so here we choose the latter measure as the quantity of interest. These are the δ_{ik} and g_{ik}, where k is the affected cluster, of the LCA and GoM models respectively. Thus, the aim is to average these values across models.

The Laplace-Gibbs approximation to the marginal likelihood and the DIC were used as model weights for Method 1 and the BIC and posterior probability were used for Model 2. Given that the aim of the examples is to demonstrate the implementation of the proposed models, and given no information to support an alternative decision, we gave equal prior probability to each model and each weighting measure. The Laplace-Gibbs method is similar to the Laplace-Metropolis approach described in Section 5.3, but estimates are derived from Gibbs rather than Metropolis-Hastings samples. The Hessian matrices required for both models are analytically derived (Appendix A.2.3); since these are almost singular, the Moore-Penrose pseudo-inverse was applied to both matrices [97]. Since both models have the form of a mixture, the DIC3 algorithm suggested by [43] was employed for estimation of the DIC.

These model evaluation approaches differ in their assumptions and approximations, their sensitivity to sample size and number of parameters, and their treatment of model complexity. For example, the DIC and BIC impose (different) penalties for increased model complexity; whereas the DIC uses the effective number of parameters, the BIC uses the observed number of parameters. In contrast, the marginal likelihood and posterior probability approaches make no such adjustment, but the marginal likelihood can exhibit much more extreme values for models when the sample size or number of parameters are large.

Given the familial pedigree and microsatellite data in the case study, QTL linkage was used to identify

important markers [112]. This identifies the linkage between the markers and disease loci by regressing the squared trait differences of sib-pairs on identity-by-descent allele-sharing. A sib-pair that shares more alleles is expected to show a similar phenotype, that is, a smaller difference in trait value. The linkage analysis was carried out using MERLIN-qtl [1].

The algorithms were implemented using the C++ programming language. Three MCMC chains were generated for each method with 20,000 iterations. The first 10,000 iterations were treated as burn-in samples and were removed from analysis.

5.4.3 Results

Simulated data- KPD

Considering that the KPD data was simulated with epistasis effects, and given that QTL linkage analysis aims to identify the dominance rather than epistasis effects, it is important to firstly evaluate the capability of MERLIN to identify the actual loci. Figure 5.2 shows the LOD scores of actual phenotypes for each of the microsatellite markers across ten chromosomes. The dotted, dashed and solid lines represent the LOD scores of Phenotypes 1, 2 and 3, respectively. Except for P3, MERLIN is able to clearly reveal the disease loci of P1 and P2 with strong LOD scores. For P3, MERLIN is able to identify three of the four major loci. When Phenotypes 1, 2 and 3 are pooled to form an affected class, MERLIN-qtl is able to clearly identify the four actual major loci linking to KPD, as shown by the LOD scores in Figure 5.3. Therefore, this result is used for evaluating the effectiveness of the proposed methods.

Figure 5.4 shows the ability of LCA and GoM to identify true phenotype classes. Based on two clusters, both models show promise in identifying the affected-like cluster using the important symptoms: the prevalence of KPD-related symptoms (symptoms *b* to *h*) is much higher in the affected cluster than the unaffected clusters. Moreover, both models are also able to identify the non-KPD related symptoms (those with minimal difference between two clusters). Although there is a moderate difference between the clusters for symptom *k* in LCA and GoM, this is mainly due to data simulation inducing this difference in the dataset (plot c of

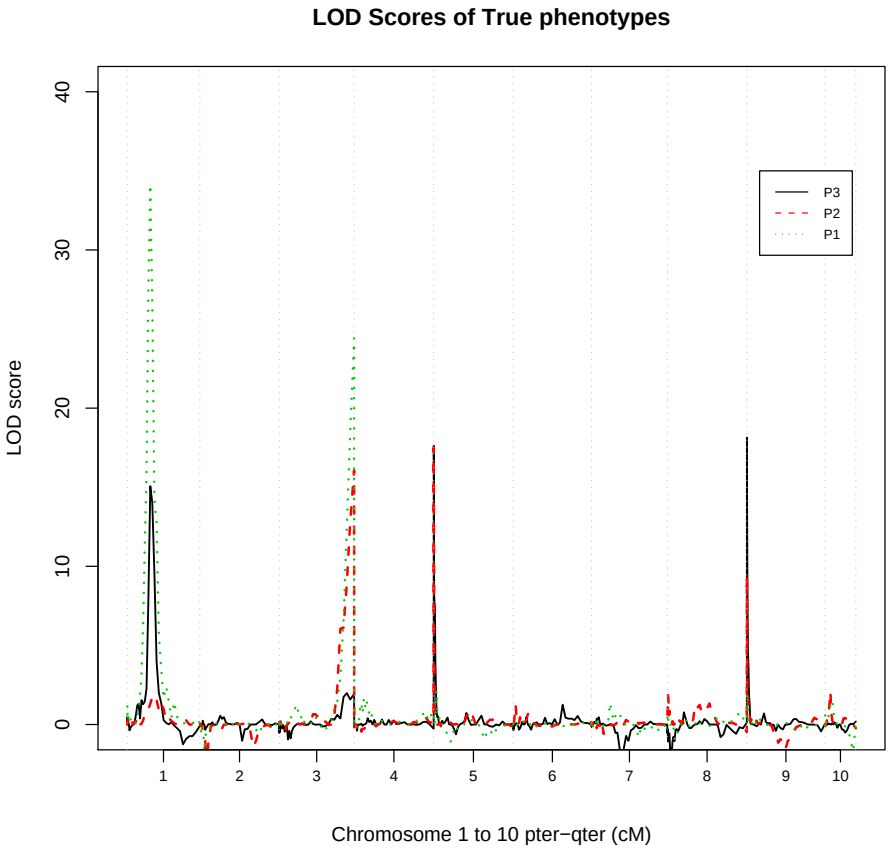

Figure 5.2: LOD scores of the actual phenotypes for each of the microsatellite markers across ten chromosomes. P1, P2 and P3 indicate the actual Phenotype 1, 2 and 3 described in Section 5.4.1. The dotted line is the LOD score of actual Phenotype 1 estimated using MERLIN-qtl; the dashed-line is the LOD score of the actual Phenotype 2 and the solid line is the LOD score of the Phenotype 3.

Figure 5.4). Although the clusters identified by GoM are more homogeneous compared with LCA when $K = 2$, the characteristics of LCA clusters actually reflect those of the true clusters.

Table 5.1: Estimated weights for each of the models using different approximations or different model selection criterion. Depending on the criterion, very different weights are given to each model.

Method	Weight	LCA (%)	GoM (%)
Method 1	Laplace-Gibbs	$\approxeq 100$	$\approxeq 0$
	DIC	$\approxeq 0$	$\approxeq 100$
Method 2	BIC	$\approxeq 100$	$\approxeq 0$
	Posterior Probability	$\approxeq 47$	$\approxeq 53$

As forecast in Section 5.4.2, the choice of model evaluation method results in very different weights for each of the models. This is clearly exhibited in Table 5.1. According to Laplace-Gibbs and BIC, LCA completely

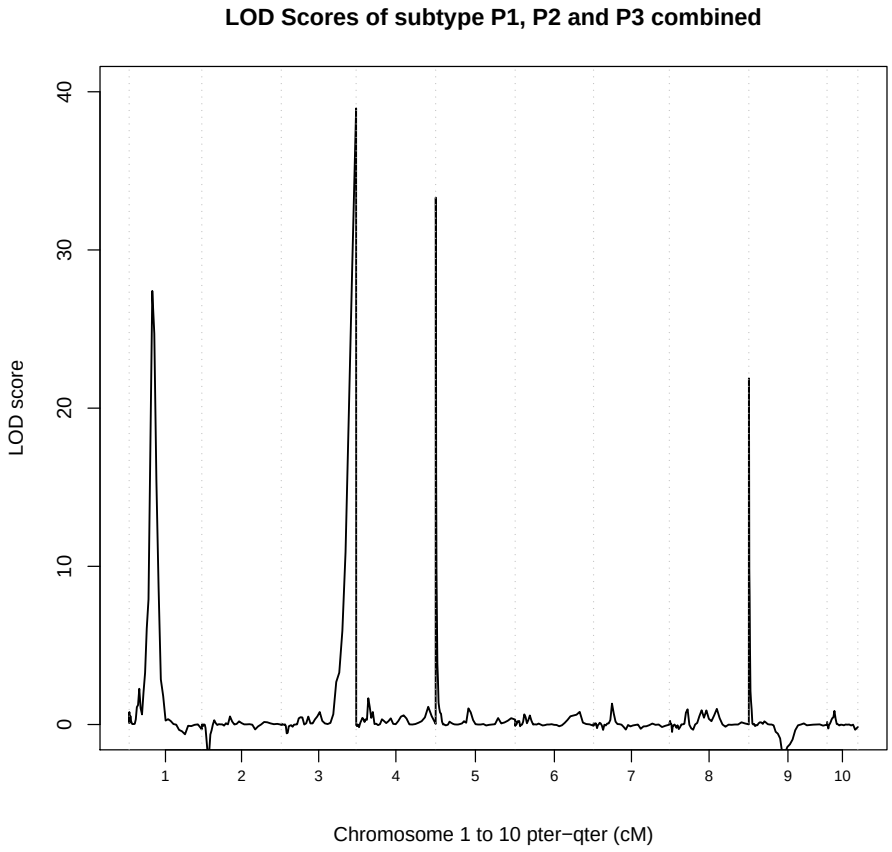

Figure 5.3: LOD scores of pooled phenotype. Four major loci are clearly identified by MERLIN; hence this is used as a benchmark for comparing the results of proposed methods.

outperforms GoM. Conversely, when models are weighted using DIC, GoM is much more preferable than its counterpart. The use of posterior probability on the other hand, gives nearly equal weight to the two models.

Under Method 1, the kernel density of the phenotype average across models using DIC and Laplace-Gibbs weights has both the features of the kernel density of LCA and GoM predictions (Figure 5.5, the solid line). As indicated in this figure, the density of the LCA prediction peaks at 0 and 1 with small variances at each peak. This reflects the more diffuse density of predictions under GoM compared with those under LCA. Moreover, the peaks of the average phenotype are shifted to the right, which is resulted from the discordance in the locations of spikes of different models.

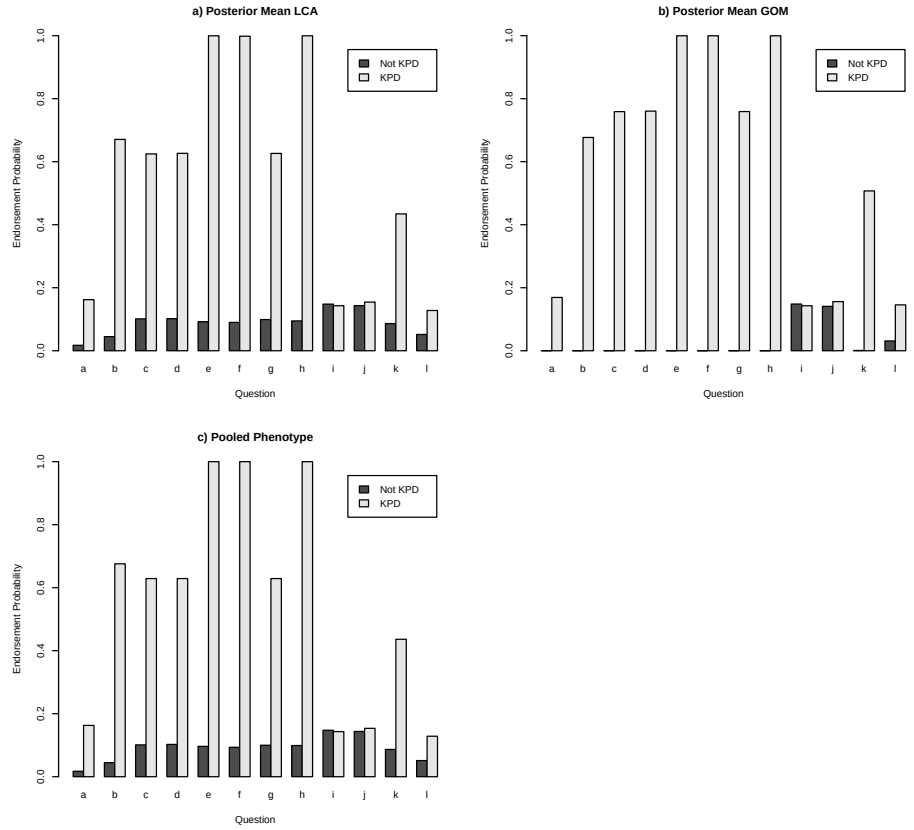

Figure 5.4: The characteristics of clusters derived from different statistical models. Figures *a* and *b* show the prevalence of symptoms in the clusters estimated by LCA and GoM; and Figure *c* shows the symptoms prevalence of true pooled phenotypes. The grey bars are the characteristics of the "affected" cluster and the black bars are the characteristics of the "unaffected" cluster.

When the input phenotype is the prediction averaged across models, the pattern of the LOD score across the chromosome is similar to those obtained with the "pooled" phenotype (Figure 5.6 vs Figure 5.3). Comparing these linkage results to those of LCA and GoM alone, the patterns of the LOD scores are also fairly consistent with the peaks located on chr 1, 3, 4 and 9. The only discordance is in the magnitude of the LOD score of peaks of chromosome 1 and 4.

Under Method 2, the phenotype of an individual is not a point estimate, but a distribution. Because it is impossible to show the densities of all individuals, we present here the results for individuals with (i) all symptoms, ii) True Phenotype 1, iii) True Phenotype 2, iv) True Phenotype 3, v) with 50% of KPD related and non-KPD symptoms, vi) 1 KPD and 1 non-KPD symptom, vii) non-KPD symptom and viii) with no symptoms (Figure 5.7 and 5.8). As indicated in these figures, at the individual level, the predicted pheno-

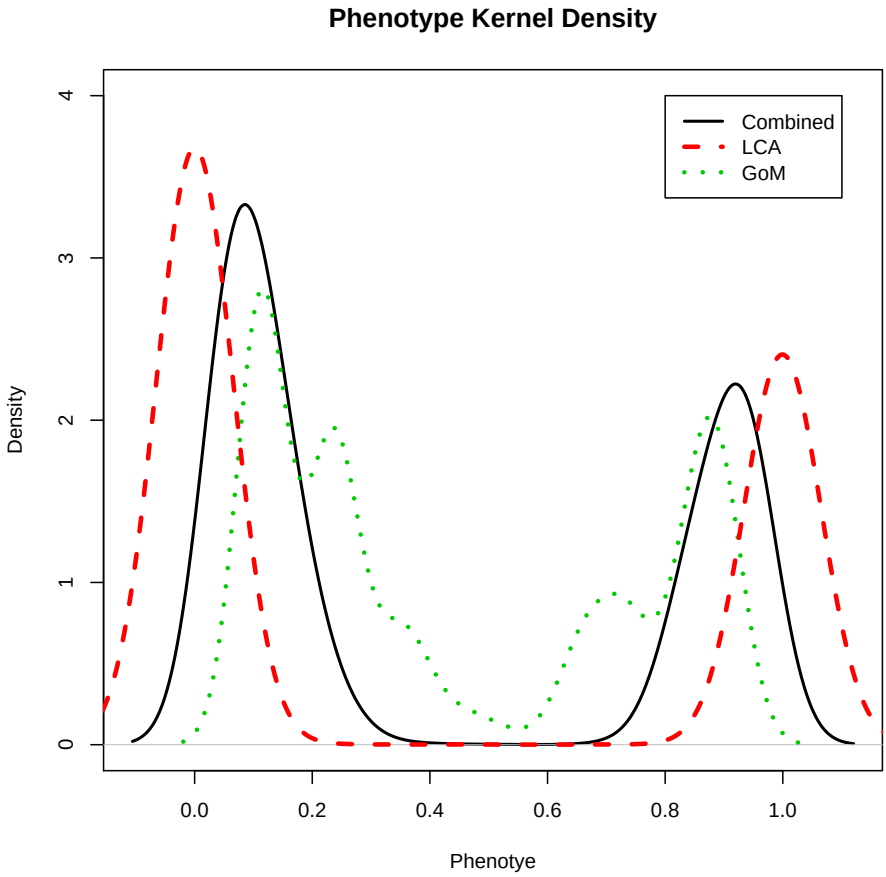

Figure 5.5: Kernel density of the estimated phenotypes. The black solid line represents the averaged phenotype weighted according to Laplace-Gibbs and DIC; dashed and dotted lines are the posterior mean of the phenotype predicted by LCA and GoM.

types under LCA are concentrated at 0 or 1 with very little variance, even when an individual has a half of KPD and a half of non-KPD related symptoms (second figure of row 1 of Figure 5.8). Conversely, prediction under GoM is more diffuse at the individual level. Therefore, averaging the predicted phenotypes across models reflects the same features, with the mode at 0 or 1 and increased variance associated with the modes.

Figure 5.9 shows the distribution of LOD score derived from Method 2 at the four major loci identified in Figure 5.6. The LOD scores at these four loci on chromosomes 1, 3, 5 and 9 are all normally distributed and the mean and credible intervals for each locus are 22.95 (CI:21.44-24.54), 44.26 (CI:42.23-46.39), 21.79

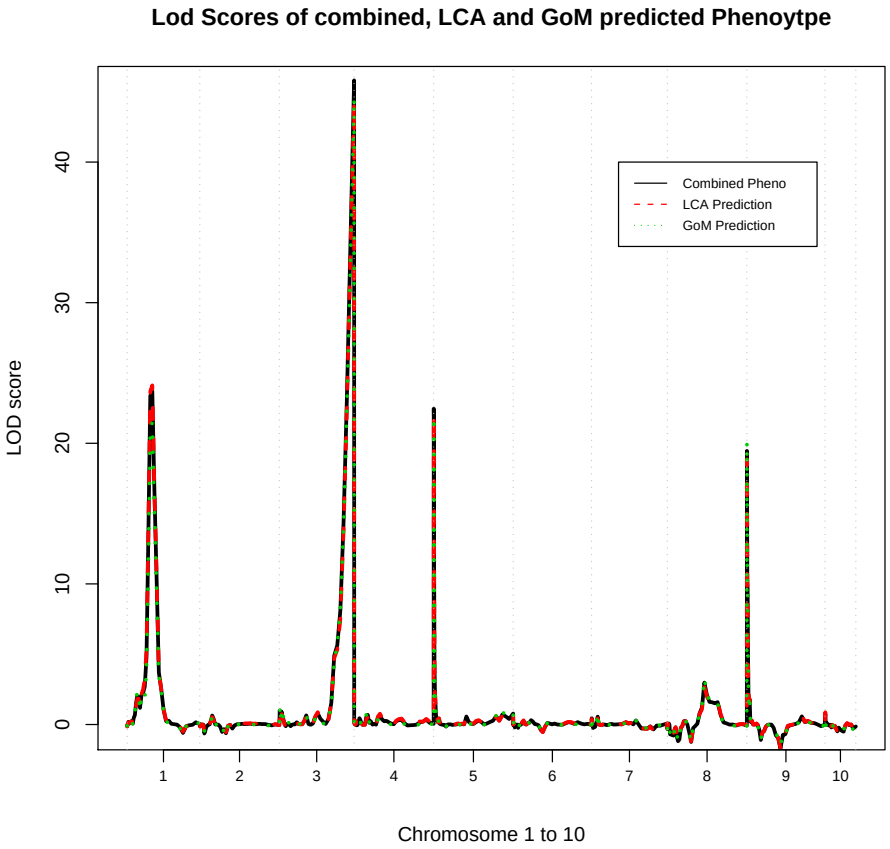

Figure 5.6: LOD scores of each satellite marker for different phenotypes. The solid line shows the LOD scores when the predictions are averaged among models; the dashed and dotted lines show the LOD score of the phenotype predicted by LCA and GoM. The LOD score pattern of the averaged phenotype is similar to the LOD score of the pooled phenotype in Figure 5.3.

(CI:20.27-23.38) and 18.95 (CI:17.53-20.39). Thus, the results show clearly strong linkage at these four loci.

Real Data-Migraine

Table 5.2 lists the weights of each model for the migraine data set. When weighting the models using the Laplace-Gibbs methods and DIC, the prediction of GoM is much better than that of LCA, with nearly 100% of weighting placed on the former model. However, when the weighting is based on BIC, all weight is placed on the prediction of LCA. The use of posterior probability, on the other hand, gives equal weight to

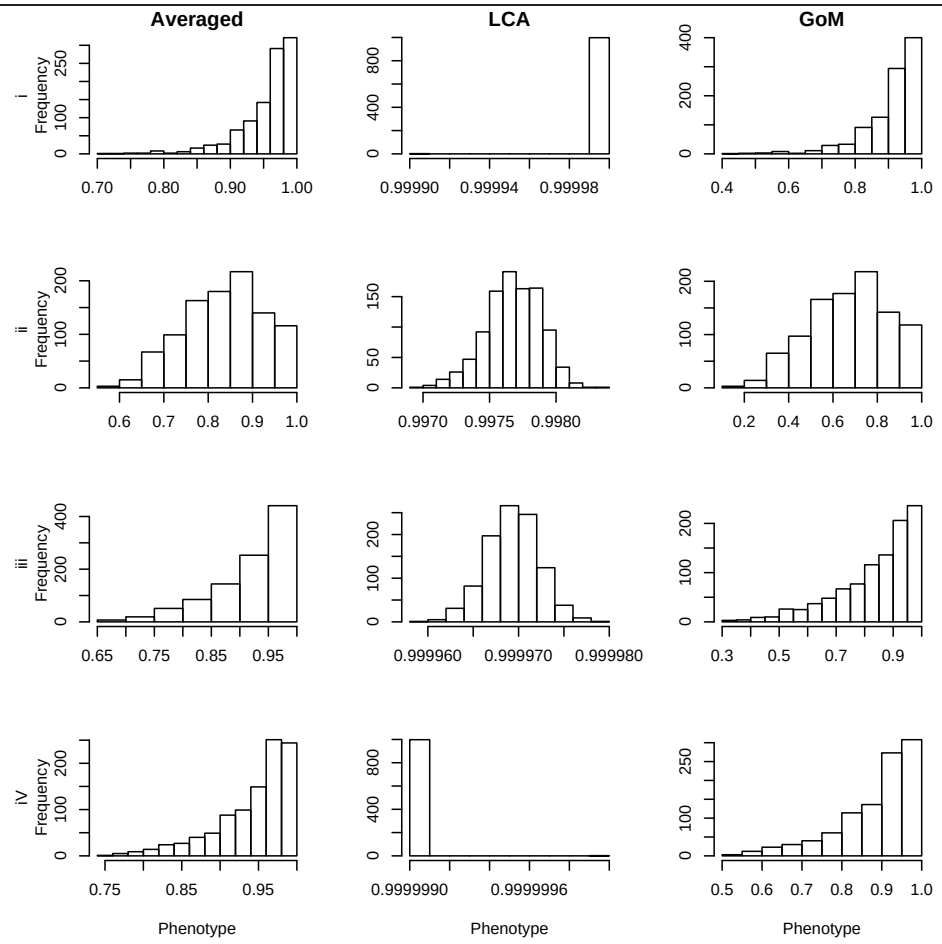

Figure 5.7: Histograms showing the phenotype distribution of cases 1 to 4, which are for individuals with i) all symptoms, ii) True Phenotype 1, iii) True Phenotype 2, iv) True Phenotype 3. The first column contains the histograms of the averaged predicted phenotype; the second and the third columns contain histograms of phenotypes predicted by LCA and GoM, respectively.

the predictions of both models.

As shown in Figure 5.10, under Method 1 the kernel density of the averaged phenotype clearly reflects a merger of the features of both LCA and GoM phenotypes. Figure 5.11 depicts the results of MERLIN-qtl genomewide linkage analysis using the phenotype of Method 1, LCA and GoM. Although these are not large in absolute magnitude (less than 3), the LOD scores of all phenotypes have peaks at chromosomes 1, 2, 7, 8 and 10. Apart from these loci, the results of LCA and GoM are quite different. The LOD score based on the LCA phenotype shows a potential linkage on chromosome 3, but the LOD score based on the GoM phenotype at the same location is below 1. Conversely, the LOD scores of the GoM phenotype

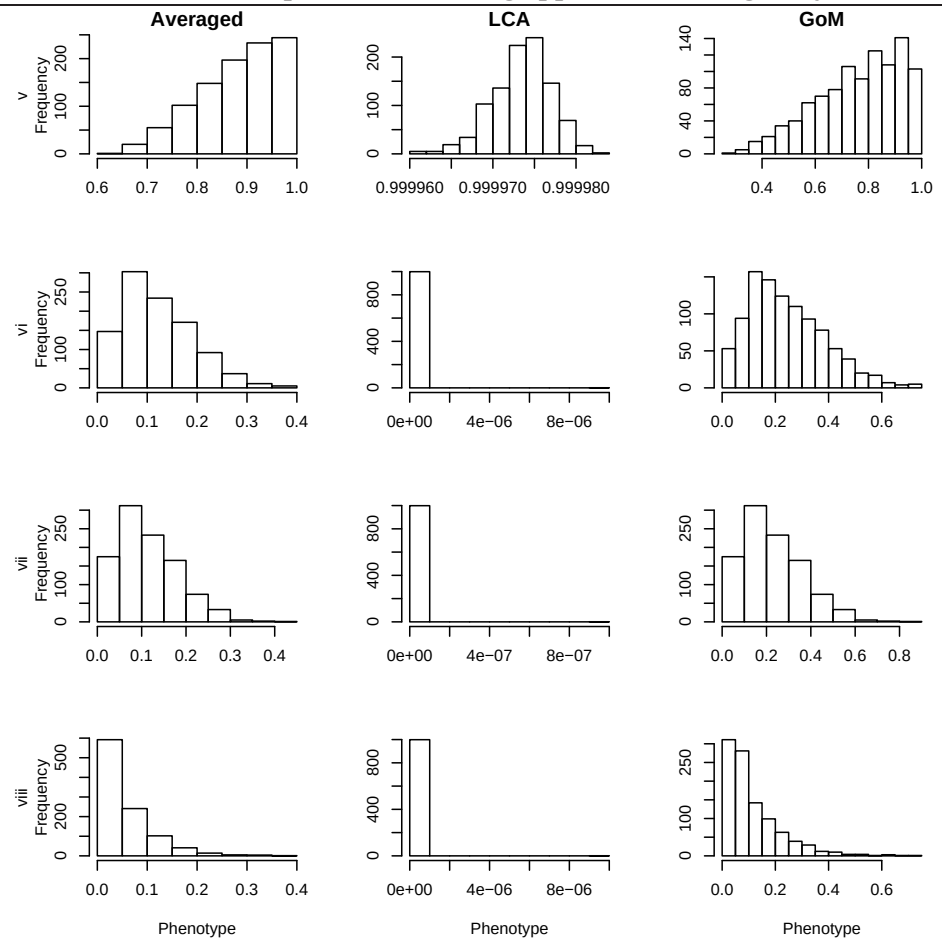

Figure 5.8: Histograms showing the phenotype distribution of cases 5 to 8, which are individuals with v) 50% of KPD symptoms, vi) 1 KPD and 1 non-KPD related symptom, vii) non-KPD related symptoms only and viii) No symptoms. The first column shows the density for averaged phenotype; the second and the third columns are the histograms of phenotypes of the predictions of LCA and GoM, respectively.

show potential linkage at chromosome 5, but this is not supported by LCA. Generally, the LOD score of the averaged phenotype is more closely allied with the LOD score of LCA than GoM. It is also interesting to note that the LOD score of the averaged phenotype is much higher than LCA or GoM alone on chromosomes 3, 7 and 8.

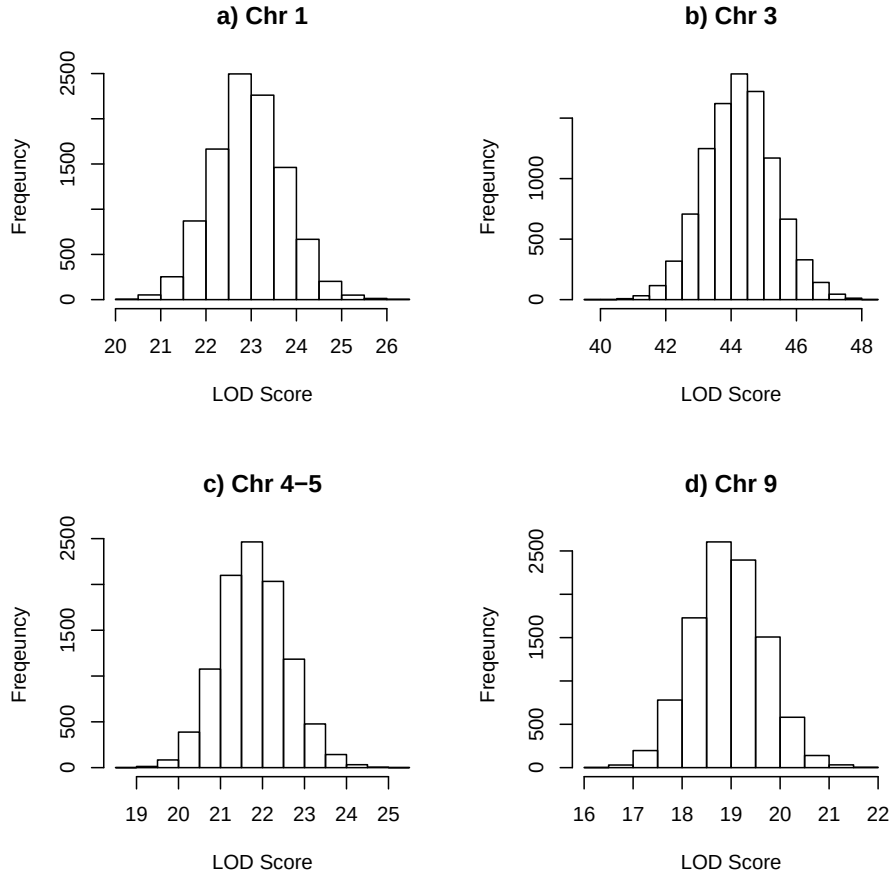

Figure 5.9: Histograms of the LOD scores for the four major peaks of Figure 5.6, located on chromosomes 1, 3, 5 (on the border) and 9.

Table 5.2: The estimated weights for each of the models using different model selection criteria for the migraine data set.

Method	Weight	LCA (%)	GoM (%)
Method 1	Laplace-Gibbs	≅ 0	≅ 100
	DIC	≅ 0	≅ 100
	BIC	≅ 100	≅ 0
Method 2	BIC	≅ 100	≅ 0
	Posterior Probability	≅ 44	≅ 56

Given the sample size is over 13,000, it is impossible to show the histograms of phenotypes derived from Method 2 for all individuals. Therefore, we selected the phenotype distribution of individuals with i) all symptoms, ii) 50% of symptoms, including unilateral, nausea and aura iii) only unilateral, nausea and aura, iv) only having more than 5 headache episodes, each headache lasted more than 4 hours and describe the

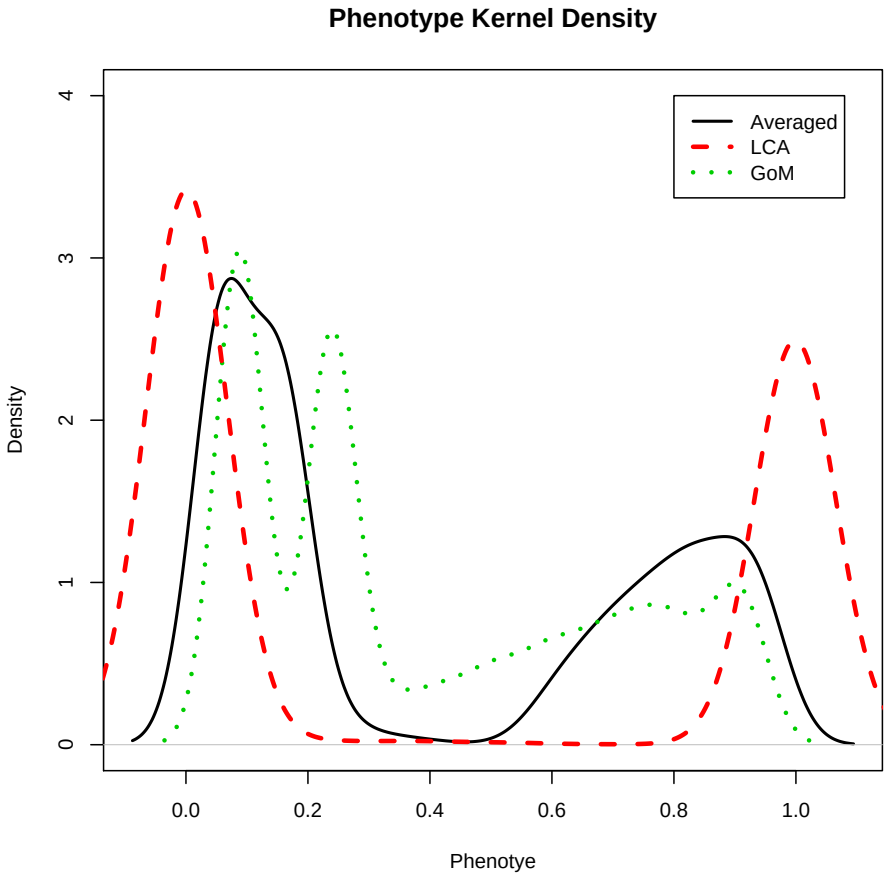

Figure 5.10: Kernel density of the estimated phenotypes of the migraine data using Method 1. The solid line is the phenotype derived from Method 1; the dashed line is the phenotype predicted only by LCA and the dotted line represents the kernel density of predicted phenotype under the GoM model.

headache as severe v) only having more than 5 headache episodes and each headache lasted more than 4 hours and vi) no symptoms (Figure 5.12). This figure reflects very similar findings as for Method 1. The phenotype estimated by LCA is more concentrated than that obtained under GoM. Except for individuals with two symptoms, which have had more than 5 headache episodes and each headache lasted more than 4 hours, the prediction of LCA is often 0 or 1. In contrast, the phenotype of GoM is more diffuse with some uncertainty in the mode. Hence, under Method 2, the distribution of the averaged phenotype has a mode near those of LCA and also incorporates the uncertainty of the GoM results.

Under Method 2, the result of the linkage analysis is no longer a point estimator, but a distribution of the LOD scores at markers accounting for the variance of the phenotype, Figure 5.13 shows the distribution of

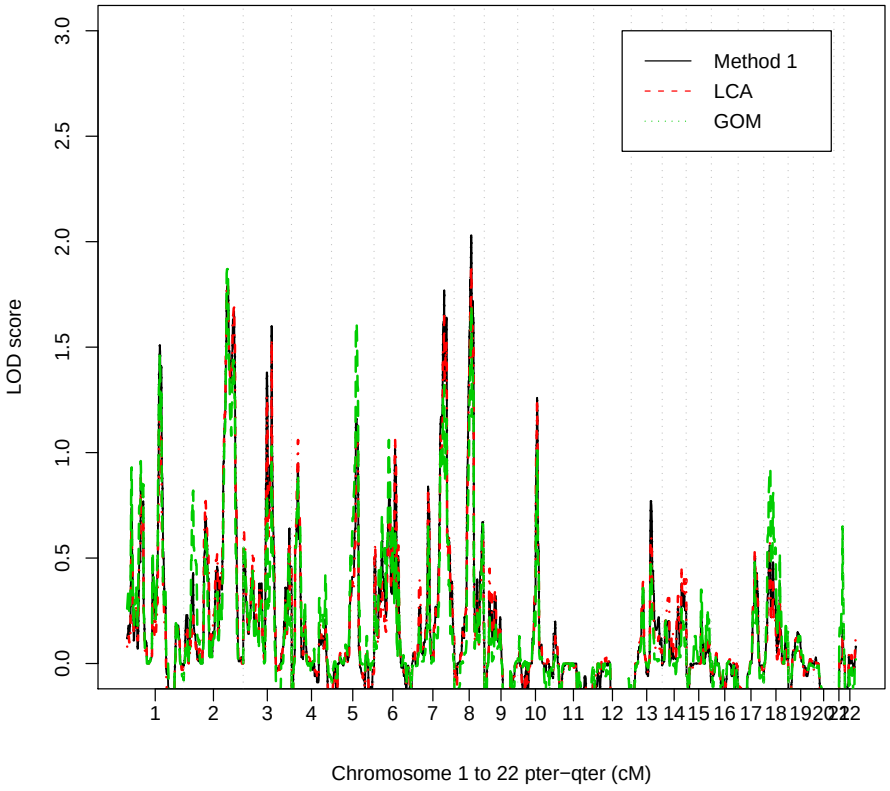

Figure 5.11: Results of MERLIN-qtl genomewide linkage analysis using the phenotype from Method 1, LCA and GoM. The solid line is the LOD score of the phenotype derived from Method 1; the dashed line is the LOD score of the LCA phenotype and the dotted line is the LOD score of the GoM phenotype. The dotted vertical lines show the boundary of each chromosome.

the LOD scores at six major peaks (LOD $\geq$ 1.5) of Figure 5.11. Although the LOD scores of these loci are not large (LOD $\leq$ 3), they are still suggestive compared with the rest of the scores. The loci with the highest LOD score of 2.04 under Method 1 is at 86.314cM of chromosome 8. Under Method 2, this is in the upper end of the distribution (plot f of Figure 5.13); the mode of this loci is around 1.8. The other interesting locus is at chromosome 5 position 122.698cM. The results of Method 1 and LCA show little evidence of linkage at this locus (LOD$\approx$ 1), but the results of GoM show some suggestive linkage at the same locus. The LOD score of 1.6 is well above the credible interval of this locus (plot d of Figure 5.13), therefore, the results of Method 2 do not support potential linkage at this locus.

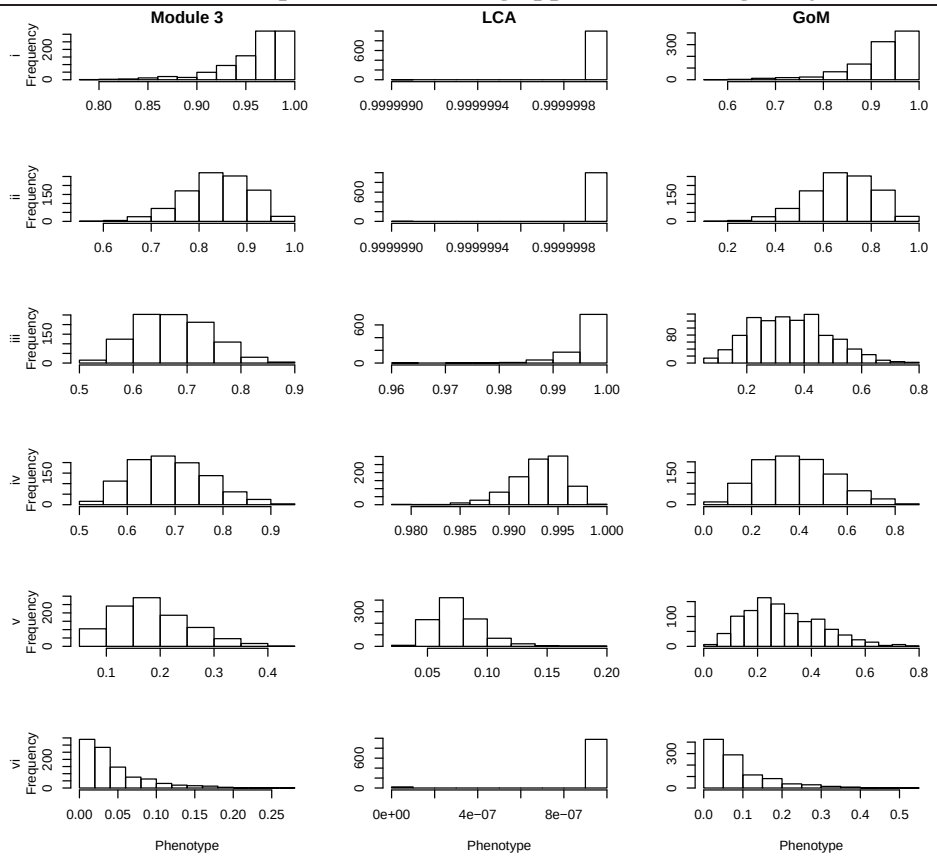

Figure 5.12: Phenotype distributions for individual with i) all symptoms, ii) 50% of symptoms, including unilateral, nausea and aura iii) only unilateral, nausea and aura, iv) only having more than 5 headache episodes, each headache lasted more than 4 hours and describe the headache as severe v) only having more than 5 headache episodes and each headache lasted more than 4 hours and vi) no symptoms. The first column contains the phenotype derived under Method 2, and the second and third columns are the phenotype distributions under LCA and GoM.

5.5 Discussion

The study of diseases with complex etiology demands a clear, statistically relevant definition of the phenotype prior to genetic analysis. Here we proposed two multi-model approaches and provided algorithms for integrating phenotypes using Bayesian model averaging as a foundation. In the examples, we selected two models which have in common a latent class framework, but are very different in terms of parameter spaces and identification of class membership (probability of belonging to phenotype clusters). Because of the substantial differences in the number of parameters between the two models, care must be taken with the choice of model selection criteria. This is reflected in the weighting of the model predictions observed in both the

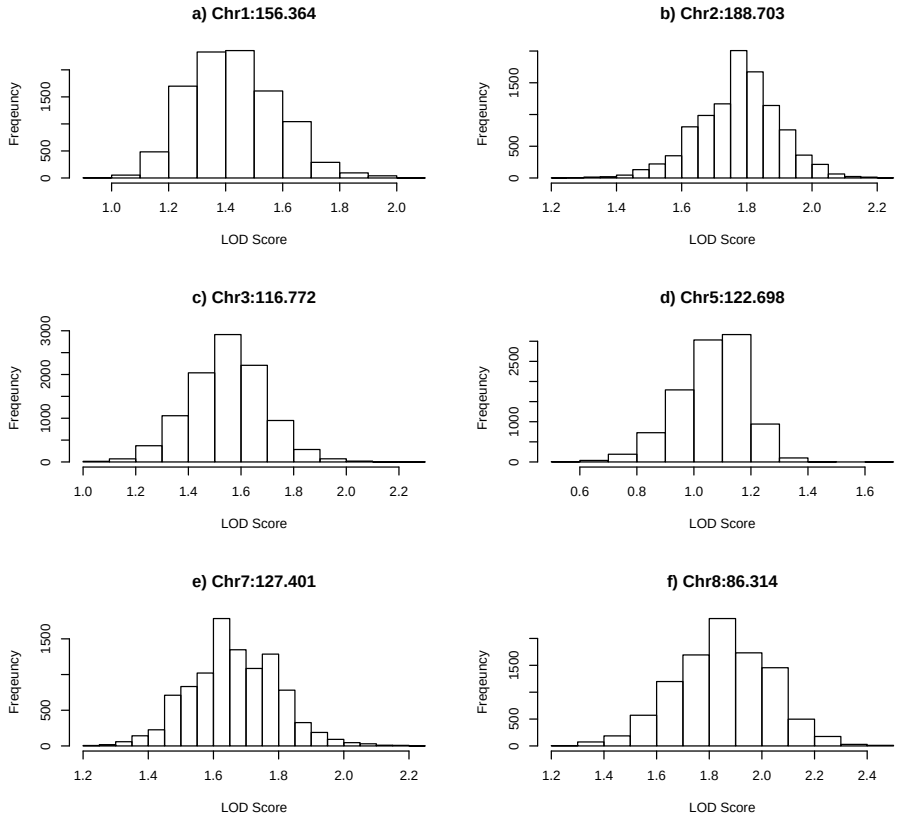

Figure 5.13: Histograms of LOD scores of the six major peaks of Figure 5.11, which are position 156.364 on chr 1, 188.703 on chr 2, 116.772 on chr 3, 122.698 on chr 5, 127.401 on chr 7 and 86.341 on chr 8.

simulated data and the real case study considered here. Although the GoM model had a larger likelihood, the model was penalized heavily by the BIC criterion due to the large number of parameters. Conversely, when using DIC for model selection, GoM strongly outperformed LCA. By proposing averaging over model selection criteria as well as over models, the methods proposed in this chapter may potentially overcome such conundrums, yielding moderate phenotypes that have the qualities of phenotypes derived from different models weighted by the posterior probability of the models.

A further advantage of model averaging is the consolidation of the cores of the clusters commonly identified under the different models and clearer reflection of the model uncertainty by increasing the fuzziness at the boundaries of the clusters. Consequently, individuals tend to be more clearly well allocated if they are in the core of a model-averaged cluster or more clearly poorly allocated if they are at a cluster periphery. Thus, in the subsequent linkage analysis, loci which are strongly differentiated at the cluster cores may have stronger

LOD scores under the combined model than under an individual model. Method 2 has the same advantages as Method 1, with the additional appeal of more completely incorporating parameter uncertainty (as well as model and model choice uncertainty) into the analysis. Consequently, false-positives arising from variation in the input phenotype may be reduced.

Of course, other approaches to combining the results of phenotype and linkage analyses may be considered. An example is running the linkage analyses for each of the separate phenotype models and combining the linkage results. In the case study, where the two models have nearly equal weight, this would result in a simple averaging of the LOD scores at each loci. Under this method, however, the LOD score of each locus will necessarily lie within the range of the LOD scores obtained under the individual models. While this may be appealing in one sense, it can be argued that the combination of methods should allow for increased inferential capability. As demonstrated here, this is possible by model averaging prior to linkage analysis.

In our examples, the maximum number of clusters was fixed to two. This is often an ideal practice because it ignores potential subclusters in the data. In the simulation dataset, the definite number of clusters is four (three subtypes of KPD and an unaffected subtype), and from our previous work and other published literature [47, 45, 211, 165], the optimum number of clusters for the migraine data is also four. However, as the dataset and models increase in complexity, LCA and GoM may not be able to identify "real" clusters. Although the results are not shown here, we analysed the KPD data with $K = 4$ using the LCA model. Three clusters were identified (P1, P3, unaffected) but P2 did not correspond to the remaining cluster. It is also interesting to note that even when the true clusters are identifiable, the linkage analysis may not always identify the important genes for each subtype (Figure 5.2). Generally, if the phenotype is monotonic in nature and if the linkage signal is genuine and strong, although the results may not pin-point the relationship between the loci and the subtypes, the loci involved in the expression of all subtypes are identifiable even when K is set to two. Thereafter, an additional analysis may be required to identify the relationship between genes and subtype. A further challenge of implementing model averaging methods for three or more clusters is the compatibility of clusters found by different models. More research is needed to develop a sound method for K greater than 2.

Further research is also warranted into the impact of different model evaluation strategies when the mod-

els are strongly contrasting with respect to number of parameters. In this chapter, a number of common approaches were considered. Other approaches may be more applicable, and other approximations to the marginal likelihoods [142, 208, 40, 93] may be investigated. The methods proposed in this chapter may be more applicable when the number of parameters in the two models are more comparable, for example, item response theory [168] and GoM or mixture models with different distributions.

There are other open questions about the methods proposed in this chapter, such as the choice of priors. The Bayes factor has been shown to be sensitive to the choice of priors [142]; thus it is important to validate the prior distribution with sensitivity analysis. Moreover, in the examples of this chapter, the subsequent analysis is restricted to genome-wide linkage analysis implemented in MERLIN-qtl. The linkage analysis by [112] assumed that the markers are independent, so lack ability to detect an interaction effect. Although linkage analysis shows great success in mapping the genes for Mendelian disorders, to detect the finer resolution of the putative risk susceptibility loci through linkage analysis is only feasible with the availability of large recombination events from large pedigrees. Therefore, the feasibility of detecting variants with low penetrance using linkage methods is questionable [274]. Furthermore, the methods may also be suitable for genetic association studies.

Part II: Methods for Identifying Epistasis Effects

6

Bayesian Method for Genome-Wide Association Studies: Review and Illustration

Chapter Summary

The following three chapters attempt to address the second main objective of this thesis, which is to develop and review methodologies for identifying the SNPs and/or SNP interactions associated with a disease or phenotype. In these chapters, we develop both model-based and non-model based approaches for the identification of potentially causal SNPs and/or SNP interactions.

In this chapter, the aim is to explore the potential of using a Bayesian logistic model with variable selection (SSVS) to identify associated SNPs or SNP interactions. We develop two models based on logistic regression and used SSVS as the method of dimension reduction. The first model includes only the SNPs joint effect, while the second model includes both SNPs joint and multiplicative effects. We also explore use of slice sampler to sample the posterior conditional distribution for parameter estimation. The approach described in this chapter is able to analyse a larger number of SNPs at once than various previously published methods.

Chapter Conclusion

The model for identifying the SNPs joint effect was tested using chromosome 6 of the diabetes data obtained from the WTCCC. We also tested the including both joint and multiplicative effects with smaller-scale data obtained from GENICA.

Setting aside the drawbacks concerning its computational inefficiency, the Bayesian logistic model with SSVS proposed in this chapter demonstrate the capability of identifying a group of SNPs that contribute to the genetic causes of disease status through joint and multiplicative effects. In the WTCCC data, only less than 25 SNPs are found to be informative and the majority of these SNPs are within the major histocompatibility complex (MHC) region, which has been previously identified for its association with Type I diabetes. The model also identified some novel SNPs with very strong signals of association. Although these SNPs have not previously been found, there is a possibility that the effect of these SNPs can only be highlighted by the presence of the SNPs of the MHC region.

The second model also demonstrated the potential for identifying the SNP interaction effects for a candidate

gene study. The same SNP interaction is also identified by other published studies.

The advantage of the logistic model is that the effect of SNPs genotype or a SNP genotype combinations can be quantified, hence tone can potentially quantify the risk of having the disease in a given genotype and/or genotype combination.

Authorship

Carla C.M. Chen, Kerrie L. Mengersen, Jonathan M. Keith

Discipline of Mathematical Sciences, Queensland University of Technology

Katja Ickstadt

Department of Statistics, Technische Universität Dortmund , Germany

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

Contributor	Statement of contribution
C C.M Chen	conception and conduct the research, write the code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature & Date:	
Mengersen KL	conception, interpretation, editing
Keith JM	conception, execution, coding, interpretation, editing

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

6.1 Abstract

Genome-wide association studies are rapidly becoming the leading technique for understanding the genetic architecture of complex diseases. One of the challenges faced by these studies is the identification of disease-related loci in data containing a large number of SNP genotypes for a relatively small number of individuals. The most predominant approach is to fit SNP-by-SNP logistic regression models, but this approach lacks the ability to detect epistatic effects, which are often present in complex genetic diseases. A potentially superior approach is to simultaneously estimate the main and interaction effects associated with a disease, for all markers. Within this paradigm, the problem becomes one of regression with variable selection, which is well handled using established Bayesian techniques. In this chapter, we apply such an approach to this problem. We demonstrate the main effect model with WTCCC data for Type 1 Diabetes and the two-way gene interaction model using a dataset on sporadic breast cancer data (GENICA).

6.2 Introduction

Genome-wide association studies (GWAs) aim to identify, from among a large number of marker loci drawn from across the genome, those markers that are in linkage disequilibrium with a locus associated with some disease or phenotype. Due to increasing knowledge of common variations in the human genome, advancements in genotyping technologies and in particular the reduction in the cost of gene chips, GWAs have become more prevalent. The current challenge faced in GWAs is to find an adequate and efficient statistical method for analyzing large Single Nucleotide Polymorphism (SNP) datasets.

The first and still most common approach for the analysis of GWAs data is to test markers individually using 2-by-2 contingency tables with χ^2-statistics or simple regression [248], for dichotomous and continuous traits respectively, then adjust for multiple hypothesis testing using Bonferroni correction or False Discovery Rate (FDR) (Crohn's disease, [69]; Type 2 diabetes, [255]). Recent work by [273] also adopted the SNP-by-SNP framework, but instead fitted a logistic model for each marker and applied Bayes Factors as a variable selection tool. [279] adopted a similar model, but implemented forward variable selection for the Type

I Diabetes data set. Furthermore, [276] fitted a classical linear regression model for individual markers to find the associations between SNPs and obesity. Questions have been raised about these single maker allelic tests. One difficulty with the analysis of individual SNPs is the problem of multiple testing. Although adjustments such as those described above may be adopted to address this problem, it has been argued [273] that multiple testing is not strictly relevant in this context anyway [273], it is nevertheless difficult to find an optimal balance between the probabilities of type I and type II error. Another criticism of SNP-by-SNP tests is their inability to detect epistatic effects, that is, gene-gene interaction effects, which can be manifested in multiple ways [50].

Consequently, several groups have proposed approaches for finding epistatic effects in whole genome data. [124] proposed using a two-stage analysis scheme. This involves selecting a subset of SNPs from the whole genome, then modeling the interactions among these markers and between marker and trait. [98] employed the same configuration to analyze markers associated with rheumatoid arthritis. Computationally, these two-stage approaches are relatively efficient, but they can easily miss epistatic interactions between genes which have no main effect. In other words, for the genes to be tested for the interaction, they have to be selected at the first stage of analysis [56].

Recent developments in GWA analysis methodologies have focused on non-parametric approaches, such as the combinatorial partitioning method (CPM, Triglyceride levels [207]), multifactorial dimension reduction (MDR, sporadic breast cancer [231]; type 2 diabetes mellitus [48]; multiple sclerosis [31]) and random forests (RF, HDL and triglycerides glucose [36] and asthma [35]). These methods have proved to be relatively efficient at finding the genes associated with the trait from the whole genome.

An alternative potentially superior approach is to regress the trait or disease status against all SNPs simultaneously with an effective variable selection algorithm. Excellent methods for the variable selection problem have been developed within a Bayesian context. The issue of multiple comparisons is also handled simply and effectively in a Bayesian context [25]. [195] introduced model selection via an assignment of prior probabilities to the various models, and subsequent updating of those probabilities in accordance with Bayes rule.

The model proposed in this chapter is able to detect both the additive and the multiplicative effects. The variable selection adopted in our model is more allied with [95], which introduced the use of a latent variable for the identification of promising subsets of variables. [96] give a detailed overview and comparison of different approaches of Bayesian variable selection.

The use of a Bayesian regression for identifying important loci is not novel. The recent paper by [122] identified subsets of important SNPs using Bayesian inspired penalized maximum likelihood. They assigned a sharp prior mode at zero to the regression coefficients and SNPs with non-zero coefficients estimates were said to have some signal of association. Apart from this, most studies to date is focused on the analysis of QTL data [292, 293, 294, 300, 297, 299]. The model by [292] was initially developed for detecting single locus effects simultaneously, and is later developed for detecting epistasis effects [293]. In [293], the empirical Bayes approach is implemented to estimate genetic epistasis effects without using variable selection, and the relative importance of effects is based on the ratio of variances.

In contrast, our model is more closed allied to [299], but our method differs in a number of aspects. Firstly, [299] partition the genome into fixed number of loci and assume that the QTL occurs at one of these sites. This partitioning is required to be specified prior to the analysis. In contrast, the SNPs data can be directly utilized in our model and the number of potential causal loci is estimated directly from the dataset without boundaries. Second, because the model by [299] is for a QTL study, a design matrix can be employed; although this is an ideal approach, it is not feasible for population studies.

In this chapter, we introduce two Bayesian models, the first for continuous traits and the second for dichotomous traits. These models are initially described in the context of main effects only and then further extended for the detection of gene-gene interaction effects.

6.3 Methods

6.3.1 Main effect models

Continuous trait model Let y_i be the observed value or realization of the dependent variable (continuous trait) for individual $i, i = 1, \ldots, n$. We model y_i as in Equation 6.1 below: dependent on a constant term μ_i, on n_c continuous-valued covariates, n_d discrete-valued covariates, and up to n_s SNPs. Let the jth continuous-valued covariate for individual i be x_{ji}. For each continuous-valued covariate, we introduce a regression parameter β_j. For the jth discrete-valued covariate, let L_j be the number of levels and let h_{jki} be 1 if the covariate has level k for individual i and 0 otherwise, for $k = 1, \ldots, L_j$. For each discrete-valued covariate and each level of that covariate we introduce a regression parameter ω_{jk}. Let z_s be an indicator variable for SNP s, taking the value 1 or 0 depending on whether SNP s is included in the model or not. Let g_{sli} be an indicator variable taking the value 1 or 0 depending on whether individual i has genotype l (where $l = 0, 1, 2$) at SNP s or not. Let ν_{sl} be the contribution to the dependent variable made by genotype l at SNP s. Let ε_i be a random error.

$$y_i = \mu_i + \sum_{j=1}^{n_c} \beta_j x_{ij} + \sum_{j=1}^{n_d} \sum_{k=1}^{L_j} \omega_{jk} h_{jki} + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{2} \nu_{sl} g_{sli} + \varepsilon_i \tag{6.1}$$

Because the SNPs are categorical variable, we arbitrarily assign the value $\nu_{s2} = 0$ for all SNPs s.

Case-Control Model (Logistic Model) For case-control data, y_i is the presence/absence of the phenotypic trait, and takes the value 1 when the phenotype is present, else 0. The model proposed in Equation 6.1 can be simply modified by introducing a logit link function, this $log(\frac{q_i}{1-q_i})$ where q_i is the probability that individual i has the trait of interest. Then Equation 6.1 follows with the same notation for the model parameters.

Prior Distributions As part of the Bayesian approach, a prior distribution is required for each of the model parameters. In our two-case studies, no prior information is available, therefore noninformative priors are

considered here. Moreover, because the indicator variable $Z, Z = (z_1, \ldots, z_{n_S})$ is not directly observed, we adopted a hierarchical approach. Details on the priors used in our case studies are described in the examples.

Parameter Estimation Model parameters are estimated using Markov Chain Monte Carlo. The Gibbs Sampler involves sampling from one-dimensional conditional distributions given other parameters and this is used for the estimation of all variables with one exception which we discuss below. Except for z_s, all other parameters possess non-standard conditional distributions; thus we used the slice sampler [205] to draw from these.

Instead of sampling from the distribution function, the slice sampler samples from the area under the density function. Despite the complexity of using the slice sampler for multivariate distributions, it is relatively simple to implement for updating a single variable. Let x denote a model parameter and x_0 and x_1 be the current and new values of x, respectively. The procedure for updating x involves three steps. First, draw a real value y uniformly from $(0, f(x))$, where $f(x)$ is some function which the density of x is proportional to, and consider the horizonal "slice" $S = \{x : y < f(x)\}$. Next, establish an interval, $I = (L, R)$, around x_0 which contains this slice. A new value is then drawn uniformly from the interval, and becomes x_1 if it is within S, else it is iteratively redrawn.

For simplicity, we used an initial interval of $(-1000, 1000)$ and used the shrinkage procedure [205] for sampling from the interval.

The estimation procedure for z_s is described in the following.

Variable Selection Variable selection is an important element of the new models, which utilize the variable inclusion indicator (z_s) to determine the importance of SNP s. At each MCMC iteration, the value of z_s depends on the ratio of the conditional posterior probabilities of including and excluding SNP s. At the first iteration, start with a randomly generated vector of length n_S, comprising 0's and 1's, denoted $z^0 = (z_1^0, \ldots, z_{n_S}^0)$. Let t denote the MCMC iteration, $t = 1, \ldots, T$, where T is the total number of iterations. Let Θ^t be a vector containing all parameters other than z at iteration t. At each t, SNP s is randomly selected from all SNPs and z_s is updated as follows

1. Estimate the conditional posterior probability with $z_s = z_s^{t-1}$, $P(z_s^{t-1}|\Theta^t, Y, z_{-s})$.

2. Estimate the conditional posterior probability with the complementary value, $P(z_s^{'}|\Theta^t, Y, z_{-s})$, $z_s^{'} = 1 - z_s^{t-1}$.

3. Determine the ratio of the values computed in Step 2 and 1.

4. Accept the proposed $z_s^{'}$ if the value of Step 2 is greater than a value randomly generated from a uniform distribution with minimum 0 and maximum 1; else retain z_s^{t-1}.

After SNP *s* is updated, the procedure is repeated for another SNP drawn randomly from the remaining SNPs. This continues until all SNPs are updated. The probability that SNP *s* is associated with the trait of interest is then estimated as the number of times SNP *s* is included in the model over the total number of iterations after burn-in.

Example 1: Case-Control of Type I diabetes We tested the performance of the proposed model using a Type I diabetes (TID) data set. The data were obtained from the Wellcome Trust Case Control consortium (WTCCC, http://www.wtccc.org.uk). In their study, the WTCCC collected 14000 cases and 3000 shared controls for 7 different familial diseases. Here we focus on Type I diabetes.

Individuals involved in this study are self-identified white Europeans who live in Great Britain. The controls are recruited from two sources: 1500 are from the 1958 British Birth Cohort and the remaining are blood donors recruited for the WTCCC project.

TID cases are recruited from three sources. The first is from approximately 8000 individuals who attend the paediatric and adult diabetes clinics of 150 National Health Service Hospitals across mainland UK. The second source of cases is voluntary members of the British Society for Paediatric Endocrinology and Diabetes. The rest are from the peripatetic nurses employed by the JDRF/WT GRID project (http://www-gene.cimr.cam.ac.uk/todd/).

Diagnosis of the TID cases is based on the participants' age of diagnosis and their insulin dependency. The cases of the TID study are required to be diagnosed with TID at age less than 17 and have been insulin

dependent for more than six months. Individuals with other forms of diabetes, such as maturity onset diabetes of the young, are excluded from the data set.

Both cases and controls were genotyped with the GeneChip 500K Mapping array (Affymetrix Chip) with 500,568 SNPs. After filtration, there was a total of 469,557 SNPs. Details on the WTCCC experimental design, data collection, data filtration and more are in [273].

The previously published results of single locus analysis indicated strong signal association of Chromosome 6 [273], in light of this, we used only the SNPs data on Chromosome 6 for this study.

In addition to the filtration methods and exclusion genotypes recommended by [273], we set CHIAMO calls with a score less than 0.9 to missing and removed all SNPs with one or more missing values to speed up the computation time. This leads to a total of 26,291 SNPs in the TID data.

As only the genotype information is presented in the data set thus obtained, the logistic regression model is simply:

$$log(\frac{q_i}{1-q_i}) = \mu + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{2} \nu_{sl} g_{sli} + \varepsilon_i \tag{6.2}$$

where $i = 1, \ldots, 4857$ and $s = 1, \ldots, 26291$ and we arbitrarily assigned $\nu_{s2} = 0$.

Non-informative priors are used for this model as follows. The prior probability distributions for both overall mean (μ) and the contribution of level l of SNP s are assumed to be normally distributed with mean 0 and precision 1. The prior distribution for the residual, ε, is assumed to be a normal distribution with mean 0 and precision τ, and the prior for τ is assumed to be a gamma distribution, with parameters set to 0.05 ($\alpha = \beta = 0.05$). For z_s, we adopted a hierarchical approach, and let the probability that $z_s = 1$ be p_z, where p_z is a hyperparameter. We assumed the prior probability of z_s follows a Bernoulli distribution.

Five independent MCMC chains were generated with 100,000 iterations each. The first 50,000 iterations of each were considered as burn-in and the remaining were extracted for building the posterior marginal distributions. The algorithm was implemented in C.

6.3.2 Main effects and interactions

The model introduced in Equation 6.1 includes the main effects only. This can be extended for detecting SNP interaction effects as follows. Using the same notation as before, let η_{jk} be the indicator parameter, $\eta_{jk} = 1$ if the interaction of SNPs j and k is included in the model, else 0 and let $\gamma_{jl_jkl_k}$ be the coefficient of the interaction between the genotype l_j of SNP j and the genotype l_k of SNP k ($l_j = 0, 1, 2$; $l_k = 0, 1, 2$ and $j \neq k$). Then the model with two-way interactions is as follows:

$$y_i = \mu + \sum_{j=1}^{n_c} \beta_j x_{ij} + \sum_{j=1}^{n_d} \sum_{k=1}^{L_j} \omega_{jk} h_{jki} + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{2} \nu_{sl} g_{sli} + \sum_{j=1}^{n_s} \sum_{k=1, j\neq k}^{n_s} \eta_{jk} \sum_{l_j=0}^{n_1} \sum_{l_k=0}^{n_2} \gamma_{jl_jkl_k} g_{jl_jkl_ki} + \varepsilon_i \tag{6.3}$$

This model can be extended in an obvious manner to include multi-way interactions. By introducing a logit link function, this model can be implemented for the case-control study.

Typically, when an interaction effect and the two corresponding main effects are included in a model, then the number of levels for the interaction is $(n_1 - 1)(n_2 - 1)$, where n_1 and n_2 are the number of levels for each of the main effects (the maximum number of level is nine). However, here we have chosen to include $n_1 \times n_2 - 1$ levels for the interaction, because one or both of the main effects may not be included in the model (that is $z_1 = 0$ and/or $z_2 = 0$).

Parameters of this model are estimated following the same procedure as described earlier. The combination of Gibbs and Slice samplers was implemented for sampling from the conditional posterior distributions. Likewise, variable η_{jk} was updated following the same procedure as for z_s.

Example 1: GENICA

We illustrate this expanded model using the GENICA data set. Although factors such as smoking history, family history of breast cancer and menopausal status were collected in the GENICA study, these variables were not available at the time of our study.

GENICA is an interdisciplinary study group on Gene ENvironmental Interaction and breast CAncer in Germany, with its main focus on the identification of both genetic and environmental effects on sporadic breast cancer. The data were collected between August 2000 and October 2002 on incident breast cancer cases and population-based controls in the Bonn region in Germany. Among the cases, 688 were first-time diagnoses of primary breast cancer, and were later histologically confirmed. There were 724 controls, matched within 5-year age classes. Samples contain only Caucasian females younger than 80 years old.

Each SNP genotype can take one of three forms: homozygous reference genotype, heterozygous variant genotype and homozygous variant genotype. The homozygous reference genotype is taken to be the genotype which has both alleles being the most frequent variant. The heterozygous variant genotype occurs when one of the base pairs is more frequent while the other base is less frequent, and the homozygous variant genotype is when both members of the pair are less frequent.

Not all genotype data are used in this study. The subset of SNPs which are related to estrogen, DNA repair or control of cell cycle pathway are tested here, with a total of 39 SNPs. From a total of 1234 females, including 609 cases and 625 controls, individuals with more than 3 genotypes missing were excluded from the analysis. The final data therefore included 1199 women and was composed of 592 cases and 607 controls. Other missing genotypes were imputed using the k-nearest neighbor method [246]. Details of data collection and genotyping procedure are in [139].

Let θ denote the parameter space. The parameters for the GENICA data are thus

$$\theta = \{z_s, \nu_{sl,jl_jkl_k}, \tau\}$$

where $s, j, k = 1, ...39$, $l, l_j, l_k = 1, 2, 3$ and z_s and η_{jk} are independent. The priors for model parameters were

similar to the ones used in Example 1, as follows

$$\varepsilon \sim N(0,\tau); \qquad \mu \sim N(0,1); \qquad \tau \sim Ga(0.01,0.01);$$

$$\nu_{sl},\gamma_{jl_jkl_k} \sim N(0,10); \qquad z_s \sim Bern(p_z); \qquad \eta_{jk} \sim Bern(p_\eta)$$

Ten MCMC chains were generated with 300,000 iterations each. Of these, the first 250,000 iterations were considered burn-in, and the remaining 50,000 cases were extracted for the construction of the marginal posterior distribution of θ. The computational algorithm was implemented in C.

6.4 Results

6.4.1 WTCCC-Type I diabetes

The results of the MCMC runs for the WTCCC Type I diabetes data indicated multiple modes in the posterior distribution. No prominent model was identified across all five chains. At each MCMC run, there are at least 13000 unique models were tested, with the most common models occupying only 1.25% to 4.5% of the post burn-in iterations. These models identified 17 to 24 SNPs of the total 26291 SNPs, with some SNPs commonly found among all models (Table 6.1).These include SNPs 1576 (rs10901001), 4073 (rs874448), 4887 (rs950877) and 6222 (rs9272723). Five additional chains were generated using SNPs listed in Table 6.1. The posterior log-likelihood was well-mixed after 150000 iterations and with log-likelihood value between -2012 and -2052.

Although all SNPs on chromosome 6 had the opportunity to enter the model at each of the MCMC iterations in the analysis, more than half (51%) of the SNPs were not selected in any of the 250,000 iterations (50,000 iterations, 5 chains). In contrast, 4% of SNPs (1143 SNPs) were included at least once in the iterations of all five chains. Of these 1143 SNPs, all five chains selected SNP 1576 (rs10901001) and 4073 (rs874448) in nearly all iterations ($\tilde{9}$7%), followed by SNP 4887 (rs950877, 76%), which is also included in the five optimal models.

The results of the MCMC runs also identified a group of SNPs with highly variant probability of inclusion across the chains. For instance, SNP 6051(rs3131631) had high probability of inclusion for chains 1, 3 and 4, but was selected in less than 1% of iterations in chains 2 and 5. This indicated the inclusion of a SNP from this group depends on other SNPs already present in the model during the variable selection procedure. This was also observed for SNPs 6232 (rs9275418) and 6233(rs9275523). SNP 6232 was selected in nearly 100% of iterations for chains 1, 2 and 4, but was not selected for chains 3 and 5; in contrast, SNP 6232 was included nearly in all iterations for chains 3 and 5, but was never included for chains 1, 2 and 4. Since these two SNPs are physically nearby, they may be in linkage disequilibrium.

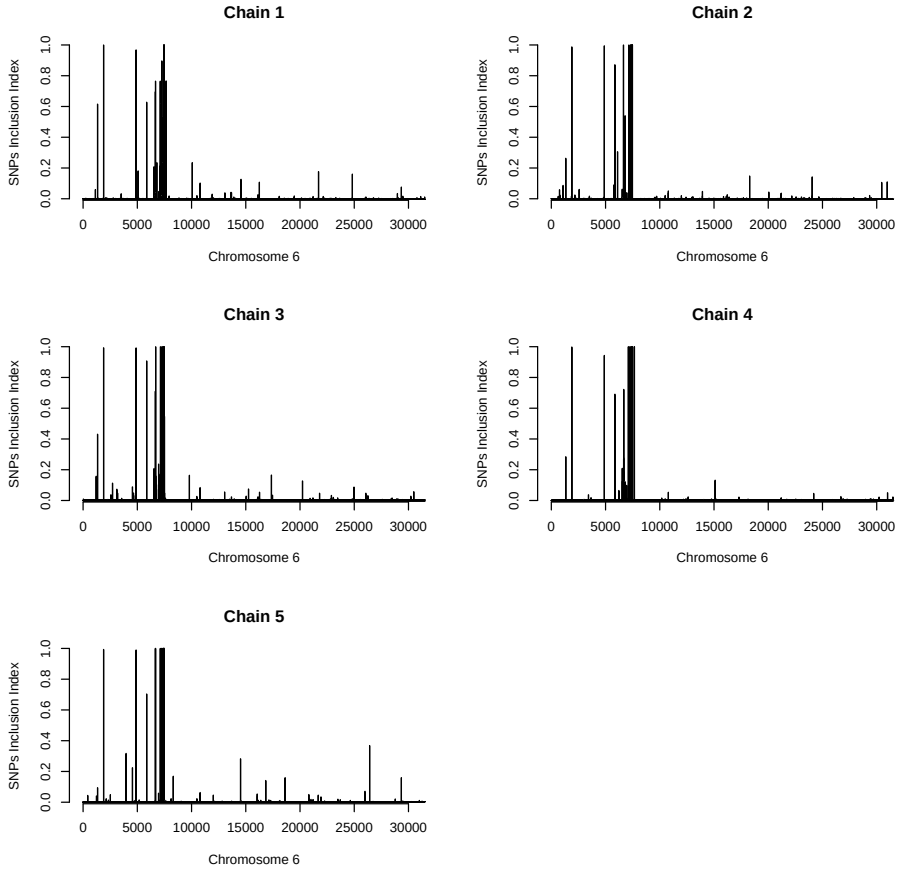

Figure 6.1: The contribution of individual SNPs on chromosome 6 to TID across five chains

Figure 6.1 shows the ranking of SNPs across Chromosome 6 for the five chains. The first two peaks correspond to SNPs 1576 and 4073. This figure also shows a strong association with TID on a region of the shorter arm of Chromosome 6 which is the major histocompatibility complexity (MHC) region (SNP 5802

to SNP 6358).

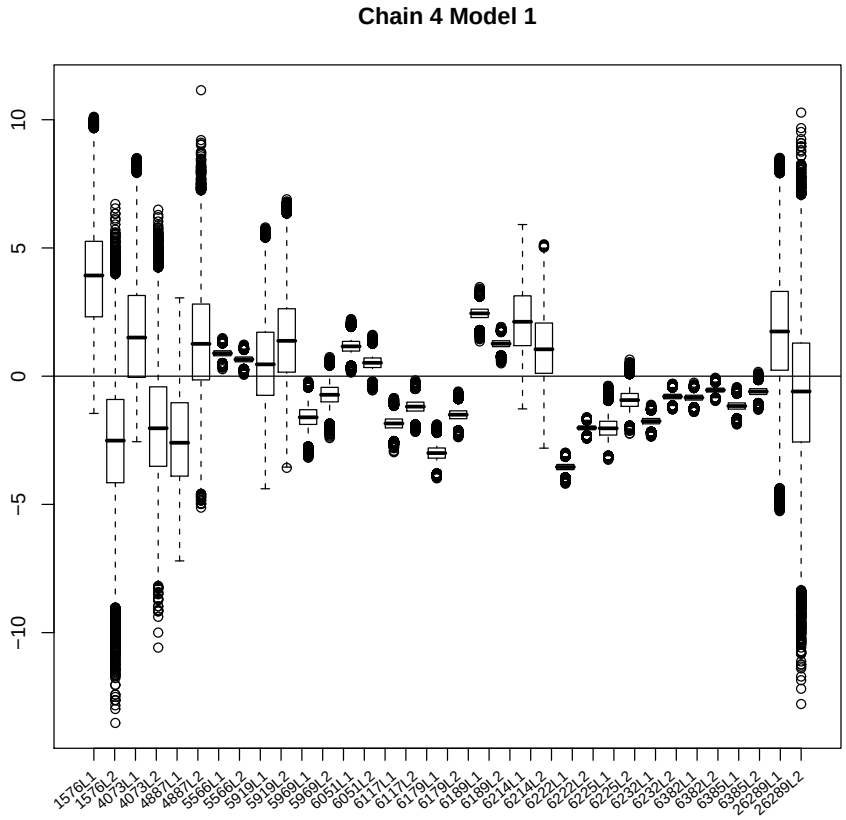

Figure 6.2: The quantified genotype type effect at SNPs selected by model 4. The x-axis shows the SNP ID and its genotype, where L1 is homozygosity reference and L2 is heterozygosity.

Another advantage of the model proposed here is the identification of the effect of genotype on the phenotype. For instance, Figure 6.2 shows the marginal distribution of the contribution to TID made by the genotype at each SNP in the model identified by chain 4 (Table 6.1); here, the homozygous variant genotype is set to zero. The figure shows that individuals with homozygous reference genotype at SNPs 1576 and 4073 are likely to have a higher chance of being TID positive than individuals with homozygous variant or heterozygous genotypes at the same SNPs. In contrast, a higher chance of being TID is observed for individuals with heterozygous variants at SNP 4887 than individuals with homozygous variants (both homozygous reference and homozygous variants). This pattern of genotype contribution is observed among all the models identified by all chains.

Table 6.1: SNPs included in the most common models from each of the five chains

Chain	No SNPs	SNPs ID
1	20	**1576**, **4073**, **4887**, 5587, 5638, 5663, *5919, 5969, 6051, 6110, 6122, 6158, 6195, 6205, 6211, 6217, 6221, 6222, 6232*, 8390
2	18	1112, **1576**, **4073**, **4887**, 5545, 5661, *5957, 6025, 6073, 6087, 6156, 6157, 6180, 6217, 6221, 6222, 6228, 6232*
3	24	1112, **1576**, **4073**, **4887**,5447, 5577, 5587, 5588, *5802, 5947, 6051, 6110, 6156, 6160, 6172, 6174, 6177, 6180, 6217, 6221, 6222, 6233*, 8169, 21883
4	17	**1576**, **4073**, **4887**, 5566, *5919, 5969, 6051, 6117, 6179, 6189, 6214, 6222, 6225, 6232, 6382, 6385*, 26289
5	23	**1576**, 3302, **4073**, **4887**,5553, 5571, *5932, 6025, 6043, 6121, 6149, 6154, 6173, 6180, 6191, 6205, 6219, 6227, 6233*, 12097, 17510, 22015, 24454

* The reference of the SNP id is supplied in Appendix A.3
* SNPs in **Bold** are the common SNPs identified across chains
* SNPs in *Italic* are the SNPs from the major histocompatibility complex region

6.4.2 Genica

As in the previous case study, the results of the MCMC runs for the GENICA breast cancer data indicated that the posterior distribution has multiple modes. Table 6.2 lists the most frequently selected model in each of the ten MCMC chains. These models were selected in each chain for at least 61% of the 50,000 post burn-in iterations. Of the ten chains, six converged to the same model (chains 2, 4, 5, 8, 9 and 10), which contains only two SNPs - SNP 20 and 21 - and both are fitted as main effects. In contrast, the remaining four models indicated the presence of interaction effects.

SNP 20 is the most prominent main effect and is observed in all models (Table 6.2). In contrast, SNP 21 is included in a model as either a main or an interaction effect. When SNP 21 is selected as an interaction effect, it interacts with a different SNP in different models. For instance, SNP 21 interacts with only SNP 23 in model 1. Besides these two SNPs, other possible SNPs and interactions are also identified as indicated in Table 6.2.

The estimated coefficient of SNP 20 is fairly consistent across models and ranges between -1.17 and -1.12 for the homozygous reference variants (level 0) and between -0.57 and -0.52 for the heterozygous genotype variants (level 1). This indicates that individuals with a homozygous variant genotype (level 2) at SNP 20 associated with a higher chance of having breast cancer, followed by individuals having a heterozygous

variant genotype (level 1) and homozygous reference variants (level 0) at SNP20. In two of the models (model 2 and 5), SNP 21 is included as a main effect and the posterior estimates of the coefficient for genotype variants are also consistent for both models. The coefficients indicate that having homozygous reference variants at SNP 21 is associated with a higher probability of breast cancer than the other two genotype variants. However, these two SNPs needed to be considered conjointly to estimate the probability of sporadic breast cancer (Model 2).

Considering SNPs 20 and 21 as additive effects, the highest chance of a sporadic breast cancer occurs when individuals have homozygous genotype variant (level 2) at SNP 20 and homozygous reference genotype variant (level 0) at SNP21, with an odds ratio of 4.17 (CI: 2.63-6.67) compared to individuals with homozygous genotype variants (level 2) at both SNP 20 and 21. The next highest probability occurs for individuals with heterozygous genotype variants (level 1) and homozygous reference variant (level 0) at SNPs 20 and 21 respectively; these individuals have an odds ratio of 2.37 (CI: 1.01-5.58). The lowest chance of sporadic breast cancer is for subjects with homozygous reference variants at SNP 20 and homozygous variants (level 2) at SNP 21.

In other models, where SNP 21 is selected as a part of an interaction effect, the effect of genotype variants at this SNP becomes more complicated. Figure 6.3 shows the posterior mean and credible intervals of the interaction terms of models 1, 3 and 4, which all involve SNP 21. In model 1, SNP 21 contributed to the probability of breast cancer by associating with SNP 23 and the genotype variants of SNP 21 in this combination are quite different from the genotype variants of SNP 21 combinations in models 3 and 4 (SNP21×SNP6, SNP21×SNP14, respectively), but some similarities are found in models 3 and 4.

Table 6.2: Unique models of ten chains

Model	Parameters	Chains	Frequency(%)
1	μ, SNP20, SNP 21×SNP 23, SNP 3 ×SNP28, SNP4×SNP28	1	61.4
2	μ, SNP20, SNP 21	2, 4, 5, 8, 9, 10	88.6-93.1
3	μ, SNP20, SNP 6×SNP 21	3	90.5
4	μ, SNP20, SNP 14×SNP 21, SNP2×SNP14, SNP3×SNP14	6	89.5
5	μ, SNP20, SNP 21, SNP 2×SNP 37, SNP3×SNP37, SNP4×SNP37	7	68.6

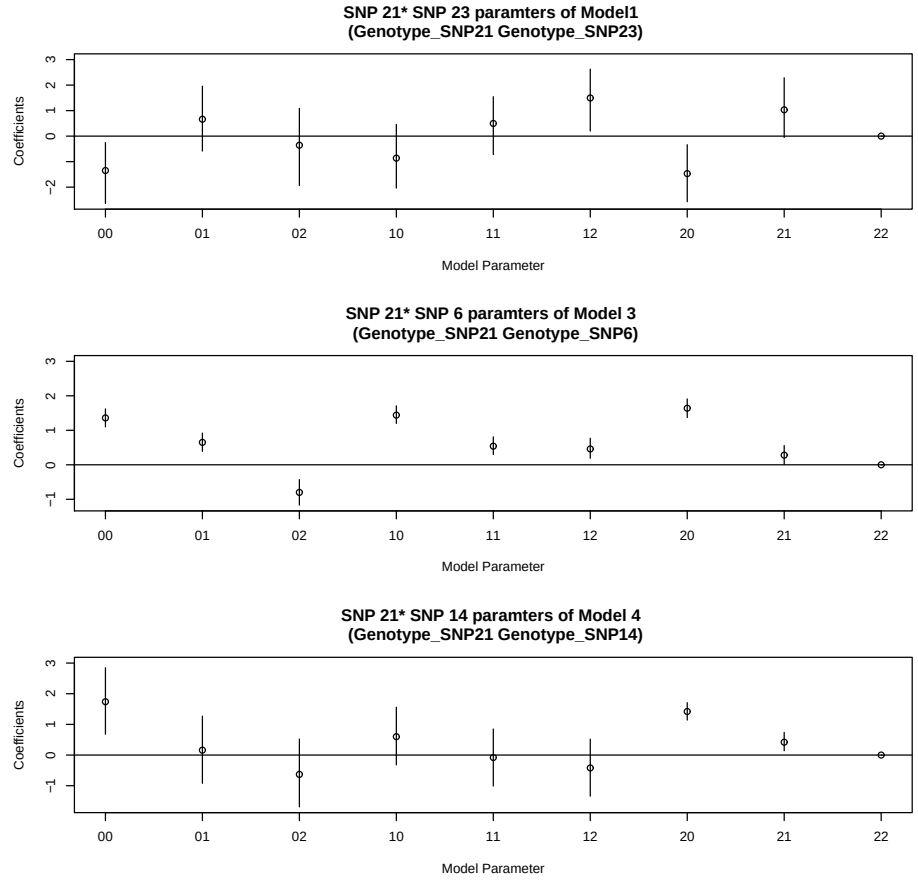

Figure 6.3: Coefficients of interaction terms with SNP 21 with credible intervals

6.5 Discussion

The aim of this chapter is to introduce a simple approach to GWA analysis which is an alternative to the current single locus analysis. This is achieved by considering a regression model with multiple SNPs, and interactions, attaching to each SNP an indicator variable representing inclusion in or exclusion from the model and performing variable selection by estimating these indicator variables. Estimation was undertaken using a novel algorithm. The model is capable of identifying a group of SNPs that contributed to the genetic causes of the diseases status through additive or interaction effects. The approach is demonstrated and evaluated using two substantive, real SNP datasets.

The results of the WTCCC analysis illustrated the ability of our model to search for main additive effects

in a relatively large data set. The model indicated that more than 50% of SNPs on Chromosome 6 do not contribute substantively to the determination of the phenotype and only less than 4% of SNPs on this chromosome are strongly informative. Of these, 17 to 24 SNPs were selected to best describe the genetic association with Type I diabetes. All four MCMC chains identified the three SNPs, rs10901001, rs874448 and rs950877, which are outside the major histocompatibility complex (MHC) region, and 12 to 13 SNPs from the MHC region. The three SNPs that are outside the MHC region all show a strong signal of association with TID. These are novel SNPs which have not been identified by other studies. In contrast, the MHC region is known to associated with a large number of infectious and autoimmune diseases [58]. The association between this region and TID has also been previously published [273] and is successfully replicated in our study.

We repeated the analysis of WTCCC data using the common SNP by SNP search algorithm. As we expected that the SNP by SNP search algorithm identified strong association signal at the MHC region, however, three novel SNPs identified by our models (rs10901001, rs874448 and rs950877) have little association to TID when tested in this manner. The unadjusted p-values for these three SNPs are 10 the power of -6, -12 and -5, respectively. It is possible that these SNPs can not be detected in isolations, but interact with SNPs on the MHC region. Further investigation on this may yield interesting information.

The aim of the analysis of the GENICA data is to exemplify the ability to detect the combination of the main and interaction effects by the described model. This study is designed for targeted gene search studies rather then running GWAs. Ten MCMC runs revealed five different models, with the most frequently identified model composed of SNP 20 and 21 as main effects. Among the models, SNP 20 is consistently selected as a main effect, but SNP 21 appears to be associated with sporadic breast cancer as either a main or an interaction effect. Analysis of the same dataset using two different types of logic regression [246] also revealed the importance of these two SNPs.

In general, according to the results from both WTCCC and GENICA analyses, SNPs identified by the models can be separated into two major categories: those present in all models and those implicated in only some of the models. The second category may be the results of correlation between SNPs entering a model and those SNPs that are already in the model during an iteration (LD between SNPs). In the

WTCCC case study, the SNPs that fall into the second category are mainly from the MHC region of Chr 6. Given that the LD structure of this region of the genome is longer and more complicated than other regions of the genome [285, 58], the SNPs of the MHC region identified by one MCMC run are potentially in linkage disequilibrium with the same region SNPs identified by other runs. However, a more complete understanding of the effect of SNPs in LD on the model requires further research.

An advantage of using the regression type of models is that the effects of genotype variants are quantified in the model. The results of the analysis of the WTCCC TID data show that individuals with homozygous genotype variants at both rs10901001 and rs874448 are more strongly associated with TID than individuals with a heterozygous genotype at the same SNPs. However, the probability of TID reduces when a heterozygous genotype is at rs950877. When a SNP is included in the model as an interaction effect, the effect of the associated variants can be quite different compared to when the SNP is included as a main effect. The results of analysis of the sporadic cancer data show that when SNP 21 is included as main effect, the homozygous reference has a strong association with the phenotype. However, when the same SNP contributes to the model by interacting with other SNPs, the homozygous reference variant is no longer the dominant effect in association with the phenotype. This finding reveals added complexity to the genetic make-up of the phenotypic trait. Although the model is able to quantify the effect of the genotype at a particular SNP, the authors are aware that the interpretation of this quantity needs to be treated with caution. The quantification of the genotype effect therefore may be more valuable for fine mapping studies.

Another advantage of the regression model is it can be easily modified for different types of phenotypic traits via using different link functions. Here, the model was developed for a binary trait, but this can be expanded in an obvious manner for more complicated phenotypes, including those with multiple subtypes.

Apart from the advantages described above, the use of a Bayesian framework overcomes the problem raised by [166]. [166] listed three drawbacks of using the logistic model in conjunction with variable selection (ie AIC, SC) under the frequentist framework. Firstly, the empty cell effect, which occurs when there is a low frequency of some genotype or genotype combination, can make the interpretation of the logistic regression result invalid. In our model, these empty cells are filtered out during the updating procedures. Although the effect of these empty cells is inconclusive, the results are not affected by these empty cells. The second

and third concerns raised by [166] are that the frequentist logistic regression demonstrated weak power for variable selection due to the correlation between variables and the problem due to the genetic heterogeneity. Although this may be true when explaining the genetic make-up with only the most prominent model, these two problems can be simply overcome by allowing multiple chains or incorporating the technique of model averaging [300].

Although it is not illustrated here with the case studies, our model has the ability to accommodate missing data by introducing extra parameters. When there are missing genotypes, the model is modified as follows. We treat the values of g_{sli} at each SNP s as unknown parameters of the model and introduce the observed genotypes g^o_{sli}. Here l can take one of four values: $0, 1, 2$ or missing. For each SNP, and each of the three possible true genotypes, we let the probability that the data is missing be ϕ, a real value in the interval (0,1). The value of ϕ is then estimated as a model parameter via the hierarchial approach.

Although the model proposed in this chapter is relatively simple, conceptually, there are some drawbacks. The first is the indecisive nature of the variable selection in each chain, indicated by the moderate contribution by various SNPs. The challenge that is admittedly only partially addressed here is how the optimally combine this information.

The second drawback is the computational burden. This problem may be overcome by adopting different MCMC algorithms, such as Reversible Jump Monte Carlo Markov Chain [104], simulated tempering [184], variational approaches [137] or population MCMC [38].

Despite the above drawbacks, the proposed model is able to detect the relevant SNPs for both TID and sporadic breast cancer. It is hoped that such investigations of alternative ways of exploring and describing the role of SNPs and their interaction in GWA studies can facilitate a better understanding of the genetics of complex disease.

7

Using gene expression programming with modified logic regression for the investigation of SNP interactions in large dimensional data

Chapter Summary

This chapter also aims to address the second main objective of the thesis. In contrast to a model based approach, here we propose a non-model based approach for detecting SNPs and/or SNP interactions. The method proposed here is a machine learning algorithm.

The method is based on logic regression, which is modified in order to speed up computation. We introduce using the gene expression programming algorithm as the searching algorithm. The method is capable of analysing a large dataset within a reasonable time frame. The model also has the flexibility to detect higher order interactions.

Chapter Conclusion

The proposed method is tested with two simulated datasets, one with 50 SNPs and the other containing 10,000 SNPs. For the smaller dataset, the methods proposed in this chapter demonstrate reasonable ability to identify the simulated SNPs and interactions. However, for the larger dataset, the results are less clear. In this chapter, we identified four areas of improvement to increase the accuracy of this method.

Authorship

Paula Macrossan, Carla C-M. Chen, Kerrie L. Mengersen

Discipline of Mathematical Sciences, Queensland University of Technology

Reference

Macrossan, P., Chen, C. C.-M., and Mengersen, K. L. (2010). Using gene expression programming with modified logic regression for the investigation of SNP interactions in large dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, submitted*

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for the associated publication is:
Macrossan P, Chen CC-M, Mengersen KL (*submitted*) Using gene expression programming with modified logic regression for the investigation of SNP interactions in large dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*

Contributor	**Statement of contribution**
Chen, CC-M, Signature & Date:	conception, interpretation, manuscript preparation, journal submission
Macrossan P	conception and conduct the research, write the code for the methods, write the manuscript
Mengersen KL	conception, interpretation, editing

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

7.1 Abstract

With the commercial availability of high throughput laboratory procedures for the identification of single nucleotide polymorphisms (SNPs), it comes the challenge of identifying genes and gene-gene interactions associated with disease in high dimensional space. In this chapter we demonstrate MLR-GEP, a non-parametric approach for identifying potentially interesting gene interactions, using a combination of Logic Regression, an adaptive classification and regression methodology that constructs predictors as Boolean combinations of the binary SNP covariates, with Gene Expression Programming, a variant of genetic programming, as the stochastic search algorithm. The performance of MLR-GEP in discovering interactions between SNPs in simulated 50 SNP and 10,000 SNP datasets is demonstrated.

7.2 Introduction

With the recent mapping of the human genome [269] and subsequent mapping of many agricultural species (e.g. The Bovine HapMap Project, `http://bfgl.anri.barc.usda.gov/`) has come the commercial availability of high throughput laboratory procedures for the identification of single nucleotide polymorphisms (SNPs) in both humans and animal species. This has produced a convergence of the focus of quantitative and qualitative geneticists in the pursuit to identify interactions between genes, which are vital in the understanding of common diseases such as diabetes, asthma and cardiovascular diseases, as well as production traits, such as growth and meat quality in livestock species.

The challenge of genome wide association studies (GWAs) is three pronged: firstly, in the development of powerful statistical and computational methods to model the relationship between combinations of SNPs and common disease and production traits; secondly, in the selection of the genetic variables to be included in the analysis, and thirdly, in the interpretation of gene-gene interaction models [198]. This chapter addresses the first and second challenges, leaving the last to the bio-sciences.

The most prominent algorithm in searching for important SNPs is SNP-by-SNP searching in which each

putative SNP is evaluated individually with respect to the strength of its association with the outcome of interest. A wide range of outcomes have been investigated using this approach, including obesity [248], Crohn's disease [69] and Type 2 Diabetes [255]. Although this is a fast algorithm and able to accommodate the analysis of large dimensional data, e.g. Affymetrix Genome-Wide Human SNP Array 6.0 with 1.8 million markers (including SNPs and the copy number variation), it is subject to the problem of multiple testing and therefore requires adjustment, such as via the Bonferroni correction or false discovery rate. Moreover, its limited focus on single SNP associations may not be suitable for complex trait, which might occur only if a particular combination of genotypes is present at different susceptible loci [123].

Various statistical methods for detecting gene-gene interactions have been recently developed. Depending on the embedded algorithm, these methods can generally be categorized into model based and non-model based approaches. The former category, including methods such as regression, often requires the estimation of model parameters; in contrast, the non-model based approaches, such as random forests [33] and neural networks [200, 201], which are sometimes also referred to as data mining approaches, are designed for detecting non-linear relationships between phenotypes and genetic markers, and may be more desirable for detecting higher order interactions. [187] reviewed different machine learning algorithms for detecting gene-gene interactions. A more comprehensive recent study reviewed both model based and non-model based approaches for detecting interactions and the computer packages available for these methods [51].

Among the different methodologies referenced by [51], logic regression (LR) is an intriguing approach. It is an admixture approach, which has a structure of the regression model, but instead of directly regressing against the predictors, the response is regressed against a combination of "logic trees" which are identified via the machine learning algorithm. A logic tree is a tree-like structure comprising Boolean expressions, such as AND, OR and NOT, and predictor variables. This method is described in Section 7.3.1. Because the method is based on the combination of the regression model and the tree structure, LR is more versatile in detecting different types of interaction, epistasis effects. This includes two different types of the epistasis effects defined by [19] and [86].

The identification of an optimal logic tree involves the use of a search algorithm. The algorithm used in the original logic regression is simulated annealing [235]. However, two limitations of the use of simulated

annealing have been noted [153]. Firstly, it aims to identify a single best model, which neglects potential competing models which fit the data almost as well. Secondly, when a SNP is included in the model, the LD SNPs of this SNP are highly likely not be selected in the model. Therefore, different searching methods have been applied for the improvement of logic regression.

Monte Carlo logic regression [MCLR, 153], which is an exploratory tool that combines a Markov Chain Monte Carlo algorithm and logic regression. MCLR has been demonstrated to be more useful than logic regression when there is a large number of LD SNPs [153]. The code for MCLR is available in the 'LogicReg' package of R [219].

Although MCLR uses MCMC methods and priors on some parameters, the coefficients of the logic trees, are estimated using maximum likelihood approach, so it is not a fully Bayes approach. Additionally, in the examples of MCLR, two parameters need to be set in advance. These pertain to the hyperparameter of the geometric prior on model size which acts as a penalty to favour parsimony and the maximum number of trees. It has been noted [91] that setting of these two parameters can have a large influence in the results. Therefore, a Full Bayesian Logic regression algorithm (FBLR) has been proposed as an alternative [91]. The reported advantages of FBLR include a prior on the coefficients, which overcomes the problem of presetting these; restricting of the Boolean parameters in the logic tree to "AND" only, which gives equal weight to all models within the same size under the uniform prior setting; and more interpretable result of FBLR [91]. Note that, by de Morgan's rule, the use of the complement can account for an "OR" operator.

Extensions of regression have also been proposed in literature. For example, a logicFS algorithm uses the simulated annealing algorithm, to perform subset selection in regression [245]. The main aim of logicFS is the identification of important SNP interactions. The method incorporates the use of bootstrapped samples and disjunctive normal form (DNF). Instead of searching for the "optimal" model over all possible model spaces, the simulated annealing algorithm is independently applied to a large number of bootstrapped samples drawn from the complete data space. Also, like FBLR, to make the results easier to interpret and the interaction more identifiable, logicFS uses only "AND" and "OR" operators in logic tree. Because of the use of bootstrapped samples, out-of-bag samples are used for the validation of variable importance [245].

All of the above methods are no doubt a great improvement on the original logic regression method. However, it is of interest to consider other potential searching algorithms which may be implemented in the logic regression setting. Therefore, the aim of this study is to introduce a different searching algorithm, namely gene expression programming (GEP) in this context.

GEP [84] is a hybrid of genetic programming (GP, [154]) and genetic algorithm (GA, [125]). It is an evolutionary algorithm based on artificial intelligence or machine learning inspired by biological evaluation, with the aim of automatically solving a problem without specifying the form of the solution. GEP, GA and GP are encompassed within a wider class of "genetic algorithm", which all generate a population of individuals, select individuals based on their *fitness*, then modify individuals using one of many genetic variants.

Because GEP is a combination of GA and GP, it also combines the advantages of both GA and GP, in such a way that the GEP eases the manipulation of the GA and has the functional complexity possible with GP. The main difference between these three algorithms is in the "individual" of the population. In GA, the individual is a symbolic string of fixed length; in GP, the individuals are nonlinear entities with different sizes and shapes and in GEP, the individuals are linear strings of fixed length which are later expressed as a nonlinear entity of different shapes and sizes.

Although specific references to using GEP are limited, there have been frequent references to GA. GA is noted as a suitable tool for the optimization of large dimensional problems [250] and has been used in a range of application including detecting outliers in linear regression models [55] and optimizing a statistical quality control problem [114].

This chapter is organized as follows. The chapter starts with an overview of logic regression (Section 7.3.1). The logic regression is then modified in order to speed up the computation and a proposed method, named *Modified Logic regression-gene expression programming (MLR-GEP)* is introduced in Section 7.3.2. The performance, specificity and sensitivity of this method is then evaluated with two simulated data sets described in Sections 7.4. Results of the evaluation are given in Section 7.6. A discussion follows in Section 7.7.

7.3 Methods

7.3.1 Logic Regression (LR)

Letting Y be a phenotypic trait, the logic regression model [235] is written as:

$$g(y) = \beta_0 + \sum_{i=1}^{K} \beta_i L_i \tag{7.1}$$

where L_i denotes the ith logic expression, $i = 1, \ldots, K$; K is total number of trees in the model, and β_0 is a constant. A logic expression is a notation of a logic tree, which is a tree-like structure comprising Boolean operators, 'AND', 'OR' AND 'NOT', and leaves, which represent the SNPs. Figure 7.1 is an example of logic tree with a logic expression given by

$$L_{\text{Figure7.1}} = (S_1 \wedge S_2^c) \vee [(S_3 \vee S_4) \wedge S_5^c]$$

Like the generalized linear model, $g(.)$ is a link function which links the random and systematic components. The choice of $g(.)$ depends on the type of trait of interest; for example, for a case-control study, the logit link function is often used. The model parameters, β_0 and β_i are usually estimated using maximum likelihood.

As mentioned in the Introduction, the search algorithm used in the logic regression is typically simulated annealing. Simulated annealing is defined on a state space, S, which is a collection of states. The states are related by a neighbourhood system, where a set of neighbour pairs in S defines a substructure, M. The elements of M are called moves. When the states are adjacent, they can be reached by a move; otherwise the states can be connected by any number of moves. There are four possible moves in the logic tree, which are 1) alternating a leaf, 2) changing operator, 3) growing and pruning and 4) splitting and deleting. Not all moves, however, are permissible at a state s. For instance, when the maximum size of tree is reached, the moves which result in adding leaf/leaves are prohibited. A move from one state to another depends on the acceptance probability which in turn depends on the ratio of two state spaces and the *temperature* of the

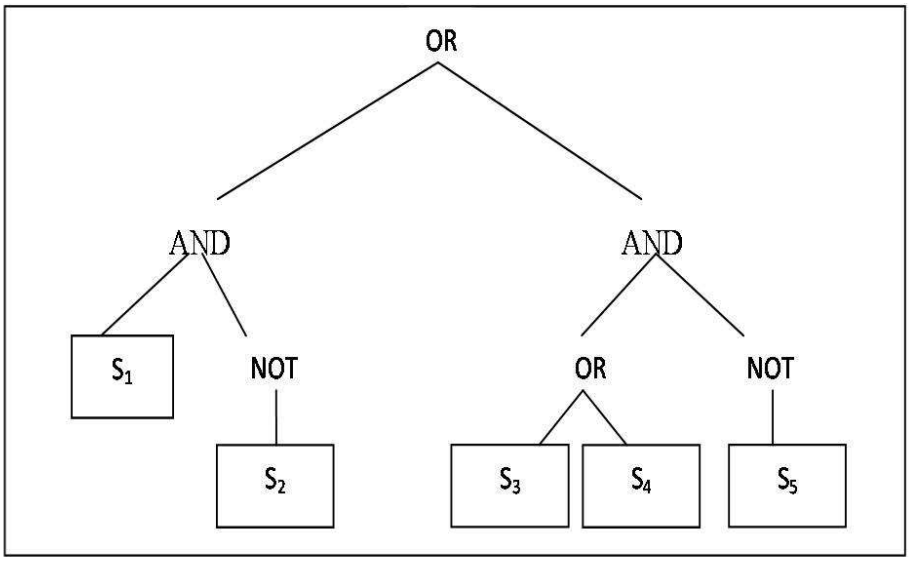

Figure 7.1: Logic tree of MLR representing the logic expression Y, where Y = L1 OR L2, and L1=(S_1 AND NOT(S_2)), L2=((S_3 OR S_4) AND NOT(S_5)).

position of the chain. When the temperature is high, there is larger chance of accepting the move than at a cooler temperature. The temperature is not constant over iterations in the logic regression. At the beginning, the temperature is set to be high so nearly all possible moves are accepted. As the iteration proceeds, the temperature reduces.

7.3.2 Modified Logic Regression with gene expression programming (MLR-GEP)

The aim of our proposed method is to identify a set of SNP interactions, that are potentially associated with the expression of a trait. Thus the model coefficients, β_i, $i \in 0, \ldots K$ are less relevant in the modified logic regression (MLR). The method proposed here is therefore similar to the ensembles approach described in [64], and less related to model fitting. Giving each tree an equal weight can substantially speed up the computational time, therefore β_i is fixed to one. Thus using the same notation as earlier, the MLR model becomes:

$$g(y) = \sum_{i=1}^{K} L_i \tag{7.2}$$

Gene expression programming (GEP) is an iterative procedure which utilizes the concept of gene, population and evolution. The changes of individual from one iteration (also called 'generation') to the next is called an evolution and the evolution repeats until a maximum number of evolutions is reached or until a desired fitness is achieved.

The linear strings of fixed length in GEP are called 'genes'. Genes are composed of 'nodes' representing either functions (i.e. Boolean- AND, OR and NOT) or 'terminals' (i.e SNPs). A number of genes can be linked by functions to form a 'chromosome'. Genetic operations such as mutation and transposition take place on genes and chromosomes, after which the latter are expressed as non-linear entities of different shapes and sizes, called 'expression trees' (ETs), which is equivalent to the logic tree of logic regression.

The GEP gene comprises a 'head' and a 'tail'. The head contains both functions and terminals, whereas the tail contains only terminals. The first head node of each gene, or 'root' node, must be a function. The tail length is a fixed function of the head length and the maximum function arity (number of function arguments). The structure of the GEP gene and the translation system from fixed length string to expression tree guarantee that all modifications arising from evolution of the individuals result in syntactically correct ETs. Despite the fixed length of GEP genes, they have the potential to code for ETs of widely differing shapes and sizes. The number and length of GEP genes are peculiar to the problem at hand.

GEP individuals are subjected to genetic operators (genetic variation) that can substantially modify their structure. The genetic operators in GEP include mutation, transposition, insertion sequence, root insertion sequence and recombination. Mutation is a change occurring in a single node of a gene. A mutation can occur at both the head and tail of a gene. When it occurs in the gene head, it may produce either a function or terminal, whereas a tail mutation must result in a terminal. Transposable elements of GEP are fragments of the genome that can relocate to another place in the chromosome. Insertion Sequence (IS) elements are short fragments with a function or terminal in the first position that may transpose to the head of genes except the root. Root Insertion Sequence (RIS) elements are short fragments with a function in the first position,

and which transpose to the root of genes. In addition, an entire gene may transpose to the beginning of the chromosome (gene transposition). Recombination in GEP may take one of three forms. In all cases, two parent chromosomes are randomly chosen and paired to exchange 'genetic' material. During one-point recombination, two parent chromosomes cross over at a randomly chosen point to form two daughter chromosomes. During two-point recombination, two parent chromosomes exchange the fragment contained between two randomly chosen points to form two daughter chromosomes. In gene recombination, an entire gene is exchanged during crossover. 'Elitism', or the survival and cloning of the best individual chromosome in each generation into the next generation, is practised.

GEP individuals (or solutions) are selected according to their fitness, where fitness is defined as the ability of the solution to predict the trait. In the problem at hand, the dependent variable is the case or control status of each datum, being a binary trait represented as 1 or 0, respectively. Selection for reproduction, mutation and crossover is based on the fitness proportionate selection roulette-wheel scheme [100], so that the chance of a potential solution participating in any of these operations is proportional to the fitness of that solution as a fraction of the total fitness.

In the current application of finding SNP interactions, the functions represented in the nodes of MLR-GEP genes are Boolean operators, and the terminals are single SNP identifiers. Boolean operators link a number of such genes to form the chromosome. The root node of the gene head must contain a Boolean operator. Otherwise, the head of a gene may contain both Boolean operators and SNPs, whereas the tail contains only SNPs. Figure 7.2 is an example of the translation of a fixed length MLR-GEP string into an expression tree and its associated logic expression.

That the tail length is a fixed function of the head length and the maximum function arity guarantees that the tail always contains enough terminals to fully satisfy any possible head arrangement. However, this also means that certain terminals in the head and the tail may not be used in the expression tree and its associated logic expression, as seen in Figure 7.2 where the last value 7 is not included in the expression tree. The extreme case occurs when the head of the gene contains the Boolean NOT operator having an arity of one, and the next node contains a SNP, giving a single NOT(S..) expression regardless of the head (and therefore tail) length. As highlighted above, it follows that a single point mutation in the gene head can lead to a

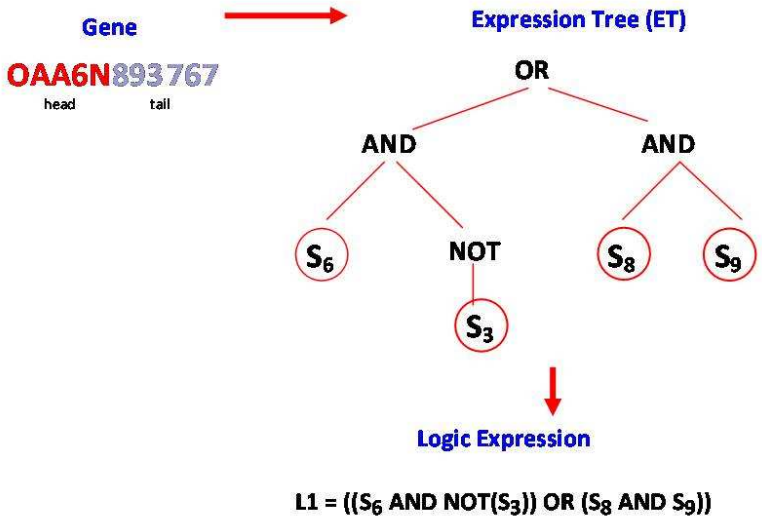

Figure 7.2: An example of the fixed length string of an MLR-GEP 'gene' and its translation to an MLR-GEP expression tree and associated logic expression. The 'head' of the gene is composed of the sequence of nodes OAA6N, representing the Boolean operators AND (A), OR (O) and NOT (N), and the SNP identifier 6. The 'tail' of the gene is composed of the sequence of nodes 893767, all representing SNP identifiers. Note that three SNP identifiers at the end of the tail, 7, 6 and 7 are not used in the ET. The ET of GEP is equivalent to the logic tree of logic regression (see Figure 7.1)

dramatic change in the associated expression tree. Figure 7.3 illustrates point mutation in MLR-GEP.

In this application, fitness is defined as the ability of the solution to predict the case/control status of each datum, which is the same as correct classification. For any GEP individual i, the fitness is

$$\text{fitness}_i = \sum_{j=1}^{C} (c_{ij} == T_j) \tag{7.3}$$

where C is the number of subjects in the data set, T_j is the case/control status for subject j, and c_{ij} is the predicted case/control status under GEP individual i for subject j.

In the current setting of the MLR-GEP, a number of parameters are required to be set in advance. These include the maximum number of iterations, the head length of the gene, the number of genes for each

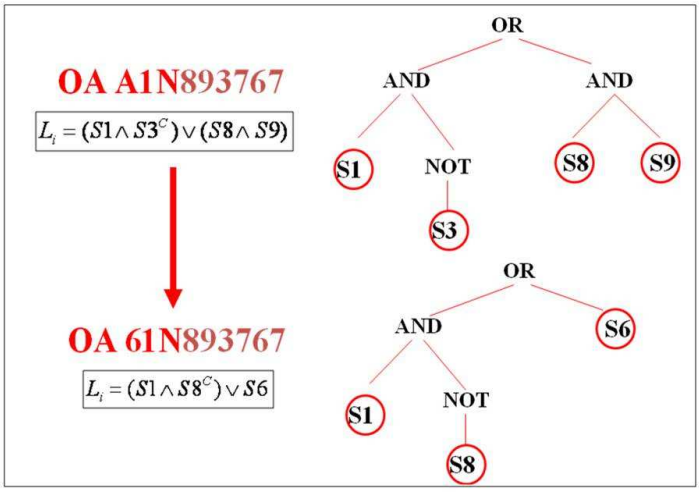

Figure 7.3: An example of point mutation of MLR-GEP 'gene' and the resultant change in the expression tree and associated logic expression. The 'head' of the gene is composed of the sequence of nodes OAA1N, representing the Boolean operators OR (O), AND (A) and NOT (N) OR (O) and the SNP identifier 1. Point mutation occurs in the third node of the gene head of the parent gene, with a change from the operator AND to the SNP identifier 6. Note that three SNP identifiers at the beginning of the tail of the parent gene, 8, 9 and 3 are used in the logic expression associated with the parent gene, whilst in the daughter gene, only the initial SNP identifier 8 in the tail is used.

chromosome, and the parameters associated with the genetic variants, including mutation rate, one and two point recombination rate, gene recombination rate, and IS, RIS and Gene transposition rate.

7.3.3 SNP Coding

SNPs can be coded as covariates in a number of ways, depending on the problem. [235] suggest coding of the *i*th SNP as two binary predictors, $X_{i,1}$ and $X_{i,2}$. Here the subscript i of X refers to the SNP number, whilst the subscripts 1 and 2 refer to the minimum number of variant (recessive or 'a') alleles at the SNP site. Thus $X_{i,1}$ and $X_{i,2}$ respectively code the dominant and recessive effects of SNP i. In contrast to the separate coding system, in our approach the SNPs are coded as a single integer and represent the genotype. Table 7.1 shows the difference between the two methods. Note that in this chapter, the terms homozygous variant and homozygous reference indicate the genotypes 'aa' and 'AA' respectively, and this is in line with Ruczinski's SNP nomenclature.

The coding method implemented here halves the number of SNPs and thus halves the solution space. However, this results in running the searching algorithm twice with two different searching rules to achieve the same search results. This is advantageous given the availability of parallel computing. The first search requires searching for only the homozygous variant (genotype 3 in Table 7.1) and the second requires searching for the homozygous variant or the heterozygote (genotypes 2 or 3 in Table 7.1).

Table 7.1: Genotype coding of SNPs using a single covariate X_i, compared with Dominant and Recessive coding using two binary covariates $X_{i,1}$ and $X_{i,2}$ demonstrated by [235].

Genotype	SNP Coding Method		
	Logic Regression [235]		MLR-GEP
	Dominant ($X_{i,1}$)	Recessive ($X_{i,2}$)	Genotype Code (X_i)
Homozygous Reference(AA)	0	0	1
Heterozygote(Aa, aA)	0	1	2
Homozygous Variant (aa)	1	1	3

Table 7.2 demonstrates the use of the logical NOT operator and the two different search goals for the two runs of the MLR-GEP taken in this study. It is noted that the genotype coding approach allows any genotype or combination of genotypes to be set as the goal of the MLR-GEP search. The SNP nomenclature used in this study to describe the interactions is also given in Table 7.2, and follows that of [245], where the subscript i of S_i refers to the SNP number, whilst the subscripts 1 and 2 refer to the minimum number of variant (recessive, 'a') alleles at the SNP site. Thus, $S_{i,1}$ refers to SNP i being either the homozygous variant or the heterozygote genotype (aa, Aa, or aA), whilst $S_{i,2}$ refers to SNP i being the homozygous variant genotype (aa).

Table 7.2: Search types used with genotype coding compared with dominant/recessive coding. The two search types and possible results of genotype searching using the logical NOT operator, compared with the equivalent coding of [235] requiring only a single search.

	Search Goal		NOT(Search Goal)		
Search Type	SNP	Genotype	SNP	Genotype	Ruczinski Coding
1	$S_{i,2}$	aa	NOT($S_{i,2}$)	AA,Aa,aA	Recessive, $X_{i,2}$
2	$S_{i,1}$	aa,Aa,aA	NOT($S_{i,1}$)	AA	Dominant, $X_{i,1}$

The MLR-GEP code was implemented in Fortran and compiled using an Intel ®Fortran Compiler Version 10.1. The code was run on a SGI Altix XE1200 Cluster 120x E5345 64 bit Intel Xeons at 2.33 GHz.

7.4 Data Description

In this chapter, MLR-GEP is tested using two datasets, which are simulated from the same settings with 50 and 10,000 SNPs respectively. The simulations are based on the allele frequency and assume full penetrance of disease in a given combination of genes. The strategies used for data simulation are the same as those described in [235], [245], [91] and [209].

7.4.1 Simulation Set 1

Fifty datasets of 1000 observations each (500 cases and 500 controls) and 50 SNPs were simulated with allele frequencies for each SNP randomly generated within the range of 0.2 to 0.4. The case/control status of each observation was based on the rules described in Experiment 1 of [245], where an observation is classified as a 'case' if one of four logic rules is true. The four logic rules, the number of cases simulated for each, and the number of controls are given in Table 7.3.

Table 7.3: The four logic rules L1 to L4 describing the simulated datasets, the number of cases simulated for each rule, the proportion of the data described by each rule, and the number of controls simulated per dataset. Each rule describes SNP combinations using Boolean AND and NOT operators for each SNP i for a minimum of one or two variant alleles (a) occurring at the SNP site, coded as $S_{i,1}$ and $S_{i,2}$ respectively

Rule	Simulated Interaction	Number of Cases	Proportion of Data
L1	$S_{1,2}$	100	10%
L2	NOT($S_{2,1}$) AND $S_{3,2}$	150	15%
L3	$S_{4,2}$ AND $S_{5,2}$ AND $S_{6,2}$	100	10%
L4	$S_{7,2}$ AND $S_{8,2}$	150	15%
No Rule	None	500 (Controls)	50%

The datasets can be fully described by further combining the four rules with Boolean OR operators to form a single logic rule Y, where $Y = L1$ OR $L2$ OR $L3$ OR $L4$. However, to achieve 'clean' data (setting of full penetrance), simulation was controlled so that each 'case' datum contained only one of the four rules possible (i.e. using exclusive OR (XOR)), giving $Y = L1$ XOR $L2$ XOR $L3$ XOR $L4$. Additionally ten of the fifty datasets were eliminated, since one of SNP4, SNP5 or SNP6 in the three-way interaction of Rule L3 was not needed to explain all the cases and controls correctly, leaving forty datasets in Simulation Set 1.

Initial statistical screening of Simulation Set 1 was undertaken using the 'direct method' of simple analysis of variance (ANOVA) testing association of each individual SNP genotype with the case/control status of the observation, using [219].

7.4.2 Simulation Set 2

Forty datasets of 1000 observations of 9950 SNPs each were simulated with allele frequencies for each SNP again generated randomly within the range of 0.2 to 0.4. These 40 datasets were then combined with Simulation Set 1 to create 40 datasets each containing 10,000 SNPs, plus the corresponding case/control status of each observation taken from Simulation Set 1. Thus, the datasets in Simulation Set 2 contained the same interactions as Simulation Set 1, the same case/control status of each observation, and the same numbers of cases and controls, as described in Table 7.3.

7.5 Settings

Table 7.4 shows the parameter settings in the analysis of Simulation Set 1. We generated 20 populations and evolved each population over 50,000 generations. The heads and tails were preset to 3 and 4. For the settings on the evolutionary rate, we used the most neutral settings that required a very limited optimization, that is 0.3 [84].

Similar settings were also used for the analysis of Simulation Set 2; however, given that the number of SNPs is much larger, the number of populations and the number of generations per population was also increased. The final settings for these two parameters were 200 populations and 150,000 generations for each population.

Table 7.4: MLR-GEP settings used in Experiment 1, and (with exceptions) for Experiment 2 and GAW14 data. Exceptions for the latter are for the number of generations per run (150,000), the population size (200), and the number of SNPs in the terminal set are SNPs 1 to 10,000 and SNPs 1 to 9,187 respectively.

Parameter	Setting
Number of runs for each dataset	20
Number of generations per run	50,000
Population size	20
Number of fitness cases	1000
Boolean function set	AND, OR, NOT
SNP Terminal set	SNPs 1 to 50
Head length	3
Number of genes	4
Boolean linking function	OR
Mutation rate	0.3
One-point recombination rate	0.3
Two-point recombination rate	0.3
Gene recombination rate	0.3
IS transposition rate	0.3
RIS transposition rate	0.3
Gene transposition rate	0.3

7.6 Results

Simulation 1 Table 7.5 shows the average percentage of times each of the four significant interactions was discovered over the 20 runs of the MLR-GEP for each of the datasets in Simulation Set 1. Table 7.5 also shows the number of times the sub-rule of L2, NOT($S_{2,1}$), was found, and the number of times variations of the Rule L3, either rules ($S_{4,2}$ AND $S_{5,2}$) or ($S_{4,2}$ AND $S_{6,2}$) or ($S_{5,2}$ AND $S_{6,2}$), were found. The average fitness of the runs is given as a percentage of times out of a total of 1000 fitness cases the correct case/control status was predicted from the MLR-GEP rule based solution.

Although SNP 1 was simulated to be associated with the case/control status, it was not involved in any interactions with other SNPs (Rule L1; Table 3). When the search goal was set to the homozygous variant (*aa*) (Search Goal 1; Table 7.2), SNP 1 was correctly found in 100% of rules in an OR association with other SNPs, consistent with Rule L1. In contrast, when the search goal was set to the homozygous or heterozygous variants (*aa,Aa,aA*) (Search Goal 2; Table 7.2), then although SNP 1 was found in over 50% of the solutions, it was invariably associated in interactions with other SNPs.

The results in Table 7.5 show that the average fitness of the outcomes of Search Goal 1 was substantially

higher than that of Search Goal 2 (95% cf. 67%). However, this superior fitness can be attributed to finding Rule L1 in 100% of runs, and finding all the other SNPs participating in the interactions, although not always the correct interactions in the case of SNP 7 and SNP 8, or the full interactions in the case of SNPs 4, 5 and 6. In the latter case, the incomplete rules of ($S_{4,2}$ AND $S_{5,2}$), ($S_{4,2}$ AND $S_{6,2}$) and ($S_{5,2}$ AND $S_{6,2}$) were found in over 60% of rules for Search Goal 1. From Table 7.2 it can be seen that ($S_{2,1}$) cannot be found using Search Goal 1; this is consistent with the results in Table 7.5 for Rule and sub-rule L2.

Table 7.5: MLR-GEP Results for Experiment 1. The mean (and range) for percentage of times each of the Rules 1 to 4, plus subsets of Rules 2 and 3, describing the simulated datasets in Simulation Set 1 (50 SNPs; see Table 7.3) were found for Search Goals 1 (homozygous variant: aa) and 2 (homozygous variant or heterozygous; aa,Aa, or aA), plus the mean (and range) of the fitnesses found.

Rule	Interaction	Mean(Range) SearchGoal 1 (aa)	Mean(Range) SearchGoal 2 (aa,Aa,aA)
L1	$S_{1,2}$	100 (100-100)	0 (0-0)
L2	NOT($S_{2,1}$) AND $S_{3,2}$	0 (0-0)	34 (0-95)
L3	$S_{4,2}$ AND $S_{5,2}$ AND $S_{6,2}$	3 (0-20)	50 (0-95)
L4	$S_{7,2}$ AND $S_{8,2}$	69 (10-100)	79 (5-100)
Sub(L2)	NOT($S_{2,1}$)	0 (0-0)	67 (0-95)
Sub(L3)	($S_{4,2}$ AND $S_{5,2}$) OR ($S_{4,2}$ AND $S_{6,2}$) OR ($S_{5,2}$ AND $S_{6,2}$)	61 (10-100)	69 (0-100)
Fitness %		95 (93-98)	67 (65-70)

For Rules L2 and L3, Search Goal 2 was greatly superior. Although in some cases the algorithm does not find these rules, the average number of times these rules were found over the 20 runs for each dataset marks these rules as significant findings. In contrast, incorrect interactions containing SNPs other than those simulated to be significant were never repeated, although in some datasets certain SNPs would reappear in other interactions in up to 20% of the runs.

Simulation Data 2 Table 7.6 shows the average percentage of times each of the four significant interactions was discovered over the 20 runs of the MLR-GEP for each of the datasets in Simulation Set 2. Table 7.6 also shows the number of times the sub-rule of L2, NOT($S_{2,1}$), was found, and the number of times variations of the Rule L3, either rules ($S_{4,2}$ AND $S_{5,2}$), ($S_{4,2}$ AND $S_{6,2}$) or ($S_{5,2}$ AND $S_{6,2}$), were found. The average fitness of the runs is also shown.

The results in Table 7.6 show that, as for Simulation Set 1, with Search Goal 1(*aa*) SNP 1 was correctly found in 100% of rules in an OR association with other SNPs, consistent with Rule L1. In contrast to

Simulation Set 1, when the search goal was set to the homozygous or heterozygous variants (*aa*, *Aa*, or *aA*) (Search Goal 2; Table 7.2), SNP 1 was never found in any rule. However, all other significant SNPs were found in either OR associations with other SNPs, or incorrect associations with other SNPs. This latter observation explains the relatively good fitness rates (average 61%; see Table 7.6) for Search Goal 2 on Simulation Set 2, in that although in most cases the rules were not found, the significant SNPs were identified (except for SNP 1, as explained above).

Table 7.6: MLR-GEP Results for Experiment 2. The mean (and range) for percentage of times each of the Rules 1 to 4, plus subsets of Rules 2 and 3, describing the simulated datasets in Simulation Set 2 (10,000 SNPs, see Table 7.3) were found for Search Goals 1 (homozygous variant: aa) and 2 (homozygous variant or heterozygote; aa, Aa, or aA), plus the mean (and range) of the fitnesses found

Rule	Interaction	Mean(Range) SearchGoal 1 (aa)	Mean(Range) SearchGoal 2 (aa,Aa,aA)
L1	$S_{1,2}$	100 (100-100)	0 (0-0)
L2	NOT($S_{2,1}$) AND $S_{3,2}$	0 (0-0)	0 (0-0)
L3	$S_{4,2}$ AND $S_{5,2}$ AND $S_{6,2}$	0 (0-0)	0 (0-0)
L4	$S_{7,2}$ AND $S_{8,2}$	31 (5-65)	7 (0-41)
Sub(L2)	NOT($S_{2,1}$)	0 (0-0)	25 (0-100)
Sub(L3)	($S_{4,2}$ AND $S_{5,2}$) OR ($S_{4,2}$ AND $S_{6,2}$) OR ($S_{5,2}$ AND $S_{6,2}$)	42 (6-70)	0 (0-6)
Fitness %		93 (90-95)	61 (59-64)

As for Simulation Set 1, the average fitness of the outcomes of Search Goal 1 was significantly higher than that of Search Goal 2 (93% cf. 61%). Again, this superior fitness can be attributed to finding Rule L1 in 100% of runs, and finding all the other SNPs participating in the interactions, although not necessarily in the correct interactions in the case of SNP 7 and SNP 8, or the full interactions in the case of SNPs 4, 5 and 6. In the latter case, the incomplete rules of ($S_{4,2}$ AND $S_{5,2}$), ($S_{4,2}$ AND $S_{6,2}$) and ($S_{5,2}$ AND $S_{6,2}$) were found in 42% of rules for Search Goal 1, but rarely for Search Goal 2. As discussed for Experiment 1, Rule L2 cannot be found using Search Goal 1, and the results in Table 7.6 are again consistent with this. Although Rule L2 was not found using Search Goal 2, the significance of SNP 2 in the homozygous reference form (AA) was found on average in 25% of searches.

Computer Running Time. Using the hardware and software specified in Section 7.3.3, the speed of the MLR-GEP was 20 generations per second for the 10,000 SNP datasets in Experiment 2. Thus the time taken for a single run of MLR-GEP over 150,000 generations was approximately 8.5 hours. However, using Goal

Search 1, average population fitness levels of over 90% were always achieved after only 1000 generations. On average, convergence had occurred by generation 50,000 or a runtime of less than 3 hours.

7.7 Discussion

In this study, we presented the use of an alternative searching algorithm for finding the logic tree under the framework of logic regression. The MLR-GEP, as anticipated, benefits from the computational efficiency of gene expression programming. A similar advantage is also found when applying genetic programming for identifying higher order SNP interactions [209].

The overall fitness measure of MLR-GEP is comparable to some of existing methods. For instance, Nunkesser et al [209] applied genetic programming (GPAS) to identify higher order SNP interactions. Under the same methods of data simulation, the reported misclassification rate of their study is around 33% which is similar to the fitness level of MLR-GEP. Nunkesser et al [209] also compared the misclassification rate of GPAS with the standard logic regression, CART, Bagging and random forests using the same simulation dataset, and found the misclassification rates of these methods were between 34 to 38% which is also comparable with our method.

For the single locus effect, such as SNP 1, Search Goal 1(aa) was 100% accurate in finding the true single-locus state of this SNP. Although the significance of SNP 1 can be demonstrated through simple ANOVA with adjustment for false discovery rates, the findings for Search Goal 1 positively attest to its performance for the single-locus case. However, for identifying SNP interactions in the smaller dataset, Search Goal 2 identified most of interactions compared with Search Goal 1, despite the lower average fitness levels. The differences between these two search goal is less obvious when the dataset is substantially large.

Even though the overall fitness is within a reasonable range, one major drawback of the MLR-GEP is that the algorithm is potentially unstable in its current state. When the dataset was small (i.e candidate gene search), the performance of the MLR-GEP in finding pre-specified interactions was just above 50% and ranged from 0 to 100%. However, when the dataset became substantially large and noisy, even though the fitness of the

model was above average, the probability of finding the 'real' interaction was less than 50% with some SNP interactions not identified. This is a concern, considering that there is possibility that expression of the trait can be the result of the interaction of many genes, each with small effect [123]. However, the drawback is not restricted to this algorithm; indeed detecting of complex interaction is an ongoing " holy grail" of current research.

MLR-GEP can potentially be improved by changing various settings. Firstly, in this study, in order to achieve computational efficiency in the algorithm, we avoided parameter estimation by assigning equal weighting to each logic tree, and predicted individual phenotypes based upon the combination of these trees. This can potentially contribute to a poor discovery rate. An improvement is thus possible by retaining the original formulation of LR and estimating the model parameters (β) of Equation 7.1. Furthermore, the model parameter, can then be utilized for establishing variable importance ranking.

A second possible contributor to the poor discovery rate relates to the tuning parameters. For the evolutionary process, various parameters are required to be set in advance, including the settings of the MLR-GEP 'chromosomes' such as head length, number of genes and Boolean linking function. Various studies have noted strong dependency between the tuning parameters of genetic algorithms (including GA, GP and GEP) [89, 254], and observed that inaccurate setting of these parameters can result in premature convergence to local optima [301]. For a complicated genetic system, an optimal parameter setting is obviously difficult to achieve. Moreover, some researchers have found even after adjust the tuning parameters according to the problem in hand, genetic algorithms can still perform below expectation. For example, [89] found surprisingly unsatisfactory results when they used a genetic algorithm ([125]) to find the maximum a posterior (MAP) estimate of a binary variable in Bayesian image analysis. Similarly, [134] found that GA performs badly in some simple optimization problems. Thus the theoretical and empirical basis of GA has been questioned [134].

To overcome the problems associated with GAs, researchers have suggested the use of hybrid algorithms [89, 301, 254]. Although the method of hybridization is different, [89] and [301] both propose to combine GA with simulated annealing to achieve effectiveness and efficiency in the optimization. Similarly, [254] suggest to combine GEP with simulated annealing to reduce the dependency of GEP on the tuning

parameters and improve the performance of GEP.

A more sophisticated fitness measure can potentially yield better results in identifying SNP interaction. In this study, MLR-GEP adopted correct classification as the measure of purity, which may potentially oversimplify the problem. For other tree like methods, such as classification and regression trees (CART, [34]) and random forests (RF, [33]), different criteria have been proposed as measures of impurity, including misclassification rate, Gini index [34], cross-entropy [113], Gain ratio [218], DKM [63] and minimum description length (MDL). Although these impurity measures are based upon the misclassification rate, some are more sensitive than others [113, 232, 233]. Moreover, MDL has been successfully applied to genetic programming as the fitness function for pattern recognition problems [128]. An alternative criterion is to use the multi-objective function. For example, [209] used multidimensional fitness value for GPAS which aims to balance the misclassification rate with the complexity of logic expression.

Although the results are not presented here and the current state of MLR-GEP has potential for further improvement, the method has been applied to a real SNP data on sporadic breast cancer [46], and yielded similar findings in SNP interaction as MClogic, logicFS, GPAS, random forests and a Bayesian regression model. This suggests that the current state of MLR-GEP is able to analyse small datasets, but for larger dimension data, it still requires some further development.

In this chapter, we introduced the use of gene expression programming as an alternative searching algorithm for modified logic regression to perform genetic data analysis, with a focus on identifying SNP interactions. Like all other machine learning algorithms, the use of GEP successfully reduced the computation time required for logic regression, but the overall ability of MLR-GEP in identifying SNP interactions falls short of expectation. In this study, we identified a few areas of development, which can potentially improve the performance of the proposed method.

8

Methods for identifying SNP interactions: A review on variations of Logic Regression, Random Forests and Bayesian logistic regression

Chapter Summary

Although the purpose of this chapter is still to address the second main objective of the thesis, the focus is on the strengths and weaknesses of the machine learning algorithm and model based approaches. In the previous two chapters we introduced model-based and a non-model based approaches for identifying associated SNPs and/or SNP interactions. Although the advantages and disadvantages of each method are documented independently in the previous two chapters, the aim here is to investigate how these methods differ when compared with each other. Moreover, during the literature review of Chapter 5 we noted there are several variants of logic regression, therefore we compare our methods also with a few of these. These variants include logic regression with feature selection (logicFS), Monte Carlo logic regression (MCLR), genetic programming for association study (GPAS) and modified logic regression with gene expression programming (MLR-GEP). Because logic regression has a tree-like feature, we further included another tree-like algorithm, random forest, in this analysis.

Chapter Conclusion

The methods included in this chapter all have their advantages and limitations. Therefore none of the methods is innately superior to the others. However, we observed some common characteristics among the similar methods. For instance, the non-model approaches, namely GPAS and MLRGEP, are the only methods that are capable of dealing with large volumes of SNP data; however, the main drawback of these methods is lack of accuracy and specificity. Among the methods included here, these two have the highest false positive rate. In contrast, the model based approaches all display high accuracy but are limited in the number of SNPs that can be efficiently analysed. However, all these methods are better at identifying SNP interactions than the SNP by SNP approach.

Authorship

Carla C.M. Chen, Paula Macrossan, Kerrie L. Mengersen, Jonathan M. Keith

School of Mathematical Sciences, Queensland University of Technology

Holger Schwender Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 USA.

Robin Nunkesser Collaborative Research Centre 475, TU Dortmund University, Germany

Reference

Chen, C. C.-M., Macrossan, P., Schwender, H., Nunkesser, R., Keith, J., and Mengersen, K. (2010). Methods for identifying SNP interactions: A Review on variation of Logic regression, Random Forests and Bayesian Logistic regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence, submitted*.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of journals or other publications, and the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Digital Thesis database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for the associated publication is:
Chen CC-M, Macrossan P, Schwender H, Nunkesser R, Keith J, Mengersen K (*submitted*) Methods for identifying SNP interactions: A Review on variation of Logic regression, Random Forests and Bayesian Logistic regression.*IEEE/ACM Transactions on Computational Biology and Bioinformatics*

Contributor	**Statement of contribution**
Chen CC-M	conception and conduct the research, write the code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature & Date:	
Macrossan P,	conception, editing, analysis
Schwender H	conception, editing, analysis
Nunkesser R	editing, analysis
Keith J	editing
Mengersen KL	conception, interpretation, editing

Principal Supervisor Confirmation – I have sighted email or other correspondence for all Co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

8.1 Abstract

Due to advancements in computational ability, enhanced technology and a reduction in the price of genotyping, more data are being generated for understanding genetic associations with diseases and disorders. However, with the availability of large data sets comes the inherent challenges of new methods of statistical analysis and modelling. Considering a complex phenotype may be the effect of a combination of multiple loci, various statistical methods have been developed for identifying genetic epistasis effects. Among these methods, logic regression (LR) is an intriguing approach incorporating tree-like structures. Various methods have built on the original LR to improve different aspects of the model. In this study, we review four variations of LR, namely Logic Feature Selection, Monte Carlo Logic Regression, Genetic Programming for Association Studies and Modified Logic Regression-Gene Expression Programming, and investigate the performance of each method using simulated and real genotype data. We contrast these with another tree-like approach, namely Random Forests, and a Bayesian logistic regression with stochastic search variable selection.

8.2 Introduction

Single nucleotide polymorphism (SNP) is the most common genetic variation among individuals and it was estimated that the human genome has approximately 10 million SNPs [158]. With the recent mapping of the human genome [269] came the availability of high throughput laboratory procedures for the identification of SNPs. Strong correlation among blocked SNPs, i.e. linkage disequilibrium, allows scientists to study the association between genetic and phenotypic variation using a subset of SNPs. Genome Wide Association Studies (GWAs) attempt the mapping of SNPs to phenotypic variation among individuals. Such procedures require a sound statistical methodology and associated computational capability to cope with the analysis of a large data set. Most studies are focused on single locus analysis, which directly tests the association between individual SNP and phenotypic variant. The most commonly implemented statistical approach for these studies is a SNP-by-SNP testing algorithm. This procedure requires an additional statistical correction

for the Type 1 error associated with multiple testings. [227] provide summaries of commonly used correction methods, including Bonferroni correction, permutation test and false discovery rate and discuss the benefits and drawbacks of each of these.

Although the SNP-by-SNP approaches are relatively fast and capable of incorporating covariates [e.g. 303], the major limitation of such approaches is the difficulty of detecting possible gene epistasis effects [124], which is often suggested as the reason for lack of success in genetic studies of complex diseases [51]. Although "epistasis" is commonly defined as the interaction of different genes, there is some confusion on the definition of epistasis in the literature owing to the existence of different types of interaction [51]. [50] and [214] provide thorough reviews on different types of epistasis. In this study, we are focused on using statistical methods to identify gene interaction, this is the "statistical epistasis" according to [214].

Various statistical methods that have been developed for searching for epistasis effects in complex diseases include Bayesian epistasis association mapping (BEAM, [302]), multifactor dimensionality reduction [231], Polymorphism Interaction Analysis [191], logic regression [235], Bayesian model selection [90] and a two stage approach that firstly selects SNPs with strong marginal effects, then identifies interactions among the SNPs [192]. [118] provide an overview and evaluation of the performance of five widely applied methods in detecting interaction effects. One of these, logic regression [LR, 235], is a hybrid method that has the structure of a generalized regression method but with a Boolean combination of variables as predictors. LR is motivated and developed for a plausible but difficult association pattern between SNPs and phenotype, which often involves using words like "AND","OR" and "NOT". For example, an individual may have a higher chance of having a specific trait when *"the homozygous variant genotype is at SNP* S_1 ***AND*** *the homozygous reference genotype is at SNP* S_2 ***OR*** *both SNP* S_3 ***AND*** S_4 *are* ***NOT*** *of the homozygous reference genotype"*

LR has been widely applied in the analysis of SNP data for various phenotypes including sporadic breast cancer [91, 245], trachoma [14], bladder cancer [8], renin-angiotensin [151] and myocardial infarction [151]. [244] indicates that LR is more preferred when compared with other tree-based approaches, such as Random Forests [RF, 33] and Classification and Regression Trees [CART, 34].

Although LR was initially developed for prediction, its capability has been extended through algorithms such as logic Feature Selection [logicFS, 245], Monte Carlo Logic regression [MCLR, 153] and Full Bayesian logic regression [FBLR, 91].

Another extension to the original LR involves variations in the searching algorithm. [153] pointed out two drawbacks with the simulated annealing algorithm implemented in original LR. Firstly, it identifies a single best model which potentially neglects competing models. Secondly, simulated annealing is not geared for the identification of SNPs in linkage disequilibrium (LD). Although the latter limitation has not yet been resolved, the former limitation can potentially be resolved by using different searching algorithms. Methods such as Reversible Jump MCMC [104], Genetic Programming for Association Studies [GPAS, 209] and Gene Expression Programming [MLR-GEP, 174] have a framework similar to logic regression but implement different searching algorithms.

The aim of this paper is to summarise these variations of LR for a case-control study and compare the performance of the methods using simple simulated examples. Due to the fact that LR is a tree-based algorithm, we also consider Random Forests [33] in this paper. Furthermore, we compare the methods with a Bayesian logistic model. Therefore, the methods included in this study include logicFS, MCLR, GPAS, MLR-GEP, RF and Bayesian logistic regression.

8.3 Methods

Logic regression Before introducing LR, it is important to note how SNPs may be coded in LR. Let allele 'A' be a disease allele; that is, having allele 'A' increases the probability of expressing a certain phenotype. Typically the SNP is coded as 0, 1, 2 which corresponds to genotypes 'aa', 'aA' and 'AA'. Alternatively, the SNPs may be coded as a binary variable, which represents the dominant and recessive effect, for instance, genotype 'Aa' or 'AA' at SNP S may be coded as $S_{i,1}$ and genotype 'AA' as $S_{i,2}$.

LR was initially developed for classification and regression, which aims to find Boolean combinations that enhance the prediction of the model. The LR thus comprises Boolean combinators such as AND- and OR-,

and variables, i.e SNPs, in a logic expression, L. Using the same example as in the Introduction, L is then

$$L = (S_{1,2} \wedge S_{2,1}^{C}) \vee (S_{3,1} \wedge S_{4,1}) \tag{8.1}$$

where $\wedge$ and $\vee$ denote the AND and the OR operator, respectively, and ^{C} denotes the complement of a boolean variable.

Logic expressions can be structured into a tree representation which is referred to as a logic tree. The terminology of the logic tree is very similar to that used in CART, although the trees of LR and CART are different structurally, as discussed later in this paper. A node is a point on the tree structure where a split occurs. In LR, a node represents one of the Boolean operators (AND-) and (OR-), and each leaf corresponds to one of the variables (SNPs). Figure 8.1 is an example of a logic tree of LR, and with a logic expression given by

$$L = (S_{1,2} \wedge S_{48,1}^{c}) \vee (S_{18,1} \vee (S_{37,2} \wedge S_{25,2})). \tag{8.2}$$

Here, the leaves include the dominant effect of SNP 18; and the recessive effect of SNP 1, 37 and 25. SNP 48 is highlighted in dark shade, representing the complement of SNP 48 (i.e NOT ($SNP_{48,1}$)).

When the number of SNPs increases, searching among all possible logic trees/expressions becomes unmanageable. This motivates the implementation of a stochastic searching algorithm. The simulated annealing algorithm proposed by [235] and [146] starts with a logic tree, L_1, consisting of randomly selected variables. At each iteration s, a new logic expression, L_{new} is proposed by randomly selecting one of six possible moves: alternate a leaf, alternate an operator, grow a branch, prune a branch, split a leaf or delete a leaf. Each move is assigned with a pre-specified probability, and not all moves are permissible at an iteration. For instance, when the maximum size of the tree is reached, moves which result in adding a leaf/leaves are prohibited. The acceptance of L_{new} depends upon the acceptance probability, given by

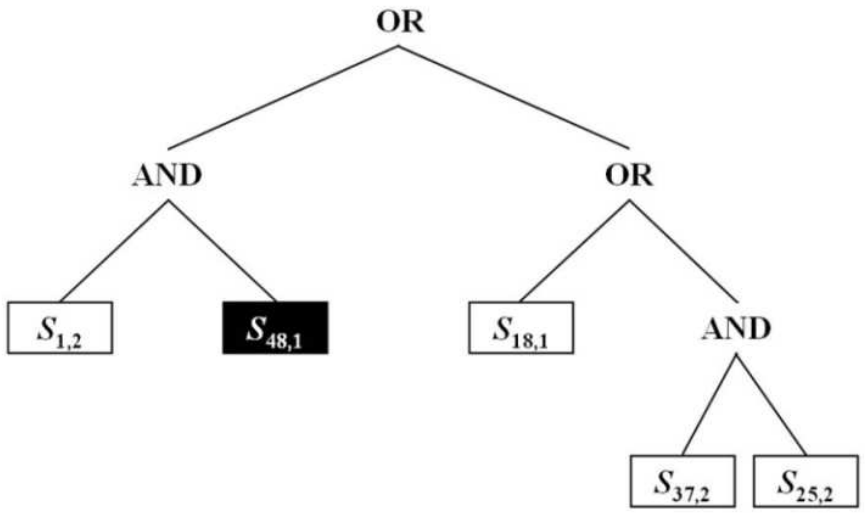

Figure 8.1: An example of a logic tree of LR.

$$a(MCR_s, MCR_{new}, T) = \min\left\{1, \frac{\exp(MCR_s - MCR_{new})}{T}\right\} \tag{8.3}$$

where MCR_s is the misclassification rate of the tree s and T denotes the 'temperature', which decreases with the duration of the annealing process. Thus, the acceptance rate of a new logic tree is much higher at the beginning of the process (when T is large) and eventually becomes almost zero at the end of the search.

For more complicated problems, multiple trees can be combined using a generalized linear model

$$g(y) = \beta_0 + \sum_{q=1}^{Q} \beta_q L_q \tag{8.4}$$

where $g(.)$ is a link function, β_0 is the intercept, $\beta_q, q = 1, \ldots, Q$, is the coefficient of the tree L_q, and Q is the maximum number of trees allowed. Using such a format increases the versatility of LR for the analysis of different types of phenotypes [235] and can be easily modified for more complicated models such as the Cox proportional hazards model.

Monte Carlo Logic Regression [153] proposed that instead of selecting a single optimal model, it is preferable to identify various competing models and combinations of covariates that are potentially associated with the phenotype. Their method incorporates Bayesian model selection techniques using Markov Chain Monte Carlo to explore a large number of models. Therefore, the model is called Monte Carlo Logic Regression (MCLR).

The main difference between MCLR and LR is in the use of priors and the searching algorithm. MCLR requires specification of a prior on the model size. The model size is defined as $\sum_{q=1}^{Q} |L_q|$, where $|L_q|$ is number of terminal nodes of the tree q. Because the model parameters of Equation 8.4 are not essential for detecting the SNP interaction, [153] adapted the maximum likelihood approaches for parameter estimation instead of using a fully Bayesian approach.

Compared with LR, the searching algorithm of MCLR is more complicated as it uses Reversible Jump MCMC [RJMCMC, 104]. At each iteration, a logic tree is selected at random and modified using the same moves as the LR. Once a new model is selected, the acceptance of the new model will depend upon the prior, posterior and likelihood ratio as described in [104].

Like other MCMC methods, a large number of iterations is required to ensure the convergence of a MCMC chain. The importance of SNPs and SNP interactions is determined from the post burn-in samples, i.e. samples after the chain has converged. For instance, the importance of a two-way SNP interaction is defined as the frequency of the pair of SNPs found in the same logic tree over all post burn-in models. The same paradigm is used for finding the interactions of three variables.

Logic Feature Selection (logicFS) LogicFS is more closely related to LR in that it follows the same paradigm as the LR and uses simulated annealing as the searching algorithm. However, instead of seeing them as two separate methods, logicFS improves the variable selection of LR by repetitively fitting logic regression models to different bootstrap samples. This is achieved by employing bagging [32] with the base learner LR.

LogicFS draws a bootstrap sample from the original samples, i.e. n samples are randomly drawn with

replacement from the original samples, and then applies logic regression to the bootstrap sample. This process is repeated several times (typically 50-100 times). LogicFS also improves the interpretation of the logic expression by transforming the expression into a disjunctive normal form (DNF). This makes the SNP interactions directly identifiable. For example, assume a standard logic expression

$$L = (S_{1,1} \wedge S_{2,1}^c) \vee (S_{3,2} \vee S_{4,2}) \wedge S_{5,1}^c, \tag{8.5}$$

of an original LR. L is then transformed into a DNF, which becomes

$$L = (S_{1,1} \wedge S_{2,1}^c) \vee (S_{3,2} \wedge S_{5,1}^c) \vee (S_{4,2} \wedge S_{5,1}^c), \tag{8.6}$$

Compared with Equation 8.5, the identification of interactions is much easier in Equation 8.6. The two way SNP interactions are SNPs connected by 'AND' operators, which are $S_{1,1}$ AND $S_{2,1}^c$, $S_{3,2}$ AND $S_{5,1}^c$, and $S_{4,2}$ AND $S_{5,1}^c$. This representation can then be used to estimate the importance of any interactions based on its predictability, which is essential for distinguishing a 'real' influential interaction from noise. Moreover, transforming the logic expression into a DNF pools the AND-combination and makes some variables redundant. For example, if both $S_{1,1} \wedge S_{2,1} \wedge S_{3,1}$ and $S_{1,1} \wedge S_{2,1} \wedge S_{3,1}^c$ are in the logic expression, logicFS shortens the logic expression by removing $S_{3,1}$ and the expression becomes $S_{1,1} \wedge S_{2,1}$.

The importance of each interaction is estimated using the out-of-bag (OOB) approach, which is similar to that used in Random Forests. During each iteration, about 60-65% of the subjects are drawn to become the bootstrap samples for the construction of a logic tree. The remaining subjects which are not included in the construction are called out-of-bag (OOB) samples. In the case-control study, the importance of an interaction P is estimated as the value of the variable importance measure (VIM) which is the average difference in the misclassification rate of OOB samples with and without the interaction P in the logic regression model over all iterations of logicFS, i.e.

$$\mathrm{VIM}_{\mathrm{single}} = \frac{1}{b}\left(\sum_{b:P\in I_b}(N_b - N_b^-) + \sum_{b:P\notin I_b}(N_b^+ - N_b^-)\right) \tag{8.7}$$

where I_b is a set of all interactions identified in the bth iteration, $b = 1,\ldots,B$, N_b is the number of OOB samples that are correctly classified with P in the model and N_b^- is the number of OOB samples that are correctly classified without P in the logic expression. Similarly, N_b^+ is the number of OOB samples that are correctly classified when P is added to the logic expression when P was not originally included in the expression.

Genetic Programming for Association Studies (GPAS) Genetic Programming for Association Studies [GPAS, 209] is, as the name suggests, a genetic programming [GP, 154] approach for genome-wide association studies. Unlike all methods discussed so far, GPAS does not require the fitting of Equation 8.4, but directly searches for logic expressions in DNF using the GP method.

Figure 8.2 is an example of an individual (tree) in GPAS. Although there are some similarities between Figures 8.1 and 8.2, these are essentially quite different. Firstly, in contrast to other methods, variables in GPAS can be polytomous. Thus SNPs can be coded as 0, 1 and 2, and consequently, when applied to GWAs, it is not necessary to recode the genotypes.

Because GPAS is based on the concept of genetic programming, the terminology used in this approach is more aligned with biological evolutionary terminology than that of LR. For example, the logic tree of the LR is referred to as an "individual" in GPAS and the combination of many individuals becomes a "population". Moreover, the "literal" of GPAS is the same as a leaf of a tree in LR, and a "monomial" refers to a case where two or more literals are connected with an AND-operator, which is similar to the interaction of two SNPs. For example, there are five literals and two monomials in Figure 8.2. For the consistency of this paper, we converted the GPAS terminology into comparable terms of the LR.

Like other searching methods, GPAS is also an iterative approach. The algorithm starts with a random population of two individuals, each consisting of randomly selected SNPs. A new set of individuals is generated as candidates for the next iteration (or so-called generation). These candidates are generated in

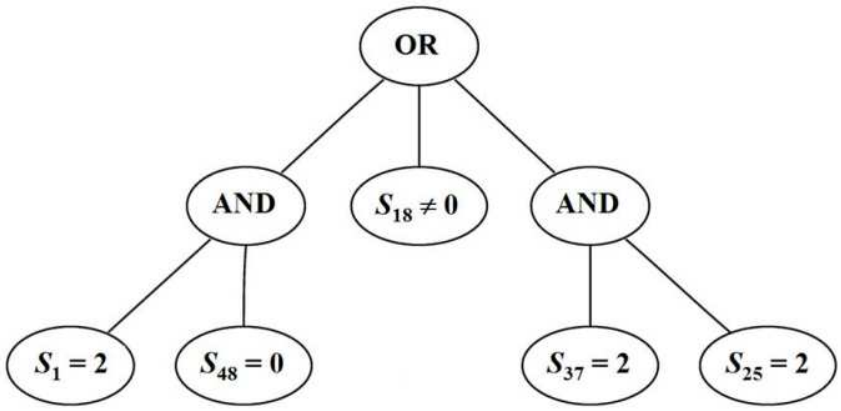

Figure 8.2: An example of an individual in the GPAS algorithm. There are 5 literals and two monomials. $S_1 = 2$ indicated SNP 1 is AA (or aa, depending on user's preference), and it is called a literal. An example of a monomial is $S_1 = 2$ AND $S_{48} = 0$.

three different ways. Firstly, all individuals of the current generation automatically become candidates for the next generation. Secondly, two individuals are randomly selected from the population and a 'crossover' is performed by randomly selecting a part of an individual (namly monomial) and attaching the selected part to the other individual to form a new individual. Thirdly, five different moves (mutation or alteration) are applied to randomly selected individuals. The moves (mutation) in GPAS include inserting a literal (adding a SNP), deleting a literal (removing a SNP), replacing a literal with another literal, inserting a new monomial (adding a new "AND" combination) and deleting a monomial (deleting a SNP_x AND SNP_y). These additions and deletions are performed at random, meaning that the locations of deletion/insertion are chosen at random and items to be inserted are also chosen at random.

After having generated a pool of candidates, a set of individuals is then selected from the pool to form the next generation. The selection criterion used in GPAS is called 'fitness', which aims to balance the number of correct classifications (NCR) of both cases and controls and to also penalize the size of the classifier, s. The fitness of the GPAS tree in the ith iteration of GPAS is expressed as a set of objectives

$$\text{fitness}_i = (\text{NCR}_i^{\text{Cases}}, \text{NCR}_i^{\text{Controls}}, s_i). \tag{8.8}$$

An individual is said to be *dominant* to others if at least one of the objectives is superior and none of the objectives is inferior. Only the dominant individuals are then selected for the next generation. This selection process is called *domination selection* [209]. The iteration repeats until either the number of generations reaches the predetermined number of generations, or the desired fitness level is achieved.

The size of an individual is restricted in GPAS, although it is possible to have more monomials in an individual. [209] limited the individual to only one monomial.

Modified Logic Regression - Gene Expression Programming (MLR-GEP) Although MLR-GEP [174] is based on LR, it is actually more closely aligned with GPAS. Since MLR-GEP has the aim of identifying SNP interactions, the model parameters of Equation 8.4 are considered to be less relevant and are thus ignored. The advantage of this approach is it increases the computational efficiency, thereby making it more capable of accommodating the computational burden of GWAs. Using the same notation as earlier, the MLR model becomes:

$$g(y) = \sum_{i=1}^{K} L_i \tag{8.9}$$

where $g(.)$ is a link function. For a case-control study, the most commonly used link is logit. The stochastic searching algorithm used in MLR-GEP is the Gene Expression Programming [GEP, 84], which is a hybrid of genetic algorithms [GA, 125] and genetic programming [GP, 154].

The terminologies of GPAS and MLR-GEP are interchangeable with a key difference in the definition of an “individual”. In GA, individuals are linear strings with fixed length, whereas in GP, individuals are non-linear objects with different sizes and shapes. GEP combines the features of individuals of GP and GA, leading to individuals of GEP encoded as strings with fixed length, which can be later expressed as non-linear objects with different shapes and sizes. Therefore, GEP has the advantages of both GA and GP, with

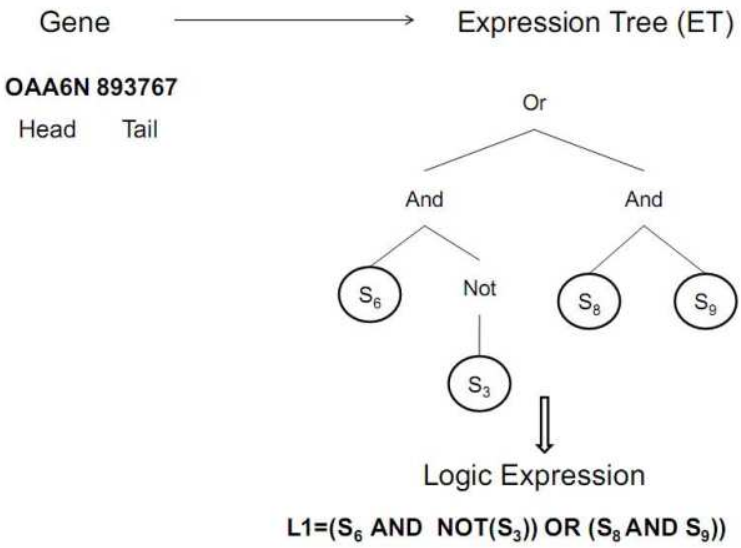

Figure 8.3: An example of an individual in MLR-GEP, showing the translation of single string to an object of shape and size. The length of the gene is fixed, therefore node 767 at the end of the gene tail is redundant.

the ease of manipulation of GA and the functional complexity of GP.

The linear string in GEP is referred to as a 'gene', and a gene is composed of 'nodes' representing either functions (i.e. Boolean- AND, OR and NOT) or 'terminals' (i.e SNPs). A number of genes can be linked by functions to form a 'chromosome'. The structure of the GEP gene is divided into a 'head' and a 'tail' (Figure 8.3). The head contains both functions and terminals, whereas the tail contains only terminals. The first head node of each gene, or 'root' node, must be a function. The tail length is a fixed function of the head length and the maximum function arity (number of function arguments). The structure of the GEP gene and the translation system from a fixed length string to an expression tree guarantees that all modifications arising from evolution of the individuals result in syntactically correct expression trees (ETs). Despite the fixed length of the GEP genes, they have the potential to code for ETs of widely differing shapes and sizes. The number and length of GEP genes is peculiar to the problem at hand.

The moves (also called mutations or genetic operations) of MLR-GEP can take place at genes and chromosomes, and include mutation, transposition, insertion of sequence, root insertion of sequence and recombi-

nation. Mutation is a change occurring in a single node of a gene and can occur at both the head and tail of a gene. When it occurs in the gene head (other than at the root node) it may produce either a function or terminal, whereas tail mutation must result in a terminal. Transposable elements of GEP are fragments of the genome that can relocate to another place in the chromosome. Insertion Sequence (IS) elements are short fragments with a function or terminal in the first position that may transpose to the head of genes except the root. Root Insertion Sequence (RIS) elements are short fragments with a function in the first position, and which transpose to the root of genes. In addition, an entire gene may transpose to the beginning of the chromosome (gene transposition). Recombination in GEP is similar to crossover in GPAS. It may take one of three forms. In all cases, two parent chromosomes are randomly chosen and paired to exchange 'genetic' material. During one-point recombination, two parent chromosomes cross over at a randomly chosen point to form two daughter chromosomes. During two-point recombination, two parent chromosomes exchange the fragment contained between two randomly chosen points to form two daughter chromosomes. In gene recombination, an entire gene is exchanged during crossover. 'Elitism', or the survival and cloning of the best individual chromosome in each generation into the next generation, is practised.

Like GPAS, GEP individuals are selected according to their fitness. In contrast to GPAS, the fitness here is defined as the ability of the solution to predict the case/control status of each datum. This is the same as the correct classification. For any GEP individual i, the fitness is

$$\text{fitness}_i = \sum_{j=1}^{J} (c_{ij} = T_j) \tag{8.10}$$

where J is the number of subjects j in the data set, T_j is the case/control status for the subject j, and c_{ij} is the predicted case/control status under GEP individual i for subject j.

Like GPAS, MLR-GEP starts with randomly generated individuals (not limited to two), then evaluates the fitness of all individuals. Each individual is altered with one of the moves described earlier. The fitness of altered individuals is evaluated. Individuals with reasonable fitness then evolve into the next generation. Like GPAS, the process continues until a pre-determined number of generations is achieved, or until a desired fitness is achieved. Finally, the interactions of SNPs are identified from the surviving expressions

where SNPs are connected by Boolean operator 'AND'.

Random Forests Random Forests [RF, 33] is a method which involves a collection of numerous classification or regression trees [CART, 34]. CART is a simple statistical tool applying recursive binary partitioning of the feature space. CART is well known for its efficiency in coping with large data sets. However, as the data become noisier, and less information is contained in each variable, the predictive ability of CART diminishes. RF overcomes this problem by introducing random elements into the model by which subsets of variables are chosen at random and bootstrap samples are selected with replacement for tree growing.

Although the Boolean operators are not physically present in the actual CART structure, the CART tree can be translated into a combination of SNPs, AND- and OR- operators. For example. Figure 8.4 is an example of a classification tree. Following the far right path of this figure, it is equivalent to "when an individual has genotype AA at SNP 7 and genotype AA at SNP 8, this individual is more likely to have the phenotype". Moreover, in contrast to LR, CART trees aim to predict both affected and non-affected individuals. Because variables of RF can have more than two levels, the coding of SNP can remain in the original genotype forms, i.e. 'aa', 'aA' and 'AA'.

A binary split is denoted as a node, and is defined as a parent or a child. For instance, in Figure 8.4, SNPs 1 and 8 are the children nodes of SNP7. A leaf is where the splitting terminates (also called terminal nodes). The training dataset is first split into two subsets using the criteria which resulted in the lowest misclassification rate, i.e. genotype 'aa' at SNP7 in the example tree shown in Figure 8.4. The binary splitting continues until the child nodes have a reasonable level of homogeneity, or the sample sizes (n) of the child nodes are smaller than a prespecified value. In the standard CART, the trees are required to be pruned/shrunk to avoid overfitting; however, this is not required in RF.

The error rate of RF depends on the correlation between any two trees in the forest and the strength of individual trees. Higher correlation between trees in the forest results in a higher error rate, and greater strength of trees reduces the error rate. These two indicators are affected by the size of the subset of variables used in tree building. Reducing the size of the subset also reduces both correlation and strength. The optimal size of the subset is not directly estimated from the data, but determined by users [33].

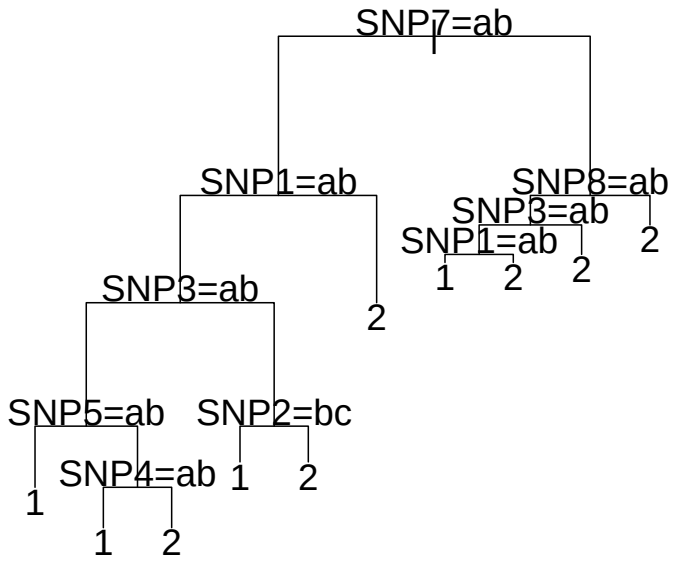

Figure 8.4: An example of a classification tree in RF, where 1 and 2 are the disease status. This tree contains 10 terminal nodes and 9 binary splits. Code *a*, *b* and *c* represent genotype *aa*, *aA* and *AA*.

The prediction error of RF is estimated using out-of-bag (OOB) samples, which are the same as described in LogidFS. At each bootstrap iteration, the prediction of OOB samples is estimated from the tree grown in that iteration. The OOB error is the average of the ratio of the number of times that OOB cases are misclassified to the number of times the respective case is an OOB sample, across the entire forest.

RF provides a variable importance ranking via the variable predictive importance, which is estimated also using the OOB cases. The importance of variable j is estimated as the average difference between the correct classification rate of OOB cases, and the correct classification rate of OOB cases with the value of the variable of interest (j, in our example) replaced with a randomly permuted value over all trees.

Variables, j and k, say, are defined here as interacting if, when one variable is used for a split, the other

variable is systematically more or less likely to be used for another split. The measurement used for the interaction importance ranking is the gini index. The gini value is calculated and ranked for each tree and each pair of variables within the tree. The absolute difference between the rank of the tree and the rank of a pair of variables is the gini measure for that pair of variables, which is then averaged across the forest.

Bayesian logistic regression with stochastic search variable selection (BV) The last method included in this paper is different from methods described so far. This model does not have a tree-like structure, but is instead based on logistic regression of a dichotomous phenotype [185] in conjunction with a stochastic search algorithm for variable selection. Stochastic search variable selection (SSVS) using MCMC [95, 96] is a commonly used model for variable selection in the Bayesian framework. The earliest implementation of this model for genetic research was for the identification of multiple quantitative trait loci for complex traits [297]. Similar methods have also been applied to SNP data [49, 90]. BV is different from the Bayesian epistasis association mapping (BEAM) proposed by Zhang and Liu (2007), which detects epistasis effects by applying a Bayesian method to partition the markers into three groups: markers unlinked to the disease risk, markers contributing independently to the disease risk and markers jointly influencing the disease risk, and then confirms the association using a frequentist approach. In contrast, BV assumes both independent and epistasis SNP effects can be modeled in a linear framework. Letting Y_i denote the phenotype of individual i and q_i be the probability of individual i having the phenotype, the typical logistic model is

$$\log\left(\frac{q_i}{1-q_i}\right) = \mu + \sum_{s=1}^{n_s} \nu_s x_{is} + \varepsilon_i \tag{8.11}$$

where μ is the population mean, x_{is} is the genotype of SNP s for individual i, ν_s is the coefficient of x_{is} and n_s is the total number of SNPs. Instead of using SSVS proposed in [95], we implement a variation of SSVS, which is more closely aligned with the one discussed in [49]. Let z_s be a latent indicator variable, where $z_s = 0$ indicates that SNP s is not in the model, conversely, $z_s = 1$ indicates that SNP s is included in the model. Assuming that genotypes are diallelic, $x_s \in \{0, 1, 2\}$, the model then becomes

$$\log(\frac{q_i}{1-q_i}) = \mu + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{l=2} \nu_{s_l} g_{is_l} + \varepsilon_i \tag{8.12}$$

where g_{is_l} is an indicator variable taking the value of 0 or 1 depending on whether individual i has genotype l at SNP s. The parameter ν_{sl} is the contribution of genotype l at SNP s to the expression of the phenotype and ε_i is the residual. This single model can be easily built upon to incorporate two-way interaction effects, so that

$$\log\left(\frac{q_i}{1-q_i}\right) = \mu + \sum_{s=1}^{n_s} z_s \sum_{l=0}^{2} \nu_{s_l} g_{is_l} + \sum_{j=1}^{n_s} \sum_{k=1, j\neq k}^{n_s} \eta_{jk} \sum_{l_j=0}^{2} \sum_{l_k=0}^{2} \gamma_{jl_jkl_k} g_{ijl_jkl_k} + \varepsilon_i \tag{8.13}$$

where η_{jk} is an indicator variable, with $\eta_{jk} = 1$ if the SNP $j \times k$ is included in the model, else 0. The parameter $\gamma_{jl_jkl_k}$ is the contribution due to the interaction between genotype l of SNP j and genotype l of SNP k. Similarly, $g_{ijl_jkl_k}$ is an indicator variable taking the value of 0 or 1 depending on whether individual i has genotype l at SNP l_j and genotype l_k at SNP k.

The importance of SNP s is measured as the number of times that SNP s is included in the iterations after burn-in over the total number of post burn-in iterations. The importance measure is thus confined between 0 and 1. The importance of SNP interactions is also estimated following the same paradigm.

In the following examples, we used non-informative priors for all parameters, as follows:

$$\begin{aligned} \varepsilon &\sim \text{Normal}(0, \tau^{-1}); & \tau &\sim \text{InverseGamma}(0.05, 0.05); \\ z &\sim \text{Bernuolli}(p_z); & \mu, \nu, \gamma &\sim \text{Normal}(0, 1); \\ \eta &\sim \text{Bernuolli}(p_\eta); & p_z, p_\eta &\sim \text{Uniform}(0, 1) \end{aligned} \tag{8.14}$$

Model parameters were estimated using a Gibbs sampling algorithm. With the exception of z and η, all parameters have non-standard conditional distributions, so a slice sampler [204] was used. The estimation of z and η was based on a combination of Gibbs and Metropolis-Hasting algorithms [46]. At each MCMC iteration, the value of z and η depend on the ratio of the conditional posterior probability of the model including and excluding a SNP. For example, if the condition posterior probability of the model with SNP i is larger than the model without SNP i if the ratio exceeds a random value drawn uniformly between 0 and 1, then z_i is assigned with value 1, else 0.

Ten independent chains were generated with 100,000 iterations each. The first half of the iterations of each of the chains were treated as the burn-in and the variable importance measures were derived from the last 50,000 samples, that is the number of times the SNP or the SNP interaction is included in the model at each of the remaining 50,000 iterations. The convergence of MCMC chains was assessed by comparing the model likelihoods of different simulation sequences, all of which started from different points.

Data

We use two data sets to evaluate the performance of the six methods described in the previous sections. These comprised a simulated dataset and a real data set obtained from the GENICA study [139]..

Simulated Data

For each of these fifty data sets, 500 cases and 500 controls are generated so that for each case exactly one of the conjunctions $P_1, \ldots, P_4$, summarized in Table 8.1, is true, and none of these conjunctions is true for any of the controls. Thus, employing the logic expression

$$L = P_1 \vee P_2 \vee P_3 \vee P_4$$

as classification rule leads to a correct classification of all 500 cases and 500 controls in each of the 50 data sets. Apart from the values of the informative SNPs, i.e. the SNPs forming $P_1, \ldots, P_4$, the genotypes of

Table 8.1: The four conjunctions $P_1, \ldots, P_4$ used in the first simulation. These represent SNP interactions responsible for the presence of the phenotype. The number of cases simulated for each conjunction and the proportion of the observations described by each of these conjunctions are summarized in the third and fourth column. The last row indicates the number of controls included in the data set, which made up half of the total population.

Conjunction	Interaction	Number of Cases (Controls)	Proportion of Data
P_1	$S_{1,2}$	100(0)	10%
P_2	$S^c_{2,1}$ and $S_{3,2}$	150(0)	15%
P_3	$S_{4,2}$ and $S_{5,2}$ and $S_{6,2}$	100 (0)	10%
P_4	$S_{7,2}$ and $S_{8,2}$	150(0)	15%
No	None	0 (500)	50%

the non-informative SNPs are randomly drawn with a minor allele frequency randomly selected in the range from 0.2 to 0.4.

Similar methods of simulation were also implemented by [246] and [209].

Real Data: GENICA

The GENICA study is an age-matched and population based case-control study that has been carried out by the Interdisciplinary Study Group on Gene ENvironment Interaction and Breast CAncer in Germany, a joint initiative of researchers dedicated to the identification of genetic and environmental factors associated with sporadic breast cancer. Further details on the GENICA study, such as data collection and cleaning, are in [139].

In this paper, we focus on a subset of the genotype data from the GENICA study. More precisely, data of 1,234 women (609 cases and 625 controls) and 39 SNPs belonging to the estrogen, the DNA repair, or the control of cell cycle pathways are considered in the analyses.

Because a few of the women show a large number of missing genotypes, all observations with more than three missing values are removed from the analysis leading to a total of 1,199 women (including 592 cases and 607 controls). The remaining missing genotypes are imputed by a weighted *k* nearest neighbours approach described in [246] and implemented in the R package *scrime*.

8.4 Results

Table 8.2 provides a parallel comparison of features of all the methods included in this study. The comparison is mainly focused on the difference in structure of the methods, genetic implementation, alterations allowed from one state to another and tree structures. Among all methods, even though the structure of RF and BV does not directly utilize boolean operators, the tree of RF can potentially be interpreted as a combination of 'OR', 'AND' and SNPs, while the addition (+) of BV is similar to 'AND'.

To prevent a local maximum, all methods required adaptation of some form. For logicFS, GPA and MLR-GEP, this is achieved by repeating the analysis a number of times. For methods utilising a form of MCMC (MCLR and BV), this is done by using multiple chains. RF avoids a local maximum by generating multiple trees in the forest and basing inferences on the results of the forest. We present here the results after these types of repetition, i.e. after applying each of the approaches once to each of the fifty simulated data sets, and fifty times to the GENICA data sets with different starting points of the search.

In the simulated dataset, although the methods compared here are somewhat different, except for the RF, all other methods are able to identify at least some of the pre-specified SNPs. Among six methods compared in this paper, only logicFS, MCLR, RF and BV provide rankings for the variable importance. Of all these methods, logicFS most successfully identifies all four SNP interactions in each of the fifty data sets with relatively large importance (usually, shown in the Top 4 rankings). For MCLR, only one of the four interactions is always detected, namely P_2. The other interactions, P_1, P_3 and P_4, are identified in 90%, 50% and 80% of the fifty samples respectively. RF, on the other hand, did not identify any of these conjunctions in its interaction rankings. However, when considering individual SNPs separately, SNPs involved in the interactions all appeared with high rankings.

After 50,000 iterations, BV is able to identify two-way interactions, namely P_2 and P_4, in all fifty data sets. Because the BV model we used here is designed for detecting only the main and/or two-way interaction effects, it is not possible to identify the three-way interaction (P_3 of Table 8.1). However, the effects of the three-way interaction can be identified by BV as subsets of three-way interactions, i.e. S_4 AND S_5, S_4 AND

S_6 and S_5 AND S_6. The conjunction, P_1 on the other hand, is often identified as a part of an interaction effect rather than a solitary effect.

Similar to the results of logicFS, GPAS detects all four SNP interactions explaining the cases in each of the fifty data sets. However many non-related SNPs are also identified.

MLR-GEP is limited in identifying many conjunctions. Of all interactions listed in Table 8.1, the only conjunction consistently identified is when the SNP is a main effect, namely P_1. The conjunction with the second highest chance of detection is P_4 with an average of over 50%; however, the chance of detecting this interaction varies from 10% to 100%. The other two conjunctions, P_2 and P_3, on the other hand were not found under the MLR-GEP approach.

When applying the methods to the GENICA data, except for RF, all other methods identify a probable association of the interaction of ERCC2_18880 and ERCC2_6465 with sporadic breast cancer. These two SNPs are from the Excision Repairs Cross-Complementing group 2 region (ERCC2, formerly XPD). LogicFS, MCLR, GPAS and MLR-GEP all indicate that having the homozygous reference genotype at ERCC2_6465 and either heterozygous or homozygous genotype at ERCC2_18880 is likely to increase the chance of breast cancer. This result is also supported by BV with more detail. According to the results of BV, the highest chance of developing sporadic breast cancer is when individuals show the homozygous genotype at ERCC2_18880 and homozygous reference genotype at ERCC2_6465 with an odds ratio of 4.17 (CI: 2.63-6.67), followed by individuals with heterozygous genotype at ERCC2_18880 and homozygous reference genotype at ERCC2_6465 with an odds ratio of 2.37(CI: 1.01-5.58).

Another interesting finding which is identified only by the BV approach is the functionality of ERCC2_6465. The results of BV show that ERCC2_6465 is potentially associated with the sporadic breast cancer in two different ways, by acting as a solitary additive effect or by interacting with SNPs other than ERCC2_18880.

8.5 Discussion

In this study, we review different variations of logic regression, Random Forest and Bayesian logistic regression with stochastic search variable selection, for their ability to identify SNP interactions. The methods are then discussed and compared using simulated and real datasets.

In the simulated evaluation, because the data are simulated with the conditions closely aligned with logic regression, i.e. using Boolean expression,"AND", "OR", "NOT", it is not surprising that the overall results are better for logic tree-based approaches. GPAS and logicFS both identified all expected SNPs interaction of the simulation data. In contrast, BV is a regression type approach which does not use Boolean operators and the level of interactions between variables is required to be specified prior to analysis (i.e. the current coding of BV was only designed to detect up to two-way interactions). However, considering all these potential constraints, BV showed better results in detecting the conjunctions compared with RF and MLR-GEP.

Among the different methods, the results of the RF analysis of the simulation data are the most unexpected. Although the RF is a tree-based method, it did not identify any conjunctions listed in Table 8.1. However, when considering SNPs at an individual level, these SNPs involved in the interactions were all successfully identified by RF with relatively high importance measures. The same pattern was also found in the results of the analysis of the GENICA data: even though RF did not find the interaction of ERCC2_18880 and ERCC2_6465 to be important, these two SNPs were the top two ranking SNPs when SNPs were considered individually.

These findings reflect the problem with the definition/measurement of interaction importance that is currently implemented in the RF code. The program we used for carrying out the analysis is not the *randomForest* package of R, but the Fortran code available from the author's website[*]. In this version of RF, the importance of a pair of variables is defined as the absolute difference between the ranking of the pair and the ranking of the tree which is averaged across the forest. Although developers of this code stated that "caution" is required for the interpretation of the interaction effects, the results confirm the problem of us-

[*] http://www.stat.berkeley.edu/~breiman/RandomForests/

ing such criteria. This criterion is only useful for detecting the interaction of a pair of SNPs, say A and B, when these two SNPs are often selected jointly in the random selection of the potential predictors used for tree growing. Furthermore, this criterion is easily obscured due to the nature of recursive partitioning embedded in CART. For example, using the dummy example of Figure 8.4, at the root node, the training samples are split into two subgroups, one group with genotype *aa* and *aA* at SNP 7, while the other group has the complement genotype at the same SNP. The further splitting of these two subgroups depends only on the structure embedded within each subgroup, i.e. the splitting which resulted in the most reduction of the impurity measure within that subgroup. Therefore, unless the interaction of SNP A and B is prominent in the subsets, the importance of these two SNP interactions is likely to be overlooked using current criteria.

Although the interaction cannot be identified directly under current settings, the interaction effects are captured by the solitary variable importance measured using the permutation methods and OOB samples. The assertion is confirmed in [169]. Therefore, with some improvements, RF can be a useful tool for identifying SNP interactions. For instance, [135] suggest the use of a sliding window sequential forward feature selection in conjunction with statistical testing to find epistasis effects.

The detection of false informative SNPs is commonly observed across all methods; however, it is difficult to compare the false positive and false negative rates of these methods. GPAS and MLR-GEP identify a set of SNPs showing possible association without giving a quantitative measure, such as variable importance ranking, to show the degree of association between a SNP and disease. In this study, the set of possible models according to GPAS is exponentially large, and without the variable importance ranking, it is more difficult to identify the false informative SNPs in the real data. Despite the fact that the ranking of variable (interaction) importance is available in other methods, an appropriate threshold point for these measures is still not well understood. This is because a threshold point may potentially depend on the underlying genetic model and the ratio of the causal and noise SNPs, which is often impossible to know prior to the analysis (Lunetta et al., 2004). Therefore, instead of basing conclusions on the results of a single method, a more sensible approach is to analyse data with different methods and to compare the results. Further investigation on how to integrate the results of different methods would be beneficial.

Methods incorporating tree-based structures are more robust in identifying the higher order interactions (e.g

3 or more way interactions). In the tree-based methods, higher order interactions are directly identified from a tree or a collection of trees. In contrast, to find higher order interactions using regression models, the order of interactions needs to be specified a priori. Moreover, as the number of terms increases in a regression model, the parameter space increases exponentially and consequently reduces the computational feasibility which is especially difficult in a genome-wide association study.

BV, on the other hand, gives better results for understanding the allele effects on the expression of the phenotype. This information is available from the magnitude of the coefficients of the different terms. For example, the coefficient of g_{sl} gives the relative measure of the effect of the genotype l of the SNP s. BV also provides a quantified measure of the risk of having the phenotype for different genotype combinations at causal loci.

Among all methods, GPAS and MLR-GEP are the only methods capable of coping with the intensity and computational power required for the analysis of large data sets. This is because these algorithms are based on a machine learning algorithm (i.e. GP and GEP). LogicFS and MCLR, on the other hand, are limited to a maximum of 1000 SNPs in the written code. It is noted that BV has been used for finding individual SNP additive effects (but not for two-way or higher interactions) for up to 23,000 SNPs. Unless more effective programming or a fast searching algorithm is adopted, most of the methods described here are only suitable for candidate gene search or fine mapping.

The major drawbacks of GPAS and MLR-GEP are in the accuracy and specificity of the identification of important interactions. Both of these methods implemented a machine learning algorithm, and although fast, the results are less reliable. This problem is especially noticeable in MLR-GEP. The performance of MLR-GEP can be improved in various ways, such as paying greater attention to the parameter setting in the evolutionary process, incorporating model parameters and use of more sophisticated fitness measures [174].

The most relevant genetic questions for such models concern their ability to detect genetic heterogeneity and linkage disequilibrium (LD) SNPs, and the effect of LD SNPs on the model. Of all methods, logicFS is expected to be less capable of identifying any of these effects given that it is highly related to logic regression and has therefore inherited the same shortcomings identified in [153]. However, this problem can arguably

be solved by applying logicFS for several repetitions to several subsets of the data sets thereby identifying a large number of different models.

All other methods potentially have strategies for detecting genetic heterogeneity. Bayesian methods (MCLR and BV) identify heterogeneity from a collection of multiple models [153] and/or the use of various Markov chains [46]. In GPAS and MLR-GEP, by repeating the analysis with different starting populations, the heterogeneities are potentially identifiable from a collection of tree structures. In these two methods, trees are connected by the "OR" operator and the sub-tree therefore represents different possible genetic pathways. Similarly, in RF, genetic heterogeneity can be determined from trees nested within the full tree.

When LD SNPs are in the data sets, Bayesian approaches again have the advantage of multiple chains. When two SNPs are highly correlated, if one SNP is selected in the model, although the chance of the other SNP being selected is very small, it does have an equal chance of being selected in the model. When the number of chains (or models) is large enough, the LD SNPs are identified. In RF, LD SNPs are identified as surrogate variables. However, as noted by [169], correlated SNPs can diminish the variable importance ranking.

Although some of the methods included in this study have the same foundations, they are manifestly different in various facets. Each method has its advantages, and conversely some limitations. Even so, the methods included in this study, in general, are superior in identifying SNPs in which the effect of the SNP is highlighted by the presence of other SNPs. For instance, although the results of the analysis are not included here, we tested the SNP effect of the GENICA data using SNP-by-SNP Fisher's exact test and found the p-value of ERCC2_18880 is far from significant (p-value =0.106, prior to power adjustment).

None of the methods included in this study, exhibits distinct superiority over another. In conclusion, the GPAS and MLR-GEP may be preferred for searching through large dimensional spaces; logicFS, MCLR, RF and BV may be preferred for candidate gene/region searches, and BV may be preferred for providing detail on the allele effects.

Table 8.2: Parallel comparisons of features, genetic implementation, alteration (move) and tree structures of LR, logicFS, MCLR, GPAS, MLR-GEP, RF and BV.

Methods	LR	logicFS	MCLR	GPAS	MLR-GEP	RF	BV
Features							
Model based	y	y	y	n	n*	n	y
Iterative searching Algorithm							
Require (y/n)	y	y	y	y	y	n	y
Algorithm	Simulated Annealing	Simulated Annealing	RJMCMC	Genetic Programming	Gene Expression Programming	NA	MCMC (Gibbs+MH)
Iterative/Evoluationary†	I	I	I	E	E		I
Quantify Interactions	n	y	y	n	n	y	y
Use Boolean	y	y	y	y	y	n‡	n‡
Boolean Operators	AND, OR	AND,OR	AND, OR	AND, OR	AND, OR	OR, AND‡	AND‡
Genetic Implementation							
SNP Coding	R/D§	R/D	R/D	A/F	A/F	A/F	A/F
LD	y¶	y¶	y	y¶	y¶	y	y
Max SNPs	1000	1000	1000	GWAs‖	at least 23,000	*	at least 23,000**

[1]Although it is based on LR, the parameters are ignored
[2]Iterative (I) indicates a state depends immediate previous state only, Evolutionary (E) indicates a state depends previous states.
[3]Strictly, RF and BV do not have Boolean operators, however, the trees of RF can be interpreted as combination of OR and AND. Similarly, the additive of BV model is like AND operator
[4]RD: Recessive/Dominance; A/F: Allele Frequency
[5]Although LD is not directly considered in the method, LD can be detected via runs with different starting points.
[6] [209] stated that GPAS is able to analyse the GWA data, however it is yet to be verified
[7]Considering the additive effect only
*Unclear

Table 8.2: Parallel comparisons of features, genetic implementation, alteration (move) and tree structures of LR, logicFS, MCLR, GPAS, MLR-GEP, RF and BV.

Methods	LR	logicFS	MCLR	GPAS	MLR-GEP	RF	BV
Tree Structure							
Have Tree Structure	y	y	y	y	y	y	n
Boolean*	y	y	y	y	y	n	
Operators	y	y	y	y	y	n	
Node	B	B	B	B	B	S	
Terminal Node	S	S	S	S	S	P	
Binary/Mutiple Split	Binary	Binary	Binary	Multiple	Binary	Binary	
Fitness Measure	MCR	MCR	MCR	Multiple NCR	NCR	OOB MCR	
Moving between States	Acceptance Prob	Acceptance Prob	RJMCMC	Fitness	Fitness	NA	
Alteration							
Allow Alteration†	y	y	y	y	y	n	n‡‡
No. Alterations	6	6	6	7	5	2	
Method of Alteration							
Change SNP‡	√	√	√	√	√		
Change Boolean§	√	√	√		√		
Grow Branch¶	√	√	√	√	√	√	
Prune Branch‖	√	√	√	√	√		
Split leaf	√	√	√	√	√	√	
Delete leaf	√	√	√	√			
Crossover**				√	√		
Insert new split at Root node				√			
Require Pre-setting††	√	√	√		√		

[1] Tree structure
[2] Changes made to the tree of current state
[3] Change SNP with another SNP
[4] Change Boolean with another Boolean
[5] Adding a part to existing tree
[6] Deleting a part of existing tree
[7] Exchange parts between two trees
[8] Strictly, the model does not have these alterations. However, some alterations are equivalent to the addition and deletion embedded in BV.
[9] Need to assign the probability to each alterations prior to analysis
B-Boolean operators, S-SNPs, P-Prediction (case or controls)

9

Conclusions and Future Work

Dissecting the genetics of complex diseases in the human genome is a challenging and somewhat daunting task. In this thesis, we have investigated the problem from a statistical perspective and focused on two main areas of challenge, defining phenotypes and detecting epistasis effects from large scale SNP data.

In the first part of this thesis, we illustrated the effect of phenotyping on subsequent genetic analysis, and demonstrated the effects using four different models, three of which are latent models. Although the results were not overly surprising, they illustrate the sensitivity of genetic analysis to phenotype estimation. This process, however, is often ignored in the current practice of genetic research of complex diseases. It is difficult to determine whether the phenotype derived from a statistical method is accurate, given that the "true"

phenotype is unobservable, and thus can not be easily validated. Therefore, the models were compared using a parsimonious measure. In Chapter 3, three frequentist models were compared using a common parsimonious measure, namely BIC. Under this criterion, GoM is heavily disadvantaged due to the large number of parameters in the model. However, phenotypes estimated using GoM had the highest heritability compared with the two other methods.

In Chapter 4, we compared two different latent models, with the subsequent linkage results being nearly identical. The models are introduced in the Bayesian context and were therefore compared using DIC3, proposed in [43]. DIC3 is well suited for latent models. Because DIC uses the number of effective parameters instead of the number of stated parameters, IRT is not as heavily penalised and had a comparable value of DIC.

In light of the variety of model selection criteria, and difficulties in validating the estimated phenotype, in Chapter 5 we developed two methods for consolidating estimates of different models using Bayesian model averaging as the foundation. These methods show promise in enhancing individuals at the cores of clusters (individuals with/without all symptoms), as well as increasing the fuzziness (ambiguity) of individuals at the borders of the clusters. Consequently, loci with 'true' signals are amplified and the signals of the 'false' loci are reduced. Furthermore, due to the use of Bayesian methods, the uncertainty occurring at the phenotyping level is easily incorporated into the subsequent analysis. This provides some measure of confidence in the findings.

These methods have so far been tested on two models, namely LCA and GoM. The next step is to test on other models and assess the stability and validity of the methods. If uncensored phenotypic data is available, a further stage would be to compare the phenotypes derived using the models developed in chapter 5 with the phenotypes derived using the IHS criteria, then investigate any subsequent variations. In Chapter 2, we noted the limitations of genome-wide linkage analysis in identifying the loci linking to a complex disease. Therefore, for better understanding of the genetics of complex diseases, it may be more beneficial to substitute the linkage analysis with an association study. This, however, cannot be achieved at the current point of time for this particular cohort because the SNP data is unavailable, but the methods developed in Chapter 5 can be used for other types of complex traits, such as schizophrenia and Parkinson's disease,

where the data is freely available.

Although migraine is often considered to be a dichotomous phenotype (affected/not affected), in this thesis, we have decided to treat it as a continuous measure for various reasons. Firstly, when the symptom data was analyzed using LCA and GoM, various goodness-of-fit measures indicated there are more than two clusters in the data. This suggested treating the migraine as dichotomous data may underestimate the true underlying structure of migraine. Secondly, even when analyzing the data using different models and assessing the models with different goodness-of-fit criteria, the results all suggested the optimum number of clusters for the migraine data is four. Because no linkage analysis to date is capable of analyzing data with tetrachotomous phenotypes, the predicted phenotype will require some forms of adjustment before the mapping can be carried out. Here, there are two options, one is to aggregate the number of clusters into two or three clusters and the other is to convert the multinomial phenotype into a continuous measure.

[210] found similar results in their migraine study and they choose the former approach by combing two lower clusters (clusters with lesser prevalence in symptoms) and two higher clusters (clusters with higher prevalence in symptoms) and assigning them the value of 0 and 1 respectively. In the same study, they also compared the effect of collapsing the clusters to two and collapsing the clusters to three, and found little difference in the LOD scores between the two approaches.

From the results of model fitting, we also observed a trend in the symptom prevalence of the four clusters. The individuals in the intermediate clusters have either more or less symptoms than individuals in two extreme clusters and the symptom prevalence for two extreme clusters is zero and nearly 100%. However, there is no dramatic reduction in symptom prevalence between the two intermediate clusters or two higher clusters. Furthermore, because the symptoms of migraine often overlap with other forms of headaches [115], it is possible that migraine is a severe form of headache. Therefore, using a continuous measure to represent the degree of severity in migrainous headache is plausible. We also show that the use of the continuous estimates for migraine does not have a large effect on the subsequent linkage results, which was affirmed by replicating the results of [212].

The major challenge for the multilocus approach in identifying epistasis effects in GWAs is that the possi-

ble SNP combinations are excessively large, and therefore computationally demanding. In Chapter 7, we proposed a machine learning algorithm, namely gene expression programming, to overcome this inherent difficulty. This method has shown some promising ability in improving computational efficiency. We believe that the accuracy of the method can benefit from further development and improvement in the evolutionary and optimization processes. In contrast, Bayesian logistic regression with stochastic search variable selection, as covered in Chapter 6, has demonstrated better accuracy in identifying subsets of important SNPs. The model has been tested on approximately 26,000 SNPs of chromosome 6 for SNP additive effects. The next logical step is to test the approach in a whole genome scale study, i.e. at least 500,000 SNPs.

Detecting epistasis effects using the model described in Chapter 6 is still problematic, at least in its current form. However, such problems may be overcome by improving various aspects of algorithm such as

- parameter estimation

  In the current algorithm, parameters are estimated using MCMC. To improve the speed of parameter estimation, other algorithms such as Approximate Bayesian Computation [21, ABC] and Variational Bayes [137] could be considered. Moreover, instead of a fully Bayesian approach, an empirical Bayes approach has demonstrated efficiency in parameter estimations [293]. [122] suggested a Bayesian inspired penalised maximum likelihood approach to overcome the computational burden; that is, instead of using the MCMC approach, the EM algorithm is used for optimisation. Similar methods are also proposed by [293] and [296] for estimating epistasis effects of Quantitative Trait Loci (QTL).

- sampling distribution of the SNP coefficient

  When a SNP is not included in the model, the coefficient of the SNP is currently sampled from the prior distributions. Alternatively, it could be directly assigned a value of zero, which may potentially reduce the computation time. This has been implemented by [49] for QTL analysis.

In Chapter 8, we compared various subtypes of logic regression (LR) with random forests (RF) and Bayesian logistic regression with the stochastic search variable selection algorithm. Even though there are different subtypes of logic trees, these methods nevertheless show similar ability in identifying subsets of SNPs and SNP interactions, but with LR being more versatile and better suited for higher order interactions. When

comparing different trees, RF has a similar ability to identify the subset of SNPs; however, in the current setting of the Fortran Code, the ability to detect interaction effects is less effective. Two other machine learning algorithms, namely GPAS and MLE-GEP, are the only algorithms capable of coping with large scale GWAs. However, the reduced accuracy of these algorithms makes them less preferable.

Although the data simulation procedure implemented in Chapter 7 is the same as the procedure implemented in [245], it does not reflect the reality of the complexity in the genetic data. For example, the simulation assumes a full penetrance of the disease and it does not take into account factors such as disease prevalence, LD structure and recombination fraction between genes. However, when comparing MLR-GEP with other types of logic tree, RF and Bayesian logistic regression with stochastic search algorithm on a common small scale dataset (GENICA), it is able to identify the same set of interactions as other methods (as shown in Chapter 8). However, this is only limited to a small scale study. The performance of MLR-GEP is less than satisfactory, and requires substantial improvement in its accuracy and requires testing on a more realistically simulated data.

At present the methods introduced and discussed in the second part of this thesis are not satisfactory in detecting the epistasis effects of large scale GWAs and other genetic aspects such as detecting genetic heterogeneity, the effect of linkage disequilibrium (LD) and imputation of missing genotypes. Future work should focus on incorporating these factors into the model and improving the computational efficiency of the model without losing the accuracy of prediction and estimation. This may involve developing a hybrid algorithm that merges the accuracy of the model-based approaches with the efficiency of machine learning algorithms. Moreover, the algorithm may be guided with available knowledge in molecular genetics.

One advantage of the Bayesian framework is the use of priors. When priori knowledge about the parameters is available, it can be easily incorporated in a Bayesian model. For genetic research this can be especially useful because the advancement in molecular genetics can have substantial input into quantitative genetics. For instance, although the gene network for human genome is still far from completion, the knowledge of the interlocking network can help in identifying the epistasis for complex traits. For example, if two genes are known to exhibit functional epistasis from the gene network, this information can be included in the models developed in Chapter 6 by adjusting the prior weights on these genes, so when one gene is selected for

model fitting, the other gene will have a higher probability to be included in the model. This can potentially reduce the computational time required for GWA.

Due to the near completion of the 1000 Genome project, higher quality and more detailed information about the human genome will become available. Thus the need for sound statistical methods with which to examine this information shall remain an integral part of genetic research.

A

Appendix

A.1 Chapter 4

A.1.1 Deviance information criteria for LCA and GoM

Deviance information criteria is the difference between twice the posterior mean deviance and the deviance of estimated η

$$DIC = 2\overline{D(\eta)} - D(\tilde{\eta}) \tag{A.1}$$

In the third DIC proposed by [43] (DIC3) when the likelihood has a closed form, the first term can be approximated using M simulated values, $\eta^{(1)}, ..., \eta^{(M)}$, where $\eta^{(m)} = (p^m, \lambda^m)$ from an MCMC chain.

$$\begin{aligned}\overline{D(\eta)} &= \mathbb{E}_\eta[-2\log f(y|\eta)|y] \\ &\approx -\frac{2}{M}\sum_{m=1}^{M}\log f(y|\eta^{(m)}) \end{aligned} \tag{A.2}$$

The second term of equation 4.3 we used here is the posterior expectation, $\mathbb{E}[f(y|\eta)|y]$ which is also approximated using the parameters of an MCMC chain.

$$\begin{aligned} D(\tilde{\eta}) = -2\log\hat{f}(y) &= -2\log\mathbb{E}_\theta[f(y|\eta)|y] \\ &\approx -2\log\frac{1}{M}\sum_{m=1}^{M} f(y|\eta^{(m)}) \end{aligned} \tag{A.3}$$

From equation A.2 and A.3, equation 4.3 is the expanded form of equation A.1. In the Bayesian LCA model, $f(y|\eta^{(m)})$ is

$$f(y|\lambda^{(m)}, p^{(m)}) = \sum_{k=1}^{K} p_k^{(m)} \prod_i^n \prod_j^J (\lambda_{kj}^{(m)})^{y_{ij}}(1-\lambda_{kj}^{(m)})^{1-y_{ij}}$$

and the posterior mean deviance is

$$\overline{D(p,\lambda)} = -\frac{2}{M}\sum_{m=1}^{M}\log\sum_{k=1}^{K} p_k^{(m)} \prod_i^n \prod_j^J (\lambda_{kj}^{(m)})^{y_{ij}}(1-\lambda_{kj}^{(m)})^{1-y_{ij}}$$

and $D(\hat{\eta})$ is

$$D(\hat{p},\hat{\lambda}) = -2\log\{\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} p_k^{(m)} \prod_i^n \prod_j^J (\lambda_{kj}^{(m)})^{y_{ij}}(1-\lambda_{kj}^{(m)})^{1-y_{ij}}\}.$$

For the Bayesian IRT model, the likelihood is

$$f(y|\theta,a,b) = \prod_i^n \prod_{j=1}^J [\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}]^{yij}[1-\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}]^{1-yij}$$

therefore $\overline{D(\eta)}$ is

$$\overline{D(\theta,a,b)} = -\frac{2}{M}\sum_{m=1}^{M}\log\prod_i^n\prod_{j=1}^J[\frac{e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}{1+e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}]^{yij}[1-\frac{e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}{1+e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}]^{1-yij}$$

and $D(\hat{\eta})$ is

$$D(\hat{\theta},\hat{a},\hat{b}) = -2\log\{\frac{1}{M}\sum_{m=1}^{M}\prod_i^n\prod_{j=1}^J[\frac{e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}{1+e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}]^{yij}[1-\frac{e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}{1+e^{a_j^{(m)}(\theta_i^{(m)}-b_j^{(m)})}}]^{1-yij}\}.$$

A.2 Chapter 5

A.2.1 Symptom description of Migraine data

Table A.1: The IHS diagnostic criteria for migraine without aura (MO).

Item	Description
A	At least five attacks fulfilling B-D
B	Headache attacks lasting 4-72 hours
C	Headache has at least two of the following characteristics:
	Unilateral Locations
	Pulsating quality
	Moderate or severe intensity(inhibits or prohibits daily activities)
	Aggravation by walking stairs or similar routine physical activity
D	During headaches at lease one of the following:
	Nausea and (or) vomiting
	Photophobia and phonophobia

Table A.2: The IHS diagnostic criteria for migraine with aura (MA).

Item	Description
A	Headache fulfilling criteria B-D list in Table A.1
B	At least five attacks fulfilling B-D
C	Aura consisting of at least one of the following but no motor sickness
	Fully reversible visual symptoms including positive features
	(*ie* flicking of lights) and (or) negative features (*ie* loss of vision)
	Fully reversible sensory symptoms including positive (*ie* pins and needles)
	and (or) negative features (*ie* numbness)
	Fully reversible dysphasic speech disturbance
D	At lease two of the following:
	Homonymous visual symptoms and (or) unilateral sensory symptoms
	At lease one of the aura symptom develops gradually over ≥5 minutes
	Each symptoms lasts ≥5 minutes and ≤60 minutes.

A.2.2 Full Symptom description of KPD data

Table A.3: Clinical characteristics of KPD. This is the Kofendred Research Assessment Protocol for testing affected/unaffected status.

Indices	Description
a	Joining/founding cult
b	Fear/discomfort with strangers
c	Dislike of jokes told face to face
d	Obsession with entertainers
e	Humor impairment
f	Fascination with automobiles
g	Aversion to walking
h	Uncommunicative, contentless speech pattern
i	Fiscal irresponsibility
j	Morbid anger/fear/terror concerning rain/snow
k	Reluctance to wear clothing appropriate for subjective temperature
l	Body-image concerns/mild body dysmorphic disorder

A.2.3 Hessian Matrix

LCA The Hessian matrices for both LCA and GoM are derived analytically. The posterior probability for LCA is

$$h(f) = P(\mathbf{Y}|\mathbf{p}, \Lambda)\pi(\mathbf{p}, \Lambda) = \sum_{i}^{n}\sum_{j}^{J} log(\sum_{k}^{K} p_k \lambda_{kj}^{y_{ij}}(1-\lambda_{kj})^{1-y_{ij}}). \tag{A.4}$$

The Hessian matrix is a square matrix of second-order partial derivatives of $h(f)$, and for LCA the Hessian Matrix is

$$\begin{pmatrix} \frac{\partial^2 h(f)}{\partial {p_1}^2} & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial p_1 \partial \lambda_{KJ}} \\ \frac{\partial^2 h(f)}{\partial p_2 \partial p_1} & \frac{\partial^2 h(f)}{\partial p_2^2} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial p_2 \partial \lambda_{KJ}} \\ \frac{\partial^2 h(f)}{\partial \lambda_{11} \partial p_1} & \frac{\partial^2 h(f)}{\partial \lambda_{11} \partial p_2} & \frac{\partial^2 h(f)}{\partial \lambda_{11}^2} & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial \lambda_{11} \partial \lambda_{KJ}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h(f)}{\partial \lambda_{kj} \partial p_1} & \frac{\partial^2 h(f)}{\partial \lambda_{kj} \partial p_2} & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial^2 \lambda_{kj}} & \cdots & \frac{\partial^2 h(f)}{\partial \lambda_{kj} \partial \lambda_{KJ}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h(f)}{\partial \lambda_{KJ} \partial p_1} & \frac{\partial^2 h(f)}{\partial \lambda_{KJ} \partial p_2} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial^2 \lambda_{KJ}} \end{pmatrix}$$

The computation of the second order partial derivatives can be grouped into eight different combinations as shown in Table A.5. With the eight possible combinations, because j are assumed to be independent, the covariance for cases where js are not equal is zero.

Table A.4: Table showing 8 parameter combinations in Hessian matrix

θ_1	θ_1	Combination	Equation
p_{k_1}	p_{k_2}	$k_1 = k_2$	A.5
		$k_1 \neq k_2$	A.6
p_{k_1}	$\lambda_{k_2 j}$	$k_1 = k_2$	A.7
		$k_1 \neq k_2$	A.8
$\lambda_{k_1 j_1}$	$\lambda_{k_2 j_2}$	$k_1 = k_2,\ j_1 = j_2$	A.9
		$k_1 \neq k_2,\ j_1 = j_2$	A.10
		$k_1 = k_2,\ j_1 \neq j_2$	**
		$k_1 \neq k_2,\ j_1 \neq j_2$	**

** indicates the second-order partial derivative for such combination is zero.

When differentiating Equation A.4 w.r.t. p_{k_1} and p_{k_2}, and $k_1 = k_2$, let l denote k_1, then

$$\frac{\partial^2 h(f)}{\partial {p_l}^2} = -\sum_i \sum_j [\frac{\lambda_{lj}^{y_{ij}}(1-\lambda_{lj})^{(1-y_{ij})}}{\sum_k p_k \lambda_{kj}^{y_{ij}} . (1-\lambda_{kj})^{(1-y_{ij})}}]^2 \tag{A.5}$$

.

Similarly, when $k_1 \neq k_2$ then

$$\frac{\partial^2 h(f)}{\partial p_{k_2} \partial p_{k_1}} = -\sum_i \sum_j [\frac{\lambda_{k_1 j}^{y_{ij}}(1-\lambda_{k_1 j})^{(1-y_{ij})} \lambda_{k_2 j}^{y_{ij}}(1-\lambda_{k_2 j})^{(1-y_{ij})}}{[\sum_k p_k \lambda_{kj}^{y_{ij}}(1-\lambda_{kj})^{(1-y_{ij})}]^2}]. \tag{A.6}$$

When differentiating the same equation w.r.t. p_{k_1} and $\lambda_{k_2 j}$, and $k_1 = k_2$, using similar notation as earlier, let l and m denote k_1 and any symptom, then

$$\frac{\partial^2 h(f)}{\partial p_l \partial \lambda_{lm}} = \sum_i \frac{(y_{im}\lambda_{lm}^{(y_{im}-1)} - \lambda_{lm}^{y_{im}})[(1-\lambda_{lm})^{y_{im}}(\sum_k p_k \lambda_{km}^{y_{im}}(1-\lambda_{km})^{(1-y_{im})}) - p_l \lambda_{lm}^{y_{im}}(1-\lambda_{lm})]}{\left[(1-\lambda_{lm})^{y_{im}}[\sum_k p_k \lambda_{km}^{y_{im}}(1-\lambda_{km})^{(1-y_{im})}]\right]^2}. \tag{A.7}$$

.

On the other hand, when $k_1 \neq k_2$ and , $m = 1, \ldots J$, then

$$\frac{\partial^2 h(f)}{\partial p_{k_1} \partial \lambda_{k_2 m}} = -\sum_i \frac{p_{k_2}\lambda_{k_1 m}^{y_{im}}(1-\lambda_{k_1 m})^{(1-y_{im})}[y_{im}\lambda_{k_2 m}^{y_{im}-1} - \lambda_{k_2 m}]}{(1-\lambda_{k_2 m})^{y_{im}}(\sum_k p_k \lambda_{km}^{y_{im}}(1-\lambda_{km})^{(1-y_{im})})^2}. \tag{A.8}$$

The last two combinations are the partial derivative w.r.t the λ parameters, $\lambda_{k_1 j_1}$ and $\lambda_{k_2 j_2}$. As indicated earlier, due to the assumption of independence, $j_1 = j_2$ which is denoted by m. When k_1 is the same as k_2, denote by l then the second order partial derivative w.r.t these parameters becomes,

$$\frac{\partial^2 h(f)}{\partial {\lambda_{lm}}^2} = -\sum_i \Big[\frac{p_l(y_{im}\lambda_{lm}^{y_{im}-1}(y_{im}-\lambda_{lm})]}{(1-\lambda_{lm})^{y_{im}}\sum_k p_k \lambda_{km}^{y_{im}}(1-\lambda_{km})^{(1-y_{im})}}\Big]^2. \tag{A.9}$$

When k_1 and k_2 are not equal, the second order partial derivative becomes,

$$\frac{\partial^2 h(f)}{\partial \lambda_{k_1 m} \partial \lambda_{k2m}} = -\sum_i \Big[\frac{p_{k_1} p_{k_2}(\lambda_{k_1 m}\lambda_{k_2 m})^{(y_{im}-1)}(y_{im}-\lambda_{k_1 m})(y_{im}-\lambda_{k_2 m})}{(1-\lambda_{k_1 m})^{y_{im}}(1-\lambda_{k_2 m})^{y_{im}}[\sum_k p_k \lambda_{km}^{y_{im}}(1-\lambda_{km})^{(1-y_{im})}]^2}\Big]. \tag{A.10}$$

GoM The posterior probability of GoM has the form:

$$h(f) = P(\mathbf{Y}|\mathbf{g}, \Lambda)\pi(\mathbf{g}, \Lambda) = \sum_i^n \sum_j^J log(\sum_k^K g_{ik}\lambda_{kj}^{y_{ij}}(1-\lambda_{kj})^{1-y_{ij}}). \tag{A.11}$$

Because of the large number of parameters in GoM, the Hessian matrix is a $(n * k + k * j) \times (n * k + k * j)$ square matrix,

$$\begin{pmatrix}
\frac{\partial^2 h(f)}{\partial g_{11}^2} & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial g_{11}\partial \lambda_{KJ}} \\
\frac{\partial^2 h(f)}{\partial g_{12}\partial g_{11}} & \frac{\partial^2 h(f)}{\partial g_{12}^2} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial g_{12}\partial \lambda_{KJ}} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
\frac{\partial^2 h(f)}{\partial g_{nK}\partial g_{11}} & \frac{\partial^2 h(f)}{\partial g_{nK}\partial g_{22}} & \cdots & \frac{\partial^2 h(f)}{\partial g_{nK}^2} & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial g_{nK}\partial \lambda_{KJ}} \\
\frac{\partial^2 h(f)}{\partial \lambda_{11}\partial g_{11}} & \frac{\partial^2 h(f)}{\partial \lambda_{11}\partial g_{12}} & \cdots & \frac{\partial^2 h(f)}{\partial \lambda_{11}\partial g_{n2}} & \frac{\partial^2 h(f)}{\partial {\lambda_{11}}^2} & \cdots & \frac{\partial^2 h(f)}{\partial \lambda_{11}\partial \lambda_{KJ}} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 h(f)}{\partial \lambda_{KJ}\partial g_{11}} & \frac{\partial^2 h(f)}{\partial \lambda_{KJ}\partial g_{22}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 h(f)}{\partial^2 \lambda_{KJ}}
\end{pmatrix}$$

Using a similar approach to LCA, we observed ten possible combinations of model parameters. These are:

Table A.5: Table showing 10 parameter combinations in Hessian matrix of GoM

θ_1	θ_1	Combination	Equation
$g_{i_1k_1}$	$g_{i_2k_2}$	$i_1 = i_2$ and $k_1 = k_2$	A.12
		$i_1 = i_2$ and $k_1 \neq k_2$	A.13
		$i_1 \neq i_2$ and $k_1 = k_2$	**
		$i_1 \neq i_2$ and $k_1 \neq k_2$	**
g_{ik_1}	$\lambda_{k_2 j}$	$k_1 = k_2$	C
		$k_1 \neq k_2$	D
$\lambda_{k_1 j_1}$	$\lambda_{k_2 j_2}$	$k_1 = k_2$, $j_1 = j_2$	E
		$k_1 \neq k_2$, $j_1 = j_2$	F
		$k_1 = k_2$, $j_1 \neq j_2$	**
		$k_1 \neq k_2$, $j_1 \neq j_2$	**

** indicates the second-order partial derivative for such combination is zero.

Like LCA, the assumption of independence has resulted in the covariance of some combinations; this includes the independence among the subjects and among the symptoms. When differentiation the posterior probability w.r.t g_{ik_1} and g_{ik_2}, if $k_1 = k_2$, let ℓ denote k_1 and k_2, and ι denote any subject, the second order partial derivative becomes

$$\frac{\partial^2 h(f)}{\partial g_{i\ell}^2} = -\sum_j \left[\frac{\lambda_{\ell j}^{y_{\iota j}} (1-\lambda_{\ell j})^{(1-y_{\iota j})}}{\sum_k g_{ik} \lambda_{kj}^{y_{\iota j}} (1-\lambda_{kj})^{(1-y_{\iota j})}} \right]^2 . \tag{A.12}$$

.

When $k_1 \neq k_2$, using similar notation as earlier, the second order second order partial derivative becomes

$$\frac{\partial^2 h(f)}{\partial g_{\imath k_1} \partial g_{\imath k_2}} = -\sum_j \left[\frac{\lambda_{k_1 j}^{y_{\imath j}} (1-\lambda_{k_1 j})^{(1-y_{\imath j})} \lambda_{k_2 j}^{y_{\imath j}} (1-\lambda_{k_2 j})^{(1-y_{\imath j})}}{\sum_k g_{\imath k} \lambda_{kj}^{y_{\imath j}} (1-\lambda_{kj})^{(1-y_{\imath j})}} \right]^2. \tag{A.13}$$

.

When differentiating the posterior probability w.r.t $g_{i k_1}$ and $\lambda_{k_2 j}$ and k_1 equals k_2. Let ℓ denote both k_1 and k_2, the second order partial derivative for any subject $\imath$ and symptom $\jmath$ is as follows,

$$\frac{\partial^2 h(f)}{\partial g_{\imath \ell} \partial \lambda_{\ell \jmath}} = \frac{(y_{\imath \jmath} \lambda_{\ell \jmath}^{(y_{\imath \jmath}-1)} - \lambda_{\ell \jmath}^{y_{\imath \jmath}})[(1-\lambda_{\ell \jmath})^{y_{\imath \jmath}} (\sum_k g_{\imath k} \lambda_{k \jmath}^{y_{\imath \jmath}} (1-\lambda_{k \jmath})^{(1-y_{\imath \jmath})}) - \lambda_{\ell \jmath}^{y_{\imath \jmath}} (1-\lambda_{\ell \jmath}) p_\ell]}{\left[(\sum_k g_{\imath k} \lambda_{k \jmath}^{y_{\imath \jmath}} (1-\lambda_{k \jmath})^{(1-y_{\imath \jmath})}) (1-\lambda_{\ell \jmath})^{y_{\imath \jmath}} \right]^2}. \tag{A.14}$$

.

When k_1 and k_2 are not equal, the second order partial derivative becomes

$$\frac{\partial^2 h(f)}{\partial g_{\imath k_1} \partial \lambda_{k_2 \jmath}} = \frac{g_{\imath k_2} (y_{\imath \jmath} \lambda_{k_2 \jmath}^{(y_{\imath \jmath}-1)} - \lambda_{k_2 \jmath}^{y_{\imath \jmath}})[\lambda_{k_1 \jmath}^{y_{\imath \jmath}} (1-\lambda_{k_1})^{(1-y_{\imath \jmath})}]}{(1-\lambda_{k_1 \jmath})^{y_{\imath \jmath}} (\sum_k g_{\imath k} \lambda_{k \jmath}^{y_{\imath \jmath}} (1-\lambda_{k \jmath})^{(1-y_{\imath \jmath})})}. \tag{A.15}$$

.

The last combinations are the variance and covariance of $\lambda_{k_1 j}$ and $\lambda_{k_2 j}$. Again letting ℓ denote k_1 and k_2 where these are equal, for any symptom $\jmath$, the second order partial derivative is

$$\frac{\partial^2 h(f)}{\partial \lambda_{\ell \jmath}{}^2} = -\sum_i \left[\frac{p_\ell (y_{\imath \jmath} \lambda_{\ell \jmath}^{y_{\imath \jmath}-1} (y_{\imath \jmath} - \lambda_{\ell \jmath})]}{(1-\lambda_{\ell \jmath})^{y_{\imath \jmath}} \sum_k p_k \lambda_{k \jmath}^{y_{\imath \jmath}} (1-\lambda_{k \jmath})^{(1-y_{\imath \jmath})}} \right]^2. \tag{A.16}$$

When k_1 is not equal to k_2, the second order partial derivative becomes

$$\frac{\partial^2 h(f)}{\partial \lambda_{k_1 \jmath} \partial \lambda_{k_2 \jmath}} = -\sum_i \left[\frac{g_{i k_1} g_{i k_2} (\lambda_{k_1 \jmath} \lambda_{k_2 \jmath})^{(y_{i \jmath}-1)} (y_{i \jmath} - \lambda_{k_1 \jmath})(y_{i \jmath} - \lambda_{k_2 \jmath})}{(1-\lambda_{k_1 \jmath})^{y_{i \jmath}} (1-\lambda_{k_2 \jmath})^{y_{i \jmath}} [\sum_k g_{i k} \lambda_{k j}^{y_{i \jmath}} (1-\lambda_{k \jmath})^{(1-y_{i \jmath})}]^2} \right]. \tag{A.17}$$

A.3 Chapter 6

Table A.6: The SNP ID referenced in this study

SNPID	SNP names	SNPID	SNP names
1112	rs4959334	6157	rs9268403
1576	rs10901001	6158	rs12201454
3302	rs7749556	6160	rs12528797
4073	rs874448	6172	rs3806156
4887	rs950877	6173	rs3763307
5447	rs16894900	6174	rs3763308
5545	rs9258205	6177	rs2001097
5553	rs9258223	6179	rs3135378
5566	rs1633030	6180	rs3135377
5571	rs1632973	6189	rs9268560
5577	rs9258466	6191	rs3135342
5587	rs1233320	6195	rs9268645
5588	rs16896081	6205	rs9268858
5638	rs1150743	6211	rs9268877
5661	rs9261389	6214	rs9270986
5663	rs9261394	6217	rs4530903
5802	rs2394390	6219	rs9272219
5919	rs9263702	6221	rs9272723
5932	rs2073724	6222	rs9273363
5947	rs3095238	6225	rs7775228
5957	rs3130531	6227	rs6457617
5969	rs7382297	6228	rs6457620
6025	rs16899646	6232	rs9275418
6043	rs2523650	6233	rs9275523
6051	rs3131631	6382	rs3129207
6073	rs2242655	6385	rs7382464
6087	rs480092	8169	rs16872971
6110	rs408359	8390	rs2028542
6117	rs438475	12097	rs9343272
6121	SNP_A.2064274	17510	rs6938123
6122	rs377763	21883	rs9497148
6149	rs9268302	22015	rs3763239
6154	rs9268402	24454	rs16891392
6156	rs9391858	26289	rs16901461

Full Reference List

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30:97–101, 2002.

[2] G. R. Abecasis, S. S. Cherny, W. O. C. Cookson, and L. R. Cardon. *GRR: graphical representation of relationship errors*, volume 17. Oxford University Press, 2001.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*, 62(5):1198–1211, 1998.

[5] J. Altmüller, L. J. Palmer, G. Fischer, H. Scherb, and M. Wjst. Genomewide scans of complex human diseases: true linkage is hard to find. *The American Journal of Human Genetics*, 69(5):936–950, 2001.

[6] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *The American Journal of Human Genetics*, 54(3):535–543, 1994.

[7] C. I. Amos, D. K. Zhu, and E. Boerwinkle. Assessing genetic linkage and association with robust components of variance approaches. *Annals of Human Genetics*, 60(2):143–160, 1996.

[8] A. S. Andrew, M. R. Karagas, H. H. Nelson, S. Guarrera, S. Polidoro, S. Gamberini, C. Sacerdote, J. H. Moore, K. T. Kelsey, and E. Demidenko. Dna repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy. *Human Heredity*, 65(2):105–118, 2008.

[9] Amalia Annest, Roger Bumgarner., Adrian Raftery., and Ka Yee Yeung. Iterative bayesian model averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*, 10:17, 2009.

[10] V. Anttila, M. Kallela, G. Oswell, M. A. Kaunisto, D. R. Nyholt, E. Hamalainen, H. Havanka, M. Ilmavirta, J. Terwilliger, and E. Sobel. Trait components provide tools to dissect the genetic susceptibility of migraine. *The American Journal of Human Genetics*, 79(1):85–99, 2006.

[11] V. Anttila, D. R. Nyholt, M. Kallela, V. Artto, S. Vepsäläinen, E. Jakkula, A. Wennerström, P. Tikka-Kleemola, M. A. Kaunisto, and E. Hämäläinen. Consistently replicating locus linked to migraine on 10q22-q23. *The American Journal of Human Genetics*, 82(5):1051–1063, 2008.

[12] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[13] K. Asano, T. Matsushita, J. Umeno, N. Hosono, A. Takahashi, T. Kawaguchi, T. Matsumoto, T. Matsui, Y. Kakuta, and Y. Kinouchi. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the japanese population. *Nature Genetics*, 41(12):1325–1329, 2009.

[14] B. Atik, T. A. Skwor, R. P. Kandel, B. Sharma, H. K. Adhikari, L. Steiner, H. Erlich, and D. Dean. Identification of novel single nucleotide polymorphisms in inflammatory genes as risk factors associated with trachomatous trichiasis. *PLoS ONE*, 3(10), 2008.

[15] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–792, 2006.

[16] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[17] 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, the NIDDK 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, the Belgian-French IBD Consortium, the 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 crohns disease. *Nature Genetics*, 40(8):955, 2008.

[18] L. Bastone, M. Reilly, D. J. Rader, and A. S. Foulkes. Mdr and prp: a comparison of methods for high-order genotype-phenotype associations. *Human Heredity*, 58(2):82–92, 2004.

[19] W. Bateson. *Mendel's principles of heredity*. Cambridge University Press. Cambridge, 1909.

[20] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear programming: theory and algorithms*. Wiley-Interscience, 2006.

[21] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025, 2002.

[22] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[23] J. Berkhof, I. van Mechelen, and A. Gelman. A bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442, 2003.

[24] J. M. Bernado and A. F. M. Smith. *Bayesian Theory*. Wiley, New York, 1994.

[25] Donald A. Berry and Yosef Hochberg. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82:215–227, 1999.

[26] L. J. Bierut, P. A. F. Madden, N. Breslau, E. O. Johnson, D. Hatsukami, O. F. Pomerleau, G. E. Swan, J. Rutter, S. Bertelsen, and L. Fox. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human Molecular Genetics*, 16(1):24, 2007.

[27] A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F.M Lord

and M.R. Novick, editors, *Statistical theories of mental test scores*, page 395479. Addison-Wesley, Reading, MA, 1968.

[28] Ásgeir Björnsson, Grétar Gudmundsson, Einar Gudfinnsson, Mara Hrafnsdóttir, John Benedikz, Svanhildur Skúladóttir, Kristleifur Kristjánsson, Michael L. Frigge, Augustine Kong, Kári Stefánsson, and Jeffrey R. Gulcher. Localization of a gene for migraine without aura to chromosome 4q21. *The American Journal of Human Genetics*, 73(5):986–993, 2003.

[29] William C. Blackwelder, Robert C. Elston, and D. C. Rao. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genetic Epidemiology*, 2(1):85–97, 1985.

[30] John Blangero and Laura Almasy. Multipoint oligogenic linkage analysis of quantitative traits. *Genetic Epidemiology*, 14(6):959–964, 1997.

[31] D. Brassat, A. A. Motsinger, S. J. Caillier, H. A. Erlich, K. Walker, L. L. Steiner, B. A. C. Cree, L. F. Barcellos, M. A. Pericak-Vance, and S. Schmidt. Multifactor dimensionality reduction reveals genegene interactions associated with multiple sclerosis susceptibility in african americans. *Genes and Immunity*, 7:310–315, 2006.

[32] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123 – 140, 1996.

[33] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[34] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapmand and Hall CRC, New York, 1984.

[35] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, 28(2):171–182, 2005.

[36] A. Bureau, J. Dupuis, B. Hayward, K. Falls, and P. Van Eerdewegh. Mapping complex traits using random forests. *BMC Genetics*, 4 Suppl 1:S64, 2003.

[37] Zameel M. Cader, Sandra Noble-Topham, David A. Dyment, Stacey S. Cherny, John D. Brown,

George P. A. Rice, and George C. Ebers. Significant linkage to migraine with aura on chromosome 11q24. *Human Molecular Genetics*, 12(19):2511–2517, 2003.

[38] O. Cappe, A. Guillin, J. M. Marin, and C. P. Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[39] O Carlborg and C. S. Haley. Epistasis: too often neglected in complex trait studies? *NATURE REVIEWS GENETICS*, 5(8):618–625, 2004.

[40] B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484, 1995.

[41] A. Carlsson, L. Forsgren, P. O. Nylander, U. Hellman, K. Forsman-Semb, G. Holmgren, D. Holmberg, and M. Holmberg. Identification of a susceptibility locus for migraine with and without aura on 6p12.2-p21.1. *Neurology*, 59(11):1804–1807, 2002.

[42] F. Cassidy, C. F. Pieper, and B. J. Carroll. Subtypes of mania determined by grade of membership analysis. *Neuropsychopharmacology*, 25(3):373–83, 2001.

[43] G. Celeux, F. Forbes, C. Robert, and M. Titterington. Deviance information criteria for missing data models. *Bayesian Statistics*, page 06, 2006.

[44] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466, 1995.

[45] Carla C-M. Chen, Jonathan M. Keith, Dale R. Nyholt, Nicholas G. Martin, and Kerrie L. Mengersen. Bayesian latent trait modeling of migraine symptom data. *Human Genetics*, 126:277 – 288, 2009.

[46] Carla Chia-Ming Chen, Kerrie Mengersen, and Jonathan M. Keith. Bayesian method for genome-wide association studies: Review and illustration, 2010.

[47] Carla Chia-Ming Chen, Kerrie L. Mengersen, Jonathan M. Keith, Nicholas G. Martin, and Dale R Nyholt. Linkage and heritability analysis of migraine symptom groupings: a comparison of three different clustering methods on twin data. *Human Genetics*, 125(5):591 – 604, 2009.

[48] Y. M. Cho, M. D. Ritchie, J. H. Moore, J. Y. Park, K. U. Lee, H. D. Shin, H. K. Lee, and K. S. Park. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia*, 47(3):549–554, 2004.

[49] D. V. Conti and W. J. Gauderman. Snps, haplotypes, and model selection in a candidate gene region: the simple analysis for multilocus data. *Genetic Epidemiology*, 27(4):429–441, 2004.

[50] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463, 2002.

[51] H. J. Cordell. Detecting genegene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[52] H. J. Cordell and D. G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. *The American Journal of Human Genetics*, 70(1):124–141, 2002.

[53] E. H. Corder and M. A. Woodbury. Genetic heterogeneity in alzheimer's disease: A grade of membership analysis. *Genetic Epidemiology*, 10:495–499, 1993.

[54] B. K. Cornes, S. E. Medland, M. A. R. Ferreira, K. I. Morley, D. L. Duffy, B. T. Heijmans, G. W. Montgomery, and N. G. Martin. Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 australian twin families. *Twin Research Human Genetics*, 8(6):616–632, 2005.

[55] K. D. Crawford and R. L. Wainwright. Applying genetic algorithms to outlier detection. In L.J. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 546–550, San Mateo, California, 1995. Proceedings of the Sixth International Conference on Genetic Algorithms.

[56] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich. A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.

[57] Mariza De Andrade, Ren Guguen, Sophie Visvikis, Catherine Sass, Grard Siest, and Christopher I. Amos. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genetic Epidemiology*, 22(3):221–232, 2002.

[58] P. I. W. de Bakker, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, and M. Delgado. A high-resolution hla and snp haplotype map for disease association studies in the extended human mhc. *Nature Genetics*, 38(10):1166–1172, 2006.

[59] M. De Fusco, R. Marconi, L. Silvestri, L. Atorino, L. Rampoldi, L. Morgante, A. Ballabio, P. Aridon, and G. Casari. Haploinsufficiency of atp1a2 encoding the na+/k+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. *Nature Genetics*, 33(2):192–6, 2003.

[60] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[61] B. Devlin, S. A. Bacanu, K. L. Klump, C. M. Bulik, M. M. Fichter, K. A. Halmi, A. S. Kaplan, M. Strober, J. Treasure, and D. B. Woodside. Linkage analysis of anorexia nervosa incorporating behavioral covariates. *Human Molecular Genetics*, 11(6):689–696, 2002.

[62] M. Dichgans, T. Freilinger, G. Eckstein, E. Babini, B. Lorenz-Depiereux, S. Biskup, M. D. Ferrari, J. Herzog, Amjm van den Maagdenberg, and M. Pusch. Mutation in the neuronal voltage-gated sodium channel scn1a in familial hemiplegic migraine. *The Lancet*, 366(9483):371–377, 2005.

[63] T. Dietterich, M. Kearns, and Y. Mansour. Applying the weak learning framework to understand and improve c4. 5. In Lorenza Saitta, editor, *Machine Learning: Proceedsing of the Seventeenth International Conference on Machine Learning*, pages 96–103, San Francisco, 1996. Morgan Kaufmann.

[64] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, pages 1–15, 2000.

[65] T. K. Dijkstra. *On Model Uncertainty and its statistical implications*. Springer-Verlag Berlin, 1988.

[66] P. Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728731, 2008.

[67] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97, 1995.

[68] A. Drewnowski and C. L. Rock. The influence of genetic taste markers on food acceptance. *American Journal of Clinical Nutrition*, 62(3):506, 1995.

[69] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, and A. Griffiths. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463, 2006.

[70] D. L. Duffy. Sib-pair version 0.99. 9. *Queensland Institute of Medical Research, Brisbane, Australia*, 2002.

[71] W. L. Duren, M. P. Epstein, M. Li, and M. Boehnke. Relpair: A program that infers the relationships of pairs of individuals based on marker data, 2003.

[72] D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, and R. Luben. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007.

[73] L. Eaves and A. Erkanli. Markov chain monte carlo approaches to analysis of genetic and environmental components of human developmental change and g e interaction. *Behavior Genetics*, 33(3):279–299, 2003.

[74] L. Eaves, A. Erkanli, J. Silberg, A. Angold, H. H. Maes, and D. Foley. Application of bayesian inference using gibbs sampling to item-response theory modeling of multi-symptom genetic data. *Behavior Genetics*, 35(6):765–780, 2005.

[75] L. Eaves, J. Silberg, D. Foley, C. Bulik, H. Maes, A. Erkanli, A. Angold, E. J. Costello, and C. Worthman. Genetic and environmental influences on the relative timing of pubertal change. *Twin Research*, 7(5):471–481, 2004.

[76] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, pages 407–451, 2004.

[77] R. C. Elston and M. Anne Spence. Advances in statistical human genetics over the last 25 years. *Statistics in Medicine*, 25(18):3049, 2006.

[78] M. P. Epstein, W. L. Duren, and M. Boehnke. Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, 67(5):1219–1231, 2000.

[79] Elena A. Erosheva. Comparing latent structures of the grade of membership, rasch, and latent class models. *Psychometrika*, 70(4):619, 2005.

[80] Elena Aleksandrovna Erosheva. *Grade of membership and latent structure models with application to disability survey data*. Ph.d., Carnegie Mellon University, 2002.

[81] Elena Aleksandrovna Erosheva. Partial membership models with application to disability survey data. In H. Bozdogan, editor, *Statistical data mining and knowledge discovery*, pages 117–133. Chapman and Hall/CRC, Boca Raton, FL, 2002.

[82] D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genetics*, 2(9):e157, 2006.

[83] M. Evans and T. Swartz. Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical Science*, 10(3):254–272, 1995.

[84] C. Ferreira. Gene expression programming: A new adaptive algorithm for solving problems. *Arxiv preprint cs.AI/0102027*, 2001.

[85] G. G. Fillenbaum. Typology of alzheimer's disease: findings from cerad data. *Aging and Mental Health*, 2(2):105–127, 1998.

[86] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[87] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, 97:611–631, 2002.

[88] Chris Fraley. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

[89] L. Franconi and C. Jennison. Comparison of a genetic algorithm and simulated annealing in an application to statistical image reconstruction. *Statistics and Computing*, 7(3):193–207, 1997.

[90] L. Fridley, Brooke. Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology*, 33(1):27–37, 2009.

[91] A. Fritsch and K. Ickstadt. Comparing logic regression based methods for identifying snp interactions. *Lecture Notes in Computer Science*, 4414:90, 2007.

[92] S. K. Ganesh, N. A. Zakai, F. J. A. van Rooij, N. Soranzo, A. V. Smith, M. A. Nalls, M. H. Chen, A. Kottgen, N. L. Glazer, and A. Dehghan. Multiple loci influence erythrocyte phenotypes in the charge consortium. *Nature Genetics*, 41(11):1191–1200, 2009.

[93] A.E. Gelfand. Gibbs sampling. In *Encyclopedia of the Statistical Science 1*, pages 283–292. 1997.

[94] A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

[95] E. I. George and R. E. McCulloch. Variable selection via a gibbs sampling. *Journal of the American Statistical Association*, 88(423):881 – 889, 1993.

[96] E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–374, 1997.

[97] J. Gill and G. King. What to do when your hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological Methods and Research*, 33(1):54, 2004.

[98] B. Glaser, I. Nikolov, D. Chubb, M. L. Hamshere, R. Segurado, V. Moskvina, and P. Holmans. Analyses of single marker and pairwise effects of candidate loci for rheumatoid arthritis using logistic regression and random forests. *BMC Proceedings*, 1(Suppl 1):S54, 2007.

[99] S. J. Godsill. On the relationship between markov chain monte carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

[100] D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[101] L. A. Goodman. The analysis of systems of qualitative variables when some of the variables are unobservable. part ia modified latent structure approach. *American Journal of Sociology*, 79(5):1179, 1974.

[102] Leo A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[103] Leo A Goodman. Latent class analysis. In Jacques A. Hagenaars and Allan L. McCutcheon, editors, *Applied latent class analysis*, pages 3–55. Cambridge University Press, Cambridge, 2002. 2001037649 edited by Jacques A. Hagenaars, Allan L. McCutcheon. ill. ; 24 cm. Includes bibliographical references and index.

[104] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[105] D. Greenberg, J. Zhang, D. Shmulewitz, L. Strug, R. Zimmerman, V. Singh, and S. Marathe. Construction of the model for the genetic analysis workshop 14 simulated data: genotype-phenotype relationships, gene interaction, linkage, association, disequilibrium, and ascertainment effects for a complex phenotype. *BMC Genetics*, 6(Suppl 1):S3, 2005.

[106] A. Griffiths, S. Wessler, R. Lewontin, W. Gelbart, D. Suzuki, and J. Miller. *Introduction to Genetic Analysis*. W. H. Freeman and Co, New York, 2005.

[107] S. W. Guo and E. A. Thompson. A monte carlo method for combined segregation and linkage analysis. *The American Journal of Human Genetics*, 51(5):1111–26, 1992.

[108] Jonathan L. Haines, Margaret Ann Pericak-Vance, John Wiley, and Sons. *Genetic analysis of complex diseases*. Wiley-Liss, Hoboken, N.J., 2nd edition, 2006.

[109] J. F. Hallmayer, A. Jablensky, P. Michie, M. Woodbury, B. Salmon, J. Combrinck, H. Wichmann, D. Rock, M. D Ercole, S. Howell, M. Dragovic, and A. Kent. Linkage analysis of candidate regions using a composite neurocognitive phenotype correlated with schizophrenia. *Molecular Psychiatry*, 8(5):511, 2003.

[110] J. F. Hallmayer, L. Kalaydjieva, J. Badcock, M. Dragovic, S. Howell, P. T. Michie, D. Rock, D. Vile, R. Williams, and E. H. Corder. Genetic evidence for a distinct subtype of schizophrenia characterized by pervasive cognitive deficit. *The American Journal of Human Genetics*, 77(3):468–476, 2005.

[111] J. W. Han, H. F. Zheng, Y. Cui, L. D. Sun, D. Q. Ye, Z. Hu, J. H. Xu, Z. M. Cai, W. Huang, and G. P. Zhao. Genome-wide association study in a chinese han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genetics*, 41(1234-1237), 2009.

[112] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19, 1972.

[113] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[114] A. T. Hatjimihail and T. T. Hatjimihail. Design of statistical quality control procedures using genetic algorithms. In L.J. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 551–557, San Mateo, California, 2002.

[115] Headache Classification Committee of the International Headache Society. Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalagia*, 8:1–96, 1988.

[116] A. C. Heath, W. Howells, K. M. Kirk, P. A. Madden, K. K. Bucholz, E. C. Nelson, W. S. Slutske, D. J. Statham, and N. G. Martin. Predictors of non-response to a questionnaire survey of a volunteer twin panel: findings from the australian 1989 twin cohort. *Twin Research*, 4(2):73–80, 2001.

[117] S. C. Heath. Markov chain monte carlo segregation and linkage analysis for oligogenic models. *The American Journal of Human Genetics*, 61(3):748–760, 1997.

[118] A. G. Heidema, J. M. A. Boer, N. Nagelkerke, and E. C. M. Mariman. The challenge for genetic epidemiologists: how to analyze large numbers of snps in relation to complex diseases. *BMC Genetics*, 7:23, 2006.

[119] J. N. Hirschhorn and M. J. Daly. Genome-wide assocation studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95, 2005.

[120] J. A. Hoeting. Methodology for bayesian model averaging: an update. pages 231–240. Citeseer, 2002. Proceedings-Manuscripts of Invited Paper Presentations, International Biometric Conference.

[121] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999.

[122] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7), 2008.

[123] J. Hoh and J. Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4(9):701–709, 2003.

[124] Josephine Hoh, Anja Wille, and Jurg Ott. Trimming, weighting, and grouping snps in human case-control association studies. *Genome Research*, 11(12):2115–2119, 2001.

[125] J. H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence.* Ann Arbor MI: University of Michigan Press. 1975.

[126] J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Czard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, and M. Gassull. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*, 411(6837):599–603, 2001.

[127] K. P. Hummel, M. M. Dickie, and D. L. Coleman. Diabetes, a new mutation in the mouse. *Science*, 153(740):1127, 1966.

[128] H. Iba, H. De Garis, and T. Sato. Genetic programming using a minimum description length principle. *Advances in Genetic Programming*, 1:265–284, 1994.

[129] G. Imperatore, R. L. Hanson, D. J. Pettitt, S. Kobes, P. H. Bennett, and W. C. Knowler. Sib-pair linkage analysis for susceptibility genes for microvascular complications among pima indians with type 2 diabetes. pima diabetes genes group. *Diabetes*, 47(5):821, 1998.

[130] International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851861, 2007.

[131] J. Ioannidis. Effect of formal statistical significance on the credibility of observational associations. *American Journal of Epidemiology*, 168(4):374, 2008.

[132] A. Jablensky. Subtyping schizophrenia: implications for genetic research. *Molecular Psychiatry*, 11:815–836, 2006.

[133] R. C. Jansen. A general monte carlo method for mapping multiple quantitative trait loci. *Genetics*, 142(1):305–311, 1996.

[134] C. Jennison and N. Sheehan. Theoretical and empirical properties of the genetic algorithm as a numerical optimizer. *Journal of Computational and Graphical Statistics*, 4(4):296–318, 1995.

[135] R. Jiang, W. Tang, X. Wu, and W. Fu. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 10(Suppl 1):S65, 2009.

[136] Keith W. Jones, Margaret G. Ehm, Margaret A. Pericak-Vance, Jonathan L. Haines, Peter R. Boyd, and Stephen J. Peroutka. Migraine with aura susceptibility locus on chromosome 19p13 is distinct from the familial hemiplegic migraine locus. *Genomics*, 78(3):150–154, 2001.

[137] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[138] A. Julia, J. Moore, L. Miquel, C. Alegre, P. Barcel, M. Ritchie, and S. Marsal. Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics*, 90(1):6–13, 2007.

[139] Christina Justenhoven, Ute Hamann, Beate Pesch, Volker Harth, Sylvia Rabstein, Christian Baisch, Caren Vollmert, Thomas Illig, Yon-Dschun Ko, Thomas Bruning, Hiltrud Brauch, for the Interdisciplinary Study Group on Gene Environment Interactions, and Network Breast Cancer in Germany. Ercc2 genotypes and a corresponding haplotype are linked with breast cancer risk in a german population. *Cancer Epidemiol Biomarkers Prev*, 13(12):2059–2064, 2004.

[140] B. Kaabi and Robert C. Elston. New multivariate test for linkage, with application to pleiotropy: Fuzzy haseman-elston. *Genetic Epidemiology*, 24(4):253–264, 2003.

[141] B. Kaabi, J. Gelernter, S. W. Woods, A. Goddard, G. P. Page, and R. C. Elston. Genome scan for loci predisposing to anxiety disorders using a novel multivariate approach: Strong evidence for a chromosome 4 risk locus. *The American Journal of Human Genetics*, 78:543, 2006.

[142] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430), 1995.

[143] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, New York, 1990.

[144] K. S. Kendler, L. J. Eaves, E. E. Walters, M. C. Neale, A. C. Heath, and R. C. Kessler. The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Archives of General Psychiatry*, 53(5):391–399, 1996.

[145] K. S. Kendler and D. Walsh. The structure of psychosis: syndromes and dimensions. *Archives of General Psychiatry*, 55(6):508–9, 1998.

[146] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[147] B. Knox, P. Ladiges, and B. Evans. Biology, 1994.

[148] N. A. Kolchanov, E. A. Ananko, F. A. Kolpakov, O. A. Podkolodnaya, E. V. Ignateva, T. N. Goryachkovskaya, and I. L. Stepanenko. Gene networks. *Molecular Biology*, 34(4):449–460, 2000.

[149] A. Kong and N. J. Cox. Allele-sharing models: Lod scores and accurate linkage tests. *The American Journal of Human Genetics*, 61(5):1179–1188, 1997.

[150] X. Kong, K. Murphy, T. Raj, C. He, P. S. White, and T. C. Matise. A combined linkage-physical map of the human genome. *The American Journal of Human Genetics*, 75(6):1143–8, 2004.

[151] C. Kooperberg, J. C. Bis, K. D. Marciante, S. R. Heckbert, T. Lumley, and B. M. Psaty. Logic regression for analysis of the association between genetic variation in the renin-angiotensin system and myocardial infarction or stroke. *American Journal of Epidemiology*, 165(3):334, 2007.

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

[153] C. Kooperberg and I. Ruczinski. Identifying interacting snps using monte carlo logic regression. *Genetic Epidemiology*, 28(2):157–170, 2005.

[154] J. R. Koza and J. P. Rice. *Genetic programming*. Springer, 1992.

[155] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics*, 58(6):1347–1363, 1996.

[156] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc. Nat. Acad. Sci. USA*, 84:2363–2367, 1987.

[157] E. S. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11:241–247, 1995.

[158] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, and W. FitzHugh. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[159] K. Lange, D. Weeks, and M. Boehnke. Programs for pedigree analysis: Mendel, fisher, and dgene. *Genetic Epidemiology*, 5(6):471–2, 1988.

[160] L. J. Launer, G. M. Terwindt, and M. D. Ferrari. The prevalence and characteristics of migraine in a population-based cohort the gem study. *Neurology*, 53(3):537–537, 1999.

[161] P. F. Lazersfeld. The logical and mathematical foundations of latent structure analysis. In S. S. Stouffer, editor, *Measurement and Prediction*, pages 362–412. Princeton University Press, Princeton, NJ, 1950.

[162] R. A. Lea, D. R. Nyholt, R. P. Curtain, M. Ovcaric, R. Sciascia, C. Bellis, J. MacMillan, S. Quinlan, R. A. Gibson, and L. C. McCarthy. A genome-wide scan provides evidence for loci influencing a severe heritable form of common migraine. *Neurogenetics*, 6(2):67–72, 2005.

[163] S. M. Lewis and A. E. Raftery. Estimating bayes factors via posterior stimulation with the laplace-metropolis estimator. *Journal of the American Statistical Association*, 92(438), 1997.

[164] L. Ligthart, D. I. Boomsma, N. G. Martin, J. H. Stubbe, and D. R. Nyholt. Migraine with aura and migraine without aura are not distinct entities: Further evidence from a large dutch population study. *Twin Research and Human Genetics*, 9(1):54–63, 2006.

[165] L. Ligthart, D. R. Nyholt, J. J. Hottenga, M. A. Distel, G. Willemsen, and D. I. Boomsma. A genome-wide linkage scan provides evidence for both new and previously reported loci influencing common migraine. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 2008.

[166] H. Y. Lin, R. Desmond, S. Louis Bridges, and S. Soong. Variable selection in logistic regression for detecting snpsnp interactions: the rheumatoid arthritis example. *European Journal of Human Genetics*, 16(6):735, 2008.

[167] Drew A. Linzer and Jeffrey Lewis. polca: Polytomous variable latent class analysis, 2007.

[168] F. M. Lord. Applications of item response theory. *Applied Psychological Measurement*, 5(4), 1981.

[169] Kathryn 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.

[170] D. J. Lunn, N. Best, and J. C. Whittaker. Generic reversible jump mcmc using graphical models. *Statistics and Computing*, 19:395–408, 2009.

[171] D. J. Lunn, J. C. Whittaker, and N. Best. A bayesian toolkit for genetic association studies. *Genetic Epidemiology*, 30(3):231, 2006.

[172] Constantine G. Lyketsos, John C. S. Breitner, and Peter V. Rabins. An evidence-based proposal for the classification of neuropsychiatric disturbance in alzheimer's disease. *International Journal of Geriatric Psychiatry*, 16(11):1037–1042, 2001.

[173] J. MacQueen. Some methods for classification and analysis of multivariate observation. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[174] Paula Macrossan, Carla C-M Chen, and Kerrie L Mengersen. Using gene expression programming with modified logic regression for the investigation of snp interactions in large dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010.

[175] D. Madigan, J. Gavrin, and A. E. Raftery. Enhancing the predictive performance of bayesian graphical models. *Communications in Statistics-Theory and Methods*, 24:2271–2292, 1995.

[176] M. M. Maechler and M. Hubert. The cluster package, 2008.

[177] C. L. Mallows. Some comments on cp. *Technometrics*, (15):661–675, 1973.

[178] K. G. Manton, A. Korten, M. A. Woodbury, M. Anker, and A. Jablensky. Symptom profiles of psychiatric disorders based on graded disease classes: an illustration using data from the who international pilot study of schizophrenia. *Psychological Medicine*, 24(1):133–44, 1994.

[179] K. G. Manton, M. A. Woodbury, and H. D. Tolley. Statistical applications using fuzzy sets. page 68. Wiley, 1994.

[180] Kenneth G. Manton, Gu Xiliang, Huang Hai, and Mikhail Kovtun. Fuzzy set analyses of genetic determinants of health and disability status. *Statistical Methods in Medical Research*, 13(5):395–408, 2004.

[181] J. Marchini, P. Donnelly, and L. R. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, 2005.

[182] J. M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25:459507, 2005.

[183] J. M. Marin and C. P. Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Verlag, 2007.

[184] E. Marinari and G. Parisi. Simulated tempering: a new monte carlo scheme. *Europhysics Letters*, 19(6):451–458, 1992.

[185] P. McCullagh and J. A. Nelder. Generalized linear models, 1983.

[186] Allan L McCutcheon. *Latent Class Analysis*. Quantitative Applications in the Social Science. Sage Publications, Newbury Park, 1987.

[187] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore. Machine learning for detecting gene-gene interactions: a review. *Applied Bioinformatics*, 5(2):77–88, 2006.

[188] G. J. McLachlan and S. U. Chang. Mixture modelling for cluster analysis. *Statistical Methods in Medical Research*, 13(5):347–361, 2004.

[189] G. J. McLachlan, K. A. Do, and C. Ambroise. *Analyzing microarray gene expression data*. Wiley-Interscience, 2005.

[190] G. J. McLachlan, D. Peel, K. E. Basford, and P. Adams. The emmix software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4(2), 1999.

[191] L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris. Polymorphism interaction analysis (pia): a method for investigating complex gene-gene interactions. *BMC Bioinformatics*, 9:146, 2008.

[192] Yan Meng, Qiong Yang, Karen T Cuenco, L Adrienne Cupples, Anita L DeStefano, and Kathryn L Lunette. Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and bayesian networks. *BMC Proceedings*, 1(Suppl 1):S56, 2007.

[193] Barnett S. Meyers. The evolving typology of neuropsychiatric complications of alzheimer's disease: the use of latent trait analysis. *International Journal of Geriatric Psychiatry*, 16(11):1030–1032, 2001.

[194] Braxton D. Mitchell, Soumitra Ghosh, Jennifer L. Schneider, Gunther Birznieks, and John Blangero. Power of variance component linkage analysis to detect epistasis. *Genetic Epidemiology*, 14(6):1017–1022, 1997.

[195] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, pages 1023–1032, 1988.

[196] M. Mochi, S. Sangiorgi, P. Cortelli, V. Carelli, C. Scapoli, M. Crisci, L. Monari, G. Pierangeli, and P. Montagna. Testing models for genetics determination in migraine. *Cephalagia*, 13:389–394, 1993.

[197] G. Montana. Statistical methods in genetics. *Briefings in Bioinformatics*, 7(3):297, 2006.

[198] J. H. Moore and M. D. Ritchie. The challenges of whole-genome approaches to common diseases. *The Journal Of the American Medical Association*, 291(13):1642–1643, 2004.

[199] Maria Moran, C. Walsh, A. Lynch, R. F. Coen, D. Coakley, and B. A. Lawlor. Syndromes of behavioural and psychological symptoms in mild alzheimer's disease. *International Journal of Geriatric Psychiatry*, 19(4):359–364, 2004.

[200] A. A. Motsinger, S. L. Lee, G. Mellick, and M. D. Ritchie. Gpnn: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC bioinformatics*, 7(1):39, 2006.

[201] A. A. Motsinger-Reif, S. M. Dudek, L. W. Hahn, and M. D. Ritchie. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, 32(4):325–340, 2008.

[202] E. J. Mulder, C. Van Baal, D. Gaist, M. Kallela, J. Kaprio, D. A. Svensson, D. R. Nyholt, N. G. Martin, A. J. MacGregor, and L. F. Cherkas. Genetic and environmental influences on migraine: a twin study across six countries. *Twin Research*, 6(5):422–31, 2003.

[203] S. R. Narum. Beyond bonferroni: less conservative analyses for conservation genetics. *Conservation Genetics*, 7(5):783–787, 2006.

[204] R. M. Neal. Markov chain monte carlo methods based on 'slicing' the density function. Technical Report Technical Report No. 9722,, Department of Statistics, University of Toronto, 1997.

[205] R. M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

[206] M. C. Neale, Virginia Institute for Psychiatric, Genetics Behavioral, Psychiatry Department of, and Virginia Medical College of. *MX: Statistical Modeling*. Department of Psychiatry, Medical College of Virginia, 1997.

[207] M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, 11(3):458 – 470, 2001.

[208] M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994.

[209] R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt, and I. Wegener. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23(24):3280, 2007.

[210] D. R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

[211] Dale R. Nyholt, Nathan G. Gillespie, Andrew C. Heath, Kathleen R. Merikangas, David L. Duffy, and Nicholas G. Matrin. Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities. *Genetic Epidemiology*, 26:231–244, 2004.

[212] Dale R. Nyholt, Katherine I. Morley, Manuel A. R. Ferreira, Sarah E. Medland, Dorret I. Boomsma, Andrew C. Heath, Kathleen R. Merikangas, Grant W. Montgomery, and Nicholas G. Matrin. Genomewide significant linkage to migrainous headache on chromosome 5q21. *American Journal of Human Genetics*, 77:500–512, 2005.

[213] J. Olesen and T. J. Steiner. The international classification of headache disorders, 2nd edn (icdh-ii). *British Medical Journal*, 75(6):808, 2004.

[214] P. C. Phillips. Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *NATURE REVIEWS GENETICS*, 9(11):855–867, 2008.

[215] D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793, 2004.

[216] R. F. Potthoff, K. G. Manton, and M. A. Woodbury. Dirichlet generalizations of latent-class models. *Journal of Classification*, 17(2):315–353, 2000.

[217] Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *The American Journal of Psychiatry*, 166(5):540–556, 2009.

[218] J. R. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[219] R Development Core Team. R 2.4.1 a language and development, 2006.

[220] A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[221] A. E. Raftery. Hypothesis testing and model selection via posterior simulation. In W.R. Gilks, D. J. Spiegelhalter, and S. Richardson, editors, *Practical Markov Chain Monte Carlo*, pages 163–188. Chapman and Hall, London, 1995.

[222] A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[223] A. E. Raftery, D. Madigan, and C. T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics*, 5:323–349, 1996.

[224] D. C. Rao. An overview of the genetic dissection of complex trait. In D. C. Rao and C. C. Gu, editors, *Genetic Dissection of Complex Traits*. Academic Press, Burlington, 2008.

[225] G. Rasch. *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research, Copenhagen, Denmark, 1960.

[226] E. Repapi, I. Sayers, L. V. Wain, P. R. Burton, T. Johnson, M. Obeidat, J. H. Zhao, A. Ramasamy, G. Zhai, and V. Vitart. Genome-wide association study identifies five loci associated with lung function. *Nature Genetics*, 42(1):36–45, 2009.

[227] Treva K. Rice, Nicholas J Schork, and D. C. Rao. Methods for handling multiple testing. In D. C. Rao and C. C. Gu, editors, *Genetic dissection of complex traits*. Academic Press, 2008.

[228] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):731–792, 1997.

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

[230] N. J. Risch. Searching for genetic determination for the new millennium. *Nature*, 405:847–856, 2000.

[231] 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. *The American Journal of Human Genetics*, 69(1):138 – 147, 2001.

[232] M. Robnik-Šikonja. Experiments with cost-sensitive feature evaluation. *Machine Learning: ECML 2003*, pages 325–336, 2003.

[233] M. Robnik-Šikonja. Improving random forests. *Machine Learning: ECML 2004*, pages 359–370, 2004.

[234] K. Roeder, S. A. Bacanu, L. Wasserman, and B. Devlin. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics*, 78, 2006.

[235] I. Ruczinski, C. Kooperberg, and M. Leblanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[236] M. B. Russell. Genetics of migraine without aura, migraine with aura, migrainous disorder, head trauma migraine without aura and tension-type headache. *Cephalalgia*, 21(7):778–780, 2001.

[237] M. B. Russell, B. K. Rasmussen, K. Fenger, and J. Olesen. Migraine without aura and migraine with aura are distinct clinical entities: a study of four hundred and eighty-four male and female migraineurs from the general population. *Cephalalgia*, 16(4):239–45, 1996.

[238] Michael BjÃrn Russell, Vibeke Ulrich, Morten Gervil, and Jes Olesen. Migraine without aura and migraine with aura are distinct disorders. a population-based twin survey. *Headache: The Journal of Head and Face Pain*, 42(5):332–336, 2002.

[239] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, and H. E. Wichmann. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*, 2007.

[240] P. D. Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261, 1997.

[241] E. Schouwenberg, H. Houweling, M. J. W. Jansen, J. Kros, and J. P. Mol-Dijkstra. Uncertainty propagation in model chains: a case study in nature conservancy. *Alterra report*, 1, 2000.

[242] D. F. Schwarz, S. Szymczak, A. Ziegler, and I. R. Knig. Picking single-nucleotide polymorphisms in forests. *BMC Proceedings*, I(suppl I):S59, 2007.

[243] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1979.

[244] H. Schwender. *Statistical Analysis of Genotype and Gene Expression Data*. PhD thesis, Department of Statistics, TU Dortmund University, 2007.

[245] H. Schwender and K. Ickstadt. Identification of snp interactions using logic regression. *Biostatistics*, 9(1):187, 2008.

[246] H. Schwender and K. Ickstadt. Imputing missing genotypes with k nearest neighbors. Technical report, Tech. rep., Collaborative Research Center 475, Department of Statistics, University of Dortmund 2008, 2008.

[247] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397, 1971.

[248] A. Scuteri, S. Sanna, W. M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orr, and G. Usala. Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS Genetics*, 3(7):e115, 2007.

[249] T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[250] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5):335–347, 1989.

[251] S. Silberstein, J. Olesen, M. G. Bousser, H. C. Diener, D. Dodick, M. First, P. Goadsby, H. Gobel, M. Lainez, and J. Lance. The international classification of headache disorders, (ichd-ii)-revision of criteria for 8.2 medication-overuse headache. *Cephalalgia*, 25(6):460–465, 2005.

[252] M. J. Sillanpaa and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics*, 151(4):1605–1619, 1999.

[253] J. Simón-Snchez, C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S. W. Scholz, and D. G. Hernandez. Genome-wide association study reveals genetic risk underlying parkinson's disease. *Nature Genetics*, 41:1308–1312, 2009.

[254] J. Siwei, C. Zhihua, Z. Dan, L. Yadong, and L. Qu. Gene expression programming based on simulated annealing. In *2005 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1264–1267, 2005. 2005 International Conference on Wireless Communications, Networking and Mobile Computing, 2005. Proceedings.

[255] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, and S. Hadjadj. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[256] S. L. Slager and D. J. Schaid. Case-control studies of genetic markers: Power and sample size approximations for armitages test for trend. *Human Heredity*, 52(3):149–153, 2000.

[257] D. Soragna, A. Vettori, G. Carraro, E. Marchioni, G. Vazza, S. Bellini, R. Tupler, F. Savoldi, and M. L. Mostacciuolo. A locus for migraine without aura maps on chromosome 14q21.2-q22.3. *American Journal of Human Genetics*, 72(1):161, 2003.

[258] N. Soranzo, T. D. Spector, M. Mangino, B. Kühnel, A. Rendon, A. Teumer, C. Willenborg, B. Wright, L. Chen, and M. Li. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the haemgen consortium. *Nature Genetics*, 41:1182–1190, 2009.

[259] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. Winbugs 1.4. 1. bayesian inference using gibbs sampling, 2006.

[260] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[261] H. Stefansson, H. Petursson, E. Sigurdsson, V. Steinthorsdottir, S. Bjornsdottir, T. Sigmundsson, S. Ghosh, J. Brynjolfsson, S. Gunnarsdottir, and O. Ivarsson. Neuregulin 1 and susceptibility to schizophrenia. *The American Journal of Human Genetics*, 71(4):877–892, 2002.

[262] Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10:681–690, 2009.

[263] J. A. C. Sterne, G. D. Smith, and D. R. Cox. Sifting the evidence–what's wrong with significance tests? *Physical Therapy*, 81(8):1464, 2001.

[264] Dan A. Svensson, Bo Larsson, Elisabet Waldenlind, and Nancy L. Pedersen. Shared rearing environment in migraine: Results from twins reared apart and twins reared together. *Headache: The Journal of Head and Face Pain*, 43(3):235–244, 2003.

[265] Dan A. Svensson, Elisabet Waldenlind, Karl Ekbom, and Nancy L. Pedersen. Heritability of migraine as a function of definition. *The Journal of Headache and Pain*, 5(3):171, 2004.

[266] E. Szadoczky, S. Rozsa, S. Patten, M. Arato, and J. Furedi. Lifetime patterns of depressive symptoms in the community and among primary care attenders: an application of grade of membership analysis. *Journal of Affective Disorders*, 77(1):31–9, 2003.

[267] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. Gonzlez-Recio, I. R. Knig, H. Zhang, and Y. V. Sun. Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33(1):S51–S57, 2009.

[268] B. G. Tabachnick, L. S. Fidell, and S. J. Osterlind. Using multivariate statistics. 2001.

[269] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[270] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[271] The International Human Genome Mapping Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[272] The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature*, 409(6822):934–941, 2001.

[273] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature*, 447(7), 2007.

[274] D. C. Thomas, R. W. Haile, and D. Duggan. Recent developments in genomewide association scans: a workshop summary and review. *The American Journal of Human Genetics*, 77(3):337–345, 2005.

[275] Duncan C. Thomas. *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford, 2004.

[276] G. Thorleifsson, G. B. Walters, D. F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, and I. Jonsdottir. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1):6, 2009.

[277] C. Tian, R. P. Stokowski, D. Kershenobich, D. G. Ballinger, and D. A. Hinds. Variant in pnpla3 is associated with alcoholic liver disease. *Nature Genetics*, 42:21–23, 2009.

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

[279] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, and F. Payne. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, 39(7):857, 2008.

[280] K. R. Vanmolkot, E. E. Kors, J. J. Hottenga, G. M. Terwindt, J. Haan, W. A. Hoefnagels, D. F. Black, L. A. Sandkuijl, R. R. Frants, and M. D. Ferrari. Novel mutations in the na+, k+-atpase pump gene

atp1a2 associated with familial hemiplegic migraine and benign familial infantile convulsions. *Annals of Neurology*, 54(3):360–6, 2003.

[281] W. N. Venables and B. D. Ripley. Exploratory multivariate analysis. In *Modern applied statistics with S*. Spinger, New York, 2002.

[282] Heather E. Volk, Rosalind J. Neuman, and Richard D. Todd. A systematic evaluation of adhd and comorbid psychopathology in a population-based twin sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44(8):768(8), 2005.

[283] J. Wakefield and J. Bennett. The bayesian modeling of covariates for population pharmacokinetic models. *Journal of the American Statistical Association*, 91(435):917–927, 1996.

[284] C. Wallace, D. J. Smyth, M. Maisuria-Armer, N. M. Walker, J. A. Todd, and D. G. Clayton. The imprinted dlk1-meg3 gene region on chromosome 14q32. 2 alters susceptibility to type 1 diabetes. *Nature Genetics*, 42:68–71, 2009.

[285] E. C. Walsh, K. A. Mather, S. F. Schaffner, L. Farwell, M. J. Daly, N. Patterson, M. Cullen, M. Carrington, T. L. Bugawan, and H. Erlich. An integrated haplotype map of the human major histocompatibility complex. *The American Journal of Human Genetics*, 73(3):580–590, 2003.

[286] Maija Wessman, Mikko Kallela, Mari A. Kunisto, Pia Marttila, Eric Sobel, Jaana Hartiala, Greg Oswell, Suzanne M. Leal, Jeanette C. Papp, Eija H m l inen, Petra Broas, Geoffrey Joslyn, Iiris Hovatta, Tero Hiekkalinna, Jaakko Kaprio, J. uuml rg Ott, Rita M. Cantor, John-Anker Zwart, and Matti Ilmavirta. A susceptibility locus for migraine with aura, on chromosome 4q24. *American Journal of Human Genetics*, 70(3):652, 2002.

[287] Maija Wessman, Gisela M. Terwindt, Mari A. Kaunisto, Aarno Palotie, and Roel A. Ophoff. Migraine: a complex genetic disorder. *The Lancet Neurology*, 6(6):521–532, 2007.

[288] Alice S. Whittemore and Jerry Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50(1):118–127, 1994.

[289] M. A. Woodbury, J. Clive, and A. Garson Jr. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11(3):277–98, 1978.

[290] N. R. Wray, W. L. Coventry, M. R. James, G. W. Montgomery, L. J. Eaves, and N. G. Martin. Use of monozygotic twins to investigate the relationship between 5httlpr genotype, depression and stressful life events: an application of item response theory. In Michael Rutter, editor, *Novartis Foundation Sympoisum*, volume 293 of *Genetic Effects on Environmental Vulnerability to Disease*, pages 48–68. Novartis Foundation Sympoisum, 2008. Novartis Foundation symposium.

[291] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714, 2009.

[292] Shizhong Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163:789–801, 2003.

[293] Shizhong Xu. An empirical bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 63(2):513–521, 2007.

[294] Shizhong Xu and Zhenyu Jia. Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics*, 175:1955–1963, 2007.

[295] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics-Oxford*, 21(10):2394–2402, 2005.

[296] N. Yi and S. Banerjee. Hierarchical generalized linear models for multiple qtl mapping. *Genetics*, 181:1101–1113, 2009.

[297] N. Yi, V. George, and D. B. Allison. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, 164:1129 – 1138, 2003.

[298] N. Yi and S. Xu. Bayesian lasso for quantitative trait loci mapping. *Genetics*, 179(2):1045, 2008.

[299] N. Yi, B. S. Yandell, G. A. Churchill, D. B. Allison, E. J. Eisen, and D. Pomp. Bayesian model

selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344, 2005.

[300] Nengjun Yi and Shizhong Xu. Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics*, 155(3):1391–1403, 2000.

[301] M. Yoshikawa, H. Yamauchi, and H. Terai. Hybrid architecture of genetic algorithm and simulated annealing. *Engineering Letters*, 16(3):339–345, 2008.

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

[303] L. J. Zhao, X. G. Liu, Y. Z. Liu, Y. J. Liu, C. J. Papasian, B. Y. Sha, F. Pan, Y. F. Guo, L. Wang, and H. Yan. Genome-wide association study for femoral neck bone geometry. *Journal of Bone and Mineral Research*, (0):1–34, 2009.

[304] G. Zheng and J. L. Gastwirth. On estimation of the variance in cochran-armitage trend tests for genetic association using case-control studies. *Statistics in Medicine*, 25(18):3150, 2006.

[305] G. Zhu, D. M. Evans, D. L. Duffy, G. W. Montgomery, S. E. Medland, N. A. Gillespie, K. R. Ewen, M. Jewell, Y. W. Liew, and N. K. Hayward. A genome scan for eye color in 502 twin families: most variation is due to a qtl on chromosome 15q. *Twin Research*, 7(2):197–210, 2004.

[306] Dewey K. Ziegler, Yoon-Mi Hur, Thomas J. Bouchard, Ruth S. Hassanein, and Ruth Barter. Migraine in twins raised together and apart. *Headache: The Journal of Head and Face Pain*, 38(6):417–422, 1998.