Modelling, Data Mining and Visualisation of Genetic Variation Data

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Ahmad A. Aloqaily

in

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
2012

CERTIFICATE

Date: **2012**

Author: **Ahmad A. Aloqaily**

Title: **Modelling, Data Mining and Visualisation of Genetic Variation Data**

Degree: **Ph.D.**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Author

Acknowledgements

I wish to express my gratitude to many people who have inspired me both personally and professionally. First and foremost, I would like to thank my supervisors, Dr. Paul Kennedy, Prof. Simeon Simoff and Dr. Daniel Catchpoole. They introduced me to the fields of bioinformatics. I never would have had the courage to go forward with the methodology and theoretical research without their tremendous confidence in me, much more than I had in myself.

They always encouraged and challenged me to think bigger and better. I will always be grateful for all the time and attention that they has invested in me. I can hardly think of anything that has been achieved without their help, as they offer me the freedom to explore new ideas independently.

I would also like to acknowledge the Faculty of Engineering and IT at UTS for offering me a good research environment. I am also grateful to my fellow students, for their encouragement and friendships. I would also like to thank Nicholas Ho, computational biologist at the Children's Cancer Research Unit at the Children's Hospital at Westmead, for assisting me with some of the work on biological insights.

I gratefully acknowledge the funding sources that made my PhD work possible. I was sponsored by the Faculty of IT, Hashemite University - Jordan. My work was also supported by the Australian Rotary Health Research Fund (ARHRF). In terms of external datasets that have been used in this research study, I acknowledge the use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. I also acknowledge the use of a genome-wide SNP dataset generated at the St Jude Children's Research Hospital, the dataset includes 242 patients with acute lymphoblastic leukaemia treated at St Jude Hospital, USA.

Lastly, my special thanks go to my parents, brother and sisters for their constant support and encouragement, and more importantly their faith in me over many years. My parents, I can never thank you enough for whatever you have done for me. Last, but not least, I wish to send personal thanks to

my beloved wife Malak, who has been with me for nearly 5 years already and sacrificed a lot for helping me to pursue my academic pathway. Thank you for your patience and care, and for accompanying me to go through highs and lows accentuated by my research studies.

To My Family,

My Wife and Our Angel, Tala.

Table of Contents

List of Tables

List of Figures

Abstract

Data mining and knowledge discovery have been applied to datasets in various industries including biomedical informatics. The major challenges in data mining in the area stem from the fact that biomedical data comes in many forms with a highly dimensional nature. This research thesis focuses on one specific biomedical dataset, termed as genetic variation data in the form of genome-wide single nucleotide polymorphisms (SNPs) datasets.

Advances in single nucleotide polymorphism genotyping technologies have revolutionised our ability to explore the genetic architecture and models underlying complex diseases by conducting studies based on the whole genome. These studies are called genome-wide association studies. The basic strategy used in these studies is to examine the relationship between the disease of interest and genetic markers across the whole genome.

Many association studies have led to the discovery of single genetic variants associated with common diseases. However, complex diseases are not caused by single genes acting alone but are the result of complex non-linear interactions among genetic factors, with each gene having a small effect on disease risk. For this reason there is a critical need to implement new approaches that can take into account non-linear gene-gene interactions in searching for markers that jointly cause complex diseases.

Several computational methods have been developed to deal with the genetic complexity of complex diseases. However, testing each SNP for main effects and different

orders of gene-gene interaction is computationally infeasible for such high-dimensional data. Also, these methods do not scale well. Therefore, there is growing interest in applying non-parametric predictive models including data mining and machine learning approaches to understand genetic variation data.

This thesis constructs models which incorporate genetic variation data in a manner that will alleviate the error induced by the high dimensionality of such data. Data mining approaches, specifically non-parametric ones, are developed for the modelling, exploration and visualization of patient-to-patient relationships based on genome-wide SNP data. This thesis focuses on three main issues in genetic variation studies: (1) feature selection and distance calculations, (2) framework for the task of disease diagnosis and prognosis, and (3) models for the comparison and visualisation of patient-to-patient relationships based on genome-wide SNP profiles.

This thesis proposes efficient feature selection approaches to find an optimal subset of markers with the highest predictive power for the disease of interest, while managing the large search space required. The proposed approaches select genetic markers for marginal effects as well as gene-gene interaction effects. Markers with marginal effects are selected with an iterative random forest (RF) based procedure, called RF-RFE. The importance measure generated by random forest was chosen for estimating the importance of each SNP (weighting) and facilitates the selection of an appropriate set of SNPs. To deal with the large search space involved in detecting gene-gene interactions, putative markers are prioritized in the search using a new measure, called Interaction Effect (IE), that quantifies the potential for a SNP to be involved in gene-gene interaction. This measure can also be used as a splitting criterion in random forest construction to define a cut-off value of a ranked list of SNPs. The prioritized SNP set is used to construct new combined features, which carry the information to account for gene-gene interactions.

This thesis proposes three new methods for calculating distances between genotype

profiles based on a kernel-based weighting function including: RFK, using the RF variable importance measure; MAFK, based on the minor allele frequency measure and EK, using the entropy measure. The distances can be subsequently incorporated for the purpose of disease classifications, cluster analyses and visualizations.

The feasibility of using genetic variation data for disease diagnosis and prognosis is explored with a new computational framework. The framework demonstrates the use of different phases of data processing and modelling to build reliable disease diagnostic and prognostic models using genetic variation data. The proposed feature selection approaches are incorporated in the framework to select an optimal subset of SNPs with the highest predictive power.

The proposed framework is empirically evaluated using two case studies of acute lymphoblastic leukaemia. The results demonstrate that the framework can produce highly accurate diagnosis and prognosis models. This thesis shows that a significant improvement of models' performance requires including interaction markers. The results are consistent with known biology while the accuracy of the produced models is also high.

Finally, several data reduction methods are used to visualize genetic variation data. For unsupervised-based visualization, they are compared based on the trustworthiness metric. For the supervised-based visualization, the performance is compared based on class discrimination. This thesis finds that the Neighbour Retrieval Visualizer method shows the best results for unsupervised-based visualization. Furthermore, in the supervised-based approach, the results highlight the importance of using feature selection to remove insignificant features. The visualization has the potential to assist clinicians and biomedical researchers in understanding relationships between patients and has the potential to lead to delivery of advanced personalized medicine.

The methodologies and approaches presented in this thesis emphasise the critical role that genetic variation data plays in understanding complex disease. The availability of a flexible framework for the task of disease diagnosis and prognosis, as proposed in this thesis, will play an important role in understanding the genetic basis to common complex diseases. A comprehensive validation of the methods and approaches embedded in the framework is a matter of applying this framework to other complex diseases.

Publications

Al-Oqaily, A., Tafavogh, S., Catchpoole, D., and Kennedy, P. 'A new computational framework for the task of disease diagnosis and prognosis', In preparation for submission to BMC Bioinformatics journal.

Al-Oqaily, A., Kennedy, P., Catchpoole, D., and S. Simoff, S., (2008). 'Comparison of visualization methods of genome-wide SNP profiles in childhood acute lymphoblastic leukemia', *Data Mining and Analytics 2008: proceedings of the Seventh Australasian Data Mining Conference (AusDM'08)*, **87**, pp. 111-121, Australian Computer Society.

Kennedy, P., Simoff, S., Catchpoole, D., Skillicorn, D., Ubaudi, F. and **Al-Oqaily, A.** (2008). 'Integrative visual data mining of biomedical data: Investigating cases in chronic fatigue syndrome and acute lymphoblastic leukaemia'. In S. J. Simoff, M. H. Boehlen, and A. Mazeika, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer-Verlag New York Inc, pp. 367-388.

The work in chapters 3 and 4 was turned into a project plan for a successful grant application: 'Implementing personalized medicine using global genomic similarity', Cancer Institute NSW Research Innovation Grant 2011, 10/RFG/2-23, Daniel Catchpoole, Paul Kennedy, $50,000.

Chapter 1

Introduction

Data mining and knowledge discovery have been applied to datasets in various industries including biomedical informatics. Modelling, data mining and visualisation in biomedical informatics address the problem of extracting knowledge from data originating from multiple sources, encoded in different formats or protocols, and processed by multiple systems. The major challenges in data mining in the area stem from the fact that biomedical data comes in many forms with a highly dimensional nature. The intention of this thesis is to construct models which incorporate large amounts of biomedical data in a manner which will alleviate the error induced by the high dimensional aspect of such data.

To develop an informed clinical management system, a systems biology approach can be utilized to deliver the best treatment strategy in the form of personalized medicine. Such an approach relies on information extracted from a range of biomedical datasets including (i) clinical data such as age, demographics, pathology test

results, etc.; (ii) patient outcome data (including results from drug trials and treatment protocols); (iii) gene expression data which represents the activity levels of a particular gene under different conditions; (iv) domain ontologies and other public databases freely available over the Internet; (v) genetic variation data, and others. This thesis focuses on genetic variation data in the form of genome-wide Single Nucleotide Polymorphism (SNP) datasets.

The human genome was found to contain a large amount of genetic variation in the form of sequence polymorphisms. Polymorphism (from Greek: poly "many", morph "form") is a variation of a DNA nucleotide sequence that has an allele frequency in at least 1% of the population (Cavalli-Sforza 1974). There are several types of polymorphism in the human genome: SNPs, repeated polymorphisms and insertions or deletions, ranging from a single base-pair to thousands of base-pairs in size (Tabor et al. 2002). Single Nucleotide Polymorphisms are the simplest but most abundant type of genetic variation among individuals which represents less than 0.1% of the whole genome. The genetic variations (SNPs) data utilized in this thesis are germline mutations[1].

Genetic variations, especially SNPs, are known to be a key feature in discovering disease-causing genes. The current interest in genetic variation studies is focused on

[1]The germline mutation of an individual is any detectable and heritable variation in the lineage of germ cells (eggs or sperms), this type of mutations can be transmitted from parents to offspring (NCI Dictionary of Cancer Terms 2011).

disease-gene association analyses. Such analyses are important in identifying which variants are associated with a specific disease. In the case of complex disease, such as cancer, identifying multiple genetic variants would be possible by conducting association analysis between a variant or a set of variants and a given disease. This association involves examining all genetic variations (differences) in large scale population samples of affected and unaffected individuals (Hirschhorn & Daly 2005).

Therefore a comprehensive evaluation of common genetic variations, through association of SNPs with common complex diseases at the genome-wide scale, is currently a promising area of human genome research. Identification of genetic variants that contribute to susceptibility of diseases, such as cancer, will provide fundamental tools into pathogenesis, diagnosis and treatment of human diseases (Carlson, Eberle, Kruglyak & Nickerson 2004).

Recent advances in biomedical science such as the completion of sequencing the human genome sequence (Venter et al. 2001), the deposition of millions of Single Nucleotide Polymorphisms into public databases (Sachidanandam et al. 2001), rapid improvement in SNP genotype technology and the initiation of the international HapMap Project (Gibbs et al. 2003, Manolio et al. 2008) have made association studies a powerful approach for mapping complex-disease genes by conducting studies based on the whole genome. These studies are called Genome-Wide Association

(GWA) Studies, in which a dense set of SNPs across the genome is genotyped to survey genetic variations for their role in disease or to identify the heritable quantitative traits that are risk factors for disease (Hirschhorn & Daly 2005, Reich et al. 2003).

Therefore, GWA studies have the potential to reveal the underlying mechanisms of many common complex diseases, without the need for any prior hypothesis about the possible causes of different complex diseases, thus, leading us into new insights of the etiology [2] of complex diseases. As a result there is a need for dual improvements in advanced technology for detecting and genotyping SNP data and for computational modelling to enable SNP data to be incorporated into genetics studies of complex diseases. This thesis addresses the latter problem.

Indeed in this research study data mining approaches are used to develop methods and techniques for the modelling, exploration and visualization of patient-to-patient relationships based on genome-wide SNP data. Eventually other data mining techniques such as case-based reasoning systems can be combined with the proposed approaches to assist clinicians in understanding the biological basis of diseases and may assist in clinical decision making.

[2] The causes of diseases or pathologies

1.1 Genetic Variation Studies

One of the fundamental subjects in genetic variation studies is to find an optimal subset of SNPs with the highest predicting power for different disease models. Most of the current genetic variation studies have been conducted to identify gene susceptibility markers that marginally (independently) contribute to common complex diseases such as diabetes, cancer, cardiovascular and other complex diseases (Rampersaud et al. 2007, Rioux et al. 2007, Saxena et al. 2007, Thomas et al. 2008, Yamauchi et al. 2010, Craddock et al. 2010, Feero et al. 2010, Zhernakova et al. 2011).

These studies essentially rely on evaluating one marker at a time based on the assumption that disease susceptibility genes can be identified through their independent contribution to disease variability. Although GWA studies have achieved great success in identifying more than 600 common genetic variants associated with common diseases, for most diseases the proportion of genetic variation explained by those variants is limited (The National Human Genome Research Institute 2003). Genome-wide association studies are mainly conducted using statistical-based methods and are used to discover genetic factors that contribute to susceptibility to diseases. SNPs are selected based on a particular p-value threshold that shows a significant or potential association with the targeted disease. Factors that show a high statistically significant level of association are chosen for further analyses.

In contrast there is an evidence that complex diseases are caused not by single genes acting alone but are the result of complex non-linear interactions among genetic factors, with each gene having a small effect on disease risk (Musani et al. 2007, Wu et al. 2010, Wang et al. 2011). For that reason there is a critical need to implement new approaches that can take into account non-linear gene-gene interactions in searching gene susceptibility markers that jointly cause complex diseases. Hence this thesis is dedicated towards proposing new approaches to resolve the given difficulties.

1.2 Motivation and Challenges

To deal with the genetic complexity of complex diseases, several analytical and computational methods have been developed to detect and model disease susceptibility genes accounting for the main effects and the gene-gene interaction effects of complex diseases. However, these studies were conducted on relatively small datasets with only hundreds of SNPs or using small simulated data (Oh et al. 2011, Cordell 2009, Motsinger-Reif et al. 2008, Millstein et al. 2006, Sha et al. 2006, Wu et al. 2010). Considering the large scale of genetic variation data, which is now available in genome-wide experiments, testing each marker for the main effect and different orders of gene-gene interactions is computationally infeasible. The dimensionality involved in such studies containing large numbers of SNPs is such that traditional statistical-based methods cannot be used without having a prohibitively large sample size of

individuals (Pagano et al. 2000, Storey & Tibshirani 2003, Balding 2006, Todd 2006). This problem is referred to as the "curse of dimensionality": as the number of SNPs increases the number of possible interactions between genes increases exponentially and the produced models become unstable.

Therefore, there is growing interest in applying non-parametric predictive models to understand genetic variation data. Data mining and machine learning approaches are examples of these models. These approaches are generally model-free and able to detect non-linear interactions in such high-dimensional data (Wu et al. 2010). These approaches can be used in a specific framework that can handle the computational complexity of genetic variation data and the large number of interactions to be examined.

Although current data mining and machine learning approaches used in GWA studies have been promising for discovering the genetic basis for complex diseases, the scalability issue is still problematic. Searching SNP data with a large number of markers, that is, more than 10,000 SNPs is the main limitation of these approaches. A flexible framework to deal with these difficulties would be of great interest for genome-wide studies (McKinney et al. 2006, Liang et al. 2007). Such a framework would use a multi-step procedure to reduce the computational complexity of genome-wide studies, especially for the task of extracting markers involved in gene-gene interactions. Indeed, this thesis is mainly dedicated to developing such a framework.

1.3 Problem Statement

This thesis hypothesizes that genetic variation data, as assessed by genome-wide SNP profiles, will be effective for modelling, data mining and visualization of patient-to-patient relationships and eventual clinical outcome.

For modelling different aspects of any complex disease, there is a need to reliably identify patients at greater risk of not responding to current treatment and who can then undergo modified therapy. The ultimate goal of such work, is to develop computational tools which will allow doctors and researchers to examine patients' genetic background in order to answer several research questions including identifying those patients who have a high chance of relapse using established therapy, to learn why and to help choose treatment strategies which best suit each individual patient, more specifically, translating genomic knowledge into public health and medical care in the form of personalized medicine. Based on that, this research study endeavours to answer the following research questions:

Q1: Are individual genetic risk factors, identified in the context of genome-wide SNP data, considered highly accurate for predicting a disease status?

Q2: Is there a computationally feasible approach for identifying multi-way interactions between SNPs that are contributing to a given disease status?

Q3: Can genome-wide SNP data be utilized to calculate a distance between any two genetic profiles to reflect the key points of difference and similarity?

Q4: Can genome-wide SNP data be used to accurately classify genetic profiles into different disease classes/subtypes?

Q5 Can genome-wide SNP data be utilized for the comparison, clustering and visualisation of patient-to-patient relationships?

The answers to these questions will be validated in the context of complex datasets from real genetic-based studies. Although the domain application of this research study is the Childhood Acute Lymphoblastic Leukaemia (ALL), approaches and methods proposed in this thesis have also the flexibility to be extended to other complex diseases such as heart diseases, diabetes and inflammatory diseases.

1.4 Scope and Contributions of the Thesis

Whilst the main goal of GWA studies is to identify the optimal set of SNPs that can be used to explain the genetic basis of different diseases, an exhaustive search of all possible SNP subsets to perform feature selection is computationally infeasible for such high-dimensional data. Therefore, employment of different feature selection strategies for developing an accurate disease model is a challenging task. In this work, non-parametric data mining approaches have been chosen to deal with the

"curse of dimensionality" problem induced by the data and non-linear interactions between multiple markers and the disease outcome. The proposed approaches are used for resolving different disease models including disease diagnosis, prognosis, and visualizing patient-to-patient relationships.

The approaches developed in this thesis use feature selection methods to select genetic markers that account for marginal effects as well as gene-gene interactions associated with disease status. Applying only feature selection methods based on traditional methods will result in choosing markers that marginally contribute to disease and will ignore the gene-gene interactions, which may be useful for producing highly predictive models for large scale genome-wide studies.

By analysing the main challenges of the given domain, this thesis identifies five contributions to knowledge. These contributions are first listed and then explained below.

1. A *feature selection approach, based on a random forest procedure*, is proposed. The proposed approach is used for the task of feature selection of SNPs and to determine the importance of each marker towards a given disease status (feature weighting).

2. An *interaction effect measure* for selecting markers of potential gene-gene interaction is proposed. The measure can be incorporated with other data mining

methods to reduce the computational complexity of detecting gene-gene interaction markers.

3. New *distance measures* for computing similarities between genetic profiles is proposed.

4. A new flexible *framework for the task of disease diagnosis and prognosis* is introduced. Methods and techniques proposed in the previous three contributions are incorporated in the proposed framework.

5. *Models for the comparison and visualisation* of patient-to-patient relationships, based on genome-wide SNP profiles, are employed.

Each of these contributions is validated by application to real-world problems of leukaemia studies.

1. A feature selection approach for SNP selection and weighting

The critical question that comes to mind in dealing with high-dimensionality aspect of genetic variation datasets is how to select an optimal set of SNPs with the highest discriminative power. In principle, many feature selection methods can be applied to do that. However, because of the "curse-of-dimensionality" problem with genetic variation datasets, most classical feature selection approaches either will not be effective for selecting an optimal set or are not computationally feasible. To deal

with this problem, non-parametric machine learning techniques have been chosen for the purpose of feature selection, weighting, and prioritizing of SNPs as well as for calculating distance measures between patients. Each of these approaches will be jointly considered for building different disease models using data from genome-wide studies.

Indeed, Chapter 3 proposes new feature selection methods and approaches in the context of genomic studies. The proposed methods and approaches are used to deal with three different tasks, namely, feature selection, weighting and detecting gene-gene interaction effects. First, to deal with the task of detecting marginal-based association markers, a Random Forest (RF) based approach, named RF-RFE, is proposed for the task of feature selection and weighting. In this method, an iterative RF-based procedure is introduced to select SNPs with marginal contribution to a given disease model. The proposed method is also used to weight the importance of markers to the disease of interest. The weights can then be used to calculate similarities between different genetic profiles.

2. A measure for detecting gene-gene interaction effects

It is computationally infeasible to search for all high-order interactions among SNPs in a genome-wide association study. One approach is to filter out a subset of genetic variations with high quality (i.e. those likely to have non-linear interaction) that

can then be efficiently analysed for the task of detecting gene-gene interactions. For this task, a new approach for SNP prioritization is defined, called Interaction Effect (IE). The IE measure can be used to search for a set of SNPs that has the potential to be involved in gene-gene interaction. The set of SNPs selected, based on the IE measure, is used to construct new combined features which carry the information that account for gene-gene interactions. A feature induction approach is used to construct the new combined features. Based on the constructed features set, a new approach is proposed to obtain an optimal set explaining a given disease model. The proposed RF-RFE approach was used for this purpose.

3. Distance measures for genetic profiles

For calculating distances between genotype profiles, kernel-based approaches are proposed. The proposed distance measures use SNPs that have been selected, weighted and constructed based on the proposed approaches to calculate distances between different genetic profiles. Three distance measures are defined using a kernel-based weighting function. These measures include random forest, minor allele frequency and entropy based kernel functions.

4. A framework for disease diagnosis and prognosis

The past decade has reported many new genetic associations that have been identified by GWA studies (Nolte et al. 2010, Craddock et al. 2010). Despite the exponential

increase of such studies, few reported approaches have been developed for the analysis of microarray analysis for disease diagnosis and prognosis (Kruglyak 2008, Frazer et al. 2009). Genome-wide association studies are primarily limited to the fact that the individual effect sizes of the reported associations are mostly small (McCarthy et al. 2008). Furthermore, there are also arguments that many disease-associated markers have not yet been identified, and the process of generating reliable prediction models may improve as more markers are included (Zhao & Liu 2009, Rosenberg et al. 2010).

The main idea of generating prediction models is to combine markers, some having direct predictive (associations) power and others low association but having non-linear (indirect) relationships, in mapping genotype data to a given phenotype. Therefore, combining multiple markers in a single model can be stronger, especially by including markers that have non-linear associations. One of the main focal points of chapter 4 is the detection and characterisation of gene-gene interactions in genetic studies. Detecting and characterizing gene-gene interactions in such studies will strongly contribute to a better understanding of etiology, prediction, discovery and prevention of most common complex diseases.

For this purpose, methods and techniques, developed in chapter 3, are incorporated to propose a new modelling framework for the task of disease diagnosis and prognosis, in chapter 4. The proposed framework utilizes machine learning and data mining techniques to perform these tasks. The proposed framework is applied to

biomedical data, more specifically SNP data, derived from the high-throughput genotype technology through GWA studies.

The proposed framework, consisting of four phases, can be seen as a flexible computational approach for understanding complex diseases. The process starts in a data preparation phase, where data from different genetic models is generated and filtered using a quality control check. Then, data proceeds into phase 2, where feature selection, prioritizing and construction are used to generate an important list of features that can be used to classify patients to different disease classes/subtypes. Methods and approaches proposed in chapter 3 are used for this purpose. In phase 3, the combined subset of selected features is used to build classification models. The Support Vector Machine (SVM) methods are used for this purpose. Finally in phase 4, the models are evaluated using several validation algorithms.

The proposed multi-phase framework is evaluated to build reliable disease diagnostic and prognostic classification models. Several SNP microarray datasets are employed to show the feasibility of the proposed framework. The results suggest that the proposed prediction approaches can effectively define important markers, which are consistent with known biological findings while the accuracy of the produced models is also high.

In fact, the heart of the proposed framework is the feature selection phase, which endeavours to use methods for selecting disease markers that account for marginal

effects as well as gene-gene interactions. Therefore, both marginal and interaction markers are included in the final model building. Applying traditional feature selection methods will result in choosing markers that marginally contribute to disease and ignore gene-gene interactions that would be useful for producing highly predictive models for large scale of genome-wide studies.

5. Visualizing Genome-wide SNP Profiles

Finally, for the task of visualizing patient-to-patient relationships, chapter 5 employs several data reduction methods for visualizing genetic variation data. Information visualization is considered as a direct way to help browse the datasets. The visualization results can be seen as an important tool that can be used to assist clinicians and biomedical researchers in understanding the different structure of patients and to compare different clusters in the visualization.

The main challenge in visualizing genetic variation datasets stems from the high dimensionality of the data. To deal with large amounts of genetic variation data, different dimensionality reduction methods were applied to genome-wide SNP profiles of leukaemia patients to determine the best method for visualizing this type of data. Visualization approaches were compared based on measures such as trustworthiness and continuity of the resultant visualizations. These measures evaluate how much a given visualization preserves the local relationships (neighbourhoods) of data points.

Visualization results of multi-class SNP datasets showed the importance of using feature selection methods for removing insignificant features. The resultant visualizations were more accurate in discriminating the major characteristics of the utilized dataset. In particular, prior knowledge or domain-driven dissimilarity measures may improve the performance of visualization using data reduction methods.

1.5 Thesis Outline

This thesis is organized as follows: Chapter 2 gives an extensive review of the genetic variation domain, which includes an overview of genetic variation as genetic-based association studies and the computational analysis of conducting GWA studies. This chapter also reviews the current direction in genetic variation studies.

In chapter 3, new methods and approaches are proposed to deal with different feature selection tasks of the given domain. The proposed approaches are used for the purpose of feature selection, weighting, and prioritizing of SNPs, as well as distance measure calculations. Each of these tasks will be jointly considered for building different disease models in dealing with data from genome-wide studies. Methods are also introduced for calculating distances between different genotype profiles.

Chapter 4 proposes and validates a new multi-phase computational framework for disease diagnosis and prognosis. Methods and techniques proposed in chapter 3 are embedded in the proposed framework. Several SNP microarray datasets are employed

to show the feasibility of the proposed framework.

Chapter 5 applies several dimensionality reduction methods to visualize genetic variation data. Measures such as trustworthiness and discontinuity, are used to compare and evaluate the resultant visualizations. A supervised-based visualization is also demonstrated using the proposed feature selection measures prior to data reduction procedures.

Finally, in chapter 6, the thesis is concluded including a summary of the work in this thesis and proposing a number of future research directions

Chapter 2

Literature Review and Background

Common diseases such as cancer, diabetes, inflammatory and heart diseases are caused by combinations of multiple genetic and environmental factors (King et al. 2002). Discovering these genetic markers will provide fundamental tools into pathogenesis, diagnosis and treatment of human diseases. Many studies have endeavoured to discover single gene disorders based on linkage analysis[1]. Such analyses have primarily been used for searching for causative variants in chromosome regions. By contrast, linkage studies have not been very successful for discovering many genetic variants that affect common complex diseases, where each variant has a small contribution to disease risk (Botstein & Risch 2003, Risch 2000).

In the case of complex diseases, identifying multiple genetic variants would be possible by conducting association analysis between a specific variant and a disease. This association involves examining all genetic differences in a large number of affected

[1]Genes are mapped by typing genetic markers in families to identify regions that are associated with disease within pedigrees more often than are expected by chance. Such linked regions are more likely to contain a causal genetic variant.

individuals with unaffected controls (Risch et al. 1996). A number of association studies, focused on candidate genes, have led to the discovery of genetic risk factors associated with common disease (Ozaki et al. 2002, Altshuler et al. 2000, Pennacchio et al. 2001, Hugot et al. 2001).

However, recent advances in biomedical science such as the completion of sequencing the human genome sequence (Venter et al. 2001), the deposition of millions of Single Nucleotide Polymorphism (SNP) into public databases (Sachidanandam et al. 2001), rapid improvement in SNP genotype technology and the initiation of the international HapMap Project (Gibbs et al. 2003, Manolio et al. 2008) have made association studies a powerful approach for mapping complex-disease genes by conducting studies based on the whole genome. These studies are called Genome-Wide Association Studies, in which a dense set of SNPs across the genome is genotyped to survey genetic variations for a role in disease or to identify the heritable quantitative trait that is risk factor for disease (Hirschhorn & Daly 2005, Reich et al. 2003).

The National Human Genome Research Institute (2003) confirmed that 99.9% of the three billion base-pairs of the human genome are identical between any two individuals on the planet. The remaining 0.1% of differences comprises more than 10 million base-pairs (or genetic variations) scattered across the whole genome. The human genome was found to contain a large amount of genetic variation in the form of sequence polymorphisms. A polymorphism is a variation of DNA sequence that has

an allele frequency of at least 1% of the population (Cavalli-Sforza 1974). There are several types of polymorphism in the human genome: SNPs; repeated polymorphisms; and insertions or deletions, ranging from a single base-pair to thousands of base-pairs in size (Tabor et al. 2002). Single Nucleotide Polymorphisms are the simplest but most abundant type of genetic variation among individuals with between 1 to 10 million existing in the human genome (Donnelly 2004). These common SNPs are thought to account for around 90% of human polymorphisms (Carlson et al. 2003, Reich et al. 2003).

It is thought that each SNP arose from a single historical mutational event (Palmer & Cardon 2005), as the most recent common ancestor of any two humans is around 10^4 generations (Gibbs et al. 2003). It is important to clarify that the genetic variations (SNPs) data used in this research study are germline mutations. The germline mutation is any detectable and heritable variation in the lineage of germ cells (eggs or sperms). This type of mutation can be transmitted from parents to offspring. A germline mutation gives rise to a constitutional mutation in the offspring. That is, a mutation that is present in the DNA of every cell in the body of the offspring (NCI Dictionary of Cancer Terms 2011). In contrast, other types of mutations, such as somatic or epigenetic mutations, are changes in the genetic structure that are neither inherited nor passed to offspring (Bird 2007). Epigenetic variations, for example, are changes in gene expression or cellular phenotype caused by mechanisms other than

changes in the underlying DNA sequence. It refers to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence (Bird 2007).

There are a number of reasons for using SNPs rather than other types of genetic polymorphisms for studying complex disease (Risch 2000), such as the fact that SNPs are spread throughout the genome and that some of these polymorphisms might themselves be functional (Collins et al. 1997, Kruglyak 1997); the existence of correlations between a group of adjacent SNPs could be used to enhance gene-mapping and to highlight recombination hotspots and the lower mutability (stability) of SNPs compared to other types of genetic polymorphisms (Chakravarti 1998). These factors could allow more consistent estimation of gene-disease association studies (Brookes 1999, Stallings et al. 1991).

There are a number of fields that utilize SNP technologies in improving our understanding of complex diseases and in formalizing the future of health care. Some of these fields extend from association-based candidate polymorphism testing; diagnostics and risk profile; SNPs involved in functional proteomics and gene therapy; pharmacogenomics and predication of response to non-pharmacological environmental stimuli (Palmer & Cardon 2005).

As a result, there is a need for dual improvements in advanced technology for detecting and genotyping SNP data and for computational modelling to enable SNP

data to be incorporated into genetic studies of complex diseases. This chapter is devoted for reviewing the domain of genetic variation studies including the type of data used in this domain. Current research studies that have been applied to this domain are also reviewed including GWA studies. The second part of this chapter reviews data mining and machine learning methods and approaches (i.e. computational methods) that can be applied to this domain, in particular, supervised and unsupervised-based data mining methods. In fact, the reviewed methods and approaches have been directly and indirectly used in this thesis to propose new methods. The proposed methods were utilized to deal with different tasks in the given domain.

The reminder of this chapter is organized as follows: Section 2.1 introduces the main concepts of the genetic variation of the human genome including basic concepts, GWA analyses, GWA approaches, Markers for GWA studies and the efficiency of conducting GWA studies. Section 2.2 describes in detail the computational part of genetic variation studies, more specifically, genome-wide association analysis. Section 2.3 discusses the relationship between the current direction of genetic variation studies and the research direction of this thesis. Section 2.4 reviews the current data mining and machine learning methods and approaches that have been applied to GWA studies. These methods include supervised and unsupervised based methods. Finally, section 2.5 describes the case study that has been used in this thesis.

2.1 Genetic Variations

2.1.1 Basic Concepts

Since the goal of genetic variation analysis is disease-gene association analysis, the following subsection provides an introduction into basic concepts of population genetics and the motivation they provide for the current research of association analysis and more broadly genome-wide association analysis.

Haplotype, Genotype and Phenotypes

In the case of diploid organisms (such as human), there are two copies of each chromosome. Each chromosome is from a different origin, one is inherited from the father (paternal chromosome) and the other is inherited from the mother (maternal chromosome). Thus, each individual has two alleles of a given SNP (Brookes 1999, Sadava et al. 2006). When the two alleles of a given SNP are similar, the SNP is called **homozygous** and when they are different the SNP is called **heterozygous**. For a single SNP, one is designated the "major" allele and the other the "minor" allele, based on their observed frequency in the general population (Crawford & Nickerson 2005).

Genotypes and haplotypes are the most important and basic concepts related to population genetics of genetic variation studies. It is important to have a clear understanding of these terms and the process used to induce genotype and haplotype

data. To illustrate these terms, a simple example of chromosome segments from three individuals (i.e. three pairs of chromosomes) is given in figure 2.1. Figure 2.1A shows these segments and their corresponding haplotype and genotype data. As Figure 2.1B shows, a set of SNPs present on one chromosome is referred to as a **haplotype** (**Haplo**id geno**type**). At a specific SNP, a person can have one of the several **genotypes**: homozygous for the major allele, heterozygous when a SNP has different alleles or homozygous for the minor allele (see Fig. 2.1C).

While haplotypes and genotypes represent the allele information of a target SNP on chromosomes, a **phenotype** is the observable properties of an individual as they develop under the combined influence of the individual's genotype and the effect of environmental factors (Sadava et al. 2006). For example, In the case study used in this thesis the phenotype of an individual is either leukaemia or no leukaemia. In general, the individuals with disease are referred to as *cases*, while the ones with no disease are *controls*.

The fundamental difference between the haplotype and genotype of an individual is that the allele of a SNP is assigned to a specific chromosome. Essentially, each individual has two haplotypes of its genome, representing the maternal and paternal chromosomes (Crawford & Nickerson 2005). While haplotype represents the set of SNP's alleles along the same chromosome (see Fig. 2.1B), the genotype is the combined information of alleles on the two chromosomes (see Fig. 2.1C). The experimental

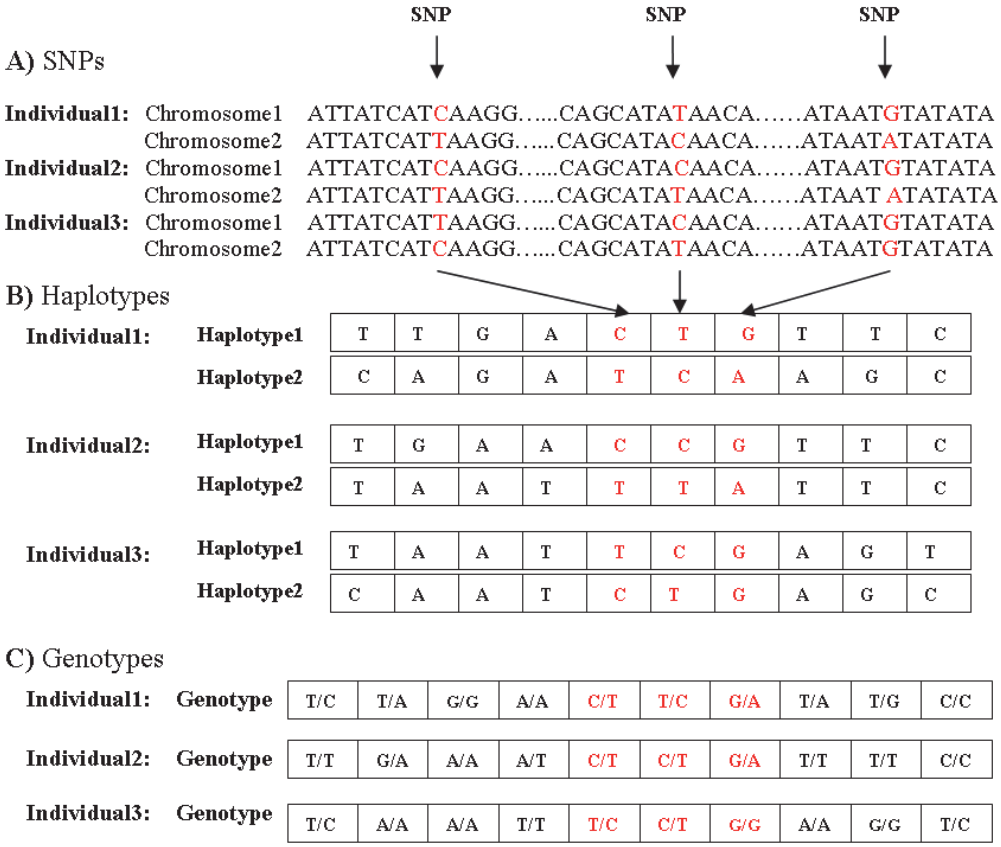

Figure 2.1: Haplotypes and Genotypes

procedure that is used to detect genotype information of individual's SNPs is called genotyping. The following subsection will describe some information associated with haplotype and genotype concepts.

Linkage Disequilibrium and Block Structure of Human Genome

As mentioned above, the specific set of alleles observed on a single chromosome or part of a chromosome is referred to as a haplotype. When meiosis[2] takes place, new haplotypes are formed by additional mutation or by recombination when the paternal

[2]The production of germ cells

and maternal chromosomes exchange corresponding segments of DNA, resulting in chromosomes of the descendent (Pääbo 2003). One feature of the new generated haplotypes is non-random associations between the SNPs in these haplotypes, known as Linkage Disequilibrium (LD) (Gibbs et al. 2003).

Linkage disequilibrium is the correlation between nearby markers (SNPs) such that the alleles at neighbouring markers (observed in the same haplotype) are associated within a population more often than if they were unlinked (Hirschhorn & Daly 2005). Such an association between SNPs declines with distance, due to the increases of the likelihood of recombination between two SNPs if they are distant. Linkage disequilibrium is considered to be the key property for disease association studies across the human genome.

Many empirical studies have been conducted to understand the LD structure of the human genome (Reich & Lander 2001, Qin et al. 2006, Dawson et al. 2002). These studies have shown regions of high significant level of LD within SNPs in these regions. This result support the hypothesis that the human genome can be partitioned into discrete regions, known as "blocks" (Gabriel et al. 2002, Daly et al. 2001, Patil et al. 2001), so that the SNPs have high LD within a block and low LD between blocks. The strong association in many chromosome regions means that there are only a few haplotypes that account for most haplotype diversity among people in those regions (Gibbs et al. 2003).

In general, the human genome has a block-like structure (Schulze et al. 2004, Ding et al. 2005) and the strong associations between SNPs in these blocks have a practical value. Genotyping only a few, carefully chosen SNPs in these regions will provide enough information to capture much of the information about the common haplotypes in these regions (Gibbs et al. 2003). As a result, genotyping a few SNPs (known as Tag-SNPs) is required to identify the common haplotypes in a region (Daly et al. 2001, Johnson et al. 2001, Carlson et al. 2003, Goldstein et al. 2003).

Based on the above assumption, where the human genome is considered as block-like and the extent of the association between nearby markers varies across the genome, it is not efficient to use markers selected randomly or even spaced in the genome (Gibbs et al. 2003). Instead, the pattern of association can be used for efficient selection of Tag-SNPs. Based on empirical studies, it has been estimated that 200,000-1,000,000 Tag-SNPs out of 10 million common SNPs can be used to capture most of the haplotype and genotype diversity in the human genome (Gabriel et al. 2002, Patil et al. 2001, Carlson et al. 2003). The datasets used in this thesis are generated based on SNP chips that have been designed using tagSNP analyses.

2.1.2 Genome-wide Association Analysis

There are many possible approaches for discovering genes that underlie common diseases. These approaches fall broadly into two main categories: candidate-gene studies

which use either association or resequencing methods and genome-wide studies that include linkage mapping and genome-wide association studies (Hirschhorn & Daly 2005).

Linkage Analysis

Linkage analysis was the first method used to identify disease genes, and has been successful for identifying genes that underlie monogenic-disease (Mendelian disease), where the disease is caused by highly penetrant[3] markers. Furthermore, linkage analysis has been used to identify genes that underlie a number of common diseases. In some cases, genome regions that show significant linkage to the disease have been identified, leading to the discovery of some genes that contribute to a small part of the disease-causing genes (Hirschhorn & Daly 2005), such as, inflammatory bowel disease (Hugot et al. 2001, Stoll et al. 2004), schizophrenia (Stefansson et al. 2002), and type 1 diabetes (Nistico et al. 1996). However, for most common diseases, linkage analysis has not been that successful for discovering disease-causing genes for complex forms of the disease (Altmüller et al. 2001), where each gene explains only a small fraction of the overall heritability of a given disease (Hirschhorn & Daly 2005).

In general, linkage analysis is not a powerful approach for identifying common genetic variants that have modest effects on disease (Risch 2000, Cardon & Bell 2001,

[3]The proportion of individuals with a specific genotype who manifest the genotype at the phenotypic level. For example, if all individuals with a specific disease genotype show the disease phenotype, then the genotype is said to be "completely penetrant".

Tabor et al. 2002). Most common diseases have a complex architecture, for which the disease is determined by the total sum of interactions between multiple genetic and environmental factors. There are probably hundreds of susceptibility markers that increase the risk of each complex disease (Wang et al. 2005).

The variant frequencies of common disease are largely unknown. Some studies have supported the hypothesis that common variants influence common disease with allele frequency of $> 1\%$. This is called the common disease/common variant (CDCV) hypothesis (Chakravarti 1999, Reich & Lander 2001, Lander et al. 2001). The extreme alternative to CDCV is the heterogeneity hypothesis (multiple rare hypotheses), in which variant frequencies have low population frequencies with allele frequency of $< 1\%$ (Smith & Lusis 2002). Nevertheless, Wang et al. (2005) suggest that the allelic spectra of most common diseases probably fall between these two hypotheses. In addition, because it is acceptable to suppose that common alleles as well as rare alleles will contribute to common disease, Hirschhorn & Daly (2005) has claimed that a complementary method to linkage analysis is desirable, as linkage analysis has less power for detecting common alleles with low penetrance.

Candidate-gene Studies

Due to above limitations of linkage-based analyses, researchers have begun to apply other approaches. One of these alternatives is candidate-gene studies which use either resequencing or association methods. Using these approaches, genes are selected for

further study, on the basis of their location in regions of linkage, or on the basis of other evidence that might have a role in the etiology of complex diseases (Cardon & Bell 2001, Tabor et al. 2002). The first proposed and comprehensive candidate gene approach was obtained by resequencing the candidate genes in a population-based sample of affected and unaffected individuals (case-control study), and searching for markers that were enriched or depleted in complex diseases (Cohen et al. 2004). This method was successfully applied to test the variation in plasma levels of high-density lipoprotein cholesterol. However, because such studies are still laboratory expensive, they have been limited to the coding region of only one or a few genes. In addition, the inability of this method to interpret or associate some results of the rare or no-coding discovered variants has limited the relevance of such a method (Hirschhorn & Altshuler 2002).

Alternatively, the candidate-gene method can use an association approach to study common allelic variants. This method is simpler and cheaper than complete resequencing of candidate genes (Hirschhorn & Daly 2005). Several studies have recommended this method to identify common variants underlying complex diseases (Risch et al. 1996, Tabor et al. 2002, Carlson, Eberle, Rieder, Yi, Kruglyak & Nickerson 2004). Using association studies, the allele frequencies of a particular variant are compared in a population sample of affected and unaffected individuals.

Candidate-gene association studies have been applied on a number of studies,

and have identified several genes that are known to contribute to susceptibility to common diseases (Cardon & Bell 2001, Lohmueller et al. 2003, Hirschhorn et al. 2002). This kind of study is facilitated by using the feature of indirect association of LD. Although candidate-gene association studies have been successful for identifying some cases of common-disease genes, there are some factors limiting this method. Candidate-gene studies rely on the fact that predicting the identity of a correct gene or genes, usually on the basis of biological hypothesis or location of the candidate gene within previously determined region of linkage. Such information is not always available. The other limitation is that the candidate-gene approach will clearly be inadequate to fully explain the genetic basis of some diseases, especially when the physiological defect of a disease is unknown (Hirschhorn & Altshuler 2002).

Genome-wide Association Studies

As a result, a genetic study having a complete coverage of the human genome is required for mapping genes that underlie common complex diseases. This type of study is called a genome-wide association study. Using this approach, most of the genome is surveyed for causal genetic variants. In most cases, where no assumptions are made about the genomic location of causal variants, this approach could exploit the strength of association studies without having knowledge about causal variants. Therefore, even in the absence of evidence about the function or location of causal

genes, genome-wide association studies could be an unbiased and comprehensive option for mapping complex diseases (Hirschhorn & Daly 2005). This approach has the advantage over candidate-gene studies in avoiding testing of only a limited number of genes which only explain a tiny fraction of the genome.

As mentioned above, recent advances in biomedical science have made GWA association studies a powerful approach for mapping complex genes by conducting studies based on the whole genome. To this end, conducting association studies based on the whole genome is not a straightforward task. Of the difficulties that influence any given association study, some are related to the study design and others to the computational part of these studies. Indeed, the applied approaches in this thesis are based on GWA studies. This thesis mainly addresses the computational part of GWA studies.

2.1.3 Genome-wide Association Study Approaches

Interest in genome-wide association studies was suggested as an alternative approach to candidate-gene approach and linkage analysis. Risch et al. (1996) noted that association studies have considerably greater power than linkage analysis to detect genetic variants with small or moderate phenotype effects. They also suggest that the number of variants can be reduced using the advantage of linkage disequilibrium. This method was later termed the indirect approach.

As discussed in section 2.1.1, variants in strong LD are likely to be inherited

together. Consequently, genotyping a subset of these markers (tagging) can be used as proxies of the entire set. The indirect approach was first applied to tag markers in a small group of genes and later expanded to include variations over the whole genome (Collins et al. 1998, Kruglyak 1999). There are two general strategies for indirect GWA studies. The first uses quasi-random that are spread across the whole genome. The second uses a subset of LD-based tag SNPs that are specifically chosen to cover the entire genome (Jorgenson & Witte 2006*b*). To make the second approach feasible, the international HapMap project was established with an initial goal of creating a set of 600,000 LD tagging SNPs (Gibbs et al. 2003, Altshuler et al. 2005). The second phase of this project was recently completed, resulting in publicly available data of more than 3.9 million validated SNPs, as well as information about the LD between SNPs, from 269 individuals from multiple populations. In addition, ten HapMap "Encyclopedia of DNA Elements" (ENCODE) regions across the human genome of 16 individuals across the HapMap populations were genotyped, providing a more complete set of SNPs that can be used to validate and evaluate the performance of genotyping sets (Manolio et al. 2008).

As a result, and based on the SNP data available through the HapMap project, various gene-based discovery companies and projects such as Illumina, Affymetrix and SeattleSNP Program for Genomic Application have identified the SNPs that can be used for gene-centric approaches to GWA studies. Such studies can use indirect

approach that focuses on markers that lie in gene-coding or cis-regularity regions which are known to be associated with the functional-part of the genes. In the next two subsections, the two main approaches of GWA studies, indirect and gene-centric approaches are described.

Indirect Approaches

As mentioned in section 2.1.1, the genetic variations of human genome were found to be partitioned into segments of high LD known as blocks. Variants in these blocks are strongly correlated with each other, and most chromosomes carry only one of a few common haplotypes (Daly et al. 2001, Patil et al. 2001, Gabriel et al. 2002).

Based on this assumption, most of the 10 million common SNPs in the genome have groups of SNPs that are all nearly perfectly correlated with each other. Genotyping of one SNP perfectly correlated with neighbouring SNPs would be enough. This SNP is known as a tag SNP. Therefore, a few tag SNPs can be chosen such that the combination of these SNPs will capture most of the common variations within the region (Johnson et al. 2001, Gabriel et al. 2002). In contrast, genomic regions that show low LD need to genotype proportionally higher density of tag SNPs to survey most of the variation in that region.

Based on previous studies and the phase II HapMap data, several hundred thousand well-chosen SNPs are adequate to cover most of the common variations in the human genome (Daly et al. 2001, Patil et al. 2001, Dawson et al. 2002, Gabriel et al.

2002). A larger number of tag SNPs is required in African populations, because this population has more variation and low LD (Gabriel et al. 2002, Crawford et al. 2004). The precise number of SNPs has not been determined and depends on the method used to select SNPs, the LD structure between blocks and the method used to select SNPs in the region of low LD (Carlson, Eberle, Rieder, Yi, Kruglyak & Nickerson 2004, Dawson et al. 2002).

Examples of commercially available products, are the Illumina HumanHap-300 set of 317,000 SNPs and the full HumanHap-500 set of 500,000 SNPs.

Gene-centric Approaches

Many theoretical and empirical studies have supported the advantage of focusing directly on the variants within genes compared to indirect approaches. Theoretically, a gene-centric approach could decrease the genotype burden and could also be more complete regarding the coverage of genes which is important for detecting causal variants (Jorgenson & Witte 2006*a*).

Based on studies that have been conducted on Mendelian disorders (monogenic diseases), Botstein & Risch (2003) suggest that association studies should only focus on missense SNPs, as most of the gene-causing mutations in these diseases are missense mutations ($\simeq$60%) and typically a gene contains one or two missense SNPs. Table 2.1 summarizes different types of SNPs. That is, genotyping 10,000-60,000 SNPs could be enough to conduct GWA studies (Jorgenson & Witte 2006*a*). Although this

assumption can be valid, there are two problems associated with it. Firstly, identifying all common missense SNPs would be prohibitively expensive and depends on the genotyping technology which might not be available until the near future (Shendure et al. 2004). Second, causal alleles (or mutations) for monogenic disease are highly penetrant and often lead to severe phenotypes, although these alleles often cause severe changes in protein function, but other mutations can also be involved in disease-causing mutation such as nonsense mutation, splicing, regularity, insertion or deletion mutation. Even for Mendelian disorders, an appreciable fraction of mutation is outside the coding region (Hirschhorn & Daly 2005). Nevertheless, the missense approach can still be a productive approach. Some missense mutations have been reported to be associated with complex diseases, because they are more likely to have functional consequences. Due to the lack of knowledge regarding common diseases risk alleles, the power of this approach is still unknown.

As a result, a convenience-based approach would be preferred. This approach is a logical tool to comprehensively survey most of the polymorphisms (or mutations) that have functional effects in complex diseases without the need to include non-functional polymorphisms that exist outside the functional regions of genes (as in the case of the indirect approach or LD-based approach). These mutations (SNPs) are not limited to missense SNPs. They also include nonsense SNPs, SNPs in 5' and 3' untranslated regions, and cis-regulatory regions that lie near the coding regions (which effect the

gene regulation) (Botstein & Risch 2003). Other SNPs that lie within introns, and particularly those that are not located near intron-exon boundaries, are less likely to have functional effects and are relatively less likely to be significant in GWA studies (Palmer & Cardon 2005, Tabor et al. 2002). Based on this set of SNPs and with the advantage of LD, this approach would be powerful in conducting GWA studies. But it is not clear whether either approach can provide a comprehensive survey of the genome.

An example of commercially available SNP-chips for GWA studies are the Illumina's Human-1 Genotyping beadchip set of 13,917 gene-centric SNPs and the MegAllel system marketed by Affymetrix which involves 12,000 SNPs. In this thesis, SNP chips that are based on both of these approaches have been used to empirically evaluate the proposed models and techniques, as described in chapters 4 and 5.

2.1.4 Markers for Genome-wide Association Studies

As discussed above, genetic variations have made association studies a powerful approach for mapping complex-disease genes by conducting studies based on the whole genome. Although conducting association studies based on genetic variation is a more practical approach than resequencing the whole genome which is laboratory expensive, it is also not practical or statistically feasible to genotype and test all SNPs in the genome to conduct association studies. Therefore, it is important to select

carefully a subset of SNPs to genotype from the whole set of SNPs. For example, it is desirable to study only those polymorphisms that affect the function of proteins or their expression, because these kinds of polymorphisms are more likely to affect the risk of disease (Tabor et al. 2002).

Although, in most cases, the information about the effect of polymorphisms on protein function is not available and is not trivial to obtain, it is most effective to evaluate all possible polymorphisms and prioritize them on the basis of their functional effects on genes. The polymorphisms of obvious molecular consequence are more likely to affect disease risk (Tabor et al. 2002). On the other hand, there should be a balance between selecting a specific set of SNPs and the level of coverage and power of conducting GWA studies.

Tabor et al. (2002) have studied the effect of different types of polymorphisms on disease risk based on candidate-gene studies and they described how these SNPs can be prioritized based on their genetic effects. Accordingly, such a prioritization is also applicable for GWA studies. Information about location and type of variants can also be used to prioritize polymorphisms.

For some polymorphisms, it is more likely that they have functional effects of a protein. For example, missense variants that alter an amino acid in a protein, or nonsense changes that result in a premature stop codon. These types of polymorphism account for most disease-causing variants, and therefore they should be

Table 2.1: Types of SNPs and their properties (from Tabor et al. (2002)).

Type of variant	**Location**	**Functional effect**	**Frequency in genome**	**Predicting relative risk of Phenotype**
Nonsense Coding	Coding sequence	Premature termination of amino-acid	Very low	Very high
Missense / non-synonymous (non conservative)	Coding sequence	Changes an amino acid in protein to one with different properties	Low	Moderate to very high, depending on location
Missense / non-synonymous (conservative)	Coding sequence	Changes an amino acid in protein to one with similar properties	Low	Low to very high, depending on location
Insertions / deletions (Frameshift)	Coding sequence	Changes the frame of the protein-coding region, usually with very negative consequences for the protein	Low	Very high, depending on location
Insertions / deletions (In frame)	Coding or non-coding	Changes an amino acid sequence	Low	Low to very high
Sense / synonymous	Coding sequence	Does not change the amino acid in the protein, but can alter splicing	Medium	Low to high
Promoter / regulatory region	Promoter, 5' UTR, 3' UTR	Does not change the amino acid into protein, but can affect the level, location or timing of gene expression	Low to Medium	Low to high
Splice sites/Intronson boundary	Within 10bp of the exon	Might change the splicing pattern or efficiency of introns	Low	Low to high
Intronic	Deep within introns	No known function, but might affect expression of mRNA stability	Medium	Very low
Intergenic	Non-coding regions between genes	No known function, but might affect expression through enhancer or other mechanisms	High	Very low

given the highest priority of genotyping in GWA studies (Tabor et al. 2002). Other polymorphisms, such as those that exist on the transcription region of genes, may directly or indirectly affect gene regulations, which may halt the process of transcription of the related gene. Therefore, it is reasonable to place a high priority on such polymorphisms.

However, even if a polymorphism in a coding region does not cause an amino acid change or if it is in the non-coding region, it can still affect gene function by altering

the stability, splicing or localization of the mRNA (Cartegni et al. 2002). Generally, the effect of non-coding polymorphisms cannot be predicted, except in some case when conserved sequence in splice sites are changed. Also, synonymous changes are less likely to affect gene function. Consequently, these types of polymorphism should be given lower priority for genotyping than other type of polymorphisms. Nevertheless, missense polymorphisms (conservative) have potential effects on mRNA stability. These types of polymorphism should have a higher priority more than polymorphisms that lie deep within introns (Risch 2000).

There are many empirical studies that have supported the strategy of prioritizing polymorphisms for association studies. Some of these studies were based on the evidence from mutational studies of Mendelian disorders. It was estimated in several of these studies that 80-90% of disease-causing mutations were due to changes in the coding regions of the disease-caused genes (e.g. Buyse et al. (2000) and (Couch et al. 1996)). Other studies were based on polymorphisms discoveries that evaluate variants and their frequency in hundreds of genes (i.e. candidate-gene based studies), such as the study conducted by Stephens, Smith & Donnelly (2001). Equally important, two other large studies found that variants in the coding regions are the least common variants, as can be seen in table 2.2, especially, nonsense and missense variants, frameshift and variants in spice sites (Cargill et al. 1999, Halushka et al. 1999). As a consequence of all of these studies, higher priority should be given for genotyping

functional-based variants, which would balance the effects of more common variants, and are more likely to discover genes that are involved in diseases.

Table 2.2: Typology of SNPs and their occurrence (from Risch (2000)).

Type	**Type Description**	**Number** (in thousands)
I	Coding, non-synonymous, non-conservative	60-100
II	Coding, non-synonymous, conservative	100-180
III	Coding, synonymous	200-240
IV	Non-coding, 5' UTR	140
IV	Non-coding, 3' UTR	300
VI	Other non-coding	$> 1,000$

As discussed above, other factors also effect the variant selection for GWA studies, such as allele frequencies of variants in the population to be tested for association. This factor will result in a statistical problem associated with the power to detect a significant association which depends on the size of sample and frequency of the allele to be tested (Risch et al. 1996, Lalouel & Rohrwasser 2002). An intermediate solution for this problem is to consider all SNPs with MAF frequencies of at least 5% for association studies (Risch 2000).

2.2 Computational Analysis of GWA Studies

2.2.1 An Overview

The ultimate goal of GWA studies is to identify a set of DNA variations that is highly associated with gene-causing diseases. Three different types of information

can be used to examine the association of genetic variations with the disease of interest. These include single-SNP, genotype and haplotype information. Single-SNP information can be used to examine the association between a specific variant and a disease such as linkage analyses. Genotype and haplotype information can be used with studies that examine the disease-gene association such as candidate-gene studies and GWA studies.

In the case of GWA studies, both genotype and haplotype information can be employed to conduct associations. However, haplotype analyses have several advantages compared to genotype analyses (Daly et al. 2001, Zhang et al. 2002). Genotypes usually do not have information about the source chromosome, known as phase. Thus, using genotype data, it could be possible to lose some information regarding the obvious associations that exist between haplotypes and a target disease.

Generally, obtaining haplotype information in large-scale association studies is prohibitively expensive and requires long operation time of biomedical technologies. Therefore, there are needs for computational methods for deducing haplotype information from genotypes. This process is called haplotype phasing, although using haplotype information that is directly obtained using biomedical technologies is more accurate than haplotype information that is computationally deduced. However, this would not be feasible until the near future when biomedical technologies become the standard techniques for obtaining haplotype information.

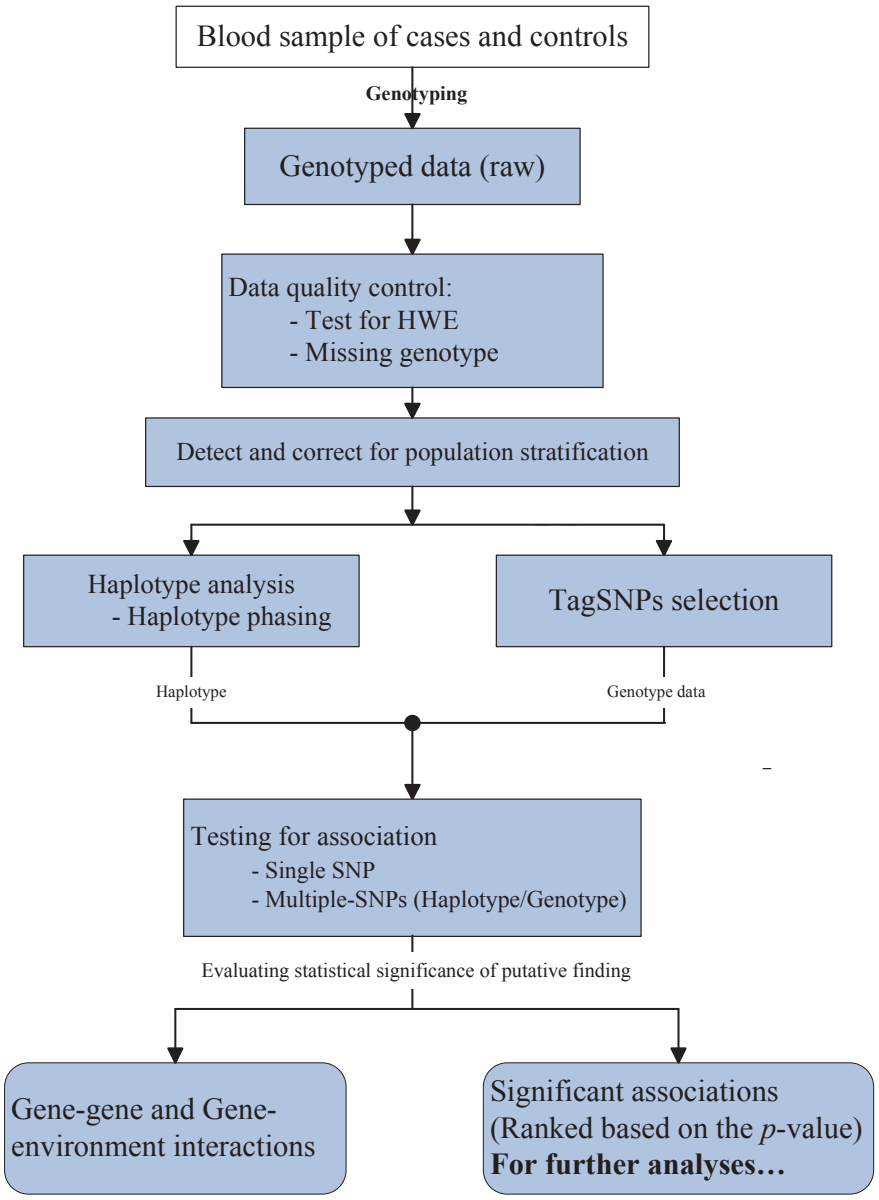

Figure 2.2: General Framework of GWA studies

Nevertheless, GWA studies primarily rely on genotype information that is usually obtained by powerful biomedical technologies. Given a large amount of genotype data, investigators are forced to follow several procedures in order to find significant associations. These procedures include: (i) preliminary analysis as a data quality control, (ii) haplotype analysis that includes haplotype phasing and Tag SNP selection, (iii) deciding which test of association to use, (vi) evaluating statistical significance of putative finding. Several statistical and computational methods are generally used along with these procedures. Figure 2.2 illustrates the general framework of conducting GWA studies based on these procedures. In addition, in the following subsections, theoretical and computational parts of these procedures will be explained. In this thesis, GWA based approaches are mainly applied to SNP datasets as data preparation and quality control checks as part of the proposed computational framework for disease diagnosis and prognosis.

2.2.2 Preliminary Analysis

As described earlier, in GWA studies, large-scale datasets must be genotyped, which might include thousands of individuals genotyped at hundreds of thousands of markers. In consequence, several statistical analysis methods must be applied before the association tests to ensure having good quality data. These methods are described as quality-control filters. That is, the markers should be checked thoroughly, and those

markers that show implausible high importance should be discarded. The following sections will describe several of these filtering criteria.

Hardy-Weinberg equilibrium

In the absence of inbreeding, population stratification, natural selection or even in some cases of genotyping errors, genotype frequencies at any locus are a simple function of allele frequencies. This phenomenon is termed as the Hardy-Weinberg equilibrium (HWE) (Wigginton et al. 2005). This type of measurement can be useful as a quality assessment of genotyping problems and for detection of disease association or population stratification in well established association studies (Nielsen et al. 1998).

This terminology and its implication is an important feature of population genetics. It has been commonly used to check whether observed genotypes conform to or deviate from Hardy-Weinberg expectation, which can be considered as a quality check measure. Loci that deviate from HWE among controls at a particular significant level (e.g. $\alpha = 10^{-3}$ or 10^{-4}) are discarded. However, in large-scale SNP data, some SNPs can have a high possibility of being involved in segmental duplications or deletions (Conrad et al. 2005, Bailey & Eichler 2006). These SNPs can be discarded as deviating from HWE, where in reality they could be important in disease causations. So, efficient implementation of HWE tests is crucial (Wigginton et al. 2005).

Consider bi-allelic markers with pair of allele A as a major allele and a as a minor allele. Under the assumption of HWE, the function of allele frequency should be

governed by the following equation: $p^2 + 2pq + q^2 = 1$, where p is defined as the allele frequency of major allele ($p = AA + 1/2AB$) and q is defined as the allele frequency of minor allele ($q = BB + 1/2AB$). In this equation, p^2 is the predicted frequency of homozygous (major allele AA) individuals in a population, $2pq$ is the predicted frequency of heterozygous (AB) individuals, and q^2 is the predicted frequency of homozygous (minor allele BB) ones.

Testing for deviation from HWE can simply be carried out using Pearson goodness-of-fit test, which is usually has a χ^2 null distribution. The χ^2 goodness-of-fit test examines how well the observed data agrees with the expectation under the null hypothesis H_0, using the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained using the HWE equation (Pagano & Gauvreau n.d.). The χ^2 statistic is defined as

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2.1}$$

where O_{ij} is the observed frequency, E_{ij} is the expected frequency under the null hypothesis, i is the individual with or without a disease, and j the distinct alleles.

Although the Pearson test is easy to compute, in some cases, when there are low genotype counts the χ^2 approximation can be poor (Balding 2006). Alternatively, another statistical test such as the Fisher exact test can be used, which does not rely on the χ^2 approximation (Wigginton et al. 2005, Maurer et al. 2007). Fisher exact

test can be more efficient in some cases of a large sample size and low genotype counts in some population. An open-source data-analysis software of these tests is available. The R software has an R genetics package that has the implementation of both tests. Also, another tool called PEDSTATS implements an efficient formula for computing the Fisher exact test (Wigginton et al. 2005).

A tool for interpreting the result of HWE test is the log quantile-quantile (QQ) p-value plot; the negative logarithm of the i^{th} smallest p-value is plotted against the $log(i/(L+1))$, where the L is the number of SNPs. Loci that deviate from the null hypothesis do not lie on the $y = x$ line (Weir et al. 2004). In this thesis, the HWE test is applied to SNP datasets as part of the "data cleaning and preparation" phase in the computational framework developed in chapter 4.

Missing Genotype data

In GWA studies, although, there are high throughput technologies that can ensure high genotyping call rates[4], but genotyping hundreds of thousands of SNPs per individual can result in a few genotypes missed. In the case of single SNP analysis, if a few genotypes are missing that may not be a problem, but missing genotypes can be more problematic if we are conducting a multi-SNP analysis because many individuals might have one or more genotypes missed (Balding 2006). Some GWA studies have considered 95% genotype call rate as a threshold to consider a particular

[4]The proportion of non-missing genotypes

marker for further analyses (e.g. Saxena et al. (2007)).

There are two factors that affect the way we can deal with missing data. The first is that the nature of the data (genetic variations) is not numerical. Thus, statistical parameters such as mean or median cannot be applied. Second, the frequency distribution of genotype data is not compatible with some statistical parameters that can be used to deal with categorical attributes. One convenient solution is data imputation: replacing missing genotypes with the most likely value based on the observed genotypes at the neighbouring SNPs. This seems inappropriate, but the LD structure between SNPs makes this method reliable.

Imputation methods can be either a single imputation or multiple imputations. Single imputation seeks the best prediction of missing genotypes, and then standard statistical procedures for complete data analysis can be used with the filled-in dataset. On the other hand, with multiple imputations, missing values will be replaced with a set of plausible values that reflect the uncertainty about the true value of missing data (Little & Rubin 1987). These imputed datasets are then analysed using the normal statistical procedures and the results from these analyses are then combined to one estimate. The advantage of the second approach is that the use of more than one imputed dataset (usually 3-5 times) can increase the performance and decrease the bias in the result (Little & Rubin 1987).

A more realistic imputation method was proposed to deal with the problem of

missing genotypes in unrelated individuals (case-controls studies); this method is based on regression models (Souverein et al. 2006). Nevertheless, most software that is used for phasing is also imputes missing genotypes (see section 2.2.3). In this thesis, tools that apply such an approach have been used as part of the "data cleaning and preparation" phase in the developed computational framework in chapter 4.

2.2.3 Computational Haplotype Analyses

As discussed in section 2.1.1, haplotype information is more powerful than genotypes because it yields information about recombination, which is the physical exchange of DNA during meiosis (Crawford & Nickerson 2005). Information about recombination is important for locating disease-causing mutation by using association studies. The use of haplotypes in disease association studies reduces the number of tests to be carried out, which it will be ultimately based on the association between the SNPs in the same haplotype.

There is much interest of developing methods for deducing haplotype information. Currently, there are two broad categories of tools that can determine haplotypes, which include directly genotyping pedigrees and using molecular methods in combination with genotyping for a population-based samples where the pedigree information is not available (Crawford & Nickerson 2005). Two widely used molecular methods include Allele-Specific Polymerase Chain Reaction (AS-PCR) and somatic

cell hybrids (Yan et al. 2000, Douglas et al. 2001). These methods have been used to determine haplotypes in a relatively small to moderate size population (e.g. Clark et al. (1998) and Patil et al. (2001)).

In the case of a large population size, both tools mentioned above would not be a good choice for deducing haplotype data because they are expensive and time consuming. However, for large-scale samples, high-throughput methods (such as microarray) can be used to identify the alleles of target locus of each individuals, but the main limitation of these methods lies in their lack of distinguishing the source chromosome of each allele, because they are mainly used for genotyping. However, there is a need for statistical and computational methods for determining haplotype information from genotype data. These methods are referred to as computational haplotype analysis.

Haplotype analyses are considered as the most central method for studying complex disease-gene association (Judson et al. 2002, Zhao et al. 2003, Shastry 2003). This has been demonstrated both through simulation and empirical studies (Morris & Kaplan 2002, Martin et al. 2000), and successfully applied to the identification of DNA variations that are involved in causing a number of complex disease (e.g. Hugot et al. (2001), Mas et al. (2005), Reif et al. (2006). There are two main procedures used in haplotype analysis namely haplotype phasing and TagSNP selection. Although haplotype phasing is an important issue in this field, it is laboratory expensive. In

this thesis, the main approaches that have been used are genotype based analyses. A full review of haplotype analyses, including haplotype phasing and TagSNP selection, is beyond the scope of this literature review. However, Lee (2006) discussed the main aspects of this problem and some solutions.

2.2.4 Tests for Association

The ultimate goal of GWA study association is to determine whether an association exists between a variation or a set of variations and a target disease. Testing for associations can be carried out using statistical analyses, which are based on either single-SNP or multiple-SNP analyses. The latter approach also includes SNP-based or Haplotype-based methods.

Single-SNP Analysis

In the case of single-SNP analysis, association between a particular genetic variation and a target disease can be determined by using genotype information. Testing for associations must be able to handle three different disease outcomes/statuses including case-control (binary outcome); quantitative (continuous outcome such as blood glucose level) and categorical outcomes such as subtype of cancers. However, this thesis concentrates on two cases: case-control and categorical outcome studies. The most straightforward approach for testing the association between SNP genotype and case-control status is to test the null hypothesis of no association between rows

and columns of the 2×3 matrix that contains the counts of the three genotypes among cases and controls. There are two general choices: the Pearson test (2-df) or the Fisher exact test. Both of these methods are used in practice. The second approach is preferable to the first one but it is computationally more demanding (Balding 2006).

For complex diseases, where it is thought that the contribution of individual SNPs to disease risk would be roughly additive, the two general tests mentioned above will not be as powerful. One solution for that is to count alleles rather than genotypes and then apply the Pearson 1-df test on a 2×2 table of allele frequencies (Balding 2006). However, Sasieni (1997) claims that such a method would not be appropriate for estimating the risk if the assumption of hardy-weinberg equilibrium does not hold in both the case and control samples together.

Another solution is called the Armitage test, which is similar to allele-count test, but is more conservative and does not rely on the assumption of HWE (Balding 2006). The Armitage test corresponds to testing the hypothesis of the goodness of fit of a line having zero slope, where the line tries to best fit the dots of the three genotype risk estimation. The dots represent the proportion of cases, among the cases and controls combined, at each of the three genotype risk estimations. The main limitation of this method is that the genotype risks, in some cases, are not additive. Lacking the knowledge about the status of the disease-predisposing variants of a disease,

which might be additive, dominant, recessive or over-dominant, makes the choice of the best test to use very difficult for researchers. Some studies suggest using an intermediate approach that takes the maximum test statistics for those designed for additive, dominant or recessive effects (Freidlin et al. 2002).

Another approach is to use linear models such as linear regression, but the main limitation of these methods is that they cannot be applied directly to case-control studies. This problem is overcome in logistic regression, in which transformations are applied to variables (e.g. the disease risk of an individual) based on the likelihood ratio test before the hypothesis testing.

In the case of continuous outcomes, linear regression and **AN**alysis **O**f **VA**riance (ANOVA) can be used. Linear regression tests the null hypothesis of no association based on the assumption of linear relationships between mean value of a disease and genotype. In either case, the test requires the disease to be normally distributed for each genotype. If normality does not hold, a transformation must be applied before testing the hypotheses. Finally in the case of categorical outcomes, regression methods such as multinomial regression can be applied. In some situations where the effects of the disease-risk outcomes are not similar, researchers might prefer to assign a different weight values based on the disease severity.

Multiple-SNPs Analysis

As discussed in section 2.2.3, haplotype information is more powerful than genotype information because it yields information about recombination. Furthermore, it is more likely that haplotypes will provide more information about the significant associations between a DNA variation and phenotypes, than using the single-SNP test (Stephens, Schneider, Tanguay, Choi, Acharya, Stanley, Jiang, Messer, Chew, Han et al. 2001, Waldron et al. 2006). This has been demonstrated both through empirical and simulation studies (Morris & Kaplan 2002, Martin et al. 2000). Multiple-SNP analyses can be conducted either using SNP-based (combination of SNPs based on the patterns of pairwise LD measure between them) or haplotype-based methods.

Firstly, multiple-SNPs analysis using the SNP-based method can be applied directly using standard statistical analyses such as regression analysis to investigate associations between continuous outcome and a set of SNPs whereas logistic regression can be used to analyse binary or categorical outcomes (Wallenstein et al. 1998). Logistic regression analyses for a set of SNPs are a natural extension of the single-SNP analyses described above.

On the other hand, multiple-SNP analysis based on haplotype-based method can be more significant in finding associations than SNP-based methods (Clark 2004). The block-like structure of the human genome is very important for conducting association analyses based on haplotypes that represent the SNP combinations within

these blocks. The main limitation of haplotype analyses is the uncertainty of haplotype phasing. Haplotypes are not observed directly; instead, they must be inferred from genotype data (see section 2.2.3), which might be hard to account for uncertainty that arises from haplotype inference. However, when LD between SNPs is high, the level of uncertainty is usually low (Schaid 2004, Balding 2006, Kelly et al. 2004, Marchini et al. 2006).

Once the haplotype estimation has been performed, the haplotype frequencies across populations of disease groups are analysed. The simplest form of analysis involves testing for independency of rows and columns in a $2 \times k$ contingency table, where k denotes the number of distinct haplotype (Sham 1998). Other type of analysis can be based on the haplotype proportions among cases and controls (Schaid 2004). The main problem of these methods is the reliance on the assumption of HWE and near-additive disease risk (Balding 2006).

A more efficient approach, suggested by Tzeng et al. (2003), is to look for the differences in haplotype distribution among cases relative to controls. In addition, a regression based model has been applied. This method treats haplotypes as categorical variables in regression analyses (Lin & Zeng 2006). Subsequently, a method was proposed that incorporates logistic regression models in a bayesian framework for detecting association (Clark et al. 2007). Other approaches such as score tests have also been applied. These methods incorporate haplotype phasing in association

analysis (Schaid et al. 2002).

There are several problems associated with haplotype-based analysis such as the large number of distinct haplotypes that may be generated from a given population (i.e. too many rare haplotypes). This will lead to a loss of power because there will be too many degrees of freedom to test the null hypothesis. Other problems are that there are no specific criterion for defining (i) block boundaries, which might vary from one population to another; (ii) the optimum sample size, and (iii) the SNP density or block definition (Ke et al. 2004).

Many methods have been proposed to overcome these problems such as cladistic-based methods (Durrant et al. 2004), clustering methods (Molitor et al. 2003, Morris 2005, Waldron et al. 2006, Tzeng et al. 2006), coalescent methods (Zollner & Pritchard 2005), and others (Toivonen et al. 2000, Beckmann et al. 2005). These methods impose the expected similarity of chromosomes with recent shared ancestry in the region flanking the disease gene based on the structure of haplotypes. They deal adequately with rare haplotypes and limit the number of tests that are required (Balding 2006). Some of these approaches are based on the evolutionary history of a given haplotypes. These approaches are often called cladistic and used to generate a tree of haplotypes that correspond to the genealogical tree of a given population. Other approaches use clustering algorithms for defining haplotype groups.

To summarize, although haplotype-based analysis seems to be the natural choice

for conducting association analyses, it might not be advantageous over analyses based on multipoint SNPs genotypes (SNP-based analysis) (Morris & Kaplan 2002, Fallin & Schork 2000). Even in some cases where recombination is entirely absent from regions, the haplotype-block model can be applied directly (Morris & Kaplan 2002, Balding 2006). Results based on a conducted study demonstrate that in some cases, SNP-based methods can be more powerful than haplotype-based methods (Clayton et al. 2004). Furthermore, tagging strategies that have been adapted by the HapMap project will limit the potential advantages of haplotype-based analyses by losing power of deducing the correct haplotype blocks. In contrast, haplotype-based analyses that are not restricted to haplotype-block boundaries – which are based on fixed length haplotype or sliding windows – hold promise for conducting association analyses (Cheng et al. 2005, Bahlo et al. 2006, Browning 2006).

To conclude, although testing for association can be carried out using SNP-based or haplotype-based analyses, for the purpose of this thesis the SNP-based analysis is mainly applied.

2.2.5 Evaluating the Statistical Significance of Putative Findings

In GWA studies, limiting the sample size will result in a reduction in power. As discussed in section 2.2.4, most of variants that cause complex diseases are likely to have

modest effects, and therefore large sample sizes are required. However, it is thought that a sample of 45 unrelated individuals should be sufficient to find 99% of haplotypes with a frequency of 5% or greater in a population (Gibbs et al. 2003). Another problem associated with large sample sizes is the large number of hypotheses to be tested, because p-values must be valid for multiple-hypothesis testing. Risch et al. (1996) propose that a p-value of 5×10^{-8} (equivalent to p-value of 0.05 after Bonferroni Correction for 1 million independent tests: $P_{corrected} = 1-(1-P_{uncorrected})^n$)' can be used as a relaxed threshold for declaring significant associations in GWA studies. In addition, allowing a smaller sample size would be possible by using a relaxed p-value threshold of 0.05, but that would result in up to 5% of the genotyped SNPs being associated by chance. For example, in a study of 500,000 SNPs this would result in a list of 25,000 false positive associations which might hide within them a true association. Therefore, a more liberal p-value threshold is crucial (Hirschhorn & Daly 2005).

Nevertheless, two of these problems (limited sample size and multiple hypothesis testing) can be overcome by using a simple procedure called a multi-stage approach (Li 2008). Using this approach a more modest threshold for passing markers as positive can be applied in multiple scans of different samples of a study. Therefore, a further increase in the efficiency can be achieved by saving in genotyping and increase the possibility of discovering true associations (limiting the false positive discovery rate),

with little loss of power (Hoh et al. 2001). This approach has been theoretically and empirically investigated by different studies (van den Oord & Sullivan 2003, Lowe et al. 2004, Satagopan & Elston 2003, Wang et al. 2006), and has been applied at the scale of genome-wide association studies by Rioux et al. (2007).

2.3 Relevance to the Thesis

In summary, genome-wide association studies are mainly conducted using statistical methods, which are used to discover genetic factors that contribute to susceptibility to disease. Factors that show a high statistically significant level of association are chosen for further analyses. However, this thesis is heading in a different direction. Data mining approaches are used to develop methods and techniques that can be applied to different disease models for the comparison and visualization of patient-to-patient relationships based on the genome-wide SNP data. The proposed data mining techniques can be further combined with more general data mining frameworks, such as case-based reasoning systems (Aamodt & Plaza 1994), to assist clinicians in understanding how a current patient compares to previous patients and hence how the current patient will react to a particular treatment protocol.

In this thesis, GWA studies will be used as a preliminary step to generate models that can be used to solve different disease problems. In this research study, GWA based studies will not be used for discovering disease-causing variants. However,

GWA studies will be adopted as a feature selection and prioritizing methods for implementing different disease models. SNPs will be selected based on a particular threshold that shows a significant or potential association with the targeted disease. The proposed methods and approaches are described in chapter 3.

In the following sections, data mining techniques and approaches that have been used in genetic variation studies for each task will be thoroughly reviewed. Then, based on the reviewed literatures, new proposed approaches are demonstrated in chapter 3, 4 and 5, and have been used for modelling and empirically evaluating different disease models.

2.4 Data Mining and Machine Learning Methods

This section describes the current direction in genetic variation domain and its connection to diseases. It is also surveys the state of art data mining and machine learning methods that have been applied to the domain of genetic variation studies. These methods include supervised and unsupervised-based methods. The reviewed methods have been used in the field of biomedical informatics, more specifically, genetic variation studies and have shown good performance results. Supervised-based methods are first reviewed including support vector machines and Random Forests. Following this, unsupervised-based methods, more specifically data reduction methods, are then reviewed.

2.4.1 Current Directions in Genetic Variation Studies

Most GWA studies have been conducted to identify susceptibility genes that marginally contribute to common complex diseases such as diabetes, cancer and cardiovascular diseases (Rampersaud et al. 2007, Rioux et al. 2007, Saxena et al. 2007, Thomas et al. 2008, Yamauchi et al. 2010, Craddock et al. 2010, Feero et al. 2010, Zhernakova et al. 2011). These studies essentially rely on evaluating one marker at a time, based on the assumption that disease susceptibility genes can be identified through their independent contribution to disease variability. In contrast, there is evidence that complex diseases are caused not by single genes acting alone, but are the result of complex non-linear interactions among genetic and environmental factors, with each gene having a small effect on the disease (Musani et al. 2007, Wu et al. 2010, Wang et al. 2011, Moore 2003). Therefore, there is a critical need to implement new approaches that can take into account gene-gene interaction in searching gene susceptibility markers that jointly cause complex diseases.

Based on the above assumption, several analytical and computational methods have been developed to detect and model disease susceptibility genes that account for main effect and gene-gene interaction effects of complex diseases, but these studies were conducted on relatively small datasets with hundreds of SNPs or using small

simulated data (McCarthy et al. 2008, Zhao et al. 2005, Millstein et al. 2006, Sha et al. 2009, Cordell 2009, Wu et al. 2010, Oh et al. 2011). However, considering the large number of SNPs that are now available in genome-wide scan data, testing each marker for main effect and all two-way, three-way and higher-order gene-gene interactions is computationally infeasible.

The dimensionality involved in such studies containing large numbers of SNPs is such that traditional statistical-based methods cannot be used without having a prohibitively large sample size of individuals (Todd 2006, Wang et al. 2011). This problem is referred to as the curse of dimensionality: as the number of SNPs increases, the number of possible interactions between genes increases exponentially and the produced models become unstable. For example, if we consider a genome-wide study of 500,000 SNPs, there are $M1 = 500{,}000$ possible one-marker statistical tests, using for example a χ^2 test, $M2 = M(M-1)/2 \sim= 12^{10}$ two-marker tests, and so on.

To overcome these difficulties, there is growing interest in applying non-parametric predictive models for understanding genetic association data. Data mining and machine learning approaches are examples of these models, which offer a powerful alternative approach to traditional statistical-based methods. These approaches can be used in a specific framework that can handle the computational complexity of the given data and the large number of interactions to be examined. An example

of one framework is to perform two or multi-stage analysis to reduce the number of interactions to be examined. In the first stage, a subset of SNPs must be identified for further analysis of interactions in the second and subsequent stages (Kang et al. 2008, lonita & Med 2006, Marchini et al. 2005, 2007, Meng et al. 2007).

Existing data mining machine learning methods of searching for a set of markers that have main effect and gene-gene interactions can be divided roughly into two groups: conditional-based approaches and exhaustive-based approaches (Zhang et al. 2008). In the conditional-based approach, a first stage is used to identify a set of markers that show significant association with the disease for further analysis of interaction in the second stage. This approach includes: pruning-based methods based on p-values (Dudbridge et al. 2006), Random Forests (Lunetta et al. 2004, Bureau et al. 2005), tree and spline based approach (Cook et al. 2004, Nonyane & Foulkes 2008), multivariate logistic regression that is used to evaluate interactions via step-wise regression (Hoh & Ott 2003), Support Vector Machines (Liu et al. 2008) and several other extensions of these methods (Chen et al. 2007). When susceptibility markers have small marginal effect but large interaction effects in a given study, conditional-based approaches will not explicitly account for the interaction of multiple genes and environmental factors in identifying markers for complex disease, especially in the case of using statistical-based methods (Meng et al. 2007).

On the other hand, exhaustive approaches search thoroughly for 1-marker, 2-marker,$\cdots$, and n-marker combinations of markers that are associated with the disease and hence account for main effect as well as gene-gene interactions. These approaches include: Multi-factor Dimensionality Reduction (MDR) (Ritchie et al. 2001, 2003), Focused Interaction Testing Framework (FITF) (Millstein et al. 2006), Combinatorial Partitioning Method (CPM) (Nelson et al. 2001), Combinatorial Searching Method (CSM) (Sha et al. 2006), Grammatical Evolution Neural Networks (GENN) (Motsinger et al. 2006, Motsinger-Reif et al. 2008) and others. These methods have shown significant results in searching for disease-susceptibility genes as well as accounting for multiple gene interactions in complex diseases such as diabetes, prostate cancer, hypertension and others (Ritchie et al. 2003, Cho et al. 2004, Williams et al. 2004, Xu et al. 2005, Brassat et al. 2006, Motsinger et al. 2007, Pattin et al. 2008). However, the scalability of searching SNP data with a large number of markers for example more than 10,000 SNPs, is the main limitation of these approaches.

Although current approaches for genome-wide association studies have been promising for discovering disease-susceptibility genes for complex disease, the scalability issue is still problematic. A flexible framework to deal with this issue would be of great interest for genome-wide studies (McKinney et al. 2006, Liang et al. 2007). Indeed, this thesis is mainly dedicated to propose such a framework.

2.4.2 Supervised-based Methods

Biomedical researchers are continuously seeking to develop and apply the most accurate classification algorithms for the formation of microarray-based biomedical data. Prior research in terms of classifying multi-category microarray-based data, suggests that among state-of-art classification techniques, Support Vector Machines (SVMs) have a predominant role, significantly outperforming k-nearest neighbours, backpropagation neural networks, probabilistic neural networks, weighted voting methods, and decision trees (Shim et al. 2009, Daemen et al. 2009, Chen et al. 2008, Chai & Domeniconi 2004, Statnikov et al. 2008).

In the last few years a significant amount of research in the bioinformatics community has been devoted in applying Random Forest (RF) algorithms for classification of microarray and other high-dimensional molecular data (Lee et al. 2005, Díaz-Uriarte & de Andrés 2006, Strobl et al. 2008, Sun et al. 2007, Meng et al. 2009).

In a recent study Meng et al. (2009) reports empirical evaluations of the RF method for classifying microarray-based datasets and concluded that RF classifiers have predictive performance comparable to that of the best performing alternatives (including SVMs). However, in a comparison study released by Statnikov et al. (2008), the authors compared the performance of RFs and SVMs for classifying cancer tissue based on microarray data. According to their results, the random forest approach is

still not as accurate as SVMs for typical microarray classification problems, despite the increasing popularity of RF in recent studies. Their results show that SVMs outperformed RF on most cases.

For the purpose of this thesis, SVMs are mainly used for classification, while RF is used as a feature selection approach. In the following sections, SVMs and RF approaches are reviewed.

Support Vector Machines

Support Vector Machines are binary classifiers that can be generalized to deal with multi-category classification problems by considering several binary SVM problems simultaneously. The main idea underlying SVM is very simple: find the best maximal margin hyperplane separating two classes of a given training data. A margin is defined as the sum of the distances from the hyperplane to the closest positive and negative correctly classified data points (support vectors), while penalizing for the number of misclassified data points. In the case of linear classification problems, linear SVMs can be used to search for the hyperplane in the original space. On the other hand, for non-linearly separable problems, the data are implicitly mapped to a higher dimensional space by means of a kernel function, where non-linear SVMs can be used to find a separating hyperplane.

For any given classification problem, if there is no hyperplane that can totally separate the two classes, a soft margin approach can be used to control the sensitivity

to outliers and allow slacks to a separating hyperplane. It chooses a hyperplane with a penalty that splits the cases as cleanly as possible, while still maximizing the distance to the nearest support vectors. Support vector machines with soft margin have been widely used and showed excellent classification results (Shim et al. 2009, Tian et al. 2007, Suykens et al. 2002, Osuna et al. 1997). A careful design and methodological approach must be taken in applying SVM algorithms.

Assume a given training data as a set of paired vectors $(\mathbf{x}_i, c_i), \mathbf{x}_i \in R^n, c_i \in \{-1, 1\}, i = 1, 2, \ldots, m$. In the simplest SVM application, the data points are linearly separable and in such a case, the SVMs aims to find the linear separating hyperplane $\mathbf{w}^T\mathbf{x}+b = 0$ with the maximal margin. The margin is defined as the sum of distances from the hyperplane to the closest positive and negative data points ($2 \parallel \mathbf{w} \parallel^{-1}$ is the distance between $\mathbf{w}^T\mathbf{x} + b = -1$ and $\mathbf{w}^T\mathbf{x} + b = 1$). Thus the classification problem can be formulated as follows:

$$\begin{cases} \mathbf{w}^T\mathbf{x} + b \geq 1, & \text{if } c_i = 1, \\ \mathbf{w}^T\mathbf{x} + b \leq -1, & \text{if } c_i = -1. \end{cases} \tag{2.4.1}$$

The two hyperplane inequalities can be combined into a single inequality:

$$c_i(\mathbf{w} \cdot \mathbf{x}_i + b) + 1 \geq 0, \forall_i \tag{2.4.2}$$

The decision function of SVMs can be formulated as

$$f(x) = sign\left(\mathbf{w}^T\mathbf{x} + b\right), \tag{2.4.3}$$

and the data points that are located on the corresponding decision region boundaries ($\mathbf{w}^T\mathbf{x} + b = -1$ or 1) are called support vectors. It should be noted that the two hyperplanes are parallel, with no learning set data points falling between them. The main feature of SVM approach compared to other classification methods is that a SVM is trained to maximize the prediction accuracy while other classifiers are trained to minimize the prediction errors. Maximizing the prediction accuracy can be accomplished through finding a separating hyperplane with the maximum margin. Intuitively, the classifier with the largest margin will give a lower expected error (i.e. better generalization). Figure 2.3 illustrates a simple 2-D example of SVM analysis model, which attempts to find a 1-dimensional hyperplane (i.e. a line) that separates the data points based on their target classes. The best hyperplane is found by minimizing the distance between hyperplane and the support vectors. Minimizing the distance between hyper-plane and the support vectors will result in maximizing the margin.

However, when a SVM algorithm is applied to linearly non-separable data, it is usually not possible to achieve a feasible separation. Therefore, some training error must be allowed on the constraints of the two hyperplanes while introducing a penalty to achieve a possible solution. An error term, $\xi_i \geq 0$, was introduced (Cortes & Vapnik 1995) to the constraints by modifying the original constraints which then become

$$\begin{cases} \mathbf{w}^T\mathbf{x} + b \geq 1 - \xi_i, & \text{if } c_i = 1, \\ \mathbf{w}^T\mathbf{x} + b \leq -1 + \xi_i, & \text{if } c_i = -1 \end{cases} \tag{2.4.4}$$

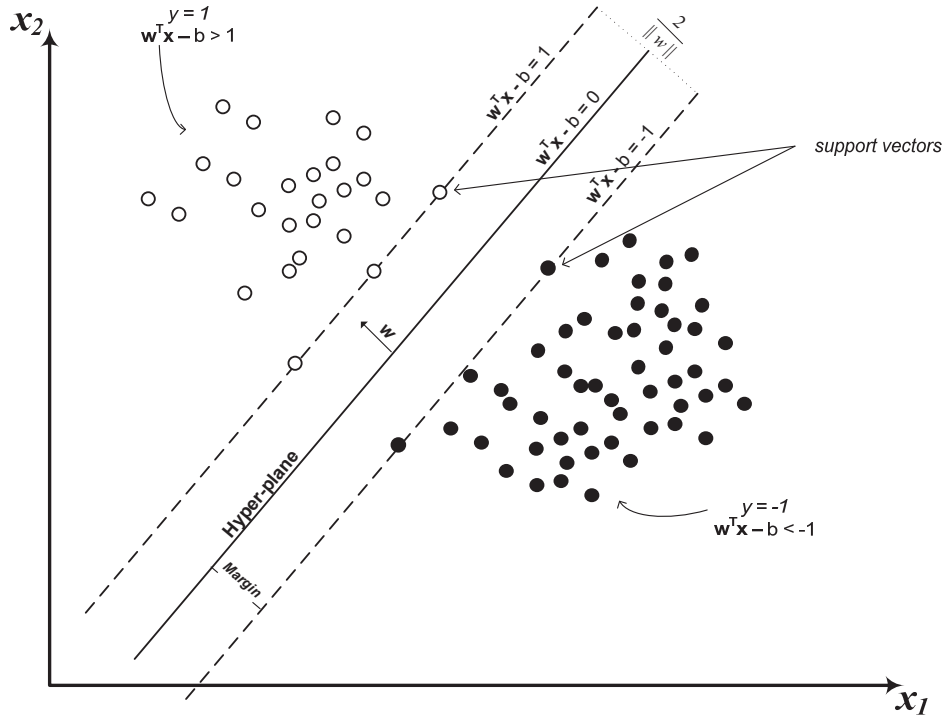

Figure 2.3: The optimal separating hyperplane of a SVM model in a linearly separable case (in the case of 2 dimension feature space). The optimal separating hyperplane is the solid line. Support vectors are the data points that lie on hyperplanes (the dashed lines) with maximal distance to the optimal separating hyperplane.

where the ξ_i are a measure of the misclassification errors, or so-called soft-margin error, the distances between those misclassified patterns with their corresponding boundary. Penalizing the objective function with $C\sum_{i=1}^{n}\xi_i$, a scalar C is a cost or penalty parameter chosen by the user. Figure 2.4 illustrates a simple 2-D example of SVM model with an optimal separating hyperplane in a linearly non-separable case.

In many situations, the decision functions of many classification problems can be non-linear functions of arbitrary complexity. In such cases, the linear SVM approach is not sufficient for performing classification. SVM approaches this problem by mapping

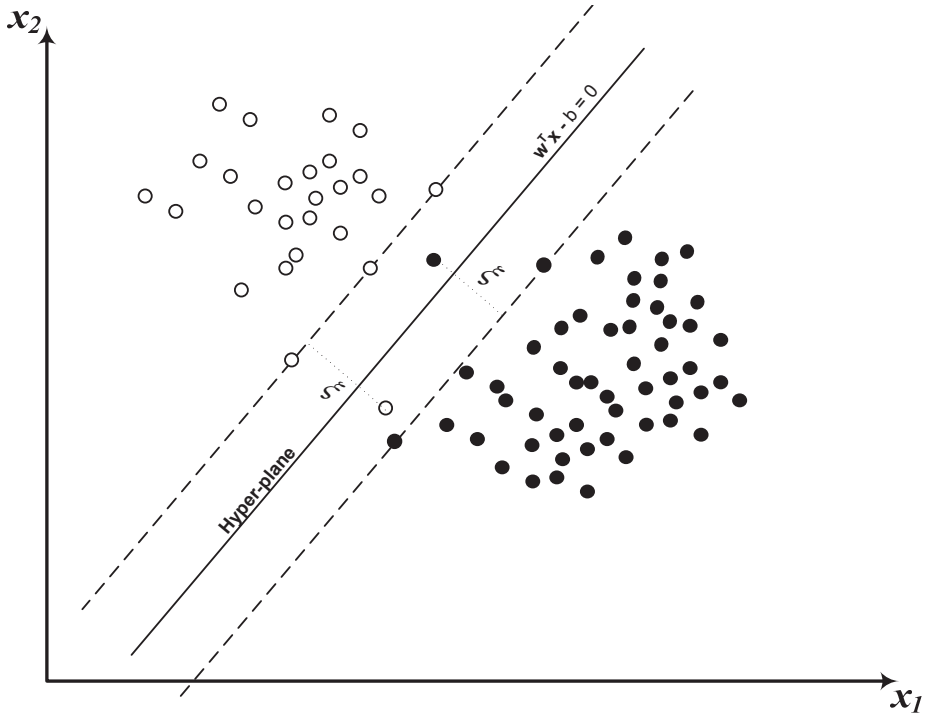

Figure 2.4: A SVM model with an optimal separating hyperplane in a linearly non-separable case.

the data in the input space, $\mathbf{x}_i$, into a high dimensional feature space, by choosing a non-linear mapping, ϕ, approach a priori. Then, a linear SVM can be used to construct an optimal separating hyperplane in the higher dimensional feature space with a maximum margin. The non-linear mapping that can be employed includes polynomials, radial basis and sigmoid functions. Together with error terms and non-linear mapping, SVMs training can be formulated as,

$$f(\mathbf{x}) = sign\left(\sum_{i \in SVs} \alpha_i c_i K(\mathbf{x}_i, \mathbf{x}) + b\right), \tag{2.4.5}$$

where $K(\mathbf{x}, \mathbf{x}')$ is the kernel function performing the non-linear mapping into the feature space, and

$$\langle \mathbf{w}^*, \mathbf{x} \rangle = \sum_{i=1}^{l} \alpha_i \tag{2.4.6}$$

$$b^* = \frac{1}{2}\left(\frac{1}{m_-^2} \sum_{\{(i,j):y_i=y_j=-1\}} K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m_+^2} \sum_{\{(i,j):y_i=y_j=1\}} K(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{2.4.7}$$

where m_+ and m_- are the number of examples with positive and negative classes, respectively. Thus, if the data are mapped using the kernel function, then the learning process only depends on data through the mapping procedures. If there is a kernel function such as $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, then we do not even need to know the explicit representation of $\phi(\cdot)$. Some commonly used kernel functions in SVMs are $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ (linear kernel), $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d$ (polynomial kernel), and $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel)$ (radial basis function) where r, d, and $\gamma > 0$ are kernel parameters. Figure 2.5 shows the basic idea of SVM kernel mapping, which maps the input space to a feature space such that the non-separable input space can be separable in the feature space. This is done by non-linear mapping to higher dimension using a kernel function and constructing a separating hyperplane in the feature space with a maximum margin.

A full details of the mapping and optimization procedures applied to SVM objective functions are beyond the scope of this review. However, a full review of SVM algorithms can be found in Burges (1998). This thesis mainly uses SVM-based methods to build disease diagnosis and prognosis classification models in the proposed computational framework, as described in chapter 4.

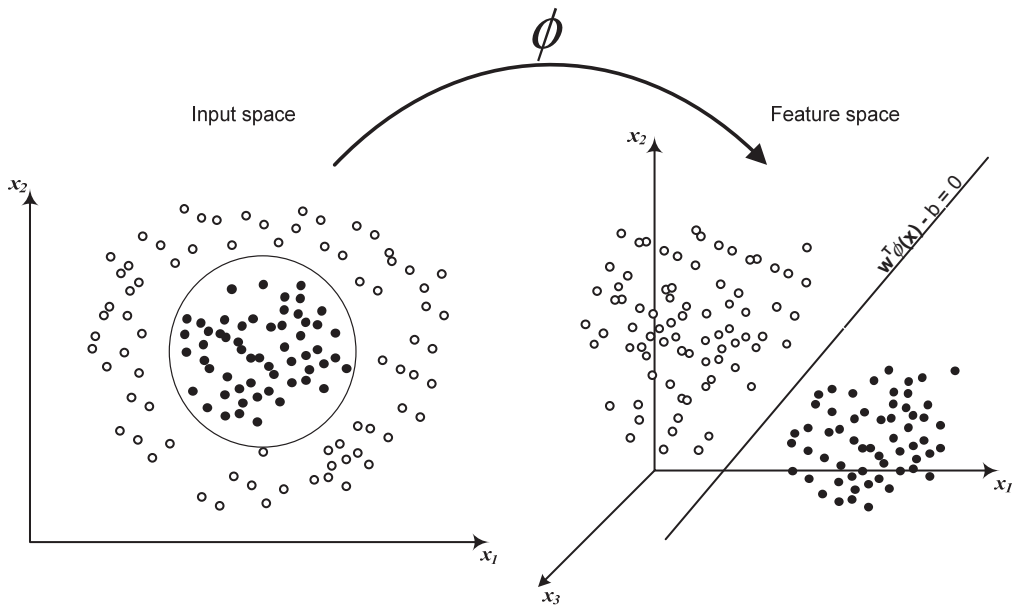

Figure 2.5: The concept of a SVM mapping procedure, which maps training data non-linearly into a higher dimensional feature space (a case of mapping 2-D input space to 3-D feature space).

Random Forests

Random Forests (RF) are a non-parametric machine learning approach that build a forest of classification or regression trees (Breiman 2001). Random forests use an ensemble of classification trees wherein each component tree is grown from a bootstrap sample of the data, and the variable at each split of a tree is selected from a random subset of variables of the data. Classification of an instance is based upon aggregate voting over all trees in the forest. There are five advantages of the random forest approach that make it a promising technique for genetic association studies. First, it

can handle large numbers of input variables compared to observations. The random forest approach is very efficient in selecting significant features from a large number of variables, such as genetic variation data, that explain a phenotype of interest. It uses both bagging and random variable selection for tree building, which are successful approaches for combining unstable learners with low correlation of the individual trees. The algorithm yields an ensemble of trees that can achieve low bias, high variance but low correlation of trees (Breiman 1996, Friedman et al. 2001).

Second, the random forests approach have good predictive performance even when most predictive variables are noise, and provide an unbiased estimate of model generalizability during the forest building process. Third, it is well-adapted to dealing with some variable heterogeneity of the data and missing variables, as separate models are automatically fit to a subset of data defined at early stages of tree building. Fourth, the learning process of random forests is very fast and computation time is modest even for a large number of variables. Finally, and most interesting, Random Forests produce an importance score for each variable that quantifies the relative importance of a variable to the prediction accuracy.

Classification algorithms that provide a metric measure for feature importance are of great interest for feature selection and feature prioritizing. Given these promising features, random forests approach has been widely applied to a large number of biomedical applications, such as microarray data and genetic variation data, and has

showed significant results (Bureau et al. 2005, Nguyen et al. 2006, Sabbagh & Darlu 2006, Chen et al. 2007, Mao & Kelly 2007, Glaser et al. 2007, Mao & Mao 2008, Strobl et al. 2008, Wang et al. 2008, Meng et al. 2009).

The random forest procedure starts by building a forest of trees, each tree in the forest is trained using a bootstrap sample of observations, where each bootstrap sample is obtained by drawing samples with replacement from the original observations. The bootstrap sample has the same number of data points as the original samples, with some data points represented multiple times. Then, at each node of a tree, random forests randomly select a subset of variables to determine the best split at that node. The trees in the forest are grown to their full extent, with no trimming or pruning. Repetition of these procedures yields a forest of trees, each of which has been trained on bootstrap samples of data points. For a given tree, some data points are left out called "out-of-bag" (OOB). Prediction error is estimated from these "out-of-bag" data points. The "out-of-bag" data points are also used to estimate the importance of a particular variable by randomly permuting the value of that variable. If randomly permuting the value of a particular variable does not affect the predictive ability of trees on out-of-bag data points, then that variable is assigned a low importance score. In contrast, if randomly permuting a value of a particular variable distorts the prediction ability of trees on out-of-bag data points, then that variable is assigned a high importance score.

Building an ensemble of trees in this manner increases the probability that some trees will capture interactions among variables with no strong main effect, thus, interactions can be taken into account when estimating variable importance. The recursive partitioning of tree building illustrates an explicit representation of variable interactions (Breiman 2001). Compared to other variable selection methods, interactions among variables do not demand a pre-specified model to explicitly test for feature interaction (Lunetta et al. 2004, McKinney et al. 2006). Thus, random forests can be considered as a natural approach for large scale analyses such as genetic association studies. The importance score of a particular feature may take into account gene-gene interaction without demanding a pre-specified model. In terms of selecting significant SNPs in genetic association studies, Random Forests outperform traditional statistical-based methods such as Fisher's Exact test in detecting markers that marginally contribute to complex diseases, and the relative superiority of random forests in detecting interacted markers (Meng et al. 2009, Wu et al. 2008, Mao & Mao 2008). Random Forests have also shown to be more robust in the presence of noise and missing data (Schwender et al. 2004, Bureau et al. 2005, Strobl et al. 2008). In this thesis, the random forest approach is mainly applied as a feature selection and weighting tool. The importance measure generated by random forest method was chosen for measuring the importance of each SNP (weighting) and selection of an appropriate set of SNPs (feature selection).

2.4.3 Unsupervised-based Methods

In this thesis several unsupervised-based methods, more specifically data reduction methods, are used to visualize genetic variation datasets. These methods include principal components analysis, multidimensional scaling, stochastic neighbour embedding, curvilinear component analysis, laplacian eigenmaps and locally linear embedding methods. In the following sections a review of these methods is given.

Principal Components Analysis

Principal components analysis (PCA) constructs a low-dimensional representation of a given data that maximally preserves as much variance in a given data as possible (Hotelling 1933). This is done by finding the linear projection or direction where the data has maximum variance.

Suppose that a dataset that is represented in terms of an $m \times n$ matrix, $\mathbf{X}$, where the n columns are the data samples (i.e. points) and the m rows are the variables. We wish to linearly transform this matrix, $\mathbf{X}$, into another matrix, $\mathbf{Y}$, also of dimension $p \times n$, where $p \ll m$.

The projection can be found by solving the eigenvalue problem of the covariance matrix $\mathbf{C}_x$ of the data using a general eigen-decomposition problem.

$$\mathbf{C}_x\mathbf{A} = \lambda\mathbf{A} \tag{2.4.8}$$

where $\mathbf{A}$ is a set of orthonormal eigenvectors of the covariance matrix $\mathbf{C}_x$ and λ

is a set of their associated eigenvalues, defined as a diagonal matrix. It can be shown that the linear projection is formed by the p principal components of the covariance matrix. The new representation of data points $\mathbf{x}$'s can then be found by projecting (or mapping) the original data using the following relation

$$\mathbf{y}_i = \mathbf{A}^T \mathbf{x}_i \tag{2.4.9}$$

Therefore, the low-dimensional data representation, $\mathbf{y}_i$, of the the data point $\mathbf{x}_i$ is computed by projecting data vector $\mathbf{X}$ using matrix $\mathbf{A}$, which contains the eigenvectors corresponding to the two or three largest eigenvalues. The new representation of the data can be visualized using the projected matrix Y.

PCA has been successfully applied in a large number of domains. However, the main limitation of PCA is that it does not work well when a given data lies in a non-linear manifold. However, PCA is advantageous when the variance of the data is mainly concentrated in a few directions.

Multidimensional Scaling

Multidimensional Scaling (MDS) represents approaches that are commonly used with non-linear mapping methods (Torgerson 1952). There are several different variants of MDS (Cox & Cox 2001), but they all share a common goal which is to find the low-dimensional representation of the data that preserves the pairwise distances of the data as much as possible. The quality of the mapping is represented by a stress function (or cost function), which tries to minimize the errors of the pairwise distances

between the low-dimensional and high-dimensional representations of the data.

The classical version of MDS is very closely related to PCA. The solution of linear MDS can be found by solving an eigen-decomposition problem. When the dimensionality of the sought space is the same and the distance measure is the Euclidean distance, the projection of the original data using PCA is similar to the configuration of points calculated by the squared Euclidean distance matrix of the data (Gower 1966).

Other variants of MDS which have a more effective stress function are the raw stress function and Sammon cost function. The raw stress function can be defined as

$$\phi(Y) = \sum_{i,j} \left(\|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{y}_i - \mathbf{y}_j\|\right)^2 \tag{2.4.10}$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in data points in the original data space, and $\|\mathbf{y}_i - \mathbf{y}_j\|$ is the Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ data points in the low-dimensional space. This cost function is able to find non-linear relationships in the data.

The Sammon cost function is slightly different to the raw stress in that it gives small distances a larger weight, which emphasises the local relationships in the data. In addition, there exist other variants of MDS, called non-metric MDS, which aim to preserve ordinal relations in data, rather than the pairwise distance (Kruskal 1964). Nevertheless, Multidimensional Scaling has been widely used for data visualization,

such as Functional Magnetic Resonance Imaging (FMRI) analysis and molecular modelling (Tagaris et al. 1998, Venkatarajan & Braun 2004).

The success of MDS has led to the proposal of new variants such as Curvilinear Component analysis (Demartines & Herault 1997) and Stochastic Neighbour Embedding (SNE) (Hinton & Roweis 2003). These methods have shown the capability of producing good quality visualizations. Extended versions of these methods will also be described in the following sections.

Stochastic Neighbour Embedding

Stochastic Neighbour Embedding (SNE), proposed by Hinton & Roweis (2003), is a probability-based embedding method. SNE tries to find the low-dimensional representation of data points that preserve neighbourhood identities. The SNE algorithm tries to preserve the probability distribution of the pairwise distances of data points in the input space, so that the probability of a data point i being a neighbour of point j in the output space is the same as in the input space.

For each data point $\mathbf{x}_i$ and its potential neighbours, $\mathbf{x}_j$, the algorithm starts by computing p_{ij}, the probability that point $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbours in the input space using

$$p_{ij} = \frac{\exp\left(-d(\mathbf{x}_i, \mathbf{x}_j)^2\right)}{\sum_{k \neq i} \exp\left(-d(\mathbf{x}_i, \mathbf{x}_k)^2\right)} \tag{2.4.11}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)^2)$ is the squared pairwise distance between data points i and j. The distance can simply be the squared Euclidean distance or it can be the scaled squared

Euclidean distance if the data is high-dimensional

$$d(\mathbf{x}_i, \mathbf{x}_j)^2 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2} \tag{2.4.12}$$

In low-dimensional output space $\mathbf{y}_i$ of data point $\mathbf{x}_i$ is defined as q_{ij}, which express the probability of the point $\mathbf{y}_i$ being a neighbour of point $\mathbf{y}_j$.

$$q_{ij} = \frac{\exp\left(-d(\mathbf{y}_i, \mathbf{y}_j)^2\right)}{\sum_{k \neq i} exp\left(-d(\mathbf{y}_i, \mathbf{y}_k)^2\right)} \tag{2.4.13}$$

The aim of the embedding is to match the two probability distributions p_{ij} and q_{ij} as well as possible. The embedding of points $\mathbf{y}_i$ can be achieved by minimizing a cost function which is the Kullback-Leibler divergence between the probability distributions of the input (p_{ij}) and output (q_{ij}) over neighbours of each data point. The cost function is

$$\mathbb{E}_i\left[D(p_i, q_i)\right] = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.4.14}$$

Stochastic Neighbour Embedding has been successfully applied to several datasets (e.g. Nguyen & Worring (2004) and Memisevic & Hinton (2005)). Results show that good optima can be achieved.

Stochastic Neighbour Embedding was originally designed as a data reduction method that tries to preserve neighbourhood identities. However, SNE can be also seen as an information retrieval algorithm. A restructured method called Neighbour Retrieval Visualizer (NeRV) was proposed by Venna et al. (2010). This method is

motivated by visual neighbour retrieval, unlike SNE, which tries to optimize recall (i.e. misses). The method balances the error caused by precision (i.e. false positives).

In information visualization, high precision is more important than recall. Minimizing precision is associated with preserving the neighbourhood of points in the output space. On the other hand, recall tries to preserve the neighbourhood of points in the input space. Neighbour Retrieval Visualizer updates the original SNE method by assigning a relative cost λ to recall and $(1-\lambda)$ to precision. Then, the total function to be optimized is

$$\begin{aligned} E &= \lambda \mathbb{E}_i \left[D(p_i, q_i) \right] + (1-\lambda) \mathbb{E}_i \left[D(q_i, p_i) \right] \\ &= \lambda \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-\lambda) \sum_i \sum_{j \neq i} q_{ij} \log \frac{q_{ij}}{p_{ij}} \end{aligned} \tag{2.4.15}$$

That is, by setting the parameter $\lambda \in [0, 1]$ the choice can be focused on either the probabilities that are high in the input space (recall) or in the output space (precision). When $\lambda = 1$ the method is equal to SNE and when $\lambda = 0$, the method focuses completely in avoiding false positives (precision). This method can be described as retrieving points based on the visualization display. In this thesis, NeRV is applied to visualize SNP datasets with a choice of λ that emphasizes the underlying structure of the data with maximum precision. In addition, SNE will be applied for comparison purposes.

Curvilinear Component Analysis

Curvilinear Component Analysis (CCA) is a variant of MDS (Demartines & Herault 1997). Whereas MDS tries to find the configuration of points that preserves the pairwise distances as much as possible, the CCA approach tries to find the configuration of points that preserve a subset of the distances that are neighbours in the output space. The cost function of CCA concentrates on preserving the distance of points in the reduced space. This can be done by introducing a weighted function F that depends on the distance between the points in the output space (or visualization), yielding a cost function

$$E = \frac{1}{2} \sum_{i} \sum_{i \neq j} \left(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j) \right)^2 F\left(d(\mathbf{y}_i, \mathbf{y}_j), \sigma_i \right) \tag{2.4.16}$$

Generally, $F(d(\mathbf{y}_i, \mathbf{y}_j), \sigma_i)$ is chosen as a bounded and monotonically decreasing function, in order to favour preserving the local geometry of the data. Decreasing exponential, sigmoid, or Lorentz functions can be suitable choices, and a simple step function can also be applied.

$$F(d(\mathbf{y}_i, \mathbf{y}_j), \sigma_i) = \begin{cases} 1 \text{ if } & Y_{ij} \leq \sigma_i \\ 0 \text{ if } & Y_{ij} > \sigma_i \end{cases} \tag{2.4.17}$$

The minimization of the cost function can be achieved using a form of stochastic gradient decent algorithm. During the optimization process, σ_i is set to cover all or at least most of the data points (as the case of MDS), and it is slowly decreased to

reach the optimal value.

Curvilinear Component Analysis has been successfully applied to various non-linear-dimensionality problems in data representation such as gene expression data and computer vision (Buchala et al. 2004, Venna et al. 2010). An extension of CCA, Curvilinear Distance Analysis (CDA), was introduced by (Lee et al. 2004). The main difference of CDA compared to CCA is to replace the Euclidean distance used by CCA with geodesic distance. Geodesic distance is based on graph theory and uses the minimum spanning tree to find the distance.

The main drawback of CCA is that the cost function may have several local optima. Although this can cause undesired results when applying CCA, solutions found by CCA have showed quite reasonable results.

A method called Local Multidimensional Scaling (LocalMDS) was proposed (Venna & Kaski 2006). This method is regarded as a derivative of CCA. Similarly to NeRV, LocalMDS has the indirect ability to control the trade-off between precision and recall, which helps for data visualization. The cost function of CCA tries to preserve the distance of points that are neighbours in the output space, by ignoring the error in distance between points that are far from each other in the reduced space. Thus, CCA could increase the errors caused by recall, which can result in lower visualization quality. In LocalMDS, a term is added to the cost function to increase recall. This

can be achieved by penalizing the errors of distance between points that are close by in the input space. The trade-off between the two types of errors helps in having a more efficient display of the local similarities of the data. The cost function of LocalMDS is defined as

$$\begin{aligned} E &= \sum_i \sum_{i \neq j} [\lambda \left(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\right)^2 F\left(d(\mathbf{x}_i, \mathbf{x}_j), \sigma_i\right) + \\ & \quad (1-\lambda) \left(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\right)^2 F\left(d(\mathbf{y}_i, \mathbf{y}_j), \sigma_i\right)] \end{aligned} \tag{2.4.18}$$

where $\lambda \in [0, 1]$ controls the trade-off between precision and recall. During the optimization the radius of the area of influence around data point $\mathbf{x}_i$, σ_i, is slowly reduced to reach the optimal value. $F(d\left(\mathbf{x}_i, \mathbf{x}_j\right), \sigma_i)$, similarly to CCA, emphasizes the local distance in the input space. F is equal to one when $d(\mathbf{x}_i, \mathbf{x}_j) < \sigma_i$ and 0 otherwise. The final radius is set equal to the distance of k-NN of a data point $\mathbf{x}_i$ in the original space.

When $\lambda = 0$ the cost function will be equivalent to that of the basic CCA method. According to Venna & Kaski (2006) a good choice of λ ranges from 0 to 0.5. The cost function can be optimized using stochastic gradient descent methods similarly to CCA. In the experiments in this thesis, LocalMDS is applied to visualize SNP datasets with choice of λ that emphasizes the underlying structure of the data to maximize precision. In addition, CCA will be applied for comparison purposes.

Laplacian Eigenmap

Laplacian Eigenmap (LE) finds a low-dimensional representation of a given data by preserving the local structure of the data (Belkin & Niyogi 2002). Laplacian Eigenmap is regarded as a geometrically motivated dimensionality reduction methods. The output space reflects the intrinsic geometric structure of the manifold. In Laplacian Eigenmap, the local structure can be preserved by keeping the local relationship between each data point and its N nearest neighbours. Therefore, the local structure of LE algorithms can be relatively insensitive to outliers and noise, and as a result the algorithm implicitly emphasizes the natural clusters in the data (Belkin & Niyogi 2002).

Laplacian Eigenmap computes a low-dimensional representation of the data in which the nearest neighbours of a data point in the original space should be mapped to nearest neighbours of that data point in the reduced space (He et al. 2005). This can be done in a weighted manner applied to graph partitioning using a weighted criterion (e.g. a heat kernel (Gaussian function)). Such a criterion allows the choice of weighting a graph in such a way that keeps the local similarity of the graph. The embedding map is constructed by computing the eigenvectors of the graph Laplacian. The algorithm's procedures are as follows.

The LE algorithm first constructs the adjacency graph G in which every node (data point) $\mathbf{x}_i$ is connected to its N nearest neighbours. For all nodes i and j in the

graph G that are connected by an edge, a weight is calculated using different methods such as a Gaussian kernel or a simple approach where $W_{ij} = 1$ if node i and j are connected by an edge. This leads to a sparse matrix W in which $W_{ij} > 0$ if node i and j are connected and $W_{ij} = 0$ otherwise.

To compute the low-dimensional representation $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n{}^T$, Laplacian Eigenmap minimizes the following objective function

$$\phi(\mathbf{Y}) = \sum_{i,j} (\mathbf{y}_i - \mathbf{y}_j)^2 w_{ij} = tr(\mathbf{Y}^T \mathbf{L} \mathbf{Y}) \tag{2.4.19}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix, $\mathbf{D}$ is diagonal matrix, with elements $D_{ii} = \sum_j W_{ij}$ being the column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$. The Laplacian matrix is symmetric and positive semidefinite.

Minimizing the objective function tries to put data points that are connected in the graph G as close together as possible. There is a trivial solution to the objective function which collapses all the new representations of the graph G into a single location. This can be prevented by adding an orthogonality constraint $\mathbf{Y}^T \mathbf{D} \mathbf{Y} = 1$.

The configuration of points in the low-dimensional space can be solved by finding the eigenvectors and eigenvalues of the generalized eigenvector problem

$$\mathbf{L}\mathbf{y} = \lambda \mathbf{D}\mathbf{y} \tag{2.4.20}$$

The low-dimensional embedding of the original data points can be formed by d eigenvectors that correspond to the smallest non-zero eigenvalues, after discarding the

smallest eigenvector that corresponds to the zero eigenvalues, which represent the case where all data points are represented by a single location.

Laplacian Eigenmap has been successfully applied to number of domains such as clustering and face recognition (Ng et al. 2002, Shi & Malik 2000, He et al. 2005). Variants of Laplacian Eigenmaps have been extended to supervised and semi-supervised data analysis (Costa & Hero 2005, Belkin & Niyogi 2004).

Laplacian Eigenmap has two main drawbacks. Firstly, in most applications it is not possible to see the structure within clusters from the visualization. Secondly, this method is mainly used for data representation or visualization and cannot compute the projection for a new test point. However, this latter problem can be solved using techniques proposed by Bengio et al. (2004) called an out-of-sample extension.

Locally Linear Embedding

The Locally Linear Embedding (LLE) algorithm is similar to Laplacian Eigenmap and tries to preserve the local geometry of the data by finding the *local linear* approximation of the manifold (Roweis & Saul 2000). This is based on the assumption that a data point and its neighbours lie in or close to a locally linear subspace on the manifold. In LLE, the local geometry of this subspace can be characterized by calculating the linear coefficients (weights) that reconstruct each data point from its N nearest neighbours. In the low-dimensional space of the data, LLE attempts to retain the reconstruction weights in the linear combination as much as possible (Van

Der Maaten et al. 2007).

The algorithm works in two stages. First, the local coordinate of each data point is calculated based on its N nearest neighbours, and the total reconstruction error to be optimized is then measured by the following cost function

$$\epsilon(W) = \sum_{i=1}^{n} \left| \mathbf{x}_i - \sum_{j=1}^{k} W_{ij} \mathbf{x}_j \right|^2 \tag{2.4.21}$$

which adds up the squared distance between all data points and their reconstruction weights. The weight W_{ij} summarizes the contribution of the jth data point to the ith reconstruction. The reconstruction error is minimized subject to the constraints that $W_{ij} = 0$ if data points i and j are not neighbours and $\sum_j W_{ij} = 1$.

In the second stage, the task is to find the low-dimensional representation $\mathbf{y}_i$ that preserves the local geometry of the data as described by the local coordinate of each data point. In other words, the reconstruction weight W_{ij} that reconstruct each data point $\mathbf{x}_i$ from its neighbours in the high-dimensional data space also reconstruct each data point $\mathbf{y}_i$ in the low-dimensional space. To do so, the p-dimensional reduced space Y can be computed based on minimizing the cost function

$$\epsilon(Y) = \sum_{i=1}^{n} \left| \mathbf{y}_i - \sum_{j=1}^{k} W_{ij} \mathbf{y}_j \right|^2 \tag{2.4.22}$$

Roweis & Saul (2000) showed that the optimization function described in (2.4.22) can be solved by the eigenvectors that correspond to the p non-zero eigenvalues of matrix $\mathbf{M}$, where $\mathbf{M} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ and $\mathbf{I}$ is the identity matrix. A linear variant

of LLE algorithm was also proposed (Kokiopoulou & Saad 2005, 2007). In this thesis, the LLE method is used to visualize SNP datasets.

2.5 Case Study

The case study of this project is childhood Acute Lymphoblastic Leukaemia (ALL), which is the most common childhood malignancy. It represents 24% of all new cancers that occurred in children between 1995 and 1999 (Coates & Tracey 2001). Nearly all children with ALL achieve an initial clinical remission, so the major obstacle to cure is patient relapse, i.e. the recurrence of evident disease. There is a need to reliably identify childhood ALL patients at greater risk of not responding to current treatment, who can then undergo modified therapy. Genetic variation data is mainly being used in this thesis. The data is used to develop computational and data mining models for building different disease models.

Chapter 3

Feature Selection, Weighting, Prioritizing and Distance Metric Measure for SNP Data

Characterizing the mapping relationships between genetic variations and susceptibility to disease has the potential to improve diagnosis, prevention and treatment of complex disease. Identifying susceptibility genes for complex disease is a major challenge for human geneticists. Genetic variation data has been regarded as instrumental to discovering disease-susceptibility genes associated with many common complex diseases in recent years.

Generation of SNP data has been facilitated by high throughput microarray based technologies. One of the fundamental subjects in genetic variation studies is to find an optimal subset of SNPs with the highest predicting power for different disease models. Most of the current interest in genetic variation studies is focused on disease gene association analyses. Such analyses are important in identifying which variants

of genes are associated with a specific disease. To identify these markers at a statistically significant level, it is necessary to obtain genetic information from a large scale population sample of affected and unaffected individuals, which is termed as a population based study.

Recent advances in biomedical technologies and genetics studies have made association studies a powerful approach for mapping complex-disease genes by conducting studies based on the whole genome. Genome-wide association studies genotype a dense set of SNPs across the genome to survey the most common genetic variations for a role in a disease or to identify the heritable quantitative traits that are risk factors of a given disease (Hirschhorn & Daly 2005).

However, most GWA studies have been conducted to identify susceptibility genes that marginally contribute to several common complex diseases. These studies essentially rely on evaluating one marker at a time, based on the assumption that disease susceptibility genes can be identified through their independent contribution to disease variability. In contrast, recent studies have shown that complex diseases are caused not by single genes acting alone, but are the result of complex non-linear interactions among genetic and environmental factors, where each gene having a small effect on the disease of interest (Musani et al. 2007, Wu et al. 2010, Wang et al. 2011, Moore 2003). Therefore, there is a critical need to implement new approaches that can take into account interaction that jointly cause complex diseases.

Several feature selection methods and approaches have been reviewed and studied in chapter 2. These methods have been used to deal with different tasks in the given domain. The reviewed studies were applied to genetic variation data of population-based studies of affected and unaffected individuals. Based on these studies, new approaches have been proposed in this chapter to deal with the tasks of SNP selection, weighting and feature construction in the domain of genetic variation studies. The proposed approaches are based on non-parametric machine learning techniques such as RF-based approaches.

The reminder of this chapter is as follows: section 3.1 describes how to deal with genetic variation data and its connection to diseases. Section 3.2 describes methods for SNP selection and weighting that have been applied in the given domain. In that section, a new approach has been proposed to deal with the tasks of SNP selection and weighting. In section 3.3, a new approach for SNP prioritization is defined. The new approach can be used to search for a set of SNPs that has the potential to be involved in gene-gene interactions. In section 3.4, the set of SNPs selected, based on the approach proposed in section 3.3, is used to construct new combined features. The new combined features carry the information that potentially account for gene-gene interactions. Section 3.5 presents methods for calculating distances between genome-wide SNP profiles. Finally, in section 3.6, the chapter is concluded.

3.1 Dealing with Genetic Variation Data: The Proposed Approaches

A review of current directions in genetic variation studies was described in chapter 2. In this section proposed approaches to deal with genetic variation data are given. Although current approaches for genome-wide association studies have been promising for discovering disease-susceptibility genes for complex disease, the scalability issue is still problematic. A flexible framework to deal with this issue would be of great interest for genome-wide studies (McKinney et al. 2006, Liang et al. 2007). Indeed, this chapter is dedicated to propose new approaches for different tasks of feature selection in the field of genetic variation studies.

Non-parametric machine learning techniques have been chosen to deal with the "curse of dimensionality" problem induced by the data and non-linear interaction between multiple markers and disease outcome. The proposed approaches use feature selection methods for selecting disease-susceptibility markers that account for marginal effects as well as gene-gene interactions that are associated with complex diseases. Applying only feature selection methods based on traditional-based approaches will result in choosing markers that marginally contribute to disease and will ignore gene-gene interactions that may be useful for producing highly predictive models for large scale genome-wide studies, where the number of SNPs would be

multiples of hundreds of thousands.

The novelty of the proposed approaches is the intensive use of non-parametric machine learning techniques for the purpose of feature selection, weighting, and prioritizing of SNPs, as well as distance measure calculations. Each of these tasks will be jointly considered for building different disease models in dealing with data from genome-wide studies. The modelling frameworks will be illustrated in chapter 4 and 5. In the following subsections, the proposed methods and approaches are described.

3.2 SNP Selection and Weighting Based on Random Forests

There are two goals for this task, selecting a set of interesting SNPs from a pool of possible candidate SNPs and weighting the importance of each feature (SNP) on a given disease model. This can be accomplished using any number of feature selection methods include parametric and non-parametric based approaches. Parametric methods include least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996); regression analysis (Thomas et al. 2001); ReliefF (Robnik-Šikonja & Kononenko 2003) and many others. Non-parametric methods include machine learning and pattern-recognition approaches (e.g. Random Forests and Neural Networks based approaches).

In this thesis, a non-parametric approach, based on Random Forests, was chosen

as a feature selection approach to genome-wide association studies. The proposed approach is used for measuring the importance of each SNP (weighting) and the selection of an appropriate set of SNPs (feature selection).

A detailed review of the random forests approach and its characteristics is described in chapter 2. Here, the use of RF to analyse genetic variation data is examined. Specifically, the random forest approach is applied as a feature selection and weighting tool. The importance measure based on random forests will be used to produce two sets of SNP importance measures: one set to search for disease-susceptibility genes that have a marginal contribution to the disease, and another set for weighting the contribution of each SNP in a given disease model. This latter measure will be used to calculate distances between different genotype profiles, as the weight will indicate the importance of each SNP in differentiating between disease classes. The weight is computed based on the whole feature set.

To search for disease-susceptibility genes that marginally contribute to a disease status, this thesis adopts an iterative process similar to the strategy previously proposed for gene expression data analyses (Jiang et al. 2004, Svetnik et al. 2004, Díaz-Uriarte & de Andrés 2006). The process iteratively runs RF after discarding those variables with the smallest variable importance measures. The selected set of variables is the one that yields the smallest OOB error rate. Svetnik et al. (2004) proposed a feature selection method by first finding the best p dimensions of the model, and

then choosing the p most important variables. This is a sound strategy when the objective is to build an accurate predicator model, without any regards to model interpretability.

On the other hand, Díaz-Uriarte & de Andrés (2006) proposed a similar approach for gene selection based on microarray data. Their algorithm iteratively builds RFs, where at each iteration it builds a new forest after discarding those variables (genes) with the lowest importance. The solution is finally chosen with the smallest number of genes whose error rate is within u standard errors of the minimum error rate of all built forests. Jiang et al. (2004) also applied a similar strategy to microarray data with the exception that the variable importance at each iteration is re-computed. More importantly, their gene selection is based on both OOB error as well as prediction error, where the forest is trained on one dataset and tested to a second, independent, dataset. However, their approach for gene selection is only feasible with small datasets. Jiang et al. (2004) also report excellent performance of variable selection using RF when applied to their studied datasets.

3.2.1 A New Proposed Approach Based on Random Forests

This thesis proposes a new RF-based iterative procedure for the domain of genetic variation studies. The proposed approach is named RF-based Recursive Feature Elimination (RF-RFE). A random forest approach is first built using a training dataset

with a large number of trees to estimate effective (stable) variable importance measures. Random forests return two main measures of variable importance. One is the mean decrease in accuracy over all classes and the other is the mean decrease in Gini index. Additional importance measures of the mean decrease in accuracy for each class, of the target attribute, are also computed. In this proposed approach, variables that show a negative contribution to disease variance, based on any of the retrieved measures, will be discarded.

The proposed feature elimination strategy depends on two criteria: the ranking procedure as well as the validation accuracy (i.e. OOB prediction errors of the built forest). The ranking procedure defines the order of features to be eliminated and the validation accuracy is used to decide whether the chosen subset of features is permanently eliminated. Then, at each iteration a random forest with a reasonable forest size is re-run and markers with negative contribution to the disease are removed.

This process is repeated for each new subset of features and the validation accuracy is estimated for the built forest at each iteration. The validation accuracy evaluates whether the selected subset of features is accepted as a final subset of features or whether more features must be eliminated. If the obtained validation accuracy, for the current subset of features, is higher than the accuracy for the previous selected subset, then more features need to be eliminated based on their ranking values. The iteration is stopped whenever the validation accuracy of the new subset of features is

lower than the one for the previous selected subset (i.e. the validation error rate starts to increase), Then, the current subset of features is considered as the final subset of features. Otherwise, the procedure is repeated again with the lowest ranked features eliminated.

The best set of features with the highest validation accuracy is chosen and is called the feature set of best prediction. The variables included in this set are chosen as the most significant SNPs that marginally contribute to a given disease status. The process of the RF-RFE approach is summarized in algorithm 1.

The main difference between the proposed approach compared to the aforementioned approaches is the strategy for discarding variables at each iteration. The proposed approach discards variables that distort the prediction error and have no power in class discrimination, instead of choosing a predefined fraction of variables to remove. Indeed, as illustrated in chapter 4, the proposed approach shows better performance than other RF-based feature selection approaches such as the VarSelRF approach (Díaz-Uriarte & de Andrés 2006). The proposed method is similar to backward elimination methods used for feature selection using SVMs (Guyon et al. 2002). Svetnik et al. (2004) reported severe over-fitting resulting from recalculating variable importance at each iteration. Their results show that when using the iterative procedure the OOB error is biased down and cannot be used to assess the overall error rate reported at each iteration. However, using error rates affected by selection bias is

not necessarily a bad procedure for selecting interesting variables (Braga-Neto et al. 2004, Jiang et al. 2004).

Algorithm 1 Algorithmic steps for the RF-RFE approach

Input : Training examples
$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)^T$
Class labels
$\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$

Output : Selected feature list S with the lowest prediction error (OOB error)

Initialize: Subset of selected feature list
S = [1, 2, ..., n] ; `//all features in the original dataset`

Train the RF classifier using all features in S
Compute the "*importance measure*" for each feature in S
$c_i = w_i$ *for all* i
Find all features that show a negative contribution to disease variance
F = $\backslash c$
Update the selected feature list
S = setdiff (S,F) ; `//remaining features`

repeat
Train the RF classifier using the remaining features in S
Estimate the OOB error for the built forest
Compute the "*importance measure*" for each feature in S
$c_i = w_i$ *for all* i
Find all features that show a negative contribution to disease variance
F = $\backslash c$
Update Subset of selected feature list
S = setdiff (S,F) ; `//remaining features`
until *OOB error rate of S starts to increase*;

3.3 Prioritizing SNPs for Evaluating Interaction Effects

As discussed in section 2.4.1, when the total number of markers is large, as in the case of genome-wide association studies, searching exhaustively for a set of markers that jointly has significant effect on complex diseases is computationally infeasible, due to the large number of possible interactions needed to be evaluated (i.e. two-way, three-way, four-way and so on). In other words, when the number of markers is very large (more than 300,000 SNPs) existing methods such as exhaustive search cannot jointly analyse more than two-way interactions. However, many interactions do not necessarily need to be evaluated (Culverhouse et al. 2004). In fact, markers with similar disease effects can fit into the same group and testing one marker from each group could be enough to test interaction effects involved with markers in other groups.

The typical way of minimizing the number of markers that need to be evaluated for interaction, is to prioritize them and markers/SNPs above a given threshold chosen for testing their interaction effects. Prioritizing features (SNPs) is usually performed using classical association test methods, such as χ^2 test. Sha et al. (2006) proposed a clustering approach to minimize the number of markers to be evaluated. However, their methods do not systematically cluster markers as it is not clear how to define

the size or number of clusters needed. Here, a new measure for prioritizing SNPs for testing interaction is proposed.

3.3.1 A New Measure for Prioritizing SNPs

The main issue in prioritizing SNPs is how to quantify the contribution of each SNP separately in gene-gene interactions. Prioritizing will then be based on the resulting quantitative measures. Here, an alternative method for prioritizing SNPs with the goal of evaluating gene-gene interactions is proposed. The main feature of the proposed method is that there is no need for evaluating all SNPs for all possible ways of interaction. To do so, an information-based measure (the entropy function) is chosen to measure the contribution of each SNP to gene-gene interactions separately. The proposed method is based on the information gain measures of patient and non-patient samples separately. According to Shannon's information theory (Shannon 1948), for a random variable $\mathbf{x}$ taking one of n possible outcomes $\{x_i : 1, 2, \cdots, n\}$, the entropy $H(X)$ is defined as

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{3.3.1}$$

where $p(x_i)$ is the probability mass function of outcome x_i.

The proposed measure is not based on the whole sample (case and control samples together). Rather it is based on the entropy of case and control separately. The Interaction Effect (IE) is defined as

$$IE(SNP_i) = H(SNP_i^{case}) + H(SNP_i^{control}) - H(SNP_i) \tag{3.3.2}$$

where IE is the interaction effect of a given SNP_i, $H(\cdot)$ is the entropy of a set of data samples. $H(SNP_i^{case})$ is the entropy of SNP_i in case data samples and $H(SNP_i^{control})$ is the entropy of SNP_i in control data samples. Thus, the interaction effect of a given SNP is measured by the information gain. That is, the sum of entropy of case and control samples separately minus the entropy of all samples (both case and control data samples).

A

SNP1

	0	1
Patient	60	60
Non-patient	60	60

B

SNP2

	0	1
Patient	118	2
Non-patient	117	3

Figure 3.1: The allelic distributions of two SNPs. Both of these SNPs are reported to have no association with phenotype

To illustrate how the proposed method works, suppose a dataset of 240 samples (120 patients and 120 non-patient individuals) is given, and suppose that the allelic distribution of a given SNP is as shown in figure 3.1A. This SNP is represented using the dominant genetic model, where 0 means that a SNP has no minor alleles (two major alleles) and 1 means it has 1 or 2 minor alleles. Based on the distribution in figure 3.1A, the SNP1 would be reported as not associated to the disease (using any

association-based tests such as χ^2 test).

However, given that the SNP1 has no association to the disease, this SNP may have a high potential to be involved in gene-gene interaction. In contrast, the distribution of SNP2, as illustrated in figure 3.1B, has nearly no association with the disease, because the number of variations are similar between patients and non-patients. However, the distribution of this SNP is skewed. Nevertheless, given that SNP1 and SNP2 are not associated to the disease, the potential of SNP1 being involved in gene-gene interaction is much higher than SNP2 as there are more individuals/samples with minor alleles to assess.

The rationale is that the probabilities of a given SNP being associated with a risk of a disease depends on the genotype at the other SNP(s). In other words, the possibility of identifying SNPs that interact to influence a disease risk depends on identifying specific combinations of genotype on tested SNPs for interaction effects. SNPs that are highly skewed are less exposed to have a specific genotype combinations with other SNPs to influence disease risk. Therefore, the IE measure tends to give SNPs that are highly skewed less importance than SNPs that are uniformly distributed. For example, SNP1 as shown in figure 3.1A could have an XOR relationship with another SNP, which is difficult for most feature selection methods to identify independently. The IE of SNP1, as shown in figure 3.1A, is equal 1. On the other hand, the IE of SNP2, shown in figure 3.1B, is equal 0.1448582. The quantitative IE results assessed

by IE measure reflect that SNP1 has a higher probability than SNP2 to be involved in gene-gene interaction effect.

There are two extreme cases of SNP that could be associated with a given disease. In the first case, SNPs that have marginal associations with a given disease may not necessarily be involved in gene-gene interactions. The second case includes a SNP set that has small marginal effect but could have potential interaction effects. The measure will be informative under both of these cases. As results show in chapter 4, based on the new measure a very small percentage of marginally associated markers were involved in gene-gene interaction.

Figure 3.2 shows an example of two SNPs that have no association to a disease. In this figure the SNPs are represented using the additive genetic model where columns represent the number of minor alleles. However, the distributions of these two SNPs have different allelic distributions. Based on the IE measure, SNP3A (rs1138294) has a much higher possibility than SNP3B (rs1724577) to be involved in gene-gene interaction. This can be due to the allelic distribution of the two SNPs. The SNP3A has a relatively uniform distribution, compared to the SNP3B which is skewed and has most of the alleles in one category. On the other hand, the SNP3A has no association to the disease, but has a higher IE value. The allelic distribution of SNP3A shows that this SNP has a higher possibility to be involved in gene-gene interactions.

Figure 3.3, on the other hand, shows examples of two SNPs that are marginally

A

SNP3A

	0	1	2
Patient	55	66	18
Non-patient	55	66	18

SNP name : rs1138294
IE = 1.421362323

B

SNP3B

	0	1	2
Patient	137	3	0
Non-patient	136	3	0

SNP name : rs1724577
IE = 0.149814003

Figure 3.2: The allelic distributions of two SNPs that have no association to a disease but with different allelic distributions.

A

SNP4A

	0	1	2
Patient	61	63	16
Non-patient	31	63	44

SNP name : rs35414
IE = 1.40160423

B

SNP4B

	0	1	2
Patient	113	22	0
Non-patient	132	5	0

SNP name : rs12334990
IE = 0.400681331

Figure 3.3: The allelic distributions of two SNPs that have associations to a disease but with different allelic distributions.

associated with a disease, as reported in experimental results (chapter 4). However, the distributions of these two SNPs are different. Based on the IE measure, SNP4A (rs35414) has much higher possibility than SNP4B (rs12334990) to be involved in gene-gene interaction. This can be due to the allelic distribution of the two SNPs. SNP4A has a relatively uniform distribution, compared to SNP4B where the most of alleles are in one category.

In this way, the proposed method for prioritizing SNPs should be informative for reducing the number of markers for evaluating the effects of gene-gene interaction

in a given disease. The proposed method should systemically outperform exhaustive search approaches for searching/identifying interaction effects on diseases as a large portion of the interaction tests do not necessarily need to be evaluated (e.g. Dong et al. (2007)). This makes the method computationally feasible and effective for choosing markers that would be involved in interaction effects on a specific disease.

Furthermore, the proposed method is different to conditional-based approaches (or two-stage approaches), which are based on selecting a set of markers using filtering methods and then evaluating interaction effects of the selected set of markers. When susceptibility markers have small marginal effect but large interaction effect in a given study, conditional-based approaches will not explicitly account for interaction of multiple genes, especially in the case of using statistical-based approaches (Meng et al. 2007).

3.3.2 Selecting Markers Involved in Gene-Gene Interactions

The proposed IE measure for quantifying the contribution of each SNP in gene-gene interactions can be informative. However, the remaining difficulty is how to define a cut-off value of the IE measure to choose which SNPs to include or not in searching for gene-gene interactions. To overcome this problem, a RF method is chosen as a way to define such a cut-off value. SNPs that are above the cut-off will be further used to search for interaction effects.

To do so, the IE measure is used as a splitting criterion instead of Gini split criterion used on the original RF method (Breiman 1996). In this way, the IE measure is employed as a variable selection criterion in RF building procedures. Specifically, in each interior node of each tree a subset of r attributes is randomly selected and evaluated with the IE heuristic measure. The attribute with the highest IE is chosen as a split in that node. Therefore, SNPs that have high IE value are chosen first in tree building. The rationale is that a variable of importance tends to appear near the top of a tree. Each tree in the forest is grown to the largest extent possible without pruning. In such a scenario, SNPs that have potential interaction will have high possibility to appear in the same tree in the forest and have a high influence in discovering their interaction.

Once the forest is built, the variable importance measures employed by RF will be used to find the importance of each feature (SNP) in tree building. Several variable importance measures are available in RF implementations: the "Gini importance" that describes the improvement in the Gini gain splitting criterion, and another variable importance measure called the "permutation accuracy importance" measure. The rationale of the latter measure is that by randomly permuting the predictor variable x_i, its original association with the response y is broken. When the permuted variable x_i, together with the remaining unpermuted predictor variables, is used to predict the response, the prediction accuracy (i.e. the number of observations classified correctly)

decreases substantially, if the original variable x_i was associated with the response. Thus, a reasonable measure for variable importance to use is the difference in prediction accuracy before and after permuting x_j. Breiman (1996) suggests the difference in prediction accuracy before and after permuting x_i, averaged over all trees, as a measure for variable importance.

In this work, the "permutation accuracy importance" measure of RF is used as an importance criterion (termed "permutation importance" hereafter). This measure is consistent with the given objective of searching for features that have interaction effects on a given outcome (i.e. disease outcome). Features that have a high permutation importance measure tend to be consistent with the structures of built trees in the forest, where the relationships between variables can be detecting using the path of nodes in the built trees. The importance of each features based on permutation importance will be used to rank features in decreasing order, then an elimination strategy based on backward elimination approach can be used to iteratively remove features that show no or negative permutation importance.

The proposed feature elimination strategy used in the proposed RF-RFE approach can also be applied for selecting (prioritizing) markers that have high interaction effects based on the structure of built trees whilst maintaining high validation accuracy. The main difference in applying RF-RFE here is the use of IE measure as splitting

criterion in forest building at each iteration. The final set of SNPs selected (prioritized) based on RF-RFE is then used, as described next, to construct a new set of features that capture multi-locus combination effects (i.e. gene-gene interactions).

3.4 Feature Construction Using Feature Induction

Once the prioritized set of SNPs is selected, as described in section 3.3.2, it can then be used in conjunction with constructive induction algorithms to generate new attributes to capture interaction information. Here, the Odds Ratio based MDR (ORMDR) algorithm was used as a constructive induction method (Hahn et al. 2003). This section will describe the MDR method first and then an extended version of this method called ORMDR (Chung et al. 2007).

Multi-factor Dimensionality Reduction method

The MDR method was developed as a non-parametric and genetic model-free data mining approach to detect gene-gene interactions in the presence or absence of main effects in population-based studies in human genetics (Ritchie et al. 2001, Hahn et al. 2003). It has been shown to have high power in detecting interactions in a wide range of simulated data and has been successfully applied to detect gene-gene interactions for a variety of common human diseases (Brassat et al. 2006, Motsinger et al. 2007). However, MDR is computationally intensive because it relies on exhaustive search algorithms, cross validation and permutation tests (Niknian 1995, Musani et al. 2007).

Prioritizing and selecting interesting SNPs to be used with such an approach would be a vital way to control the computational complexity that arises when MDR is applied to large studies such as GWA studies. The heart of the MDR approach is to find a combination of attributes associated with disease outcome using constructive induction algorithms that create new attributes by pooling multi-locus genotypes or environmental factors into high risk or low risk groups. This reduces the dimensionality of the predictors from N dimensions to one dimension (i.e. the new attribute). The newly constructed attributes can be evaluated for their ability to classify and predict disease outcome (Moore et al. 2006).

The MDR is accomplished in the following way, given a threshold T, a multi-locus genotype combination is considered as high risk if the ratio of case to control exceeds T, otherwise it is considered as low risk. When the number of case and control data points are equal, it is natural to set $T = 1$. In this way, a new one-dimensional attribute will be constructed with two classes: high and low risk. Figure 3.4 shows an example of the MDR procedures applied to construct a new multi-locus attributes. The example is shown for a two-locus model. MDR evaluates the ability of the constructed attributes to classify and predict disease outcome by multi-fold cross-validation and permutation tests. The process is completed for each k-locus combination model. Models are chosen for each level of k combinations considered, so the final set of models comprise two-locus, three-locus models, and so on. In this

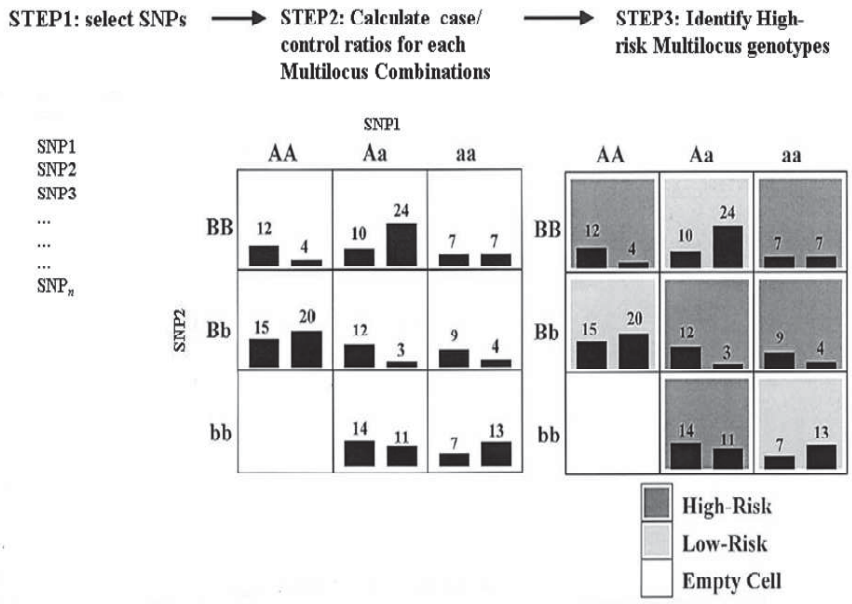

Figure 3.4: Summary of steps involved in constructing a new multi-locus attributes using the MDR method: each multi-factor cell in n-dimensional space is labelled as either "high risk" or "low risk" based on the case to control ratios. For each multi-factor combination, distributions of cases (left bars in boxes) and of controls (right bars) are shown.

thesis an open source software of MDR was used (`www.epistasis.org`).

Odd Ratio MDR

Although MDR has many advantageous features, it also has several drawbacks (Chung et al. 2007, Musani et al. 2007). The main limitation is that MDR classifies combinations of multi-locus genotypes into high risk or low risk groups in an unsystematic manner based on a simple comparison of the ratio of the number of case and control samples. Thus, the MDR is prone to false positive and negative errors when the ratio of case and control samples of a given multi-locus genotype combination is similar,

or when the frequencies of case and control samples are both small.

For this reason, Chung et al. (2007) proposed an Odds Ratio MDR (ORMDR) method that use odds ratio as a quantitative measure of disease risk. In ORMDR the combinations of genotypes are ordered from highest to lowest in term of odds ratios. The ORMDR method has been applied to a real dataset of Chronic Fatigue Syndrome disease and shows comparative results compared to the original MDR method (Huang et al. 2009).

In this thesis, the ORMDR method is used to construct a new set of multi-locus attributes using two-way, three-way and four-way genotype combinations based on the prioritized SNPs, as described in section 3.3.2. The proposed RF-RFE method is then used to select a subset of the newly constructed attributes as a set containing the most informative multi-locus combinations for the given disease. Selecting this final set is critical to account for gene-gene interaction effects of a given disease status.

3.5 Distance Measure Calculation

In the domain of genome-wide studies, several data mining and machine learning techniques require the computation of a distance matrix (e.g. kernel-based methods and unsupervised learning or visualization methods), which reflects the dissimilarities of genome-wide SNP profiles between any two individuals/patients. There are many possible metric measures that could be used to construct a similarity/distance matrix

as described in this section.

However, applying any similarity or dissimilarity metric directly to the high-dimensional input space of the given genetic data will not be informative. The high dimensional aspect of the genetic data, which may contain a great deal of irrelevant and noisy information, will distort any distance metric to reflect the exact similarity between data samples. Therefore, feature selection and extraction must be used to eliminate uncorrelated features and possible noise in the given complex data before distance calculations.

The proposed approaches for feature selection and detection of gene-gene interactions, described in earlier sections, can be used for this purpose. The proposed approaches are designed to reduce the number of features, remove irrelevant, redundant, or noisy data, thus, improving the modelling process and its performance as measured by predicting accuracy and interpretability of the results. In addition to reducing noise and improving the accuracy of classification, the selected subsets of markers may have important biological interpretations or may assist in identifying future possible research directions.

For any given disease model, once a final set of SNPs is determined, comprising the marginal-based correlated SNPs as found by methods described in section 3.2 and the interaction SNPs as constructed by methods described in section 3.4, then the combined sets can be used to calculate the similarity between any two individuals. The resultant similarities can be translated into a distance or dissimilarity matrix

using any distance measure such as the Euclidean distance. The distance matrix can be subsequently used for modelling different disease classifications, clustering and visualization models. In addition, one significant application of such distance matrices is to evaluate patient-to-patient relationships. Such relationships may lead to delivering an advanced personalized medicine, by choosing treatment strategies which best suit each individual patient compared to the most similar cases.

Most of the proposed disease modelling processes in this thesis are applied to genetic variation data and have been implemented using kernel-based methods, such as support vector machines, Multi-Dimensional Scaling (MDS) and others. The main procedure of kernel-based methods is to devise a suitable kernel function to encode similarity among data samples. Furthermore, different studies have shown that the performance of kernel-based methods is strongly related to the similarity measure used in the model building processes (Schölkopf & Smola 2002).

The aim of this section is to define new similarity measures for the analysis of genetic variation data. The main feature of the new measure is that the similarity between any two data samples is weighted according to the variable importance obtained using variable selection and/or weighting procedures, as described in section 3.2. The RF variable importance measure is used to define the weight of each attribute (i.e. genotype in this study). Another choice may be to assign weights of 1 or 0 when the feature is selected or not. Interaction-based constructed features can also be used directly in the similarity calculations.

To formulate the problem, suppose a genetic variation dataset is represented as a matrix $\mathbf{G}$ composed of m rows and n columns, where m is the number of individuals/patients and n is the number of SNPs.

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,n} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m,1} & g_{m,2} & \cdots & g_{m,n} \end{bmatrix}$$

The row vectors of $\mathbf{G}$ are represented as $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m$, where $\mathbf{g}_i$ represents the genetic profile of the i^{th} sample. The samples are genotyped and categorized into two or more classes, such as case and control. Each row $\mathbf{g}_i$ is composed of n variables, $g_{i1}, g_{i2}, \cdots, g_{in}$ representing the genotype information under study.

For this work, kernel functions that are based on the number of alleles shared between data samples being tested are proposed. The identity-by-state (IBS) method is used to compute the kernel functions (Wessel & Schork 2006). The IBS method is a widely used similarity measure in genome-wide studies and can be calculated as the fraction of alleles shared identical-by-state (IBS) for individuals i and j over all loci for which the individuals have been genotyped. The IBS kernel takes the form

$$\kappa_{IBS}(\mathbf{g}_i, \mathbf{g}_j) = \frac{\sum_{k=1}^{n} IBS(g_{ik}, g_{jk})}{2L} \tag{3.5.1}$$

where $IBS(g_{ik}, g_{jk})$ is a function mapping genotype information for individuals i and j at locus k to a particular numeric value and will take the value of 0 if individuals i and j are homozygous for different genotype alleles (e.g. $g_{ik} = AA$ and $g_{jk} = BB$);

a value of 1 if they share one allele (e.g., $g_{ik} = AA$ and $g_{jk} = AB$ or $g_{ik} = AB$ and $g_{jk} = AA$); and a value of 2 if they share both alleles (e.g. $g_{ik} = AA$ and $g_{jk} = AA$). The value L is the number of loci considered in the calculations. The denominator of equation (3.5.1) can be removed as being constant. The IBS kernel can be seen as an extension to the linear kernel function, where the dot product of the linear kernel is replaced with the IBS mapping function.

3.5.1 A Random Forest Based Kernel Function

An appealing feature of the IBS kernel is that it can be augmented to include specific weights that incorporate valuable prior information into analysis to potentially improve performance (Wessel & Schork 2006). In this thesis, the RF-based variable importance measure is incorporated into the IBS kernel as a weighting function. The proposed kernel is named RF-based Kernel (RFK) function. This kernel estimates the similarity between different genetic profiles based on the importance of each genotype of being part of a given group of subjects (i.e. a group of subjects with a specific phenotype or disease).

Define w_k as a scalar weight for genotype g_k, where w_k denotes the variable importance of g_k measured based on a RF approach. Based on equation (3.5.1), the RFK kernel can then be defined as

$$\kappa_{RFK}(\mathbf{g}_i, \mathbf{g}_j) = \frac{\sum_{k=1}^{n} w_k IBS(g_{ik}, g_{jk})}{\sum_{k=1}^{n} w_k} \tag{3.5.2}$$

3.5.2 Minor-allele Frequency and Entropy Based Kernel Functions

For unsupervised learning analyses, a distance measure can also be defined using a kernel-based weighting function. The minor-allele frequency (MAF) can be considered for such a purpose. A rare SNP with low MAF is given a higher weight compared to a more common SNP with a higher MAF. Such a weight could be valuable due to the potential information from SNPs with low MAF to be smoothed over by SNPs with more common MAF. To upweight SNPs with rare MAF, the weight $w_k = 1/\sqrt{q_k}$ is applied, where q_k denotes the MAF of SNP_k. Other MAF weights are also possible, such as $w_k = 1/q_k$, but there is concern that such stronger weights may substantially diminish the information provided by those SNPs with common MAF. Based on equation (3.5.1), a MAF-based kernel (MAFK) is defined as

$$\kappa_{MAFK}(\mathbf{g}_i, \mathbf{g}_j) = \sum_{k=1}^{n} \frac{1}{\sqrt{q_k}} IBS(g_{ik}, g_{jk}) \tag{3.5.3}$$

In addition to weights based on MAF, an information-based approach can also be used to weight each SNP. To weight SNPs based on information theory, an entropy based measure can be used, where $w_k = H(g_k)$. Such weights will give a greater weight for SNPs with higher entropy. Based on equation (3.5.1), an entropy-based kernel (EK) is defined as

$$\kappa_{EK}(\mathbf{g}_i, \mathbf{g}_j) = \sum_{k=1}^{n} H(g_k) IBS(g_{ik}, g_{jk}) \tag{3.5.4}$$

Finally, the proposed kernel based distance measures have been empirically evaluated in building different disease models, as described chapter 4 and 5.

3.6 Discussion

In this chapter, several methods and approaches have been defined and proposed to deal with the tasks of SNP selection, weighting, prioritization, feature construction and distance calculation in the domain of genetic variation studies. The novelty of this work is that the proposed approaches consist of screening genetic variation data to search for main and interaction effects of gene susceptibility markers that jointly cause complex diseases.

To deal with the problem of feature selection and weighting, the use of RF-based feature importance measure was proposed with an iterative process for the task of selecting and weighting SNP data, named the RF-RFE approach.

For the task of feature prioritization, a new measure called Interaction Effect was proposed. This measure can be used to prioritize SNPs based on their allelic distribution. This new measure uses the information gain (entropy) value of different disease statuses to calculate their distribution: highly skewed SNPs are the ones with less interaction effect. To define a cut-off value of a ranked list of SNPs based on their IE measures, the use of IE as a splitting criterion in RF trees construction was proposed. The proposed iterative feature selection approach can be used to find

the final list of SNPs that have potential interaction effects. Furthermore, a feature induction-based approach is used to construct new features based on the prioritized SNPs. The new constructed features are used to account for gene-gene interactions.

Finally, distance measures, for calculating distances between different genotype profiles, were proposed. The proposed measures use SNPs that have been selected, weighted and constructed based on the proposed approaches to calculate the similarity/distance between any two genotype profiles. A RF-based Kernel function was proposed named RFK. In this function, the RF-based variable importance measure was incorporated into the widely used IBS method as a weighting function. Other approaches were also proposed for unsupervised-based learning analyses including minor-allele frequency and entropy-based measures, named MAFK and EK kernel functions, respectively.

The proposed approaches in this chapter including feature selection, weighting, prioritization, construction as well as distance calculations address the thesis research questions (i.e. **Q1-Q3**) in section 1.3 and the thesis contributions 1-3 in section 1.4. The effectiveness of the proposed methods and approaches, presented in this chapter, is evaluated experimentally in chapters 4 and 5. Experimental results will point out the possible advantages and disadvantages of the proposed approaches.

Chapter 4

Disease Classification Models for Patient Diagnosis and Prognosis Based on Genome-wide SNP Profiles

Disease classification and prediction have become important applications of DNA microarray analysis. The last decade has shown rapid developments in our understanding of genetic etiology of various common complex diseases, such as cardiovascular diseases, type 1 and type 2 diabetes, Crohn's diseases and various cancers (Manolio et al. 2008). For a patient to receive an appropriate therapy, the clinician must identify as accurately as possible the disease type. Although traditional biomedical datasets such as clinical data or patient outcome data are still used as standard tools for disease diagnoses and prognoses, these datasets give very limited information and certainly miss much that is important about disease aspects and types.

Therefore, molecular-based diagnostic methods are needed to appropriately classify disease classes/subtypes. Most current molecular-based analyses look for DNA,

RNA, or protein markers that might be associated with a specific type of a disease and do not give biological information related to disease generation or progression. However, molecular-based technology, such as microarrays, has the advantage of containing large amounts of molecular information that can be extracted and integrated to find common patterns within different groups of samples. As this chapter shows, microarray technology, more specifically using genetic variation data, can be utilized to improve the accuracy of disease diagnosis and prognosis by looking at markers across the whole genome.

The past decade has reported many new genetic associations that have been identified by genome-wide association studies (Nolte et al. 2010, Craddock et al. 2010). Despite the exponential increase of such studies, few reported approaches have been developed for the analysis of microarray analysis for disease diagnosis and prognosis (Kruglyak 2008, Frazer et al. 2009). Genome-wide association studies are primarily limited to the fact that individual effect sizes of the reported associations are mostly small (McCarthy et al. 2008). Furthermore, there are also arguments that many disease-associated markers have not yet been identified, and the process of generating reliable prediction models may improve as more markers are included (Zhao & Liu 2009, Rosenberg et al. 2010).

The main idea of generating prediction models is to make a combination of markers, some having direct predictive (association) power, and others having low direct

associations, but instead having non-linear relationships in mapping genotype data to a given phenotype. Therefore, the approach of combining multiple markers into a single model can be stronger, especially by including markers that have non-linear associations.

Incorporating markers that have non-linear relationships is one of the most challenging issues in genetic studies. Non-linear associations can arise from phenomena such as locus heterogeneity, the dependence of genetic effects on environmental factors (i.e. gene-environment interactions) and the dependencies between genotypes at other loci (i.e. gene-gene interactions or epistasis). Epistasis is the phenomenon where interaction between two or more genes controls a single phenotype. One focus of this chapter is the detection and characterisation of gene-gene interactions in genetic studies. Detecting and characterizing the interactions has the potential to contribute to understanding the etiology of common complex diseases. However, gene-gene interactions are difficult to detect and characterize in genetic based studies due to the non-linear characteristics of markers involved in such studies. In its extreme form, interactions can occur in the absence of independent main effects of any involved markers.

This issue presents several very difficult computational and statistical challenges, especially in the context of genome-wide association studies (Moore et al. 2006). First, modelling non-linear interactions using parametric statistical approaches, such

as logistic regression, will not be effective and will have less power for detecting interactions, especially when main effects are absent (Cordell 2009). Therefore, especial analytical methods are required. Second, the implication of the high dimensional aspect of genome-wide studies with hundreds of thousands of single nucleotide polymorphisms makes the problem computationally infeasible. Exhaustive search of all possible SNP combinations (interactions) is not computationally feasible.

This chapter presents a new flexible framework for the tasks of disease diagnosis and prognosis based on methods and techniques proposed in chapter 3. Several SNP microarray datasets are employed to show the feasibility of the proposed framework. The results suggest that the proposed prediction approaches can effectively define important markers that are mostly consistent with known biological findings while the accuracy of the produced models is also high. Finally, other performance factors such as scalability and robustness properties of the methods are also investigated.

The reminder of this chapter is as follows: section 4.1 provides a practical description of the classification problem and its mathematical formulation. Section 4.2 demonstrates the proposed framework for the tasks of disease diagnosis and prognosis. Section 4.3 the proposed framework is empirically evaluated using case studies of acute leukaemia and the results are demonstrated in section 4.4. Finally, in section 4.6 the chapter is discussed and concluded.

4.1 Problem Formulation and Description

In classification tasks, we are interested in classifying whether a particular disease type is present or not. Other cases also include classifying patient samples to different disease subtypes. In a standard k-class classification problem, we are given a training dataset with m training observations. The training observations consist of n feature measurements, in the form of $(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \cdots, (\mathbf{x}_m, c_m)$, where $\mathbf{x}_i$ is an n-dimensional vector, $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})^T$ and the class label $c_i = y$ where $y = 1, 2, \cdots, k$. The m observations are aggregated into an m by n matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

For the problem we consider here, the features are genotypes, and the observations correspond to genotype profiles of samples or patients. Thus, $x_{i,j}$ is the genotype of the j^{th} SNP for the i^{th} sample. There are three possible genotype values for a given SNP, based on the minor allele frequency: homozygous in the major allele, homozygous in the minor allele and heterozygous. An ordinal scale can be used to code the SNP genotype sequences. The number 0, 1 and 2 is used for genotypes that were homozygous in the major allele, heterozygous and homozygous in the minor allele, respectively. Genotype profiles, for a given study, represent hundreds of thousands of

genotype data and it is a hard task to discover all hidden patterns in such data. For this reason, it is important to systematically find technical ways to understand such data.

This chapter is devoted to class prediction studies. Disease diagnosis and prognosis are examples of such studies. The purpose of class prediction, in genetic studies, is to understand the differences in disease markers that might be responsible for the differences between disease classes. A disease prediction model is a marker-based multivariate function that can use markers identified in class comparison to assign new patients into the correct class. However, disease prediction in genetics studies suffers from one major limitation termed as over-fitting. This means that the classification methods perform well on the training samples but poorly on independent test samples. Therefore, an important caution should be taken into account in dealing with such a problem. The high-dimensional aspect of the given data and non-linear interaction between markers are major causes of over-fitting. The methods and approach proposed in chapter 3 will be embedded in building classification methods to handle the high-dimensional aspect of the data.

Predictive models are one of the essential tools that can be applied to deliver personalized medicine. The feasibility of prediction models lies in the power of a proposed model in discriminating between individuals who will develop the disease of interest and those who will not. Methods to estimate the accuracy of prediction

models will be mentioned later in this chapter. Prediction of disease risk based on each individual genetic variant is considered as not informative, as most complex diseases result from the joint effects of multiple genetic and environmental variants, with each variant having a small contribution to the occurrence of diseases. Genome-wide based prediction of complex diseases would imply the efficient use of multiple genetic factors tested along the whole genome.

One approach to understand complex diseases is to build predictive models that able to identify genetic variants with strong effects, either on their own or in interaction with other variants or with environmental factors (i.e. gene-gene or gene-environment interactions). Yet, a feasible solution to this problem may only be achieved if we are able to understand the essential genetic factors that are jointly used to predict the risk of complex diseases. The proposed framework utilizes machine learning and data mining techniques to perform these tasks, more specifically, for the issue of gene-gene interactions.

4.2 Methods and Approaches

In the domain of biomedical research, many algorithms and approaches – in machine learning and data mining – have been designed and applied for the purpose of disease diagnosis and prognosis using different genetic-based data. However, the curse-of-dimensionality problem is the main limitation for building reliable models. To deal

with this problem, chapter 3 introduced methods and approaches for feature selection, weighting and prioritization. This section presents a multi-phase flexible framework to deal with different tasks of building reliable disease diagnostic and prognostic classification models.

4.2.1 Disease Classification: a New Computational Framework

The proposed framework, as shown in figure 4.1, consists of four main phases to accurately classify genetic variation profiles into different disease classes/subtypes. This framework can be seen as a flexible computational approach for understanding complex diseases. The process starts in a data preparation phase, where data from different disease models (i.e. additive, recessive and dominant) are generated and filtered using a quality control check. Then, data proceeds into phase 2, where feature selection, prioritizing and construction are used to generate an important list of features that can be used to classify patient samples to different disease classes/subtypes. Methods and approaches proposed in chapter 3 are used for this purpose. In phase 3, the combined subset of features selected in phase 2 is used to build classification models. The SVM methods are used for this purpose. Once the models are built, they are evaluated in phase 4. Here, the multi-phase framework is presented and its application to real biological datasets is demonstrated in section 4.3.

The proposed framework differs form other counterpart DM frameworks, such as CRoss Industry Standard Process for Data Mining (CRISP-DM) (Wirth 2000), in that the underlying procedures in each phase of the proposed framework are designed to suit the given domain and its computational complexity. In particular, the data preparation and feature selection phases are tailored to the given domain. It is important to note that a wide range of different methods and algorithms can be used at each phase. In addition, there is no one strategy that is likely to be universally optimal. The following describes each of these phases in turns and the specific algorithms and methods used.

Phase 1: *Data cleaning and preparation*

The goal of the first phase is to preprocess a given genotype dataset including data cleaning and preparation. In terms of data cleaning, in GWA studies a large volume of genotype data is generated, which might include thousands of individuals genotyped at hundreds of thousands of markers. In consequence, several statistical tests must be applied to ensure having high quality data (cleaned data).

These methods are called quality-control criteria checks and have been widely used in the literature (Zeggini et al. 2007, Wang et al. 2005). There are two types of criteria: individual exclusion criteria and marker exclusion criteria. For individual exclusion criteria, any individual with genotyping call rates less than 95% must be excluded. For marker exclusion criteria, markers should be checked thoroughly, and

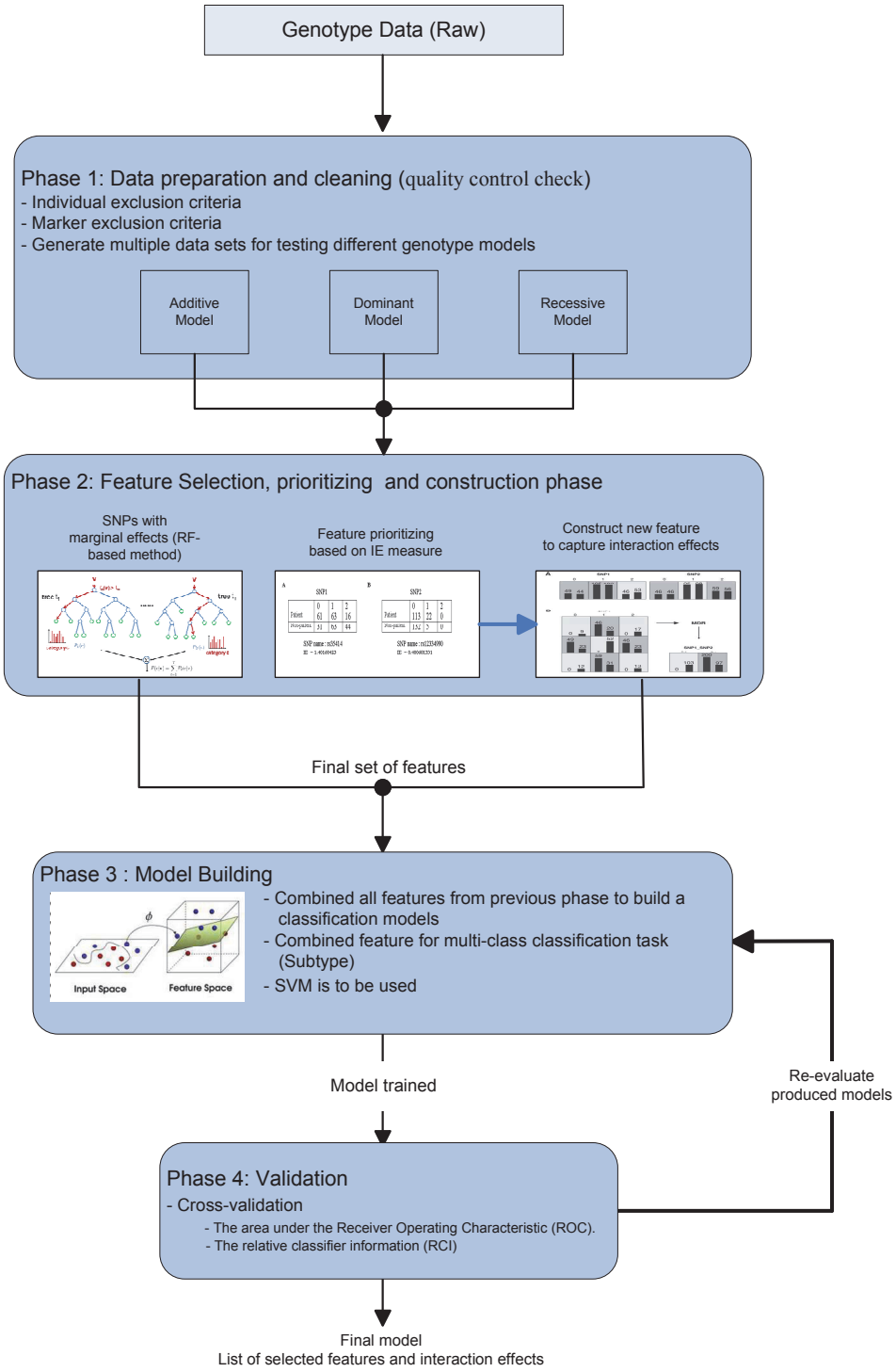

Figure 4.1: Disease diagnosis and prognosis classification Framework

those markers that show insignificant importance should be discarded. Markers must pass the following quality control criteria: (1) they did not map to multiple locations in the genome, (2) they showed a $> 95\%$ genotype call rate and a $> 90\%$ genotype call rate in both population subsets of data, (3) they have a minor allele frequency (MAF) $> 1\%$ in the case samples and $> 1\%$ in both case and control subsets of the data and (4) they demonstrated the Hardy-Weinberg equilibrium with a p-value $> 1e$-06 in controls.

In terms of data preparation, genotype data must be prepared to generate multiple datasets that can be used to represent different genetic models. These include additive, dominant and recessive models. The three genetic models are described as follows:

Additive Model: Under this model, testing is designed specifically to reveal associations that depend additively upon the minor allele. That is, where having two minor alleles rather than no minor alleles (AA) is twice as likely to affect the outcome in a certain direction as is having just one minor allele (AB) rather than no minor alleles (BB). The attribute used in this model is the count of minor alleles B, which is zero within the genotype AA, one within genotype AB, and two within genotype BB, where A is the major allele.

Dominant Model: This model specifically tests the association of having at least one minor allele B (either AB or BB) versus not having it at all (AA). The

count is zero for genotype AA and one for genotypes AB and BB.

Recessive Model: This model specifically tests the association of having the minor allele B as both alleles (BB) versus having at least one major allele (AA or AB). The count is zero for genotypes AA and AB and one for genotype BB.

In generating such data, it is important to test for marker association and/or differentiation with a given disease based on these genetic models.

Phase 2: *Feature selection, prioritizing and construction*

The goal of this phase is to select the minimum set of SNPs with the highest discriminative power. In principle, many feature selection methods can be applied to fulfil the objective of this phase. However, because of the "curse-of-dimensionality" problem with genetic variation datasets, most classical feature selection approaches are not computationally feasible. To deal with the "curse of dimensionality" problem induced by the data and non-linear interaction between multiple markers and disease outcome, non-parametric machine learning techniques have been chosen. The proposed approaches endeavour to use feature selection methods for selecting disease-susceptibility markers that account for marginal effects as well as gene-gene interactions. Performing feature selection based on traditional methods instead would result in choosing markers that marginally contribute to disease but ignore gene-gene interactions that would be useful for producing highly predictive models for the large scale genome-wide studies, where the number of SNPs is in the order of hundreds of

thousands.

This phase consists of three main tasks with different emphases. These tasks include feature selection, prioritization and construction. Chapter 3 proposed solutions for these tasks. For the tasks of feature selection and weighting, the use of the RF feature importance measure was introduced, in section 3.2, as an iterative process, called RF-RFE. This approach is used to select a set of markers that marginally correlate to a given disease status.

For the task of feature prioritization, a new measure called Interaction Effect (IE) was proposed in section 3.3. This measure prioritizes SNPs, to be tested for their interaction effects, based on their allelic distribution, thus, reducing the number of multi-locus combinations needing to be tested. To define a cut-off value of a ranked list of SNPs based on their IE measures, the use of IE as a splitting criterion in RF trees construction is also proposed. Based on the IE-based constructed forest, the proposed RF-RFE approach is used to find a list of SNPs that has potential interaction effects (prioritized SNPs).

Furthermore, in section 3.4, a feature induction-based approach is used to construct new features based on the prioritized SNPs for all two-way, three-way and four-way SNP combinations. The newly constructed features are considered to account for gene-gene interactions. The RF-RFE is also used to obtain an optimal set of features (based on the new constructed feature) that explain a given disease

model. Finally, the two subsets of feature, representing SNPs with both marginal and interaction effects, are combined to form a final set of selected features.

The identification of subtype-discriminating feature sets is performed using cases in the training set and is approached using both parallel and differential decision tree formats (Yeoh et al. 2002). In the parallel approach, class-discriminating sets are identified by defining a set of selected features that is specific to one class compared to all other cases in the training set. The class-discriminating sets for all classes are then combined and used to assign cases to different classes based on the model building process (Phase 3).

In the differential decision tree format, class-discriminating markers are first selected for one class against all other cases. The selected class cases are then removed, the selection process reapplied, and then class-discriminating genes are selected for the second class cases. The latter cases are then removed and the process repeated proceeding sequentially through all other classes. The combined identified discriminating features can then be used in the next phase for model building.

Phase 3: *Model building*

Once the final set of SNPs is determined from phase 2, they can then be used as a feature set to build classification models, using any machine learning classification method. The Support Vector Machine classifier is selected here because it has shown superior results compared to other state-of-art data mining classification methods

(Statnikov et al. 2008). a detailed review of SVM can be found in chapter 2.

Several theoretical reasons explain the superior empirical performance of SVMs in microarray data including the flexibility of choosing an appropriate kernel function for a given application (Burges 1998). Depending on whether a classification problem is linearly separable or not, the SVM can choose either a linear kernel or non-linear kernel function. Polynomial and radial basis kernel functions are the most commonly used non-linear kernel function. A non-linear kernel constructs a set of hyperplanes in a higher dimensional space to separate two classes. In addition, the introduction of soft margin approach to SVMs is another key feature of SVMs (Vapnik et al. 1996). Indeed, SVMs exhibit higher classification accuracy than random forests in different settings, even when no feature selection is performed and when several feature selection methods are used (Statnikov et al. 2008).

In the proposed framework, SVM implementations are adopted with different parameter settings in order to maintain the highest prediction accuracies. There are several user-defined parameters in SVM which affect the training/testing results. To handle multi-category data, a "one-versus-rest" SVM approach is adopted for the analysis of multi-category datasets. The "one-versus-rest" SVM approach involves building a separate SVM model to classify each class against the rest, and then predicting the class of a new sample using the SVM model with the strongest vote (Rifkin et al. 2003). To assess the effect of data transformation and mapping, different

kernel-based functions are applied and their predictive performances are compared. The kernel parameters are optimized using a nested cross-validation approach, and the ones with the best cross-validation accuracy selected (Friedman et al. 2001). Trying exponentially growing sequences of values is a practical method to identify good parameters (Chang & Lin 2011).

Finally, although the SVMs algorithm has several attractive characteristics, it is important to note that other classification methods such as k-NN can be used. To do so, extensive experimental procedures can be carried out to compare the resultant models. However, in this thesis experiments were confined to using only SVM classifiers.

Phase 4: *Validation*

In the final phase, validation of produced models can be carried out using different evaluation criteria. The validation seeks to maximize the prediction performance (e.g. accuracy) whilst assessing the generalization of the produced models. To measure classification performance, it is necessary to apply the tests on independent data samples than was used for model building. However, in terms of microarray data, obtaining an independent test data is very difficult and/or expensive, because most genetic datasets have very limited numbers of samples compared to variables.

To deal with this problem, the k-fold cross-validation technique is used (Friedman et al. 2001). The cross-validation approach performs independent tests without

requiring separate test data samples and without reducing the data samples used to build classification models. The cross-validation process involved partitioning the original data into k groups (in k-fold cross-validation), so that each group is used once as a validation set and the remaining data as a training set. Another simple cross-validation approach, called Leave-one-out approach, is also applicable. This approach involves using a single data sample from the original data as the validation data, and the remaining data as the training data. This approach is repeated so that each data sample in the original data is used once as a validation set.

Once the validation dataset is determined, a classification performance measure is applied to find the model's performance. Two classification performance metrics were used. For binary tasks, the area under the ROC curve (AUC) metric is used, which can be computed from continuous outputs of a given classifier. For multi-class tasks, where the AUC metric is inapplicable, the relative classifier information (RCI) metric is employed (Sindhwani et al. 2001, Pirooznia & Deng 2006). The RCI metric is an entropy-based measure that quantifies the amount of uncertainty of a decision problem that is reduced by a classifier relative to classifying using only the prior probabilities of each class. The RCI is similar to the area under the ROC curve (though not equivalent) in that it measures the discrimination power of a classifier. Both measures are not sensitive to unbalanced distributions, unlike the accuracy metric (proportion of correctly classified data samples). Both AUC and RCI take

values between [0, 1], where 0 denotes the worst possible classification and 1 denotes perfect classification.

An iterative process of model building and validation process must be carried out multiple times to check the generalization of the produced models. The final model is chosen based on stable generalization results.

4.3 Experimental Design

How will the proposed framework perform when applied to complex datasets from real genetic-based studies? To evaluate the performance of the proposed framework in practice, datasets from studies of Acute Lymphoblastic Leukaemia datasets are analysed and the results of the proposed framework are evaluated.

The following subsections explain different aspects of the datasets and the experimental procedures that have been applied to the given leukaemia datasets.

4.3.1 Genome-wide SNP Data

Genome-wide SNP data incorporates large scale mapping of SNPs and subsequent collation into databases. Generation of SNP data has been facilitated by high throughput microarray-based technologies (Barker et al. 2004, Leykin et al. 2005, Irving et al. 2005).

The domain of this study is Childhood Acute Lymphoblastic Leukemia. Paediatric ALL can be divided genetically into six distinct subtypes including (1) B-progenitor leukaemias with translocations t(9;22) [BCR-ABL1], (2) t(1;19) [TCF3-PBX1], (3) t(12;21) [ETV6-RUNX1], (4) rearrangements of MLL, (5) hyperdiploid and hypodiploid karyotypes and (6) T-lineage leukaemia (T-ALL). These genetic lesions are important in leukaemia initiation (Greaves & Wiemels 2003), but alone are insufficient to predict the phenotype. Other genetic factors must also be incorporated. Although additional mutations/genetic factors have been identified in a subset of cases, the full genetic basis and their distribution within the known genetic subtypes of ALL remains to be defined.

To obtain a comprehensive understanding of genetic lesions in ALL, genetic variation of the human genome can be a promising resource for studying the genetic basis of complex disease. Large numbers of genetic variations, scattered across the human genome, represent a remarkable opportunity to investigate the etiology, inter-individual differences in treatment response and outcomes of any given disease. Therefore, SNP data should be utilized to analyse genetic basis of the ALL domain.

This section describes in detail the SNP datasets used and the preprocessing steps applied to these datasets. There are two leukaemia datasets used in this study: (1) a genome-wide SNP dataset, generated in the Oncology Research Unit at the Children's Hospital at Westmead and (2) a genome-wide SNP dataset, generated at

the St Jude Children's Research Hospital (SJCRH) (Mullighan et al. 2007). The Westmead dataset is used for binary classification, while the St Jude is used for the purpose of subtype discovery (multi-class subtype classification). A summary of these ALL datasets is given in table 4.1.

Table 4.1: Summary of used ALL datasets

Dataset	Sample size	No. of attributes	Classes
Westmead	279	13917	2
St Jude	242	262,000	7

4.3.2 Acute Lymphoblastic Leukaemia Datasets

The Westmead dataset Data was obtained from the Tumour Bank at the Children's Hospital at Westmead (CHW), with approval of the hospital's Human Research Ethics Committee and the Tumour Bank Committee. DNA isolation for the Illumina BeadArray was prepared from frozen whole peripheral blood samples from ALL patients taken when the patient had reached clinical and pathological remission, normally 3 months after initial diagnosis. As few, if any, leukemic blasts would be present in these samples, it means that these samples represented the normal genetic status of the patient. DNA was isolated using the DNA Flexigene kit (Qiagen). Processing of the DNA sample on the Human NS-12 Whole-Genome assay was performed according to manufacturer's instructions. All labeling reactions proved successful and met quality criteria.

The labelled samples were hybridized to Human NS-12 Genotyping. Data was extracted using the Bead Studio software (Illumina, Inc.) with the normalized theta values for each SNP in each patient used in the analyses in preference to the SNP genotype call. Theta values were coded as 0, 1 and 2 representing AA, AB and BB, respectively.

A cohort, the "case" set, of 139 childhood ALL patients was generated with the Illumina Bead Array system using the non-synonymous bead-array chip to assess 13,917 SNPs across the genome within exon-centric loci. SNPs were scattered across the whole genome. These SNPs are classified as non-synonymous (functional) SNPs which affect the functionality of genes.

A dataset of 140 samples of individuals retrieved from the Wellcome Trust Case Control Consortium (WTCCC) was also studied, with samples coming from the "1958 British Birth Cohort study". This set was used as control individuals (Wellcome Trust Case Control Consortium 2011, Craddock et al. 2010). The sample data was generated using the same SNP-chip that was used to generate the case data. The case and control data are considered to have the same ethnic background. The combined dataset of 279 samples in total, 139 ALL samples obtained at CHW and 140 of non-patient samples (control) retrieved from WTCCC, is called the Westmead dataset hereafter.

The St Jude dataset

Two hundred and forty two patients with acute lymphoblastic leukaemia treated at St Jude Children's Research Hospital (SJCRH) were studied. These included precursor-B ALL with high hyperdiploidy (greater than 50 chromosomes on karyotyping, HD>50), TCF3-PBX1 positive (E2A-PBX1, 17 samples), ETV6-RUNX1 positive (TEL-AML1, 47 samples), BCR-ABL1 positive (9 samples), MLL rearranged (11 samples), low hyperdiploidy (47-50 chromosomes, HD47-50, 23 samples), hypodiploidy (10 samples), B-ALL with pseudodiploidy, near haploid karyotype, normal cytogenetics or non-recurring cytogenetic abnormalities (36 samples), and T-lineage ALL (T-ALL, 50 samples). The list of ALL samples and the subtype distribution of the cases used in the St Jude data are shown in table 4.2. A study approval was obtained from the SJCRH institutional review board.

The study examined DNA from the leukaemic cells (blasts) of 242 cases of different paediatric ALL subtypes genotyped using Affymetrix array. DNA was extracted from diagnostic leukaemia samples using the DNA Blood Mini Kit (Qiagen, Valencia, CA). Each sample was genotyped with Affymetrix GeneChip Human Mapping 262K Affymetrix arrays, which provide consistently high coverage across the whole genome. It is capable of genotyping approximately 262,000 SNPs.

Table 4.2: Subtype distribution of the St Jude ALL cases.

Subtype	**Sample Size**
BCR–ABL	9
E2A–PBX1	17
Hyperdiploid with more than 50 chromosomes	39
MLL	11
T–ALL	50
TEL–AMLI	47
Other	–
hypodiploidy	10
pseudodiploidy	36
Hyperdip47–50	23
Total	**242**

4.3.3 Data Preprocessing

Each SNP can have four different values (nominal): two homozygous, one heterozygous or missing. That is, the four possibilities for alleles A and B of the i^{th} SNP are two homozygous (AA or BB), one heterozygous (AB) or missing NA (not determined). An ordinal scale was used to code the SNP genotype sequences. The number 0, 1 and 2 were used for genotypes that were homozygous in the major allele, heterozygous and homozygous in the minor allele, respectively. The missing SNPs were imputed using mode of each SNP. The following describes the quality control criteria applied to both datasets.

The Westmead dataset

Genotyping of 13,917 SNPs was attempted in each sample. SNPs with a call rate less

than 99% were excluded. After applying stringent quality control criteria, high-quality genotypes for 10,750 SNPs were obtained. To advance genome-wide association and classification studies, markers that showed insignificant importance were discarded, as described in selection 4.3.3. Table 4.3 outlines the number of SNPs that have been excluded. The total number of markers considered for further analysis, after discarding 3176 SNPs, was 10,750 SNPs.

The St Jude data

Genotyping of 262,668 SNPs was attempted for each sample. SNPs with call rate less than 99% were excluded. After applying stringent quality control criteria, high-quality genotypes for 93,071 SNPs were obtained and considered for further analysis. Table 4.3 outlines the number of SNPs that have been excluded.

Table 4.3: The excluded SNPs for both the Westmead and the St Jude datasets.

Quality control criterion	Westmead	St Jude
Map to multiple locations	20	233
Genotype rate (>95%)	704	131198
MAF (>1%)	2199	38166
HWE (> 0.001)	244	-
Total	**3167**	**169597**

4.3.4 Experimental Procedures

In terms of experimental procedures, the proposed four-phase framework, as described in section 4.2, is applied to the two ALL datasets.

Application to the Westmead Dataset

The case-control dataset (the Westmead dataset) was used for the task of disease classification. After the given genotype dataset was cleaned and prepared, with the approach described in section 4.3.3, the proposed RF-RFE method was used to select SNPs that marginally contribute to the disease status. Then the proposed IE measure was used to estimate the interaction effect of each SNP. SNPs with the highest IE were prioritized and used to construct new attributes. However, it is important to note that the use of the IE measure as a splitting criterion in RF, as proposed in the framework, has not been applied to the given experiments. The SNPs with the highest IE were chosen empirically.

The ORMDR method was employed to construct a new set of multi-locus attributes using two-way, three-way and four-way SNP combinations based on the prioritized SNPs. The RF-RFE approach was then applied on each set of the new multi-locus constructed attributes individually to select the most significant attributes. The selected multi-locus constructed attributes represent the SNP combinations with high potential interaction effects. The two subsets of selected features are considered as a final set of features with this set representing SNPs with both marginal effects and interaction effects. The final set of features including the originally selected SNP features and the newly constructed features were used in building SVM classifiers for the task of disease classification.

The next step is to compare the classification accuracy of the SNPs selected for marginal effects alone and the combined SNP sets representing both marginal and interaction effects. For this purpose, two datasets were created. The first dataset consisted of the marginal effect SNPs and the second dataset consisted of both marginal and constructed interaction effect SNPs. Using 10-fold cross validation, the accuracy, sensitivity, precision and specificity of SVM classifiers of the two datasets were compared. The *mean* of each of these measures was compared using a corrected re-sampled t-test, and were considered statistically different when the p-value was less than or equal to the acceptable type I error rate of 0.05. The proposed feature selection methods have been undertaken independently for each cross-validation training subset and the results were reported on the remaining test subset. The same procedure was also applied to the St Jude dataset, as described in the next section.

A similar comparison was also applied between the combined feature set (i.e. marginal and interaction sets) and the feature set of all SNPs involved in all levels of interaction. For this purpose, two datasets were created. The first dataset consisted of both marginal effects and constructed interaction effects SNPs. The second dataset consisted of all SNPs involved in different SNP combinations. Using 10-fold cross validation, the accuracy, sensitivity, precision and specificity of SVM classifiers of the two datasets were also compared.

Application to the St Jude Dataset

The St Jude dataset was used as a case study for the task of subtype discovery models. The data description and preprocessing analyses that have been applied to this dataset are described in sections 4.3.2 and 4.3.3. The goal of this study was to identify the genetic variations that could be used to accurately diagnose the known prognostic subtypes of ALL. The identification of subtype-discriminating SNP sets was performed exclusively using cases in the training sets and was approached using both a parallel and a differential decision tree format as described in the framework, section 4.2. In the parallel approach, class-discriminating SNP sets were first identified by selecting SNPs that had a genetic variation that was specific to one class compared with all other cases in the training set.

The selected SNPs for each class-discriminating set were then combined and used in supervised learning algorithms to assign cases to one of six subtypes (T-ALL, E2A-PBX1, MLL rearrangement, BCR-ABL, TEL-AML1, hyperdiploid with more than 50 chromosomes), where unclassified cases considered as other. For the differential decision trees approach, class-discriminating SNPs were first selected for T-ALL against all other cases. The T-ALL cases were then removed, the selection procedures reapplied, and then class-discriminating SNPs were selected for E2A-PBX1 cases. The latter cases were then removed and the process repeated proceeding sequentially through TEL-AML1, BCR-ABL, MLL rearrangement, and hyperdiploid

with more than 50 chromosomes cases.

The class-discriminating SNPs were selected using the proposed RF-RFE approach. In this analysis, the proposed RF-RFE approach was applied to identify SNPs across the dataset showing variation between the specific class and non-class cases. The identified discriminating SNPs were then used to build classifiers that could accurately identify the specific genetic subtypes. As described in the proposed framework, the SVMs algorithm was used. Performance of each model was assessed by 10-fold cross-validation on a randomly selected stratified training set.

Whilst the proposed RF-RFE feature selection approach is not the main component used in the proposed framework, however, for comparison purpose the performance of the RF-RFE approach was compared with two other feature selection approaches. These include the SVM Recursive Feature Elimination (SVM-RFE) feature selection method (Guyon et al. 2002) and a RF-based feature selection method called the VarSelRF approach (Díaz-Uriarte & de Andrés 2006).

4.4 Experimental Results

4.4.1 Genetic Variation Profile as Diagnostic Tool

The Westmead dataset was used for building a diagnostic model. The data represents 279 Samples (139 cases and 140 controls) genotyped using 13,917 non-synonymous

SNPs across the whole genome. After data preparation and cleaning, as described in section 4.3.3, a SNP set of 10,750 SNPs were used for further analysis.

Feature selection results

To select SNPs that marginally contribute to the disease, the proposed RF-RFE approach was applied to the 10,750 SNPs. The RF-RFE approach was run using a forest with a large number of trees (i.e. $ntree$ = 200,000, $mtry = 5 * \sqrt{n}$, where n is number of variables/SNPs considered at each iteration) and the SNP set with the minimum OOB error was selected. The forest's size was set large due to the high dimensionality feature space of the Westmead dataset. A set of 286 SNPs was considered as the best set of class-discriminating SNPs between the two classes. Figure 4.2 shows OOB error rates of the RF-RFE procedure applied to the Westmead dataset, as a function of number of SNPs maintained at each iteration. As can be seen in figure 4.2, the minimum OOB error rate achieved on the training set was 8% (92% accuracy). Although RF-RFE approach was used as a feature selection tool, these initial results show that SNP data does not have efficient power in producing highly accurate classifiers, in addition to the fact that the given classifiers had an over-fitting problems when applied to test data points (results shown in table 4.6).

To reduce the multi-combination search space of the given high dimensional dataset the IE measure was applied to the 10,750 SNPs and the 1000 highest IE ranked SNPs were considered to construct new attributes for two-way (499,500 new attributes),

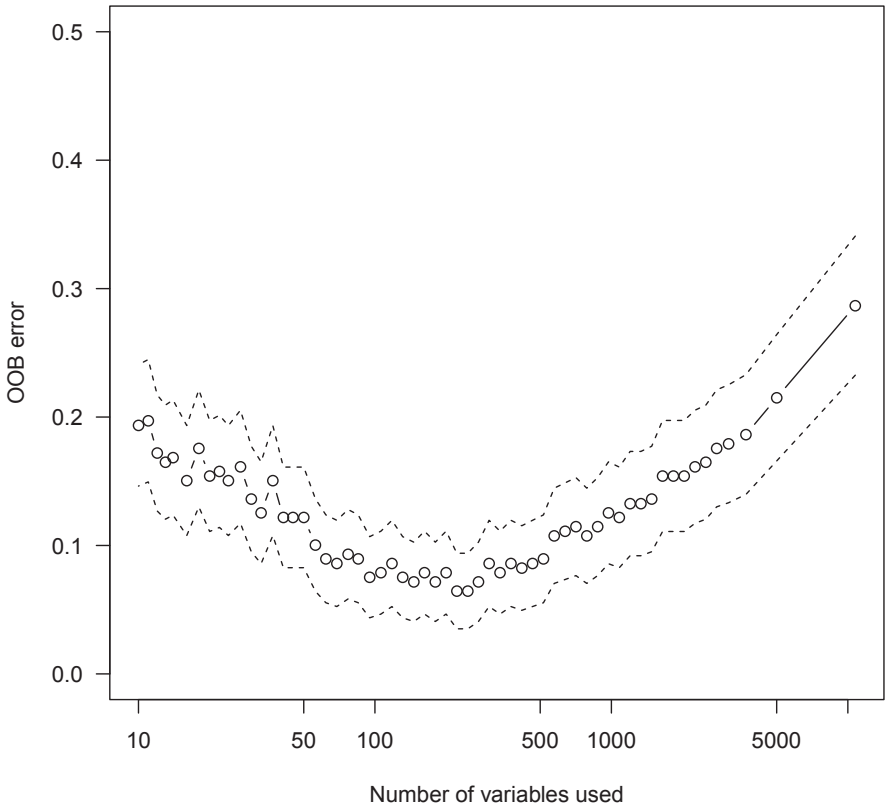

Figure 4.2: The OOB error rates of the RF-RFE procedure applied to the Westmead dataset, as a function of number of SNPs maintained at each iteration of built models

three-way (166,167,000 new attributes) and four-way (41,417,124,700 new attributes) interactions using ORMDR method. Then, given the fact that the potential interaction set that could describe a given disease status should not be large, the top 1000 highly correlated combinations of each of these new constructed attributes were considered to select a subset that accurately classify the given ALL patient from normal

individuals.

The RF-RFE approach was also used to select the best subset of new constructed attributes for each way of interaction. The RF method was set with $ntree = 10,000$ and $mtry = 5 * \sqrt{n}$, as the dimension of the attributes considered is reasonably not large. Table 4.4 shows the tests that have been applied and the number of selected attributes for each way of interactions.

Table 4.4: Feature construction and selection for each way of interactions using the highest 1000 IE SNPs.

Interaction	Number of SNPs considered for interaction	Number of new constructed attributes	Number of possible new constructed attributes considered for feature selection	Number of new constructed attributes selected using RF-RFE
Two-way	1000	499,500	1000	351
Three-way	1000	166,167,000	1000	154
Four-way	1000	41,417,124,700	1000	57

The two subsets of selected features, the selected marginal effect SNPs and interaction SNP sets for all two, three and four ways were combined into one dataset of 848 features, as can be seen in table 4.5. The new dataset is considered as a final set of features with this set representing both SNPs with marginal effects and those with interaction effects.

There are two extreme cases of SNPs that could be associated with a given disease. In the first case, SNPs that have marginal associations with a given disease may not

necessarily be involved in gene-gene interaction effects. The second case includes SNPs that have small marginal effects but could have potential interaction effects. The IE measure that has been used to prioritize SNPs to be tested for interaction effects can be informative under both of these cases. Based on the IE measure, out of the 286 SNPs that are marginally associated with the disease, there are only 15 SNPs that were involved in gene-gene interactions. On the other hand, most SNPs that were involved in interaction effects do not marginally contribute to the disease, as these SNPs have been selected based on their interaction effects to the given disease. This fact says that SNPs that have marginal effects on the given disease do not necessarily have interaction effects.

Table 4.5: Feature selected for marginal effects and each level of interaction.

Type of effect	Number of selected SNPs/attributes
Marginal SNPs	286
Two-way	351
Three-way	154
Four-way	57
Total	**848**

Classification results and comparisons

In the next step, the newly constructed attributes for different combinations of interactions were combined into one feature set defined as the interaction set. Then, the classification using the selected SNPs with marginal effects is compared with the combined feature sets of marginal and interaction sets. To avoid over-fitting, 10-fold

cross-validation is used.

Using 10 runs of 10-fold cross validation and a corrected re-sampled paired one-sided t-test, the accuracy, sensitivity, specificity and precision performance measures of SVM classifiers were compared. Parameters for the SVM classifiers were chosen by nested cross-validation procedures and the ones with the best cross-validation performance were selected, as described in the section 4.2. The SVM models were built using a polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^T \mathbf{y} + r)^p$ where $\mathbf{x}$ and $\mathbf{y}$ are samples with genotype information and p, γ and r are kernel parameters. The optimization of classifiers was performed using values of cost C, the penalty parameter of SVMs, ranging through 0.0001, 0.01, 1 and 100 and $p = \{1, 2, 3\}$. The cost function $C = 0.01$ and $p = 1$ performed the best. The kernel parameters γ and r were set to the default values $\gamma = 1/\text{number of variables}$ and $r = 0$. The SVM classifications were performed using SVMs implementation in the libSVM software library (Fan et al. 2005) (`http://www.csie.ntu.edu.tw/~cjlin/libsvm`).

Table 4.6 summarizes the results of these comparisons. The results report the mean and, in parentheses, the standard deviation of accuracy, specificity, sensitivity and precision of the two models (classifiers). The results show that the combined feature set performed better than the marginal-only set over each estimated performance measure and that these improvements are statistically significant. This suggests that

considering the interaction SNP sets between SNPs can significantly improve classification performance in the presence of non-linear gene-gene interaction and alleviate the over-fitting problem induced by the high-dimensionality of the given data.

Table 4.6: Statistical comparison of performance measures for the Westmead dataset including selected SNPs based on marginal effects and combined marginal and interaction sets. Values are mean (Standard Deviation) of each estimated performance measure over 10 runs

Measure	Marginal effects	Both marginal and interacted sets	p-value
Accuracy	0.65 (0.04)	0.93 (0.03)	$< 2.2e$-16
Sensitivity	0.66 (0.03)	0.94 (0.07)	$< 2.2e$-16
Specificity	0.90 (0.06)	0.91 (0.04)	< 0.001
Precision	0.87 (0.07)	0.93 (0.04)	$< 2.2e$-4

A similar comparison was also applied between the combined marginal and interaction feature sets and the feature set of all SNPs included in the final model (common SNP set). The common SNP set comprised of 788 SNPs involved in the marginal, two-way, three-way and four-way was considered. Using 10 runs of 10-fold cross validation, the classification performance of the two datasets was compared using a paired one-sided t-test using SVM models.

Table 4.7 summarises the results of these comparisons. The results show that searching for interacting SNPs and constructing new attributes that reflect the interactions will improve the performance of the produced models and avoid the over-fitting problem that can be introduced by the given high-dimensional datasets. Although SNPs that were involved in different interaction combinations are all considered in

building a classification model, it is very difficult for any classification algorithm to discover the non-linear interactions between the interaction SNPs.

Table 4.7: Statistical comparison of performance measures for the Westmead dataset including selected SNPs based combined marginal and interaction sets, and the same features without constructing interaction SNPs. Values are mean (standard deviation) of each estimated performance measure

Measure	Common SNP set	Both marginal and interacted sets	p-value
Accuracy	0.70 (0.03)	0.92 (0.02)	$< 2.2e$-16
Sensitivity	0.72 (0.5)	0.94 (0.65)	$< 2.2e$-4
Specificity	0.90 (0.06)	0.91 (0.04)	< 0.001
Precision	0.90 (0.05)	0.92 (0.04)	$<$0.0001

Finally, figure 4.3 shows the ROC plot of SVM models on the three selected feature sets: the marginal effect SNP, common SNP set and the combined marginal and interaction feature sets. The model built based on the marginal effect feature set has an AUC of 0.6833 compared to an AUC of 0.7735 and 0.9528 of the common and the combined marginal and interaction feature sets, respectively. Furthermore, figure 4.4 shows the box plots of AUC results of 10 cross-validation runs based on the marginal, common and combined feature sets. As can be seen in this figure, the AUC results of the combined feature set shows stable generalization results.

4.4.2 Genetic Variation Profile as Prognostic Tool

The St Jude dataset was used for building prognostic models. The dataset represents 242 ALL patients with different subtypes of leukemia genotyped using 262,668 SNPs

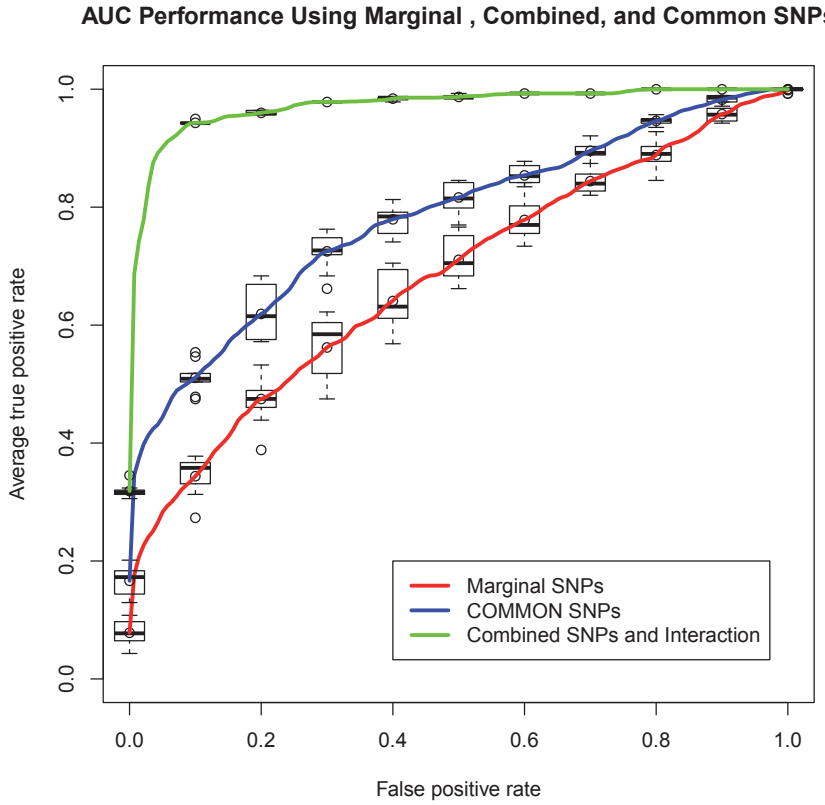

Figure 4.3: ROC curves of the Westmead dataset based on the feature selection procedures (marginal, common and combined feature sets)

across the whole genome. After data preparation and cleaning, as described in section 4.3.3, a SNP set of 93,071 SNPs was used for further analysis.

Feature selection results

Feature selection procedures were used to identify SNP sets that are the best discriminators for each individual's leukemia subtype. For this task, the proposed RF-RFE approach was applied as a feature selection method. Class classification has been

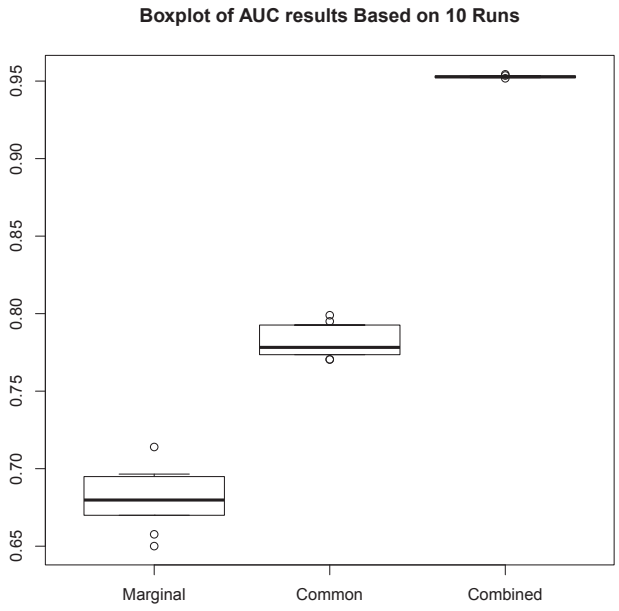

Figure 4.4: Box plots of AUC results of 10 runs based on marginal, common and combined feature sets

approached using both parallel and decision tree approaches. Since the parallel approach has shown to provide the highest level of accuracy, this approach was chosen to perform most of the analyses.

For each subtype, the RF-RFE approach was run using a forest with a large number of trees (i.e. $ntree$ = 400,000, $mtry = 5 * \sqrt{n}$, where n is number of variables/SNPs considered at each iteration) and the SNP set with the minimum OOB error was selected. The forest's size was set to be large due to the high dimensionality feature space of the St Jude dataset. A discriminating SNP set with the minimum

Table 4.8: Summary of selected SNPs for each of ALL subgroup applied to the St Jude dataset

Subgroup	Number of selected SNPs/attributes
T-ALL	47
TEL-AML1	28
Hyper>50	28
E2A-PBX1	37
MLL	33
BCR-ABL	20
Total	**193**

OOB error was selected for each subtype. The number of discriminating SNPs per leukemia subtype were T-ALL, 47; E2A-PBX1, 37; TEL-AML1, 28; BCR-ABL, 20; MLL, 33 and hyperdiploid with more than 50 chromosomes, 28. Table 4.8 shows a summary of selected SNPs and tables 4.12 to 4.16 show the selected SNPs for each of these subtypes and their genetic information.

The marked difference in the number of discriminating SNPs for various leukemia subtypes suggest that significant differences exist in genetic variation of each subtype.

Classification results

Since the goal of this study is to determine the performance of identifying the known prognostic ALL genetic subtypes using genetic variation data, the selected class-discriminating SNPs identified using the RF-RFE approach were used in a SVM classifiers. A parallel format was used for class assignment, as described in section 4.3.4. To control the over-fitting problem of the given dataset, the analyses were performed 10 times (10-fold) where each time new training and test sets were generated

based on a randomly selected stratified training and test sets. SNPs were reselected using the new training set, and then the performance was assessed on the new test set. The results were finally averaged.

For each fold, parameters for the SVM classifiers were chosen by nested cross-validation procedures and the ones with the best cross-validation performance were selected, as described in the section 4.2. The SVM models were built using a polynomial kernel. The optimization of classifiers were performed using values of cost C, the penalty parameter of SVMs, set to 0.001, 0.01, 1 and 100 and $p = \{1, 2, 3\}$. The cost function $C = 0.001$ and $p = 2$ performed the best. The kernel parameters γ and r were set to the default values $\gamma = 1/\text{number of variables}$ and $r = 0$.

Using 10-fold cross validation the accuracy performances of SVM classifiers for each subtype were reported. Table 4.9 summarizes the results of these models. The results report the mean (Standard Deviation) of accuracy for RF-RFE, SVM-RFE and VarSelRF models on each of these subtypes based on training and test datasets. For each ALL subtype, the parameter settings of the VarSelRF approach were set similarly to the RF-RFE approach except for one parameter, *vars.drop.num*, the number of variables to exclude at each iteration, which was set to 0.1 as this was reported to produce the best accuracy (Díaz-Uriarte & de Andrés 2006). For the SVM-RFE approach, the C parameter was set to 0.001 and the number of features eliminated at each iteration was set to 10.

The proposed RF-RFE approach resulted in a good accurate subtype prediction in randomly selected test sets. An overall accuracy of 84% was achieved for accurate subtype assignments, with a range from 75% to 95%. Although the number of SNPs required for optimal class assignment varied among classes, the highest classification accuracies were obtained using a range of 20 to 47 SNPs for each class. A similar level of accuracy was achieved using other supervised learning algorithms with the selected feature subsets. The combined selected SNP set of 193 SNPs had a good sensitivity to correctly classify patients into one of six known subtypes of ALL (i.e. T-ALL, TEL-AML1, hyperdiploid, E2A-PBX1, MLL, and BCR-ABL1). The 193 SNP set had median sensitivity of 95% and median classification accuracy of 98%. When applied to the independent test sets, the SVM classifier had an average sensitivity of 88% and an average accuracy of 85% to correctly classify all classes (see table 4.9).

Comparing the classification performance

The classification performance of SVM-RFE and VarSelRF methods was compared with the proposed approach. The SVM-RFE approach represents one of the state-of-art feature selection approaches (Guyon et al. 2002). Furthermore, VarSelRF method is a RF-based feature selection approach that has been recently applied to genetic based studies (Liu et al. 2011, Molinaro et al. 2010). The classification performance results obtained for each ALL subtype are shown in Table 4.9. The RF-RFE did not perform as well as SVM-RFE. Nevertheless, the majority of its results are very similar

to one other. Using 10-fold cross validation and a corrected re-sampled two-sided t-test, the accuracy of the three methods (i.e. the RF-RFE, SVM-RFE and VarSelRF) were compared. The t-test result suggests that the differences in the accuracy of RF-RFE and SVM-RFE approaches, based on test data, are not statistically significantly different (with p-value = 0.14). The classification of different ALL subtypes shows that the performance of the SVM-RFE method has an average of 85% compared to 84% using the proposed framework. On the other hand, the t-test result suggests that the differences in the accuracy of RF-RFE and VarSelRF approaches are not statistically significantly different, with p-value of 0.0631.

Figures 4.5 and 4.6 show the accuracy performances of the SVM classifiers based on the 10 runs. These figures show the results obtained based on SVM-RFE and RF-RFE feature selection methods for the six ALL subtypes. For the classification of subtypes E2A-PBX1, BCR-ABL and MLL, the RF-RFE method has good averages in all measures. For the subtypes T-ALL, TEL-AML1 and Hyperdiploid, although both RF-RFE and SVM-RFE have high precision, the accuracies obtained for these subtypes are lower than the other subtypes. As can be seen in table 4.9, the selected features using SVM-RFE approach performed very well in training datasets but the performance was much lower when applied to test datasets. This suggests that the models generated based on selected SNPs using SVM-RFE do not generalize well. However, RF-RFE methods obtained reasonably consistent accuracy performance results on both training and test datasets.

Although there was no significant improvement on the classification performance

using the RF-RFE feature selection approach compared to SVM-RFE, the novelty of the proposed RF-RFE approach is the use of non-parametric feature selection procedures implemented in the RF method. Generally, the variable importance measure used in RF method selects features that may take into account gene-gene interaction without demanding a pre-specified model. In selecting interesting SNPs, the RF-RFE performs as well as other feature selection methods such as SVM-RFE in detecting SNPs that marginally contribute to complex diseases. Furthermore, RF-RFE seems to be superior to SVM-RFE in detecting interaction SNPs.

Building an ensemble of trees in RF increases the probability that some trees will capture interactions among variables with no strong main effect. Therefore, interactions could be taken into account when estimating variable importance. The recursive partitioning of tree building in RF illustrates an explicit representation of variable interactions. Comparing the RF-based feature selection to other variable selection methods, interactions among variables do not demand a pre-specified model to explicitly test for feature interaction. The feature selection process used in RF can be considered as a natural approach for large scale datasets such as the St Jude dataset. Therefore, the comparative performance results that have been reported by RF-RFE can be attributed to the potential of the RF-based feature selection procedure to capture and identify significant features including possible interaction markers.

Table 4.9: Prediction performance of SVM models for ALL subgroups based on three feature selection methods, namely, the proposed RF-RFE, SVM-RFE and VarSelRF methods. Using 10-fold cross validation, the average accuracy based on training and test data is reported. Values are mean (standard deviation) of the reported 10-fold accuracies

	RF-RFE		*SVM-RFE*		*VarSelRF*	
Subgroup	Training	Test	Training	Test	Training	Test
T-ALL	0.87 (0.04)	0.78 (0.07)	1 (0.0)	0.81 (0.04)	0.99 (0.01)	0.77 (0.05)
TEL-AML1	0.84 (0.03)	0.75 (0.06)	1 (0.0)	0.77 (0.04)	0.96 (0.02)	0.74 (0.04)
Hyper>50	0.87 (0.04)	0.78 (0.03)	1 (0.0)	0.81 (0.02)	0.94 (0.03)	0.75 (0.043)
E2A-PBX1	0.99 (0.00)	0.90 (0.05)	1 (0.0)	0.91 (0.04)	0.98 (0.01)	0.88 (0.02)
MLL	0.97 (0.01)	0.95 (0.01)	1 (0.0)	0.93 (0.03)	0.99 (0.01)	0.91 (0.03)
BCR-ABL	0.97 (0.01)	0.95 (0.02)	1 (0.0)	0.96 (0.01)	0.99 (0.01)	0.95 (0.03)
Average	**0.91**	**0.84**	**1**	**0.85**	**0.98**	**0.83**

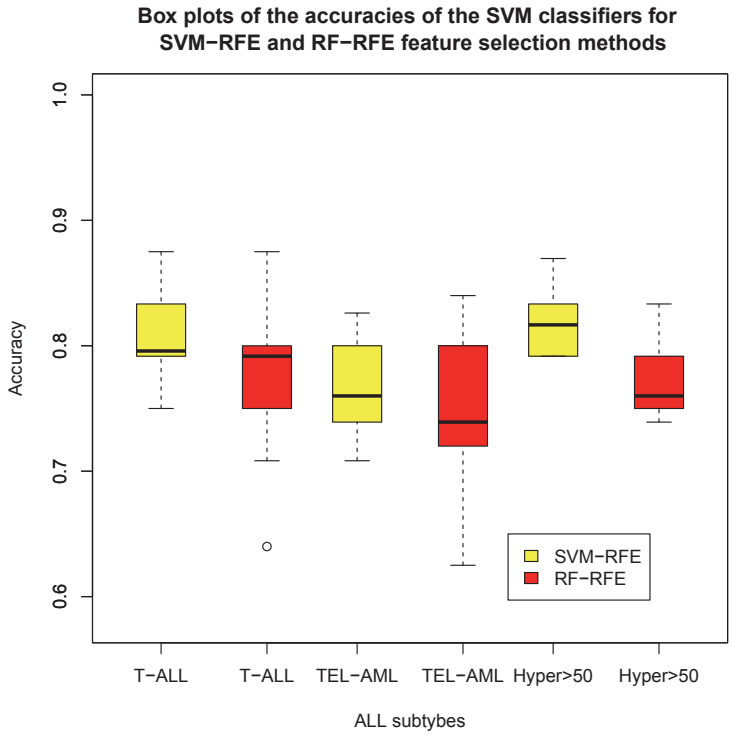

Figure 4.5: Box plots of the accuracies of the SVM classifiers based on 10 runs. The results are based on SVM-RFE and RF-RFE feature selection methods for T-ALL, TEL-AML and Hyper>50 subtypes

To conclude, even only using marginal-based selected markers, genetic variation data is a promising approach towards correctly classifying prognostically important subtypes of pediatric ALL. Compared to the reported disease diagnosis results on the Westmead dataset, the performance of disease prognosis using the St Jude data has good overall performance. However, results are less than the ones achieved on the Westmead dataset. The feature selection procedures applied to the Westmead dataset

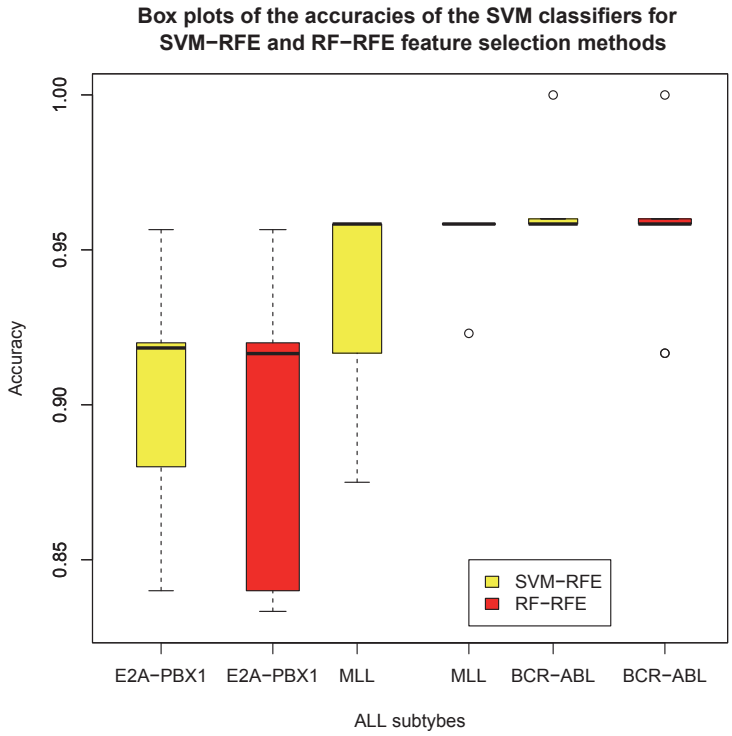

Figure 4.6: Box plots of the accuracies of the SVM classifiers based on 10 runs. The results are based on SVM-RFE and RF-RFE feature selection methods for E2A-PBX1, MLL and BCR-ABL subtypes

was the main reason for achieving such high performance, where both marginal and interaction markers were considered in model building.

4.5 Biological Insights

Significant data mining results are not complete without addressing its relationships with biological insights. Detailed studies of genes (markers) that are important in differentiating leukaemia from non-leukaemia samples as well as the genetic subtypes

of ALL might reveal more insight into the biology of each class. However, the main task of this thesis is to evaluate the feasibility of using genetic variation data to accurately classify ALL cases and subtypes. Here, initial insights for some of selected SNPs and their corresponding genes are described. The analyses have been done using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Da Wei Huang et al. 2008, Sherman et al. 2009); Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al. 1999) and Gene Ontology Biological Processes (GO-BP) (Ashburner et al. 2000) databases.

The DAVID database currently provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large lists of genes. In addition, the KEGG is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The KEGG databases are daily updated and made freely available (`http://www.genome.ad.jp/kegg/`). On the other hand, the GO-BP database, as controlled by Gene Ontology Consortium, produces a dynamic, controlled vocabulary of gene and protein roles in cells. There are three independent ontologies constructed and accessible on the Web: biological process, molecular function and cellular component (`http://www.geneontology.org`).

The most significant SNPs reported for marginal, two-way, three-way and four way interactions as well as their corresponding genes are listed in Appendix A-D.

Although the reported marginal-based SNPs did not show significant results, the initial results of the reported interaction SNPs based on GO-BP pathway analysis showed that most SNPs (i.e. genes) are related to cancer mechanisms such as cell apoptosis (programmed cell death), fertilization, response to vitamin, antigen, processing and presentation, innate immune response and others. Table 4.10 shows the summary of these results. Most of the reported SNPs (genes) did not highlight genes that are unrelated to the ALL disease (i.e. random noise). Similar analyses were also done based on KEGG database. The results are shown in table 4.11. The results verify results found based on GO-BP analysis. The list of important genes reported here warrants further investigation and clinical tests.

Table 4.10: The GO-BP analysis of genes associated with reported SNPs. Biological functions were based on edited terms from the GO-BP database

Term	**Count**	p**-value**	**Genes**
GO:0007606 sensory perception of chemical stimulus	24	2.50e-07	OR52H1, TAS2R4, OR51B2, OR52B6, TAS2R5, OR1L8, OR1N1, OR51Q1, OR6S1, ITPR3, OR13C5, OR2G2, OR4C13, TAS2R13, OR56B4, OR12D3, OR4C12, P2RX3, OR13D1, OR6K3, OR2D3, OR51I1, OR13C8, OR10P1
GO:0007600 sensory perception	32	3.06e-07	OR52H1, TAS2R4, OR51B2, TAS2R5, OR1L8, OR51Q1, OR1N1, OR6S1, OR13C5, OR4C13, OR2G2, OR56B4, TYR, PRR4, OR4C12, COL11A2, OR13C8, OR10P1, SCN10A, OPA1, MYO3A, OR52B6, SPTBN4, ITPR3, TAS2R13, OR12D3, P2RX3, OR13D1, OR6K3, USH1C, OR2D3, OR51I1

GO:0050890 cognition	33	1.22e-06	OR52H1, TAS2R4, OR51B2, TAS2R5, OR1L8, OR51Q1, OR1N1, COMT, OR6S1, OR13C5, OR4C13, OR2G2, OR56B4, TYR, PRR4, OR4C12, COL11A2, OR13C8, OR10P1, SCN10A, OPA1, MYO3A, OR52B6, SPTBN4, ITPR3, TAS2R13, OR12D3, P2RX3, OR13D1, OR6K3, USH1C, OR2D3, OR51I1
GO:0007608 sensory perception of smell	19	3.69e-05	OR52H1, OR51B2, OR52B6, OR1L8, OR1N1, OR51Q1, OR6S1, OR13C5, OR2G2, OR4C13, OR56B4, OR12D3, OR4C12, OR13D1, OR6K3, OR2D3, OR51I1, OR13C8, OR10P1
GO:0050877 neurological system process	35	6.79e-05	OR52H1, OR51B2, TAS2R4, TAS2R5, OR1L8, OR51Q1, OR1N1, COMT, OR6S1, OR13C5, OR4C13, OR2G2, OR56B4, TYR, PRR4, OR4C12, COL11A2, OR13C8, OR10P1, SCN10A, OPA1, MYO3A, OR52B6, SPTBN4, ITPR3, GRM1, TAS2R13, OR12D3, P2RX3, OR13D1, OR6K3, USH1C, OR2D3, CACNA1E, OR51I1
GO:0019882 antigen processing and presentation	8	1.63e-04	MICA, HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, CD1E, HLA-DQA1
GO:0007186 G-protein coupled receptor protein signaling pathway	31	4.58e-04	OR52H1, TAS2R4, OR51B2, TAS2R5, OR1L8, GPR109B, OR1N1, OR51Q1, LGR6, OR6S1, PKD1L2, PKD1L3, OR13C5, OR4C13, OR2G2, OR56B4, OR4C12, OR13C8, OR10P1, OR52B6, ITPR3, GRM1, FSHR, TAS2R13, OR12D3, CXCL16, OR13D1, GPR56, OR6K3, OR2D3, OR51I1
GO:0002504 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	5	0.001122	HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DQA2, HLA-DQA1
GO:0050909 sensory perception of taste	5	0.003301	TAS2R13, TAS2R4, TAS2R5, P2RX3, ITPR3
GO:0048066 pigmentation during development	4	0.0055	DCT, TYR, MITF, RAB27A

GO:0007166 cell surface receptor linked signal transduction	40	0.005896	MICA, OR52H1, OR51B2, TAS2R4, TAS2R5, OR1L8, GPR109B, MITF, EPHA10, OR51Q1, OR1N1, OR6S1, LGR6, PKD1L2, PKD1L3, OR13C5, OR2G2, OR4C13, OR56B4, PTK2B, OR4C12, NRG1, OR13C8, OR10P1, BMP4, OR52B6, SPEN, ITPR3, GRM1, FSHR, DDR1, TAS2R13, OR12D3, CXCL16, OR13D1, GPR56, OR6K3, OR2D3, OR51I1, LCP2
GO:0001539 ciliary or flagellar motility	3	0.015981	DNAH11, C14ORF104, DNAH5
GO:0010942 positive regulation of cell death	13	0.0199	BMP4, LALBA, DLC1, GPR109B, JMY, TNFRSF10A, CASP10, PLEKHG2, CASP9, RPL11, WWOX, RUNX3, RAB27A
GO:0033189 response to vitamin A	4	0.02096	BMP4, ALDH1A2, C14ORF104, TFRC
GO:0007018 microtubule based movement	6	0.0215	DNAH11, KIF14, KIFC1, OPA1, KIF7, DNAH5
GO:0006583 melanin biosynthetic process from tyrosine	2	0.027894	DCT, TYR
GO:0006917 induction of apoptosis	10	0.0367	LALBA, TNFRSF10A, DLC1, CASP10, PLEKHG2, RPL11, WWOX, RUNX3, RAB27A, JMY
GO:0012502 induction of programmed cell death	10	0.0374	LALBA, TNFRSF10A, DLC1, CASP10, PLEKHG2, RPL11, WWOX, RUNX3, RAB27A, JMY
GO:0010941 regulation of cell death	19	0.0379	DLC1, BMP4, LALBA, IFIH1, ZC3HC1, GPR109B, MITF, BIRC5, JMY, TNFRSF10A, CASP5, CASP10, PLEKHG2, CASP9, RPL11, NRG1, WWOX, RUNX3, RAB27A
GO:0043065 positive regulation of apoptosis	12	0.0404	LALBA, TNFRSF10A, DLC1, CASP10, PLEKHG2, CASP9, GPR109B, RPL11, WWOX, RUNX3, RAB27A, JMY
GO:0043068 positive regulation of programmed cell death	12	0.042156	LALBA, TNFRSF10A, DLC1, CASP10, PLEKHG2, CASP9, GPR109B, RPL11, WWOX, RUNX3, RAB27A, JMY
GO:0045087 innate immune response	6	0.04513	CRISP3, IFIH1, NCF2, CXCL16, C2, RAB27A
GO:0043473 pigmentation	4	0.05	DCT, TYR, MITF, RAB27A
GO:0048002 antigen processing and presentation of peptide antigen	3	0.058432	MICA, HLA-C, HLA-B, HLA-DOA

GO:0042981 regulation of apoptosis	18	0.0600	DLC1, LALBA, IFIH1, ZC3HC1, GPR109B, MITF, BIRC5, JMY, TNFRSF10A, CASP5, CASP10, PLEKHG2, CASP9, RPL11, NRG1, WWOX, RUNX3, RAB27A
GO:0006955 immune response	16	0.0618	CRISP3, IFIH1, MICA, NCF2, IL1F10, HLA-C, HLA-B, HLA-DQA2, HLA-DQA1, CD1E, CXCL16, HLA-DPA1, C2, HLA-DPB1, HLA-DOA, RAB27A, LCP2
GO:0043067 regulation of programmed cell death	18	0.064	DLC1, LALBA, IFIH1, ZC3HC1, GPR109B, MITF, BIRC5, JMY, TNFRSF10A, CASP5, CASP10, PLEKHG2, CASP9, RPL11, NRG1, WWOX, RUNX3, RAB27A
GO:0033273 response to vitamin	4	0.065625	BMP4, ALDH1A2, C14ORF104, TFRC
GO:0032526 response to retinoic acid	3	0.078024	BMP4, C14ORF104, TFRC
GO:0045767 regulation of anti-apoptosis	3	0.0993	OPA1, PTK2B, PIK3R2
GO:0009566 fertilization	4	0.099691	ELSPBP1, ZAN, SPTBN4, TNP2
GO:0043281 regulation of caspase activity	4	0.0996	TNFRSF10A, DLC1, CASP9, BIRC5

On the other hand, the number of selected markers that able to discriminate between the major leukaemia subtypes varied between subtypes. The reported SNPs for subtype analyses are only the marginal-based SNPs. The functional annotation analyses based on the DAVID database, of the reported SNPs for each ALL subtype are listed in tables 4.12 to 4.16. It appears that a large majority of the SNPs that have been identified do not have genes attached to them (identified as "-" in tables 4.12-4.16). This means that these SNPs are not on genes, but are actually on a non-coding part of the DNA.

Table 4.11: The KEGG analysis of genes associated with reported SNPs. Biological functions were based on edited terms from the KEGG database

Term	**Count**	***p*-value**	**Genes**
hsa04740:Olfactory transduction	19	1.94e-05	OR52H1, OR51B2, OR52B6, OR1L8, OR1N1, OR51Q1, OR6S1, OR13C5, OR2G2, OR4C13, OR56B4, OR12D3, OR4C12, OR13D1, OR6K3, OR2D3, OR51I1, OR13C8, OR10P1
hsa05416:Viral myocarditis	8	1.17e-04	CASP9, MYH4, HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa05330:Allograft rejection	6	2.32e-04	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa05332:Graft-versus-host disease	6	3.41e-04	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa04940:Type I diabetes mellitus	6	4.85e-04	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa04672:Intestinal immune network for IgA production	6	9.97e-04	IL15RA, HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa05310:Asthma	5	0.00104	HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa05320:Autoimmune thyroid disease	6	0.00119	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa05322:Systemic lupus erythematosus	7	0.0046	HIST1H2BI, HLA-DPA1, HLA-DPB1, C2, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa04612:Antigen processing and presentation	6	0.00992	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa04742:Taste transduction	4	0.0492	TAS2R13, TAS2R4, TAS2R5, ITPR3
hsa04514:Cell adhesion molecules (CAMs)	6	0.05840	HLA-C, HLA-DPA1, HLA-DPB1, HLA-B, HLA-DOA, HLA-DQA2, HLA-DQA1
hsa04650:Natural killer cell mediated cytotoxicity	6	0.059	TNFRSF10A, MICA, PTK2B, HLA-C, HLA-B, PIK3R2, LCP2

For other SNPs that do have genes attached to them, there are no pathways ontologies or biological processes which are statistically supported. That is, the pathways attached to the genes are different from each other. The one exception is rs9967367 which is found on the SETBP1 gene in the E2A-PBX1 SNP list. This

SNP is not located near where chromosomal translocations occur. There are two publications regarding leukaemia associated with SETBP1, including (i) over expression of SETBP1 being associated with poor outcome in elderly AML and (ii) NUP88 fuses with SETBP1 in paediatric T-ALL (Panagopoulos et al. 2007, Cristóbal et al. 2010). However, these publications do not concern the SNP, but rather the expression of the gene. Furthermore, there are over 4400 SNPs in SETBP1.

Consequently, it is difficult to interpret the SNPs that have been identified into something reflective of leukaemia. It may be that the identified SNPs are novel and naturally there would be no literature to support them.

In summary, the identified class and subtype-specific markers should provide a further insight into the biological functions that underlie various genetic subtypes of acute leukaemia and weather the selected markers (SNPs) cluster in the genome or not. Although the important markers have been highlighted, detailed analyses are required to explore the biological processes of the cellular pathways of these discovered markers, but this is beyond the scope of the thesis.

Table 4.12: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, T-ALL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	SNP name	Chromosome	Gene Symbol	Gene Description
1	rs10991049	9	-	-
2	rs4910689	11	-	-

3	rs9914296	17	FALZ	fetal Alzheimer antigen
4	rs10493720	1	-	-
5	rs10521125	17	-	-
6	rs10124033	9	FLJ16636	FLJ16636 protein
7	rs10454897	5	ARSB	arylsulfatase B
8	rs4469657	X	-	-
9	rs2331796	17	-	-
10	rs2862610	10	CNNM1	cyclin M1
11	rs7054479	X	-	-
12	rs7883354	X	DMD	dystrophin (muscular dystrophy, Duchenne and Becker types)
13	rs2542660	17	ACACA	acetyl-Coenzyme A carboxylase alpha
14	rs9927763	16	ZNF200	zinc finger protein 200
15	rs12324998	16	-	-
16	rs2173946	5	-	-
17	rs10959536	9	-	-
18	rs1540379	2	MYT1L	myelin transcription factor 1-like
19	rs16924153	8	TOX	thymus high mobility group box protein TOX
20	rs9572903	13	-	-
21	rs3755226	2	ADAM23	a disintegrin and metalloproteinase domain 23
22	rs683357	X	-	-
23	rs12944809	17	-	-
24	rs2646768	15	-	-
25	rs4332939	2	LOC643405	-
26	rs10811179	9	ASAH3L	N-acylsphingosine amidohydrolase 3-like
27	rs10277539	7	LMBR1	chromosome 7 open reading frame 2
28	rs11683409	2	MERTK	c-mer proto-oncogene tyrosine kinase
29	rs6702469	1	NFASC	neurofascin
30	rs10867736	9	-	-
31	rs10498205	2	-	-
32	rs4551866	13	ATP8A2	ATPase, aminophospholipid transporter-like, Class I, type 8A, member 2
33	rs7730157	5	DHX29	DEAH (Asp-Glu-Ala-His) box polypeptide 29
34	rs2806023	X	-	-
35	rs2125644	12	SLC38A1	solute carrier family 38, member 1
36	rs10419661	19	-	-
37	rs11650268	17	-	-
38	rs6704693	2	-	-
39	rs4747847	10	-	-

40	rs10205543	2	TRAF3IP1	TNF receptor-asso. factor 3 interacting protein 1
41	rs307646	9	UBAP2	ubiquitin associated protein 2
42	rs4707940	6	-	-
43	rs10483534	14	-	-
44	rs4436243	9	KIAA1529	KIAA1529
45	rs7984685	13	USP12	ubiquitin specific protease 12
46	rs4373074	3	LOC440970	similar to MUF1 protein
47	rs10794809	10	-	-

Table 4.13: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, TEL-AML, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	**SNP name**	**Chromosome**	**Gene Symbol**	**Gene Description**
1	rs1816633	18	DSC1	desmocollin 1
2	rs4577299	2	-	-
3	rs4904408	14	GALC	galactosylceramidase (Krabbe disease)
4	rs9462170	6	PNPLA1	patatin-like phospholipase domain containing 1
5	rs6995177	8	-	-
6	rs1438989	12	-	-
7	rs12649108	4	YT521	splicing factor YT521-B
8	rs10133111	14	-	-
9	rs1179697	2	-	-
10	rs561379	18	-	-
11	rs12532479	7	C7orf10	chromosome 7 open reading frame 10
12	rs1872350	X	-	-
13	rs6718607	2	STK39	serine threonine kinase 39
14	rs12035887	1	DAB1	disabled homolog 1 (Drosophila)
15	rs3808244	7	PFTK1	PFTAIRE protein kinase 1
16	rs1356755	8	-	-
17	rs6555529	5	-	-
18	rs11021127	11	-	-
19	rs1335015	10	-	-
20	rs6801238	3	-	-
21	rs12601236	17	DNAH9	dynein, axonemal, heavy polypeptide 9
22	rs6973324	7	FLJ31818	hypothetical protein FLJ31818

23	rs17782898	6	-	-
24	rs12054820	5	CNOT6	CCR4-NOT transcription complex, subunit 6
25	rs288513	3	-	-
26	rs34853285	6	-	-
27	rs10905305	10	-	-
28	rs7103334	11	-	-
29	rs12544516	8	LOC646479	-
30	rs9967367	18	SETBP1	SET binding protein 1
31	rs10404868	19	AP2A1	adaptor-related protein complex 2, alpha 1 subunit
32	rs3754658	2	CENTG2	centaurin, gamma 2
33	rs4702598	5	-	-
34	rs41405548	-	-	-
35	rs2417876	12	-	-
36	rs12540446	7	LOC221981	hypothetical protein LOC221981
37	rs16856759	2	LRP2	low density lipoprotein-related protein 2

Table 4.14: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, Hyperdiploid >50C, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	**SNP name**	**Chromosome**	**Gene Symbol**	**Gene Description**
1	rs8138460	22	-	-
2	rs7664572	4	-	-
3	rs6945438	7	LUZP5	leucine zipper protein 5
4	rs4814902	20	-	-
5	rs7671914	4	-	-
6	rs389675	21	GPXP2	glutathione peroxidase pseudogene 2
7	rs6845321	4	-	-
8	rs338228	1	-	-
9	rs158916	5	mimitin	Myc-induced mitochondria protein
10	rs4672329	2	-	-
11	rs12477902	2	LOC646832	-
12	rs2905587	5	KLHL3	kelch-like 3 (Drosophila)
13	rs430102	12	-	-
14	rs4974710	4	ZFYVE28	zinc finger, FYVE domain containing 28
15	rs6728457	2	-	-

16	rs239794	6	C6orf143	chromosome 6 open reading frame 143
17	rs11706332	3	MGC61571	hypothetical protein MGC61571
18	rs1449506	13	DGKH	diacylglycerol kinase, eta
19	rs2972419	5	GHR	growth hormone receptor
20	rs1957404	14	-	-
21	rs2618768	4	-	-
22	rs7048937	9	-	-
23	rs305567	1	CACHD1	KIAA1573 protein
24	rs2044809	4	-	-
25	rs10510947	3	MAGI1	membrane associated guanylate kinase, WW and PDZ domain containing 1
26	rs11815693	10	PHYHIPL	hydroxylase interacting protein-like
27	rs13226538	7	AUTS2	autism susceptibility candidate 2
28	rs17751326	2	-	-

Table 4.15: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, E2A-PBX1, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	**SNP name**	**Chromosome**	**Gene Symbol**	**Gene Description**
1	rs4577579	4	-	-
2	rs2344768	10	-	-
3	rs4261097	1	-	-
4	rs611829	1	-	-
5	rs16842257	1	TP53BP2	tumor protein p53 binding protein, 2
6	rs2633442	3	MKRN2	makorin, ring finger protein, 2
7	rs6597757	10	ADAM12	a disintegrin and metalloproteinase domain 12 (meltrin alpha)
8	rs6966065	7	-	-
9	rs2081593	12	-	-
10	rs10896438	11	-	-
11	rs7830390	8	LRRCC1	KIAA1764 protein
12	rs339571	1	-	-
13	rs17166514	7	-	-
14	rs4346638	4	-	-
15	rs850221	2	LOC130576	hypothetical protein LOC130576

16	rs6989950	8	DLC1	deleted in liver cancer 1
17	rs11207714	1	NFIA	nuclear factor I/A
18	rs1112963	18	-	-
19	rs2217357	1	-	-
20	rs9300745	13	-	-
21	rs17794133	9	-	-
22	rs2579750	10	-	-
23	rs11044446	12	PLEKHA5	pleckstrin homology domain containing, family A.
24	rs17718247	19	-	-
25	rs16906661	8	-	-
26	rs4521419	4	HADHSC	L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain
27	rs788985	10	NEBL	nebulette
28	rs2651515	15	-	-
29	rs2776855	1	OBSCN	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF
30	rs4706115	6	BCKDHB	branched chain keto acid dehydrogenase E1, beta polypeptide (maple syrup urine disease)
31	rs10492908	16	WWOX	WW domain containing oxidoreductase
32	rs10856984	4	HADHSC	L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain
33	rs11057310	12	EIF2B1	eukaryotic translation initiation factor 2B, subunit 1 alpha, 26kDa

Table 4.16: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, MLL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	**SNP name**	**Chromosome**	**Gene Symbol**	**Gene Description**
1	rs4883911	13	-	-
2	rs703018	10	-	-
3	rs16833719	2	-	-
4	rs2765841	10	-	-
5	rs2398652	5	-	-
6	rs771818	3	-	-
7	rs169455	18	LOC388458	hypothetical gene supported by BC040718

8	rs1375363	2	KIAA1679	KIAA1679 protein
9	rs2231963	11	TRIM68	tripartite motif-containing 68
10	rs6723261	2	NMS	neuromedin S
11	rs9942407	5	-	-
12	rs13405065	2	SLC8A1	solute carrier family 8 (sodium/calcium exchanger), member 1
13	rs1519661	2	NMS	neuromedin S
14	rs6436754	2	SKIP	SPHK1 (sphingosine kinase type 1) interacting protein
15	rs1020206	10	PCDH15	protocadherin 15
16	rs16883860	6	MAPK14	mitogen-activated protein kinase 14
17	rs3772341	3	CNTN6	contactin 6
18	rs2140239	2	-	-
19	rs1175384	13	NUFIP1	nuclear fragile X mental retardation protein interacting protein 1
20	rs154599	5	SEMA6A	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A

Table 4.17: The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, BCR-ABL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them.

Index	**SNP name**	**Chromosome**	**Gene Symbol**	**Gene Description**
1	rs1196686	2	MGC52057	hypothetical protein MGC52057
2	rs1791858	11	NELL1	NELlike 1 (chicken)
3	rs12216142	6	-	-
4	rs6981388	8	-	-
5	rs974878	1	-	-
6	rs12987254	2	EHBP1	EH domain binding protein 1
7	rs10502193	11	-	-
8	rs1934395	1	AGBL4	hypothetical protein FLJ14442
9	rs2050090	20	PLCB1	phospholipase C, beta 1 (phosphoinositidespecific)
10	rs2606642	6	-	-
11	rs2819711	10	INPP5A	inositol polyphosphate5phosphatase, 40kDa
12	rs527628	11	LOC643381	-
13	rs2197815	4	-	-

14	rs285386	1	ROR1	receptor tyrosine kinaselike orphan receptor 1
15	rs13236770	7	-	-
16	rs4275061	6	HMGCLL1	3hydroxymethyl methylglutaryl Coenzyme A lyaselike 1
17	rs4945203	11	HBXAP	hepatitis B virus x associated protein
18	rs13036801	20	-	-
19	rs305332	8	-	-
20	rs1429772	3	-	-
21	rs11166603	8	KHDRBS3	KH domain containing, RNA binding, signal transduction associated 3
22	rs9865384	3	-	-
23	rs10495544	2	-	-
24	rs9382497	6	HMGCLL1	3hydroxymethyl3methylglutarylCoenzyme A lyaselike 1
25	rs7195492	16	-	-
26	rs2402374	14	C14orf152	chromosome 14 open reading frame 152
27	rs3816279	17	-	-
28	rs12420572	11	DDX10	DEAD (AspGluAlaAsp) box polypeptide 10

4.6 Discussion and Conclusion

With the availability of high-throughput genotyping technology genome-wide association studies have been used to mine the optimal set of SNPs with the highest prediction accuracy for complex diseases. This chapter addressed the feasibility of using genetic variation data for the task of disease classification (diagnosis) and prognosis. To achieve this, a new computational framework was proposed to address different issues of the given tasks. The proposed framework demonstrated the use of different phases of data processing and modelling to build reliable disease diagnostic and prognostic models using genetic variation data.

The heart of the proposed framework is the feature selection phase, which endeavours to use feature selection methods for selecting disease-susceptibility markers that account for marginal effects as well as gene-gene interactions. The MDR approach employed in the feature selection phase considers all two, three and four-way interaction combinations, with the most significant combinations to be included in the final model. In this phase, the IE measure was used to narrow the search for all possible ways of interactions, therefore, alleviating testing large numbers of gene-gene interactions. In the proposed framework, both marginal and interaction markers are included in the final model building. Performing feature selection based on traditional methods will result in choosing markers that marginally contribute to disease and ignore gene-gene interactions that may be useful for producing highly predictive models for the large scale of genome-wide studies, where the number of SNPs are multiples of hundreds of thousands.

The power of the proposed framework was empirically tested using genetic variation datasets from studies of acute lymphoblastic leukaemia. The given datasets were used to develop disease classification and prognosis models. The results suggest that the proposed framework can produce highly accurate classification models. Furthermore, the proposed framework has the flexibility to be extended to other complex diseases. Using the framework on the given studies, the results demonstrate that genetic variation profiling is not only a robust approach for the accurate identification

of known prognostically ALL genetic subtypes, but that it also provides a possibility for new insights into the underlying biology as preliminarily investigated in section 4.5. The resultant classification models had a predictive accuracy of around 92% in the diagnosis of ALL cohort dataset using 10-fold cross validation process. Indeed, a significant improvement in the model performance required including multi-locus interaction markers.

In terms of disease prognostic tasks, the genetic variation profiling demonstrated the possibility of accurately identifying the major subtypes of ALL patients with a high level of accuracy and sensitivity. Using the proposed RF-RFE approach to identify markers significantly associated with a particular class, discriminating markers were selected for each of the six major prognostic subgroups. An average predictive accuracy of 84% for different ALL subtypes was achieved using the proposed framework. Although a higher level of accuracy is desired, a small set of SNPs was enough to predict different ALL subtypes. Including interaction markers, as the result on the Westmead dataset showed, will also improve the accuracy.

The results showed that the RF-RFE approach has provided a level of accuracy that is comparable to, although not statistically significantly better than, that obtained using the state-of-art SVM-RFE method. However, the novelty of the proposed feature selection procedures employed using the RF method could outperform SVM-RFE in detecting SNPs that marginally contribute to complex diseases, and the

relative superiority in detecting interaction SNPs. In most cases, the followed procedures of the proposed framework were unlikely to suffer from over-fitting compared to results reported using SVM-RFE feature selection method.

The analytical evaluation of the proposed framework raised the possibility of achieving more accurate diagnosis and prognosis of biologically relevant disease status/subtypes. Therefore, a single standardised diagnostic and prognostic platform can be achievable. The proposed framework addressed the thesis research question **Q4** in section 1.3 and the thesis contribution 4 in section 1.4. Indeed, the proposed multi-phase framework is evaluated to build reliable disease diagnostic and prognostic classification models. Several genetic variation datasets were employed to show the feasibility of the proposed framework. The results suggest that the proposed prediction approaches can effectively define important markers that are mostly consistent with known biological findings while the accuracy of the produced models is also high.

The use of genome-wide genetic variation data, as applied here, will eventually have the potential to lead to both a greater understanding of the revealed biology underlying the formation of the leukaemia, as well as to enhance our ability to identify leukaemia subtype-specific genetic variation differences. The ultimate goal is then to turn the identified class-specific markers into a useful therapeutic targets. As the results of this chapter demonstrated, the next step would be to produce a single diagnostic and prognostic platform.

Chapter 5

Visualizing Genome-wide SNP Profiles

Information visualization is considered as a direct way to help browse a given dataset. It is possible to combine visual exploration with other data exploration tools such as clustering analysis and data comparisons. The result of data explorations can be confirmed through the visualization. The main challenge in visualizing genetic variation datasets stems from the high dimensionality of the data, which may include tens or hundreds of thousands of SNPs. In this chapter, several data reduction methods will be applied to visualize the given genetic variation datasets.

Traditional dimensionality reduction techniques are used to find a low space representation of the high dimensionality space which preserves the global structure of the data. These methods include Principal Components Analysis (PCA) (Hotelling 1933) which tries to preserve the variance in the data, and Multidimensional Scaling (MDS) which tries to preserve pairwise distances between data points. However,

these methods are not adequate to handle high dimensional data which could have complex non-linear relationships.

Therefore, in the last decade a large number of non-linear dimensionality reduction techniques have been proposed. Some of these methods are used to find a lower dimensional manifold of the data or a non-linear embedding manifold space in the higher-dimensional data space. The main advantage of these methods is that they are able to preserve the local relationships of the data, which can be advantageous for the task of information visualization. The main difference between manifold-based methods and visualization using conventional data reduction techniques is that data visualization is limited to two or three dimensions (Saul & Roweis 2003, Fu & Huang 2010). Thus, it is difficult for manifold-based methods to know the exact number of dimensions needed to uncover the underlying structure of a given data. Therefore, different dimensionality reduction methods are needed to be applied in order to choose the most appropriate method for visualization.

In this chapter, the results of applying different dimensionality reduction methods to genome-wide SNP profiles of leukaemia patients are examined to determine which is the best method for visualizing this type of data. The results will be compared based on measures such as trustworthiness and continuity of the visualizations. The visualization results will assist clinicians and biomedical researchers in understanding the different structure of patients and how to compare different groups of patients

clustering in the visualization.

Methods and approaches applied in this chapter rely on the information extracted from biomedical datasets, derived from cancer patients. This data includes genome–wide SNP genotyping data (genetic variations). The application domain of this study is acute lymphoblastic leukaemia. The approaches and methods applied to leukaemia datasets can also be extended to other complex diseases such as heart diseases, diabetes and inflammatory diseases.

The rest of this chapter is organized as follows. Section 5.1 describes the datasets used in this study and the preprocessing steps applied. Next, in section 5.2 methods and techniques used to visualize the ALL data are described. Section 5.3 describes in detail the experimental procedures and methods used to visualize the given genetic variation data. In section 5.4 the results are presented. In section 5.5 the results are discussed and the chapter is concluded.

5.1 Datasets and Data Preprocessing

There are two datasets used in this study: (1) a genome-wide SNP dataset, generated at the Oncology Research Unit at the Children's Hospital at Westmead, and (2) a dataset of 140 samples of individuals retrieved from the WTCCC. A detailed description of these datasets was given in section 4.3.2.

The given datasets contain information about 13,917 SNPs which are scattered

across the whole genome. These SNPs are classified as non-synonymous (functional) SNPs which affect the functionality of genes. Genotyping of 13,917 SNPs was attempted in each sample. SNPs with a call rate of less than 99% were excluded. After applying stringent quality control criteria, as described in section 4.3.3, high-quality genotypes for 10,750 SNPs was obtained.

Each individual's genome has two alleles of a given SNP. At a specific SNP, a person can have one of the several genotypes. When they are the same the SNP is called homozygous and when they are different the SNP is called heterozygous. For a single SNP, one is designated the major allele and the other the minor allele, based on their observed frequency in a general population. Each SNP can have four different values (nominal): two homozygous, one heterozygous or missing. That is, the four possibilities for alleles A and B of the ith SNP are two homozygous (AA or BB), one heterozygous (AB) or missing NA (not determined). All SNPs were transformed into numerical data based on minor allele frequency, as described in Price et al. (2006). Numerical representation of the given datasets facilitates the computational process of the used methods. Let $\mathbf{G}$ be a $m \times n$ genotype matrix

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,n} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m,1} & g_{m,2} & \cdots & g_{m,n} \end{bmatrix}$$

where g_{ij} is the genotype for SNP i and individual j. The matrix $\mathbf{G}$ is centred by

subtracting the row mean, $\mu_i = (\sum_j g_{ij})/n$, from each genotype, g_{ij}, in each row g_i of $\mathbf{G}$. Missing entries in $\mathbf{G}$ are set to 0 and do not contribute to the calculation of μ_i. Each row, i, of $\mathbf{G}$ is then normalized by dividing each genotype entry by $\sqrt{p_i(1-p_i)}$ where p_i is a posterior estimate of the unobserved underlying allele frequency of SNP i defined by

$$p_i = \left(1 + \sum_j g_{ij}\right) /(2 + 2n) \tag{5.1.1}$$

with missing entries excluded from the computation. The resulting matrix is denoted as $\mathbf{G}$-normalized. The new matrix is regarded as a normalized version of the genotype data matrix and the mean of each row i is equal to 0 and the similarity of each patient with itself is equal to one.

5.2 The SNP Visualization: Problem and Approaches

Several dimensionality reduction methods for visualizing the similarity relationship between patients were reviewed in chapter 2. Firstly, the main classical methods for dimensionality reduction were reviewed including Principal Component Analysis (PCA), and Multidimensional Scaling (MDS). Some other methods based on MDS were also reviewed. Then, other recently proposed methods that focus on finding the manifold or embedding of data were illustrated.

The problem of dimensionality reduction can be defined as follows. Consider a dataset matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ consisting of m data vectors $\mathbf{x}_i$ $(i \in 1, 2, ..., m)$ with

dimensionality d which can be considered as points in a high-dimensional data space. Dimensionality reduction methods transform the dataset $\mathbf{X}$ with dimensionality d into a new dataset $\mathbf{Y} \in \mathbb{R}^{m \times p}$ with dimensionality p ($p \ll d$), while preserving the geometry (relationships) of the data as much as possible. The low-dimensional representation of $\mathbf{x}_i$ is denoted by $\mathbf{y}_i$, where $\mathbf{y}_i$ is the ith row of the p-dimensional data matrix $\mathbf{Y}$. For visualization purposes the dimensionality representation of the *output space* needs to be two or three dimensions at most, whereas the original space or *input space* can be thousands of dimensions.

Generally, the task of visualization methods is to construct a low-dimensional representation (i.e. output space) $\mathbf{y}_i$ of the input space, in such a way that the original relationships (or similarities) of the data are preserved. However, lower-dimensional representation of the data in 2 or 3 dimensions might not be able to preserve all the information of the original higher-dimensional space of a dataset. A compromise must be made by applying different data reduction methods and then selecting the best method based on how well a given method preserves the information of the original data. In the next section, measures of the quality of visualization that have been used in the experiments are described and the results are given in section 5.4.

5.2.1 Comparing Visualisations

As discussed, the first step in exploring the structure of a given dataset is to visualize it. In many previous works, visualization methods are compared through examining the produced figures. However, some quantitative criteria should be designed to compare the visualization results without considering the human as a part of evaluating a given visualizations.

One of the crucial tasks in data visualization is how to assess the quality of produced visualizations or the tools that are used. The quality measure is used to assess how well the visualization can represent the underlying relationships between data points. The local structure of the data is the most important component of the visualization. The usability of the visualization can be measured by how accurately the data is represented and how readable it is.

The first question that needs to be answered is how trustworthy is the visualization. When looking at the visualization, the first insight is how points are similar and how points group together. Looking at a visualization a user can possibly get insight into some questions such as: are the unknown data points similar to the known ones? How is the data clustered? Are there denser areas and more sparse ones? Questions like these cannot be answered without having a visualization that is capable of answering these questions.

There are a number of methods that have been implemented to assign a quantity to a visualization. Some of these methods calculate the correlation coefficient between the distance vectors (i.e. the vectors that compare the distance between all pairs of points) of the original space with that of the lower dimensional space. It was shown that this measure can provide a good measurement of the quality of the visualization procedures (Tan et al. 2005).

Others methods measure the trustworthiness of the local structure of the visualizations is (Venna et al. 2010). Based on these methods the low-dimensional representation is trustworthy if the k nearest neighbours of a point in the reduced space (or in the visualization) are also neighbours of the point in the original space. The proportion of points that are in the neighbourhood in the visualization but not in the original space is quantified as the precision (or loss of precision, i.e. 1 - precision). This number is usually not informative. However, the magnitude of the error can be used to rank the data points based on their distance instead of just counting the number of errors.

Reducing the dimensionality of data can result in losing some of the similarity relationships between data points. Two general errors can be caused by applying a reduction method. First, data points that are not neighbours in the input space can be mapped close by in the reduced space, causing points to be incorrectly identified as neighbours. These kinds of error can reduce the *precision*. Secondly, data points

that are neighbours in the input space can be mapped far away in the reduced space, causing discontinuities in the mapping and can distort the neighbour relations. This kind of error effects *recall*. The two kinds of error (i.e. precision and recall) are used in information retrieval literature in which the error is quantified based on the proportion of the points that caused the errors.

The main limitation of using *precision* and *recall*, as it is used in information retrieval, is that each of the errors is equally bad. However, in the visualization context this kind of measurement is not intuitive, whereas the distance between data points is known. Intuitively, a data point that comes into the neighbourhood of another from far away causes a larger error than one that comes from closer. By ranking data points based on their similarity, two quality measures can be defined: *trustworthiness* and *continuity* (Venna & Kaski 2007, Venna et al. 2010), which quantify the errors of a visualization tool by the neighbourhood ranks of each data point.

According to Kaski et al. (2003), the *trustworthiness* of a visualization can be defined as follows: The rank of each data sample $\mathbf{x}_j$ is ordered according to the distance from the data sample $\mathbf{x}_i$ in the input space. Let $r(\mathbf{x}_i, \mathbf{x}_j)$ define such a rank. Let $U_k(\mathbf{x}_i)$ be a set of the data samples of size k. This set represents data samples that are the in neighbourhood of sample $\mathbf{x}_i$ in the visualization space but not in the original space. Then, the measure of trustworthiness is defined as

$$M_{Tru}(k) = 1 - A(k) \sum_{i=1}^{n} \sum_{j \in U_k(i)} (r(\mathbf{x}_i, \mathbf{x}_j) - k) \tag{5.2.1}$$

where n be the number of data samples and $A(k) = 2/(Nk(2N - 3k - 1))$ is used to scale the trustworthiness measure between zero and one. The errors reach the maximum value when the ranks in the input and output space are reversed. The trustworthiness measure is closely related to the precision (as in information retrieval). However, the trustworthiness measure is a special kind of precision measure for the case where the objects are ranked based on their relevance (Kaski et al. 2003, Venna & Kaski 2006).

On the other hand, discontinuities are used to quantify whether neighbours in the original space remain neighbours in the visualization. If neighbour's points are pushed out in the displayed visualization, discontinuities arise in the visualization. The errors caused by discontinuities may be quantified similarly to the errors caused by trustworthiness.

Let $V_k(\mathbf{x}_i)$ be the set of data samples that are neighbours of the data sample $\mathbf{x}_i$ in the original space but not in the output space and $\hat{r}(\mathbf{x}_i, \mathbf{x}_j)$ be the rank of the data sample $\mathbf{x}_j$ in the ordering according to the distance from $\mathbf{x}_i$ in the visualization. Kaski et al. (2003) defined the effects of discontinuities of the mapping

$$M_{disc}(k) = 1 - A(k) \sum_{i=1}^{n} \sum_{j \in V_k(i)} (\hat{r}(\mathbf{x}_i, \mathbf{x}_j) - k) \tag{5.2.2}$$

where n be the number of data samples and $A(k) = 2/(Nk(2N - 3k - 1))$ is used to scale the discontinuities measure between zero and one. Therefore, a data sample that is mapped far away from the neighbourhood in the reduced space will cause a larger error than a data sample that was mapped just out of the neighbourhood. In the recall measure both errors are considered equally severe.

In this study the trustworthiness and continuity measures are used to assess the performance of different data reduction methods to visualize the given datasets. Based on the quality of the results, the best run method and parameters are selected. Comparison of data reduction methods can be tested on a large range of neighbourhood sizes. Small k can be important for the quality of visualization but a range of neighbourhood sizes can give an overview of the overall performance of different methods.

The performance of different data reduction methods is analysed and compared by plotting the trustworthiness and continuity measures as a function of the neighbourhood size k. Using any data reduction method, a trade-off must be made between trustworthiness and continuities. Seeking a high trustworthiness will typically lead to a lower continuity and vice versa.

5.3 Experimental Procedures

The purpose of the experiments is to gain insight and understand the behaviour of different dimensionality reduction methods in biomedical data and, more specifically,

SNP data of children with acute lymphoblastic leukaemia.

The performance of dimensionality reduction methods will be compared by visualizing the given datasets. The following methods will be included in the experiments: Principal Component Analysis (PCA), Laplacian Eigenmap (LE), Locally Linear Embedding (LLE) and methods based on Multidimensional Scaling, which include an extended version of Curvilinear Component Analysis (CCA) called Local MDS (LocalMDS) and an extended version of Stochastic Neighbour Embedding (SNE) called Neighbour Retrieval Visualizer (NeRV).

The dimensionality reduction methods were used to visualize the Westmead dataset as an unsupervised case and the visualizations were compared based on trustworthiness and discontinuities of the resultant visualizations. Furthermore, the combined dataset of 279 samples in total, 139 ALL samples obtained at CHW and 140 non-patient samples (control) from WTCCC, were visualized as a supervised case. This combined data is termed the case-control dataset. Feature selection methods used in chapter 4 were applied to this dataset before being visualized. The resultant visualizations were compared based on how well a visualization discriminates the two classes of the combined dataset.

5.3.1 Experimental Settings

All methods except PCA have a parameter N for setting the number of nearest neighbours. This parameter was set with values of N ranging from 5 to 30. The best N was selected based on the best visualization performance results. However, small neighbourhood size can be related to the data points that are most likely to be relevant. The performance of the resulting visualizations was evaluated based on the trustworthiness and continuities of the reduced dimension, as described in section 5.2.1.

Some of the applied methods including LocalMDS and NeRV that may fall into local optima were run several times (in this case 10 times) with different random initialization and the best run was selected. Lastly, to test the feasibility of all compared methods, a random-based mapping is used to randomly map the *inputspace*. Random mapping was computed based on the average of 10 different random projections (choosing two or three attributes randomly).

For unsupervised-based visualization, two types of distance metric were used to calculate the distance between data points in the input space: MAF-based kernel (MAFK) function and Entropy-based kernel (EK), as described in section 3.5. Furthermore, for the case of supervised-based visualization, the RF-based Kernel (RFK) was employed to calculate distances. In this study, all of these distance measures

were employed with different parameter settings and the dimensions of the output space were set to be equal to two for visualization.

5.4 Results

5.4.1 Unsupervised Visualization of the Westmead dataset

The visualization of the Westmead dataset is illustrated, here, as an unsupervised-based case. Data reduction methods were compared using the trustworthiness and continuity measures of the resultant visualization. Figures 5.1 and 5.2 show the trustworthiness and continuity results of the applied methods. The following subsection will discuss different aspect of the results. For visualization purposes, trustworthiness is more important than continuity, as the trustworthiness measure focuses mostly on how well a data reduction method preserves the neighbourhood of data points in the reduced space (or in the visualization). In each case the result with the best trustworthiness result was reported.

Trustworthiness and Continuity

Trustworthiness and continuity are the first aspects to be examined. In terms of exploring the result of a visualization, the local neighbourhood of each data point is the first insight a human analyst looks at. Therefore, a visualization is trustworthy if the visualization preserves small neighbourhoods as much as possible. Thus, attention

should be paid to small sizes of k (e.g. k between 5 and 15). It is clear from figure 5.1 that, in terms of trustworthiness, the NeRV method is the best. Unexpectedly, PCA is also performed quite well on this dataset. The result suggests that the dataset may have linear structure and PCA performed well in preserving the global variance of the given data points. On the other hand, state-of-art data reduction methods such as LLE and LE were not able to produce reasonable results. In fact, LE was the worst method compared in the Westmead dataset experiments and is the one with the most similar results as random mapping. The results suggests that both LLE and LE do not recover a low-dimensional embedding effectively and thus influence the final result of the embedding. Therefore, the dataset requires further scrutiny to uncover its intrinsic dimensionality. Also, the choice of the neighbourhood size should be investigated as it may have a strong influence on the results.

In this initial analysis LocalMDS was expected to perform similarly to the NeRV method as the underlying procedure used by both of these methods is similar, but the result shows slightly different behaviour. The performance of PCA method is similar to LocalMDS, even though the original cost function of LocalMDS emphasizes trustworthiness and should perform better.

In terms of continuity, as can be seen in figure 5.2, the result of NeRV method is again the best. However, unlike with trustworthiness, LocalMDS is slightly better than PCA. An explanation for this is that the cost function of LocalMDS emphasizes

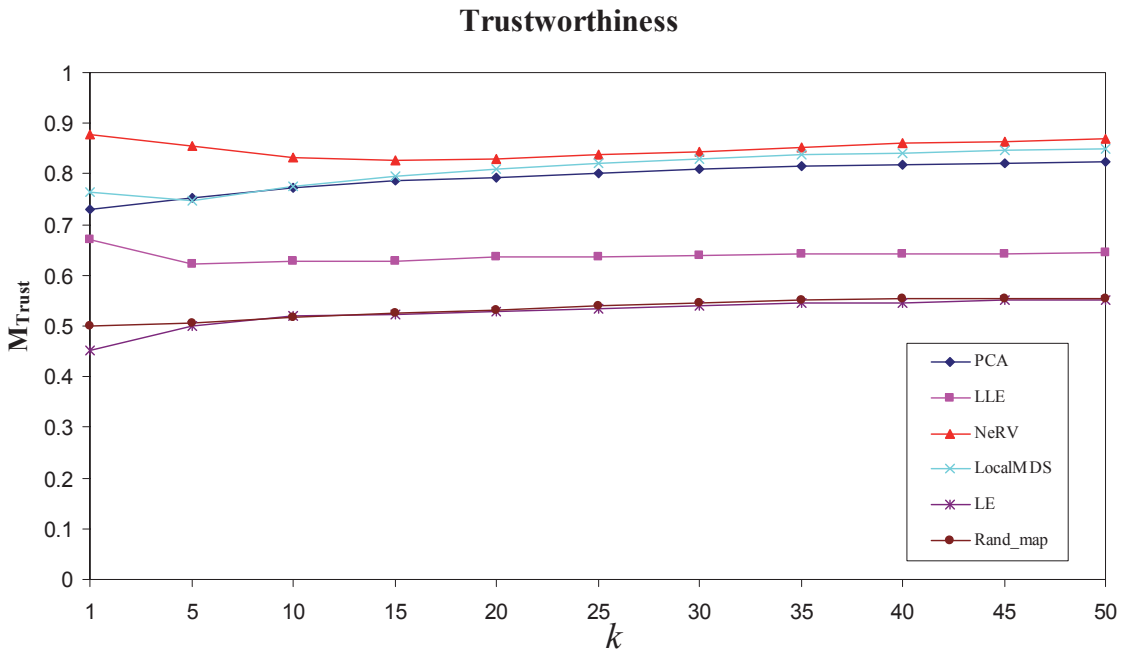

Figure 5.1: Trustworthiness of the mapping as a function of k that applied to the Westmead dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: Neighbour Retrieval Visualizer, LocalMD: Local Multidimensional scaling, LE: Laplacian Eigenmap and Rand_map: Random mapping.

the continuity measure of mapping data points. Once again, the manifold-based methods LLE and LE also perform poorly with continuity.

Sensitivity of NeRV and LocalMDS Methods

For the NeRV method, different parameters were set to explore the performance of this method on the Westmead dataset. Specifically, N the neighbourhood size and the relative cost λ between recall and precision were modified.

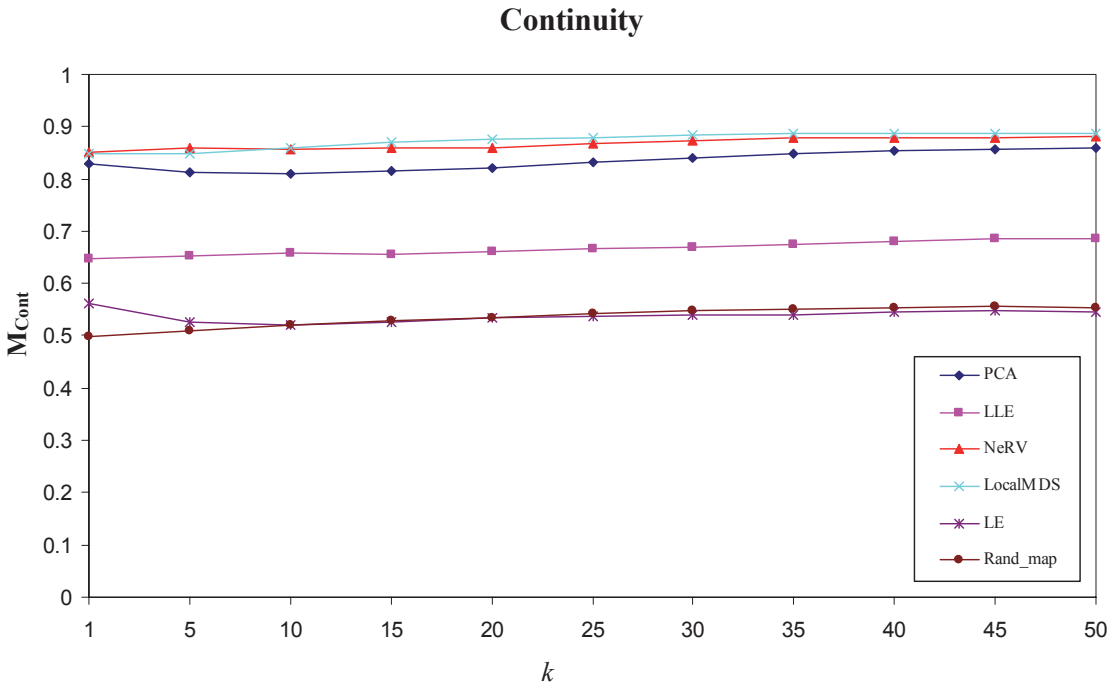

Figure 5.2: Continuity of the mapping as a function of k applied to the Westmead dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: neighbour Retrieval Visualizer, LocalMDS: Local Multidimensional scaling, LE: Laplacian Eigenmap and Rand_map: Random mapping.

As can be seen in figure 5.3, the NeRV method for different neighbourhood sizes, N, ranging from 5 to 30 were applied. If a small size of neighbourhood is considered around each point on the output space, a neighbourhood size of 30 produces the best results on the Westmead data. In contrast, the best performance of trustworthiness found by LocalMDS was a small neighbourhood size of 5, as can be seen on figure 5.4.

Next, using NeRV method the neighbourhood size k was set to be equal to 30 and was run on a range of $\lambda = 0$ to 1. The result can be seen in figure 5.5, which shows that the best trustworthiness occurs when λ equals 0.0 or 0.3. The line on figure 5.5 for λ equals 1 is the result that SNE method would give because NeRV with λ equals 1 is the case where NeRV is equivalent to SNE. This result rationalizes the high performance of NeRV compared to SNE method where the trustworthiness of SNE attained the lowest performance (i.e. when λ equals 1). Thus, NeRV shows the capability of producing a highly trustworthy visualization based on balancing the trade-off between the continuity and trustworthiness of visualization. A large value of λ, close to one, gives a lower trustworthy result and vice versa.

Distance Measures

In the given experiments two types of dissimilarity measure were used: MAF-based kernel and Entropy-based kernel function. The analyses show that the use of MAF-based kernel seems to give slightly similar results to the Entropy-based kernel (results not presented). There was no major differences in using both of these measures as they applied directly to the original high-dimensional feature space. This may be due to the similarity of their underlying approach of emphasizing significant SNPs.

Quality of Visualizations

The first perception that comes to mind is that reducing the dimensionality of the a dataset from 10750 down into just two dimensions will not show how the data

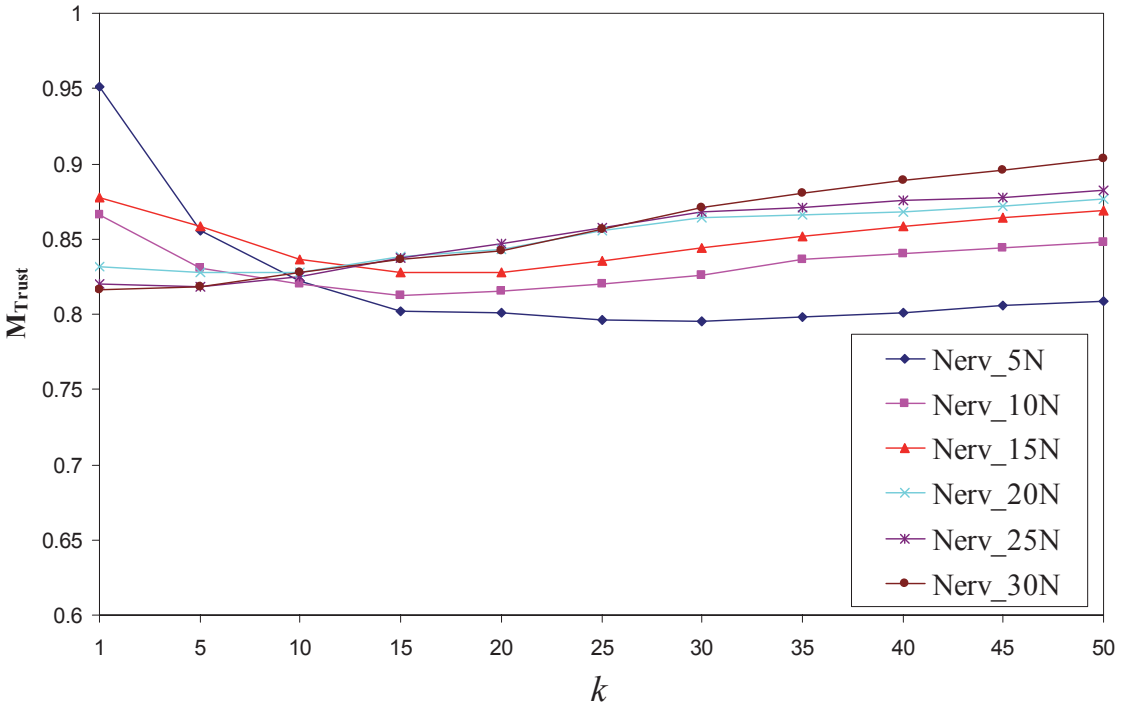

Figure 5.3: Trustworthiness of NeRV mapping as a function of k applied to the Westmead dataset, where k, the neighbourhood's size, is set by trustworthiness. The neighbourhood size, N, used by NeRV is ranging from 5 to 30.

points are similar to each other in the high-dimensional space. On the other hand, without the use of a quality measure, it is difficult to assess and compare different data reduction methods. In the previous sections, results from several data reduction methods were compared in terms of trustworthiness and continuity measures. The best method was selected based on the balance between these two measures with more emphasis put on the trustworthiness measure. Based on the results of the methods and parameter settings, the Neighbour Retrieval Visualizer method, with N nearest

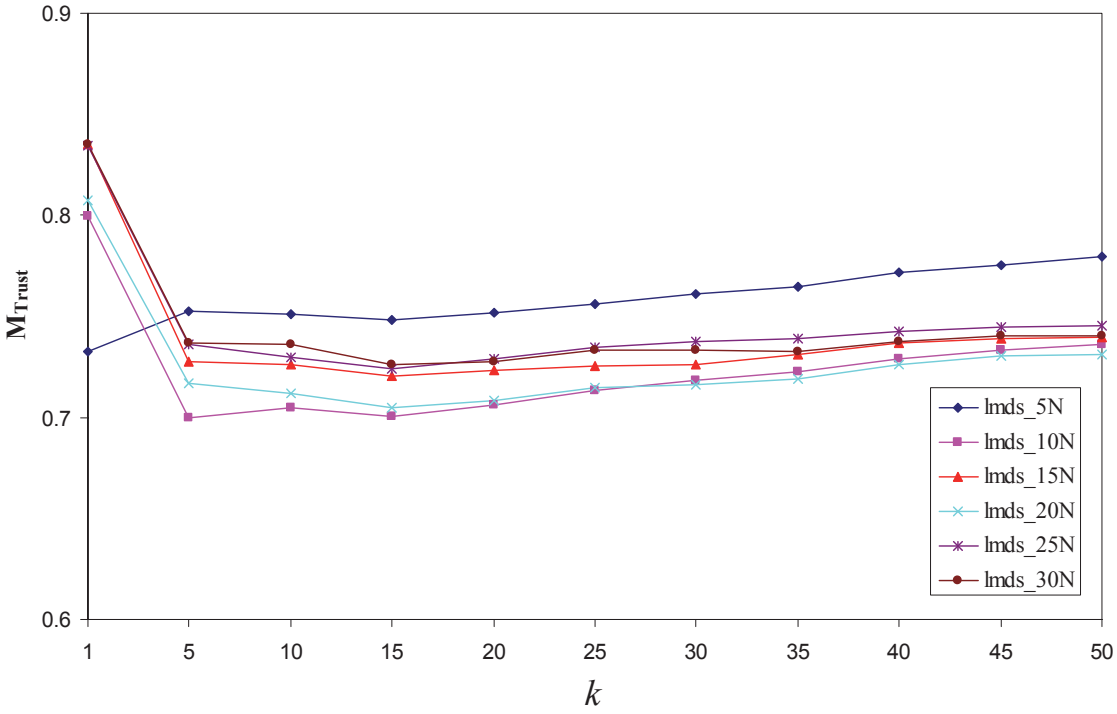

Figure 5.4: Trustworthiness of LocalMDS mapping as a function of k applied to the Westmead dataset, where k, the neighbourhood's size, set by trustworthiness. The neighbourhood size, N, used by LocalMDS is ranging from 5 to 30.

neighbours equal to 30 and $\lambda = 0.3$, produced the best result. Figure 5.6 shows the visualization of data with the NeRV method. From the visualization different clusters of data points (patients) can be seen. The leftmost point in figure 5.6 marks two outliers of patients sitting on top of one another. On the other hand, figure 5.7 shows the visualization produced by LocalMDS. As can be seen, LocalMDS does not capture the structure of the patients and not show any kind of similarity of the data.

This result confirms the ability of NeRV to produce interesting structure (clusters)

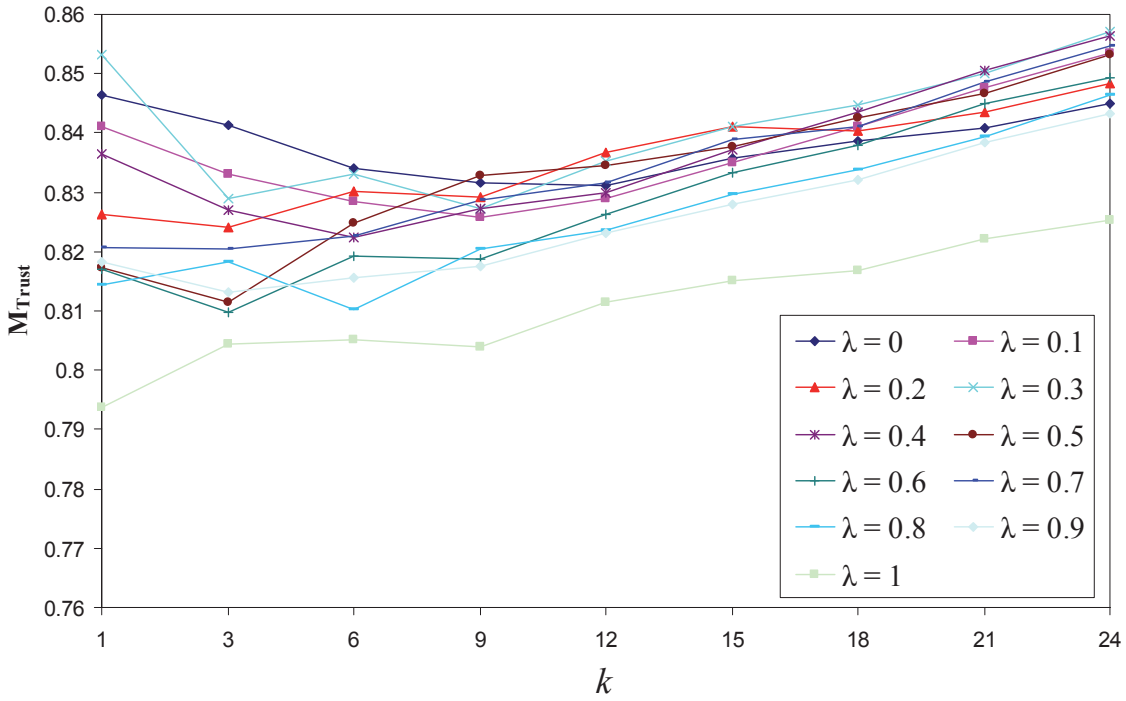

Figure 5.5: Trustworthiness of NeRV mapping as a function of k applied to the Westmead dataset, where k is the size of neighbourhood. The lambda used by NeRV is ranging from 0 to 1.

of data samples. An initial investigation (not presented) showed that the clusters seen in Figure 5.6 may be due to patients' ethnicity (i.e. country of origin). This was an expected outcome as ethnicity is the greatest characteristic of genetic variation. Furthermore, the clusters of patients require further scrutiny by domain experts to validate the results at the biological and clinical levels. At this stage, it is difficult to draw any conclusive biological interpretation. Whilst patients do cluster into groups, without the application of feature selection, the biological interpretation of the cluster is unclear.

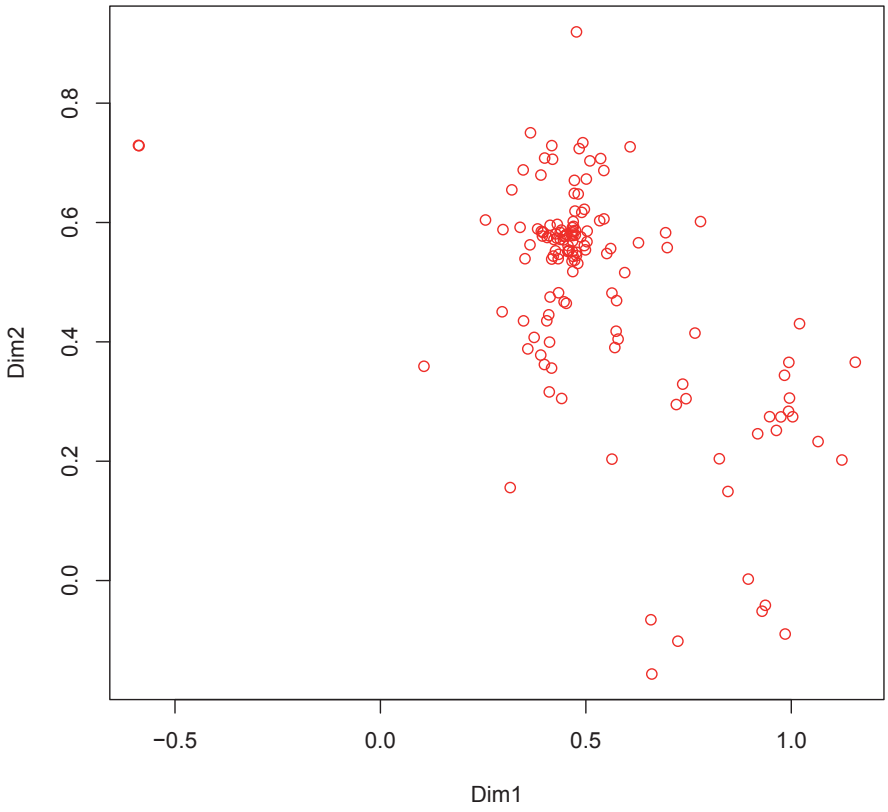

Figure 5.6: Visualization of the Westmead data using NeRV method with $N = 30$ and $\lambda = 0.3$.

5.4.2 Supervised Visualization of the Case-control Dataset

This section extends analyses that have been done in chapter 4, where disease classification and prognostic analyses were applied to different ALL datasets. Here, the visualization of the case-control dataset is illustrated as a supervised-based method. As described in section 4.2, the feature selection procedures were firstly applied to

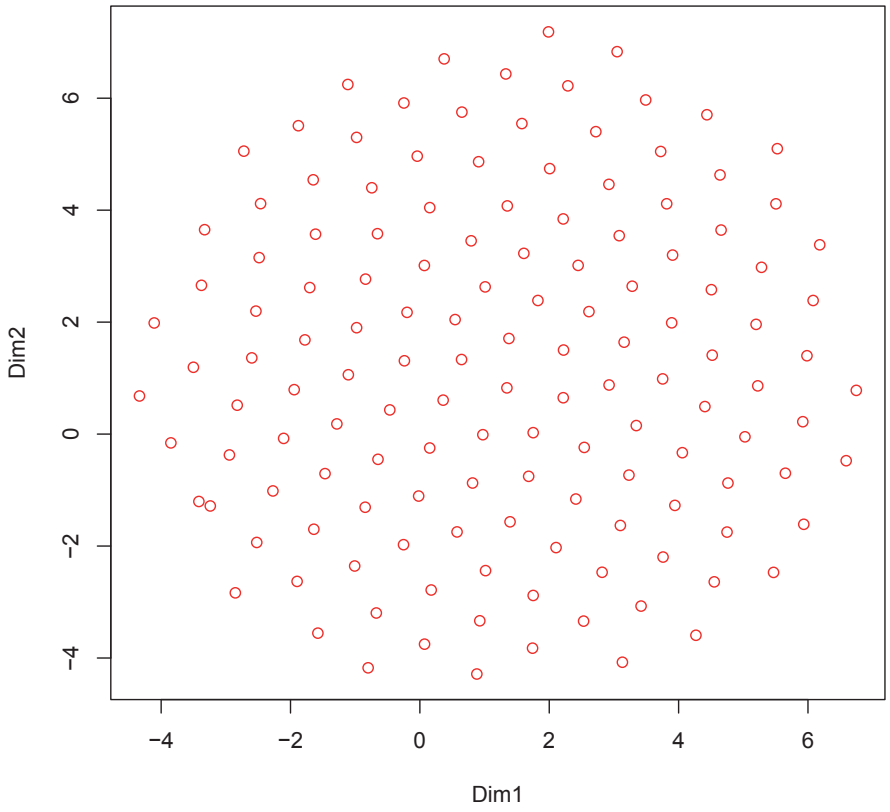

Figure 5.7: Visualization of the Westmead data using LocalMDS method with $N = 15$ and $\lambda = 0.2$

retain feature sets that could accurately differentiate between patient and non-patient data samples. The selection process was applied to select two sets of feature: features that marginally contribute to the disease and another set of features that have interaction effects (i.e. gene-gene interactions) with the disease. The features in the second set represent SNPs that do not marginally contribute to the disease but have non-linear interactions between different SNPs and the disease. Including both SNP

sets would be useful for producing highly predictive models for analysing large scale genome-wide association studies.

The MDS approach was applied to visualize the case-control dataset using three different datasets: (i) using the original SNP dataset without any feature selection process, (ii) a new dataset generated based on features that were selected based on the marginal effect SNP set and (iii) a new dataset generated based on both the selected marginal and interaction effect SNP sets.

Figure 5.8 shows the MDS visualization of case-control data using the whole feature space. The data sample was visualized using the first two dimensions of the MDS mapping of distance matrix. The MAFK-based distance was used to calculate the distance matrix. As figure 5.8 shows, the data samples are not separable. It is apparent that using the whole feature space cannot easily produce a clear picture of relationships between samples of the same class and the ones of the other class.

As the case-control dataset has two classes, feature selection procedures were applied, as described in section 4.2, to evaluate the visualization results of the same datasets based on the new selected feature spaces.

Figure 5.9 shows the visualisation of case-control data based on the marginally selected features. As can be seen, the visualization of the new dataset produced clear overall results of both class. The resultant visualisation confirms the improvement in classification performance that was achieved when feature selection was performed on

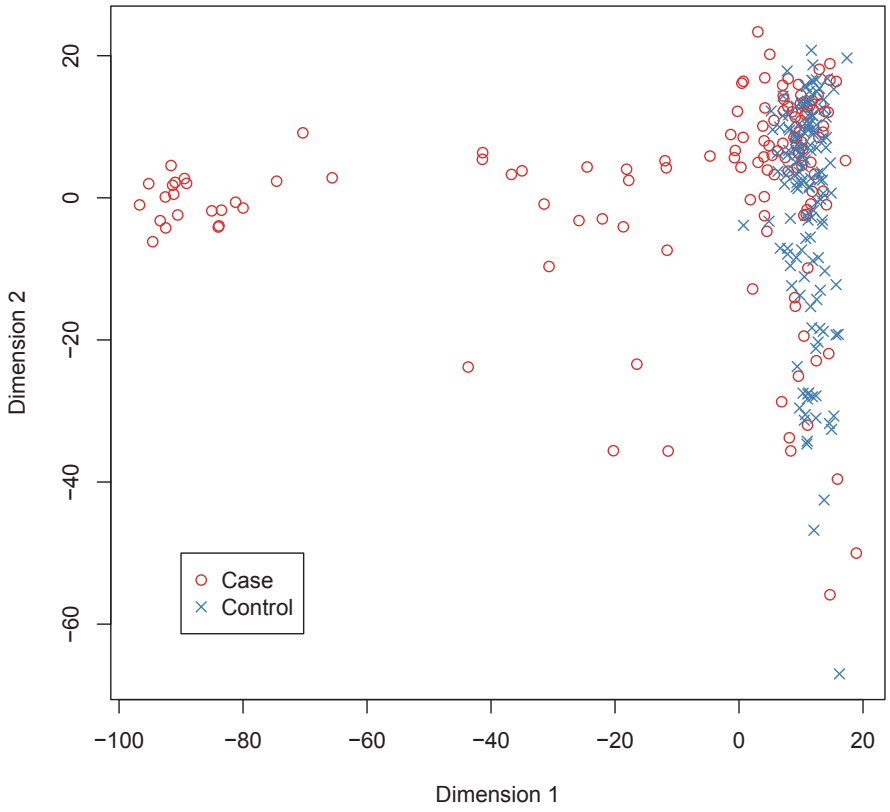

Figure 5.8: The visualization of the case-control dataset based on the whole feature set

the case-control data. However, neither the classification performance achieved nor the visualization results were superior.

The MDS visualization was, lastly, applied to the case-control data after obtaining a new dataset of features that were selected based on both marginal and interaction effect SNPs. The resultant visualization is shown in Figure 5.10, the visualization shows a clear class clustering/separation achieved based on the new selected feature

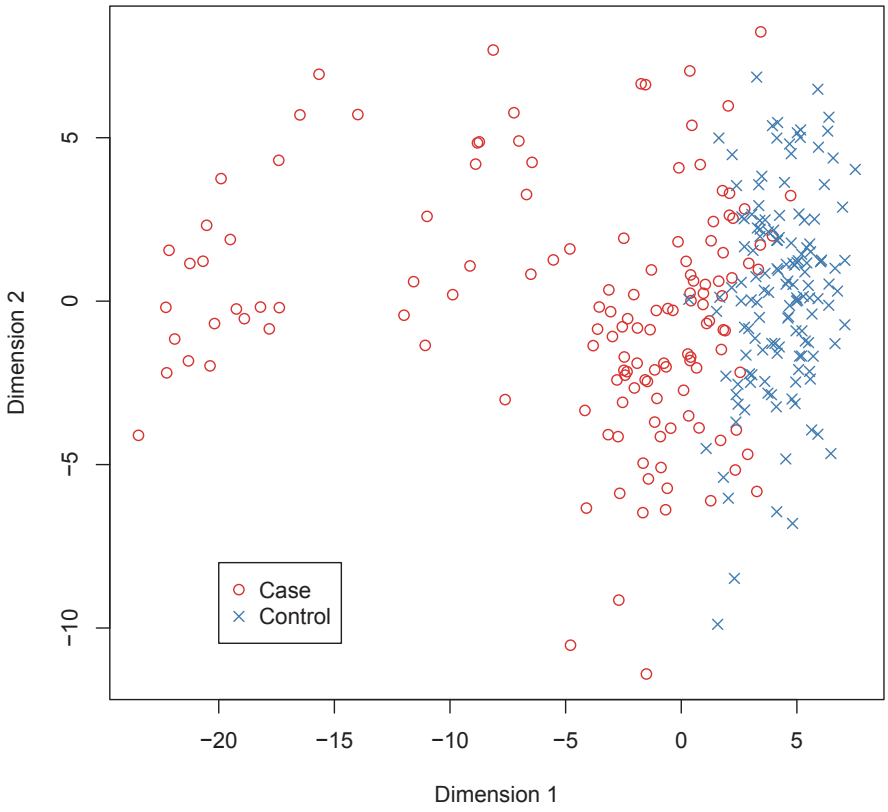

Figure 5.9: The visualization of the case-control dataset based on the marginally selected feature set

set. The resultant visualization confirms the superior improvement in classification performance when the interaction-based features were added to this data.

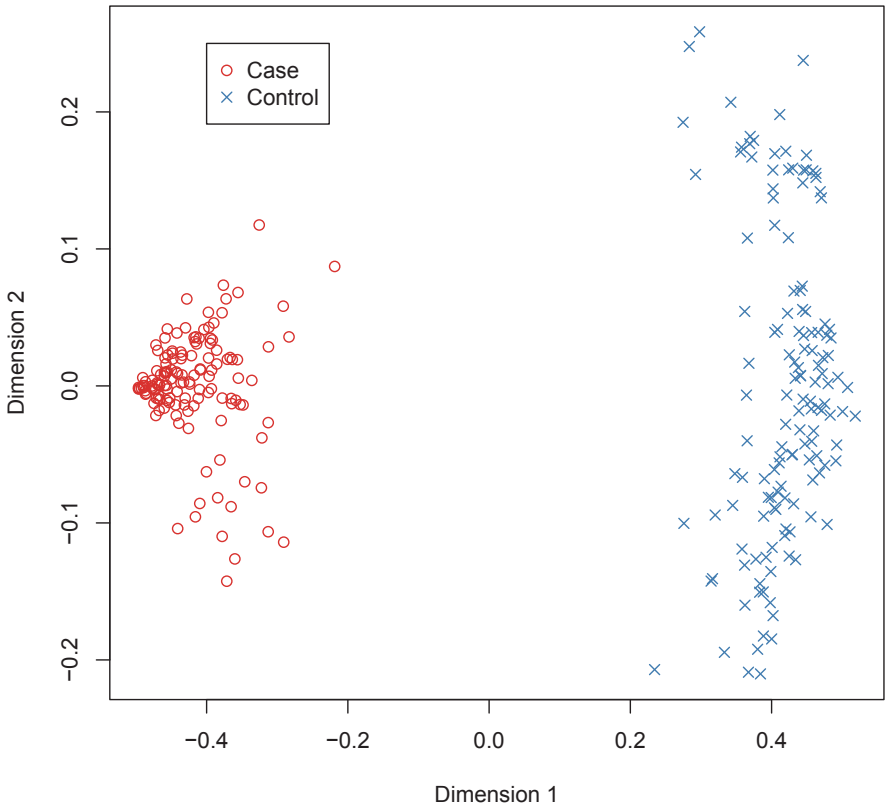

Figure 5.10: The visualization of the case-control dataset based on both selected marginal and interaction effect SNPs

5.5 Summary and Discussion

In summary, different data reduction methods were utilized for visualizing genetic variation (SNP) data as a way to discover the underlying relationships between patients. State-of-art data reduction methods, for both supervised and unsupervised-based visualization, were employed. For the unsupervised case, the result was selected

based on the trustworthiness of the visualizations. The task of visualization was formulated as an information retrieval problem where the result of the visualization describes the local structure of the data. The quality measure applied to the visualization is based on a quantitative error of the number of misses and false positives (i.e. trustworthiness and continuity measures).

Several dimensionality reduction methods were tested. These include PCA, Laplacian Eigenmap and Locally Linear Embedding that are designed to extract data manifolds. Extended versions of the Stochastic Neighbour Embedding and Curvilinear Component Analysis methods called Neighbour Retrieval Visualizer and Local MDS methods, respectively, were also applied. These data reduction methods were run on the given dataset with different parameter settings. The Neighbour Retrieval Visualizer method has shown the best performance on the Westmead dataset. This method balances the trade-off between the trustworthiness and continuity of the visualization. The result shows that a neighbourhood of size 30 was the best for the given data. A parameter λ which controls the trade-off between trustworthiness and continuity was selected to be 0.3. This parameter emphasizes the trustworthiness of the visualization which is more important for visualization.

The performance results did not show any differences using different distance matrices due to high-dimensionality aspect of the data. Specifically, the use of MAFK

and EK kernel functions did not show any differences due to their similar underlying procedures. Manifold-based data reduction methods, i.e. LLE and LE, perform surprisingly badly and the Laplacian Eigenmap method similarly performs as badly as a random mapping of the data. This result was not expected for these methods due to their high performance on other datasets. One possible reason for this may be that these methods are designed to discover the intrinsic dimensionality of the data manifold which cannot be easily unfolded. The performance of PCA was comparable to LocalMDS although the latter method is considered as a non-linear dimensionality reduction method. This behaviour suggests that the dataset has a linear relationship and it was discoverable using the first two eigenvectors.

Dealing with large scale datasets, such as genome-wide SNPs, requires the application of feature selection procedures to select a significant set of features with high discriminating power. Therefore, the visualization process would be more accurate and informative after applying feature selection to exclude insignificant (noisy) features that distort the performance of the produced models and visualizations. The visualization results of the case-control dataset, as a supervised-based case, support such an approach.

To conclude, this chapter utilized genome-wide SNP data for the comparison and visualisation of patient-to-patient relationships based on different data reduction methods. The comparison and visualisation were applied to real genetic variation

data of leukaemia studies. In the course of doing this, this chapter has addressed the defined thesis research question (i.e. **Q5**) in section 1.3 and the identified thesis contribution 5 as listed in section 1.4. The results of this chapter showed the feasibility of using the data reduction method, named NeRV, with specific parameter settings, to visualize the given genetic variation datasets. The results also showed the importance of including feature selection procedures prior to the visualization step, especially, for supervised-based cases.

Chapter 6

Conclusion

The objective of this thesis is to investigate the effectiveness of using genetic variation data, as assessed by genome-wide SNP profiles, for modelling, exploration and visualization of patient-to-patient relationships for complex diseases. Indeed, in this research study data mining and machine learning approaches are used to develop and propose methods and techniques for the modelling of different aspects of complex diseases based on the given datasets.

A major goal of genetic variation studies is to identify and characterize the genetic variants (markers) that contribute to complex diseases. One of the fundamental subjects in these studies is to find an optimal subset of variants with the highest predicting power for any given complex disease model. The basic strategy that has been used in these studies is to examine the relationship between the disease of interest and the genetic markers across the human genome in different samples of affected individuals with unaffected controls, termed as genome-wide association studies.

A large number of association studies has led to the discovery of genetic risk factors associated with common diseases. However, the proportion of genetic variation explained by those variants is limited (The National Human Genome Research Institute 2003). These studies essentially rely on evaluating one marker at a time based on the assumption that disease susceptibility genes can be identified through their independent contribution to disease variability. In contrast, complex diseases are not observed to be caused by single genes acting alone but are the result of complex non-linear interactions among genetic factors, with each gene having a small effect on the disease (Wu et al. 2010, Wang et al. 2011). For that reason there is a critical need to implement new approaches that can take into account non-linear gene-gene interactions in searching for markers that jointly cause complex diseases.

With the rapid advances in biotechnology, especially the improvement in SNP genotype technologies, genetic researchers now can genotype and analyse a large variety of genetic variants that cover most of the human genome. It is of great interest to explore the genetic architecture and models underlying complex diseases. To achieve such goals, efficient feature selection must be employed to find an optimal subset of markers with the highest predictive power for the disease of interest. Indeed, the followed approaches, in this thesis, use feature selection methods to select genetic markers that account for marginal effects as well as gene-gene interactions associated with disease status. Applying only feature selection methods based on traditional approaches will result in choosing markers that marginally contribute to disease and will

ignore the gene-gene interactions, which may be useful for producing highly predictive models for large scale genome-wide studies.

By analysing the main challenges of the given domain, the work presented in this thesis focuses on three main aspects for analysing the difficulties in genetic variation studies, including: (1) feature selection and distance calculations, (2) a framework for the task of disease diagnosis and prognosis, and (3) models for the comparison and visualisation of patient-to-patient relationships based on genome-wide SNP profiles. The novelty of the proposed approaches, analysed in this thesis, is the intensive use of non-parametric machine learning techniques for the purpose of feature selection, weighting, and prioritizing of SNPs, as well as for calculating distance measures between patients. Each of these approaches was jointly considered for building different disease models using data from genome-wide studies. In the following section, a summary of previous chapters in terms of the thesis's contributions is outlined.

6.1 Summary of Contributions

Contribution 1: Feature Selection and weighting

In chapter 3, several methods and approaches have been defined and proposed to deal with the tasks of SNP selection, weighting, prioritization, feature construction and distance calculation in the domain of genetic variation studies. The proposed approaches consist of screening genetic variation data to search for main and interaction

effects of the gene susceptibility markers that jointly cause complex diseases. The proposed approaches are based on non-parametric machine learning techniques, which offer a powerful and alternative approach to traditional statistical-based methods. These approaches are generally model-free and able to detect non-linear interactions in such high-dimensional data, such as genetic variation data.

To deal with the problem of feature selection and weighting, the use of RF-based feature importance measure was proposed, with an iterative process for the task of selecting and weighting SNP data, named the RF-RFE approach.

Contribution 2: A Measure for Detecting Gene-Gene Interaction Effects

Furthermore, a new approach for SNP prioritization was also defined in chapter 3. The new approach can be used to search for a set of SNPs that has the potential to be involved in gene-gene interaction. This new measure uses the information gain (entropy) value of different disease status to calculate their distribution: highly skewed SNPs are the ones with less interaction effect. To define a cut-off value of a ranked list of SNPs based on their IE measures, the use of IE as a splitting criterion in RF trees construction is proposed. Furthermore, a feature induction-based approach was used to construct new features based on the prioritized SNPs. The new constructed features carry the information that potentially account for gene-gene interactions. The RF-RFE approach was also used to obtain an optimal set of features (based on the new constructed feature) that explain a given disease model.

The proposed approaches are designed to reduce the number of features, remove irrelevant, redundant, or noisy data, and thus, improve the modelling process and its performance as measured by predictive accuracy and interpretability of results. The effectiveness of the proposed feature selection methods and approaches has been evaluated experimentally in chapters 4 and 5.

Contribution 3: Distance Measures for Genetic Profiles

Methods for calculating distances between different genotype profiles were given in chapter 3. Three distance measures, based on a kernel-based weighting function, were defined. These measures include random forest (RFK), minor allele frequency (MAFK) and entropy-based (EK) kernel functions. The distance measures use SNPs that have been selected, weighted and constructed based on the proposed approaches to calculate distances. The distances can be subsequently used for modelling different disease classifications, clustering and visualization models. Indeed. The effectiveness of the proposed distance measures methods and approaches has been evaluated experimentally in chapters 4 and 5. One significant application of distance calculations is to evaluate patient-to-patient relationships. Such relationships may lead to delivery of advanced personalized medicine, by choosing treatment strategies which best suit each individual patient compared to the most similar cases.

Contribution 4: A Framework for Disease Diagnosis and Prognosis

In chapter 4, a new computational framework for disease classification and subtype discovery was developed and validated. Methods and approaches used in the proposed framework have the potential to increase the performance of the given tasks without an increase in computational complexity or sample size needed. As the results showed, developing methods for detecting gene-gene interaction in large scale genome-wide genetic variation datasets is extremely desirable for characterizing the genetic basis for common complex disease. The feature selection procedures applied have emphasized the importance and need to include interaction markers in the final model assessments. At the heart of this framework is the use of constructive induction algorithms such as MDR that are capable of capturing non-linear gene-gene interactions to form new constructed attributes.

The proposed framework was empirically evaluated using two case studies of acute lymphoblastic leukaemia and as the performance results showed, a significant improvement of models' performance was only achieved by including interaction markers in the final built models. The results, further, suggest that the proposed prediction approaches can effectively define important markers that are mostly consistent with known biological findings while the accuracy of the produced models are also high.

Building reliable disease diagnostic and prognostic classification models has been the main focus in the thesis. Building classification models, based on genetic variation

data, would be one of the major steps leading to personalized medicine. The new proposed methods are innovative in terms of utilizing genetic variation profiles to obtain general and reliable models for the purpose of delivering an advanced personalized medicine, thus, choosing treatment strategies which best suit each individual patient. The generated models can be further used by clinicians and biomedical researchers for a better understanding of disease etiologies and development of future treatment strategies.

Interestingly, the power of the proposed framework can be seen in its flexibility and the ability to plug and play different methods and approaches on each phase of the framework. The ability to take multiple analytical paths through the proposed framework will facilitate knowledge and pattern discovery of results across different analytical processes, and in turn, the results can be compared and mined. Perhaps the greatest challenge is to reduce the computational complexity of discovering interaction markers. As the feature selection phase demonstrated, the use of the IE measure can be utilized to alleviate a large portion of this problem.

Contribution 5: Visualizing Genome-wide SNP Profiles

Finally, in chapter 5, several data reduction methods were employed for visualizing genetic variation datasets. For unsupervised-based visualization, data reduction methods were compared based on the trustworthiness metric of the resultant visualization. For the supervised-based visualization, the performance was compared on

how well the methods discriminated two classes of the combined dataset.

To deal with large amounts of genetic variation data, the choice was to compare the performance of different dimensionality reduction methods on the given dataset. Based on this comparison the NeRV approach showed the best results and outperformed other data reduction approaches. Even though the dimensionality of the dataset was reduced from 10750 to 2 dimensions, the quality measure of the visualization still showed excellent results. The visualization results have the potential to assist clinicians and biomedical researchers in understanding relationships between patients and to compare different clusters in the visualization. The visualization result from using the NeRV approach shows the feasibility of this method in visualizing genetic variation data.

In the supervised-based approach, the visualization of the case-control dataset showed the importance of using feature selection methods for removing insignificant features, as the resultant visualizations were more accurate and informative.

6.2 Limitations and Future Directions

The most apparent limitation associated with conducting genome-wide studies is the high cost and computational analysis required to genotype hundreds of thousands of SNPs. In fact, one of the major challenges in genomic research, in upcoming decades, is translating the new emerging genomic knowledge into public health and medical

care in the form of personalized medicine (e.g. Kasabov & Hu (2010) and Kasabov (2009)).

The main limitation of the employed methods is the feasibility and scalability of the used feature selection approaches due to the high-dimensionality of the data. Introducing the importance of including genes (markers) that have gene-gene interaction effects in models' building has shown a significant improvement in the performance of the produced models. However, assessing the feasibility of the applied feature selection approaches is a challenging task. There is a need to apply the proposed approaches to other complex diseases to test the feasibility of these approaches.

Most of the disease modelling processes, applied here, rely on distance matrix calculations, especially for the task of visualizing genetic variation profiles (data reduction methods). Therefore, the performance of built models is strongly related to the applied distance measure. The proposed feature selection approaches have shown a significant impact on improving distance calculations. The future direction of this work is to employ other distance measures that might be more appropriate in discriminating the major characteristics of genetic variation datasets. In particular, prior knowledge or domain-driven dissimilarity measures may improve the performance of the produced models in the examined datasets. In addition, the choice of the neighbourhood size, k, for different data reduction methods could have a significant impact in improving the modelling and visualizing of SNP datasets. A further investigation

in the choice of a flexible neighbourhood size, k, for each data sample separately can be optimized and evaluated (e.g. Kasabov & Hu (2010)).

The availability of a flexible framework for the task of disease diagnosis and prognosis, as proposed in this thesis, will play an important role in understanding the genetic basis to common complex human diseases. A comprehensive validation of the methods and approaches embedded in the proposed framework is a matter of applying this framework to other complex diseases. After that, the framework can be implemented in a form of user-friendly biomedical software for analysing and exploring the genetic basis of different human diseases. The analytical framework described here takes us a step closer to making this framework a real life application for use by clinicians and pathologists.

Eventually and as a future direction of this research, other data mining techniques such as case-based reasoning systems can be combined with the proposed approaches to assist clinicians in understanding the biological basis of diseases and assist in clinical decision making (e.g. Kasabov & Hu (2010) and Kasabov et al. (2008)). Therefore, and as the initial biological interpretation results showed, a number of improvements could be made to discover more meaningful biological interpretations of complex diseases.

Appendix A

SNPs that marginally Contribute to the ALL disease

SNP List

The functional annotation analyses of genes associated with reported SNPs for case-control dataset, Westmead, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them

Index	SNP name	Chromosome	Gene Symbol	Gene Description
1	rs999597	21	KRTAP13-4	keratin associated protein 13-4
2	rs9972951	17	KIAA1001	Arylsulfatase G
3	rs9944035	14	SYNE2	spectrin repeat containing, nuclear envelope 2
4	rs9936215	16	-	-
5	rs9835332	3	RAP140	retinoblastoma-associated protein 140
6	rs9652589	16	PDILT	protein disulfide isomerase-like protein of the testis
7	rs9652588	16	PDILT	protein disulfide isomerase-like protein of the testis
8	rs958755	3	-	-
9	rs950169	15	ADAMTSL3	ADAMTS-like 3
10	rs9429157	1	TOE1	target of EGR1, member 1 (nuclear)
11	rs9371533	6	RAET1E	retinoic acid early transcript 1E
12	rs9370867	6	MYLIP	myosin regulatory light chain interacting protein
13	rs9356991	6	-	-
14	rs9295676	6	SLC17A2	solute carrier family 17 (sodium phosphate), member 2
15	rs9267954	6	-	-
16	rs926487	20	ZNF337	zinc finger protein 337
17	rs9262	12	DKFZp434N2030	hypothetical protein DKFZp434N2030
18	rs9261285	6	PPP1R11	protein phosphatase 1, regulatory (inhibitor) subunit 11
19	rs910925	17	GEMIN4	gem (nuclear organelle) associated protein 4
20	rs901363	6	SYTL3	synaptotagmin-like 3
21	rs891017	19	ANKRD41	hypothetical protein FLJ39369

22	rs877834	7	C7orf9	chromosome 7 open reading frame 9
23	rs838827	9	IKBKAP	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein
24	rs8192297	15	ANPEP	alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150)
25	rs8079797	17	USP32	ubiquitin specific protease 32
26	rs806970	6	-	-
27	rs8058694	16	ABCC6	ATP-binding cassette, sub-family C (CFTR/MRP), member 6
28	rs8010699	14	SYNE2	spectrin repeat containing, nuclear envelope 2
29	rs7940667	11	FLJ45300	FLJ45300 protein
30	rs7918199	10	MKI67	antigen identified by monoclonal antibody Ki-67
31	rs785467	1	PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (p55, gamma)
32	rs7803620	7	FLJ21062	hypothetical protein FLJ21062
33	rs7799	13	ELF1	E74-like factor 1 (ets domain transcription factor)
34	rs7756521	6	-	-
35	rs7738586	6	LOC442208	similar to ribosomal protein S2
36	rs7727919	5	LOC134466	hypothetical protein LOC134466
37	rs769051	6	-	-
38	rs7593557	2	TRPM8	transient receptor potential cation channel, subfamily M, member 8
39	rs7566476	2	FIGLA	factor in the germline alpha
40	rs7494244	14	C14orf162	chromosome 14 open reading frame 162
41	rs7332388	13	KIAA1008	KIAA1008
42	rs7209474	17	CCDC57	hypothetical protein LOC284001
43	rs7206698	16	LOC400499	hypothetical gene supported by AK126539
44	rs7203179	16	-	-
45	rs7163367	15	LOC56964	hypothetical protein from EUROIMAGE 384293
46	rs709012	20	SN	sialoadhesin
47	rs708167	12	-	-
48	rs7049042	9	OR13F1	olfactory receptor, family 13, subfamily F, member 1
49	rs7038903	9	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1
50	rs6968325	7	PRKRIP1	PRKR interacting protein 1 (IL11 inducible)
51	rs6958844	7	LOC401320	hypothetical LOC401320
52	rs6932590	6	-	-
53	rs6930033	6	OR5V1	olfactory receptor, family 5, subfamily V, member 1

54	rs6929137	6	C6orf97	chromosome 6 open reading frame 97
55	rs6923761	6	GLP1R	glucagon-like peptide 1 receptor
56	rs683	9	TYRP1	tyrosinase-related protein 1
57	rs6769771	3	LOC644063	-
58	rs6710817	2	NAG	neuroblastoma-amplified protein
59	rs6661946	1	HEATR1	protein BAP28
60	rs664850	9	-	-
61	rs662515	18	MPPE1	metallophosphoesterase 1
62	rs6573560	14	C14orf50	chromosome 14 open reading frame 50
63	rs6546837	2	ALMS1	Alstrom syndrome 1
64	rs652633	20	ANKRD5	ankyrin repeat domain 5
65	rs6482626	10	PTCHD3	FLJ44037 protein
66	rs6464573	7	OR2A5	olfactory receptor, family 2, subfamily A, member 5
67	rs6442089	3	MAP4	microtubule-associated protein 4
68	rs615474	9	MGC41945	hypothetical protein MGC41945
69	rs6083	15	LIPC	lipase, hepatic
70	rs6076347	20	PSF1	DNA replication complex GINS protein PSF1
71	rs585033	18	ATP9B	ATPase, Class II, type 9B
72	rs5745549	1	MSH4	mutS homolog 4 (E. coli)
73	rs5743810	4	TLR6	toll-like receptor 6
74	rs550897	11	FAM55D	chromosome 11 open reading frame 33
75	rs543573	9	SETX	amyotrophic lateral sclerosis 4
76	rs540825	6	OPRM1	opioid receptor, mu 1
77	rs524625	20	ANKRD5	ankyrin repeat domain 5
78	rs4934281	10	MMRN2	multimerin 2
79	rs4902264	14	SYNE2	spectrin repeat containing, nuclear envelope 2
80	rs4859905	4	FRAS1	Fraser syndrome 1
81	rs482082	1	WDR78	hypothetical protein FLJ23129
82	rs4757986	11	OR52E4	olfactory receptor, family 52, subfamily E, member 4
83	rs4646626	15	ALDH1A2	aldehyde dehydrogenase 1 family, member A2
84	rs4629709	6	C6orf191	chromosome 6 open reading frame 191
85	rs4287542	15	LOC56964	hypothetical protein from EUROIMAGE 384293
86	rs4243786	2	LOC51252	hypothetical protein LOC51252
87	rs4237768	11	OR52L1	olfactory receptor, family 52, subfamily L, member 1
88	rs4151651	6	BF	B-factor, properdin
89	rs4141885	6	HIST1H1E	histone 1, H1e
90	rs3922872	22	ALG12	asparagine-linked glycosylation 12 homolog (yeast, alpha-1,6-mannosyltransferase)
91	rs3916626	14	OR4K1	olfactory receptor, family 4, subfamily K, member 1

92	rs389600	6	HLA-K	major histocompatibility complex, class I, K
93	rs3865452	19	ADCK4	aarF domain containing kinase 4
94	rs3853620	12	OVOS2	ovostatin 2
95	rs3829747	2	TTN	titin
96	rs3829405	14	OR6J1	olfactory receptor, family 6, subfamily J, member 1
97	rs3823339	6	HLA-A	major histocompatibility complex, class I, A
98	rs3815045	11	RAB3IL1	RAB3A interacting protein (rabin3)-like 1
99	rs3810510	20	JPH2	junctophilin 2
100	rs3810481	20	PRIC285	peroxisomal proliferator-activated receptor A interacting complex 285
101	rs3800855	7	PTPRN2	protein tyrosine phosphatase, receptor type, N polypeptide 2
102	rs3799277	6	TDRD6	tudor domain containing 6
103	rs3790647	1	CENPF	centromere protein F, 350/400ka (mitosin)
104	rs3770436	2	GORASP2	golgi reassembly stacking protein 2, 55kDa
105	rs3764656	19	THEG	Theg homolog (mouse)
106	rs3752302	20	DHX35	DEAH (Asp-Glu-Ala-His) box polypeptide 35
107	rs3750944	11	DKFZP566M1046	hypothetical protein DKFZp566M1046
108	rs3750050	7	PTPN12	protein tyrosine phosphatase, non-receptor type 12
109	rs3748693	1	CENPF	centromere protein F, 350/400ka (mitosin)
110	rs3748178	9	AKNA	AT-hook transcription factor
111	rs3747923	20	LOC643503	-
112	rs3745640	19	MGC24975	hypothetical protein MGC24975
113	rs3743279	15	SEMA6D	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6D
114	rs3735135	7	CDC2L5	cell division cycle 2-like 5 (cholinesterase-related cell division controller)
115	rs3732487	3	MYLK	myosin, light polypeptide kinase
116	rs369637	6	TNXB	tenascin XB
117	rs365605	11	SLC22A18AS	solute carrier family 22 (organic cation transporter), member 1-like antisense
118	rs363075	4	HD	huntingtin (Huntington disease)
119	rs35414	5	SLC45A2	membrane associated transporter
120	rs35392772	8	-	-
121	rs3177676	1	C1orf105	LOC92346
122	rs3177253	6	PECI	peroxisomal D3,D2-enoyl-CoA isomerase
123	rs3177243	22	DERL3	Der1-like domain family, member 3

124	rs3131933	6	-	-
125	rs3131085	6	LOC646260	-
126	rs3131018	6	-	-
127	rs3128982	6	-	-
128	rs3127158	6	LOC202459	similar to RIKEN cDNA 2310008M10
129	rs3124747	9	C9orf96	chromosome 9 open reading frame 96
130	rs3123101	6	SYTL3	synaptotagmin-like 3
131	rs3121478	10	NRAP	nebulin-related anchoring protein
132	rs3117299	-	-	-
133	rs3095273	6	-	-
134	rs2986671	9	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1
135	rs296091	19	ZNF91	zinc finger protein 91 (HPF7, HTF10)
136	rs2864057	6	C6orf208	chromosome 6 open reading frame 208
137	rs2839110	21	COL6A2	collagen, type VI, alpha 2
138	rs2723880	12	KIAA1853	KIAA1853 protein
139	rs2688024	17	LOC353194	keratin pseudogene
140	rs268673	19	PRX	periaxin
141	rs2578652	15	GANC	glucosidase, alpha; neutral C
142	rs2523898	6	-	-
143	rs2523072	7	C7orf31	chromosome 7 open reading frame 31
144	rs242017	12	MGC4266	hypothetical protein MGC4266
145	rs2407221	4	ESSPL	epidermis-specific serine protease-like protein
146	rs2397084	6	IL17F	interleukin 17F
147	rs2394130	6	-	-
148	rs2306590	17	MYOHD1	myosin head domain containing 1
149	rs2305962	5	LOC133993	hypothetical LOC133993
150	rs2301151	19	GPR40	G protein-coupled receptor 40
151	rs2296262	9	DFNB31	deafness, autosomal recessive 31
152	rs2294793	6	AKAP12	A kinase (PRKA) anchor protein (gravin) 12
153	rs2289030	12	LOC160313	keratin 19 pseudogene
154	rs2286612	11	NUDT22	hypothetical protein MGC13045
155	rs2280851	8	HHLA1	HERV-H LTR-associating 1
156	rs2276360	11	NADSYN1	NAD synthetase 1
157	rs2275797	1	ADORA3	adenosine A3 receptor
158	rs2273863	1	LGALS8	lectin, galactoside-binding, soluble, 8 (galectin 8)
159	rs2273566	6	SMAP1	stromal membrane-associated protein 1
160	rs2272994	1	ZNF643	zinc finger protein 643
161	rs2270581	12	NCKAP1L	hematopoietic protein 1
162	rs2250145	5	MGC23985	similar to AVLV472
163	rs2241032	16	GAS8	growth arrest-specific 8
164	rs2240229	19	OR10H3	olfactory receptor, family 10, subfamily H, member 3

165	rs2240154	19	GRIN3B	glutamate receptor, ionotropic, N-methyl-D-aspartate 3B
166	rs2236194	20	C20orf58	chromosome 20 open reading frame 58
167	rs2229333	18	TGIF	TGFB-induced factor (TALE family homeobox)
168	rs2229116	15	RYR3	ryanodine receptor 3
169	rs2227973	11	RAG1	recombination activating gene1
170	rs2227289	19	CD320	CD320 antigen
171	rs2227271	11	UBQLN3	ubiquilin 3
172	rs214830	20	TGM3	transglutaminase 3 (E polypeptide, protein-glutamine-gamma-glutamyltransferase)
173	rs2104772	9	TNC	tenascin C (hexabrachion)
174	rs2089891	1	PPP1R15B	protein phosphatase 1, regulatory (inhibitor) subunit 15B
175	rs2076185	6	C6orf105	chromosome 6 open reading frame 105
176	rs2076015	20	TXNDC13	hypothetical protein DJ971N18.2
177	rs2071789	6	-	-
178	rs2071299	6	SLC17A2	solute carrier family 17 (sodium phosphate), member 2
179	rs2069561	8	TG	thyroglobulin
180	rs2023194	16	NPM1P3	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 3
181	rs2018721	21	RIMKLP	ribosomal modification protein rimK-like (E. coli) pseudogene
182	rs2015436	6	-	-
183	rs1990313	12	AKAP3	A kinase (PRKA) anchor protein 3
184	rs1976165	17	SEZ6	seizure related 6 homolog (mouse)
185	rs197412	1	DDX20	DEAD (Asp-Glu-Ala-Asp) box polypeptide 20
186	rs1966834	11	OR1S1	olfactory receptor, family 1, subfamily S, member 1
187	rs1885986	17	C17orf31	chromosome 17 open reading frame 31
188	rs1883329	6	-	-
189	rs1825474	2	LOC645784	-
190	rs1805074	5	DMGDH	dimethylglycine dehydrogenase
191	rs1805073	5	DMGDH	dimethylglycine dehydrogenase
192	rs1805007	16	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
193	rs1801222	10	CUBN	cubilin (intrinsic factor-cobalamin receptor)
194	rs1801197	7	CALCR	calcitonin receptor
195	rs1800440	2	CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1

196	rs1799999	7	PPP1R3A	protein phosphatase 1, regulatory (inhibitor) subunit 3A (glycogen and sarcoplasmic reticulum binding subunit, skeletal muscle)
197	rs17723260	1	EFNA3	ephrin-A3
198	rs1769	3	TMEM113	hypothetical protein PRO2730
199	rs17158558	10	RET	ret proto-oncogene (multiple endocrine neoplasia and medullary thyroid carcinoma 1, Hirschsprung disease)
200	rs17118	3	XYLB	xylulokinase homolog (H. influenzae)
201	rs17101193	10	NRG3	neuregulin 3
202	rs16961689	18	DSG1	desmoglein 1
203	rs16866406	2	TTN	titin
204	rs1681596	14	OR6J1	olfactory receptor, family 6, subfamily J, member 1
205	rs1622208	1	MAST2	microtubule associated serine/threonine kinase 2
206	rs1610682	6	HCGVIII-2	HCGVIII-2 pseudogene
207	rs15967	-	-	-
208	rs157024	13	MYR8	myosin heavy chain Myr 8
209	rs1479500	12	FLJ36004	likely ortholog of mouse Pas1 candidate 1
210	rs1425917	11	LOC120379	hypothetical protein BC019238
211	rs1422698	5	ANKRD31	ankyrin repeat domain 31
212	rs138646	22	KIAA1043	KIAA1043 protein
213	rs1361754	1	FLJ32569	hypothetical protein FLJ32569
214	rs13433937	3	TNK2	tyrosine kinase, non-receptor2
215	rs13362036	5	FLJ14166	hypothetical protein FLJ14166
216	rs1320571	1	TTLL10	hypothetical protein FLJ36119
217	rs12931472	16	ABCC6	ATP-binding cassette, subfamily C (CFTR/MRP), member 6
218	rs12907196	15	NOX5	NADPH oxidase, EF hand calcium-binding domain 5
219	rs12871608	13	NUPL1	nucleoporin like 1
220	rs12660111	6	-	-
221	rs12656542	5	LOC134466	hypothetical protein LOC134466
222	rs1265096	6	PSORS1C1	psoriasis susceptibility 1 candidate 1
223	rs12604020	17	SLC5A10	solute carrier family 5 (sodium/glucose cotransporter), member 10
224	rs12595158	15	VPS13C	vacuolar protein sorting 13C (yeast)
225	rs1255953	14	SYNE2	spectrin repeat containing, nuclear envelope 2
226	rs1254319	14	C14orf39	chromosome 14 open reading frame 39
227	rs12535348	7	C7orf31	chromosome 7 open reading frame 31
228	rs12453150	17	FLJ34922	hypothetical protein FLJ34922
229	rs12437204	14	C14orf143	chromosome 14 open reading frame 143

230	rs12407003	1	-	-
231	rs12334990	8	KIAA1429	DKFZP434I116 protein
232	rs12289558	11	OVCH2	ovochymase 2
233	rs1201689	15	MAPKBP1	mouse mitogen-activated protein kinase binding protein 1-like
234	rs1165196	6	SLC17A1	solute carrier family17 (sodium phosphate), member1
235	rs11643815	16	MT4	metallothionein IV
236	rs11617984	13	-	-
237	rs11575584	9	CCL27	chemokine (C-C motif) ligand 27 IL11RA interleukin 11 receptor, alpha
238	rs11558492	1	GNPAT	glyceronephosphate O-acyltransferase
239	rs11558171	16	GOT2	glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)
240	rs11553577	10	C10orf61	chromosome 10 open reading frame 61
241	rs11550540	20	TP53RK	TP53 regulating kinase
242	rs1151687	1	OR2G2	olfactory receptor, family 2, subfamily G, member 2
243	rs1147990	1	TTC4	tetratricopeptide repeat domain 4
244	rs1131769	5	LOC340061	hypothetical protein LOC340061
245	rs1128864	4	ART3	ADP-ribosyltransferase 3
246	rs1127477	6	C6orf85	chromosome 6 open reading frame 85
247	rs11264542	1	PRCC	papillary renal cell carcinoma (translocation-associated)
248	rs11208997	1	C1orf141	hypothetical gene supported by BC047053
249	rs11190	10	C10orf89	chromosome 10 open reading frame 89
250	rs11161732	1	COL24A1	collagen, type XXIV, alpha 1
251	rs11155242	6	GPR126	G protein-coupled receptor 126
252	rs11152199	18	SDCCAG3L	serologically defined colon cancer antigen 3-like
253	rs11079	3	GUP1	chromosome 3 open reading frame 3
254	rs11057939	12	DHX37	DEAH (Asp-Glu-Ala-His) box polypeptide 37
255	rs10900862	5	LOC389333	hypothetical LOC389333
256	rs10888510	1	LCE4A	late cornified envelope 4A
257	rs10875561	5	LOC643604	-
258	rs10853953	19	LOC126520	hypothetical protein LOC126520 IL4I1 interleukin 4 induced 1
259	rs1062798	19	NUP62	nucleoporin 62kDa
260	rs1061768	19	ZNF283	zinc finger protein 283
261	rs1060575	1	TCTEX1D1	hypothetical protein FLJ40873
262	rs1060407	3		
263	rs1059307	6	C6orf160	chromosome 6 open reading frame 160 MAP4 microtubule-associated protein 4

264	rs1058930	10	CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8
265	rs1057040	17	USP36	ubiquitin specific protease 36
266	rs1053926	10	PTPLA	protein tyrosine phosphatase-like (proline instead of catalytic arginine), member a
267	rs1051931	6	PLA2G7	phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma)
268	rs10513155	5	DNAH5	dynein, axonemal, heavy polypeptide 5
269	rs10509681	10	CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8
270	rs10508578	10	C10orf112	chromosome 10 open reading frame 112
271	rs10497377	2	MAP1D	methionine aminopeptidase 1D
272	rs1049125	10	PNLIPRP1	pancreatic lipase-related protein 1
273	rs1048943	15	CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1
274	rs1048804	10	NRP1	neuropilin 1 NUDT6 nudix (nucleoside diphosphate linked moiety X)-type motif 6
275	rs1048201	4	FGF2	fibroblast growth factor 2 (basic)
276	rs1046248	14	BDKRB2	bradykinin receptor B2
277	rs1044418	6	REPS1	RALBP1 associated Eps domain containing 1
278	rs1043008	20	RPS21	ribosomal protein S21
279	rs10410943	19	MGC33407	hypothetical protein MGC33407
280	rs10407022	19	AMH	anti-Mullerian hormone
281	rs1039084	6	STXBP5	syntaxin binding protein 5 (tomosyn)
282	rs10277	5	LOC51149	hypothetical LOC51149 SQSTM1 sequestosome 1
283	rs10269582	7	DNAH11	dynein, axonemal, heavy polypeptide 11
284	rs10068763	5	BTNL9	butyrophilin-like 9
285	rs10064102	5	-	-
286	rs10009368	4	LOC132430	similar to Polyadenylate-binding protein 4 (Poly(A)-binding protein 4) (PABP 4) (Inducible poly(A)-binding protein) (iPABP) (Activated-platelet protein-1) (APP-1)

Appendix B

SNPs and Their Corresponding Genes Involved in two-way Interactions

SNP List

Index	SNP1	SNP2	Index	SNP1	SNP2
1	rs10266424	rs11786747	2	rs1361754	rs2071950
3	rs1042391	rs7727919	4	rs1361754	rs3130685
5	rs10446759	rs1050150	6	rs1361754	rs6081901
7	rs10446759	rs7203179	8	rs1462983	rs11734372
9	rs1050998	rs550897	10	rs1501940	rs394558
11	rs1051221	rs12656542	12	rs1501940	rs507879
13	rs1063193	rs6081901	14	rs1558876	rs310586
15	rs10761073	rs6081901	16	rs1673607	rs573122
17	rs10761073	rs6668857	18	rs1760904	rs5771069
19	rs10761073	rs7450342	20	rs17704679	rs7312017
21	rs10818708	rs6456880	22	rs17704679	rs9947055
23	rs10818708	rs6941946	24	rs1799841	rs1361754
25	rs10875561	rs139298	26	rs1799841	rs7206698
27	rs1126483	rs6076347	28	rs1801243	rs2276038
29	rs11604671	rs4642516	30	rs196432	rs2523898
31	rs1233627	rs9370867	32	rs2043112	rs9835332
33	rs12656542	rs139298	34	rs2072633	rs2061690
35	rs12918952	rs2523898	36	rs2072770	rs455863
37	rs13006529	rs7203179	38	rs2072770	rs9370867
39	rs1361754	rs2168749	40	rs2076109	rs11254413
41	rs1361754	rs2523898	42	rs2076109	rs211456
43	rs1361754	rs3213646	44	rs211449	rs2522833
45	rs1361754	rs3748176	46	rs211449	rs6941946
47	rs1361754	rs6076347	48	rs211456	rs4483821
49	rs1361754	rs874556	50	rs2250242	rs1361754
51	rs1378602	rs3748176	52	rs2273442	rs246074
53	rs1558876	rs10772420	54	rs2279281	rs1003582
55	rs1558876	rs2076109	56	rs2279281	rs1995641
57	rs1558876	rs3828323	58	rs2282632	rs7450342
59	rs1558876	rs6672420	60	rs2819419	rs7214723
61	rs1798192	rs211456	62	rs2857697	rs8032931
63	rs1849733	rs1864183	64	rs3130100	rs2071950
65	rs1864183	rs7203179	66	rs3213646	rs6941946
67	rs2071950	rs3178794	68	rs3735781	rs1063193
69	rs2073711	rs3803185	70	rs3735782	rs10875561
71	rs2076109	rs11734372	72	rs3736919	rs10875561
73	rs210120	rs7312017	74	rs3736919	rs7727919

75	rs2228510	rs314300	76	rs3740231	rs6729801
77	rs2248821	rs7203179	78	rs3748176	rs6941946
79	rs2255255	rs4680	80	rs3784038	rs10446759
81	rs2276038	rs11734372	82	rs3803185	rs7637449
83	rs2277680	rs550897	84	rs380421	rs958755
85	rs2279281	rs1059307	86	rs3811102	rs10875561
87	rs2279281	rs550897	88	rs3811188	rs958755
89	rs2523898	rs7203179	90	rs3811188	rs9835332
91	rs2549782	rs1042917	92	rs3812882	rs10769054
93	rs267745	rs2306985	94	rs3812882	rs1995157
95	rs2734335	rs436278	96	rs3813563	rs1361754
97	rs2846412	rs394732	98	rs3815803	rs12918952
99	rs293813	rs4642516	100	rs3817672	rs1264562
101	rs31726	rs1330811	102	rs3818717	rs7745023
103	rs3735781	rs1361754	104	rs3828323	rs2168749
105	rs3735782	rs1361754	106	rs394558	rs1059307
107	rs3735782	rs310589	108	rs394558	rs2289235
109	rs3736919	rs12656542	110	rs394732	rs1867503
111	rs3740015	rs436278	112	rs394732	rs2523898
113	rs3746471	rs1264344	114	rs436278	rs2071950
115	rs3748176	rs1059307	116	rs436278	rs958755
117	rs3784038	rs7203179	118	rs4408545	rs7312017
119	rs3795251	rs2043112	120	rs4638862	rs240780
121	rs3803185	rs958755	122	rs4642516	rs1760903
123	rs3803185	rs9835332	124	rs4642516	rs1760904
125	rs3817672	rs10146482	126	rs4642516	rs2168749
127	rs394558	rs9370867	128	rs4642516	rs6076347
129	rs394732	rs1867504	130	rs4646626	rs10941112
131	rs394732	rs7203179	132	rs4646626	rs1361754
133	rs4646626	rs2227168	134	rs4646626	rs1799908
135	rs4646626	rs2523898	136	rs4646626	rs2168749
137	rs4646626	rs9370867	138	rs4646626	rs6000172
139	rs507879	rs1799908	140	rs550897	rs10875561
141	rs540476	rs6795970	142	rs550897	rs1108842
143	rs550897	rs3213646	144	rs550897	rs12656542
145	rs5764698	rs9370867	146	rs550897	rs3617
147	rs5771069	rs10837814	148	rs550897	rs7727919
149	rs6671527	rs7312017	150	rs588997	rs3817672
151	rs6674281	rs1558876	152	rs593818	rs10875561
153	rs6687605	rs3829765	154	rs593818	rs12656542
155	rs6844637	rs12918952	156	rs593818	rs7727919
157	rs6906021	rs6076347	158	rs6007808	rs436278
159	rs704326	rs9835332	160	rs6165	rs3213646
161	rs7194136	rs3809973	162	rs621313	rs7591849
163	rs7206698	rs10838092	164	rs640573	rs1361754
165	rs7206698	rs1361754	166	rs6671527	rs10446759
167	rs7232237	rs2289235	168	rs6671527	rs1849733
169	rs7245429	rs2230660	170	rs6671527	rs6456880
171	rs7558074	rs2255255	172	rs6671527	rs6941946
173	rs7727919	rs139298	174	rs6671527	rs7203179
175	rs7727919	rs2523898	176	rs6673098	rs10875561
177	rs8050530	rs9820367	178	rs6673098	rs12656542
179	rs814501	rs2522833	180	rs6673098	rs7727919
181	rs874556	rs1633059	182	rs6923492	rs11556563
183	rs913257	rs2168749	184	rs6923492	rs9370867
185	rs9257694	rs211456	186	rs7041	rs12918952
187	rs9370867	rs11734372	188	rs7194136	rs3809970

189	rs9370867	rs9393691	190	rs7206698	rs10875561
191	rs966447	rs1361754	192	rs7206698	rs1849733
193	rs9947055	rs9835332	194	rs7206698	rs1864312
195	rs9989177	rs7745023	196	rs7206698	rs213202
197	rs1003582	rs2294478	198	rs7206698	rs9393691
199	rs1042391	rs12656542	200	rs7245429	rs2230661
201	rs1051221	rs10875561	202	rs7312017	rs1849733
203	rs1055577	rs1361754	204	rs7312017	rs268687
205	rs1063193	rs6870166	206	rs7312017	rs6081901
207	rs1063193	rs7203179	208	rs7727919	rs139302
209	rs1065457	rs12918952	210	rs7727919	rs6766410
211	rs1065457	rs4646626	212	rs7745023	rs1464890
213	rs10761073	rs1361754	214	rs7918487	rs10875561
215	rs10818708	rs5771069	216	rs7918487	rs2248490
217	rs10875561	rs139302	218	rs7975237	rs5771069
219	rs10875561	rs2255255	220	rs7990565	rs6076347
221	rs10875561	rs2523898	222	rs8050530	rs9393691
223	rs11203366	rs6941946	224	rs8050530	rs958755
225	rs11203367	rs6941946	226	rs848209	rs7214723
227	rs11238247	rs1361754	228	rs874556	rs3803530
229	rs1132528	rs12918952	230	rs913257	rs7591849
231	rs1233627	rs3748176	232	rs917361	rs4642516
233	rs12656542	rs139302	234	rs9257694	rs6671527
235	rs12656542	rs6766410	236	rs9370867	rs12102203
237	rs12708923	rs2242046	238	rs9370867	rs1849733
239	rs12918952	rs3817672	240	rs9370867	rs6931332
241	rs12918952	rs3828323	242	rs9370867	rs8500
243	rs12918952	rs957998	244	rs9639393	rs1361754
245	rs13378194	rs2076109	246	rs9678851	rs958755
247	rs1361754	rs1063635	248	rs9836841	rs13378194
249	rs1361754	rs11210490			

Their Corresponding Genes

Index	Gene1	Gene2	Index	Gene1	Gene2
1	MTERF	LOC51059	2	FLJ32569	FLJ34512
3	GMPR	ZNF300	4	FLJ32569	HLA-C
5	NR3C2	SLC22A4	6	FLJ32569	C20orf26
7	NR3C2	KLHDC4	8	OR56B4	KIAA1102
9	CXCL16	FAM55D	10	PRDM9	TATDN2
11	KIAA0143	ZNF300	12	PRDM9	CASP5
13	PRR4	C20orf26	14	KIAA1001	ASB10
15	OR13D1	C20orf26	16	Cep70	MS4A3
17	OR13D1	RHBG	18	TEP1	FLJ41993
19	OR13D1	SLC17A2	20	MGC33887	LALBA
21	OR1N1	ZNF311	22	MGC33887	C18orf26
23	OR1N1	ZNF311	24	CST5	FLJ32569
25	ZNF300	ARP10	26	CST5	LOC400499
27	ENPEP	PSF1	28	ATP7B	P2RX3
29	ANKK1	HLA-DQA1	30	DSCR1L2	C6orf205
31	RFP	MYLIP	32	AVO3	RAP140
33	ZNF300	ARP10	34	BF	PBXIP1
35	WWOX	C6orf205	36	RIBC2	IL17RE
37	CASP10	KLHDC4	38	RIBC2	MYLIP
39	FLJ32569	KLRA1	40	APOBEC3B	DNMT2
41	FLJ32569	C6orf205	42	APOBEC3B	SYNGAP1

43	FLJ32569	MGC16943	44	KIFC1	CACNA2D1
45	FLJ32569	AKNA	46	KIFC1	ZNF311
47	FLJ32569	PSF1	48	SYNGAP1	ADAMTSL3
49	FLJ32569	HKDC1	50	AKNA	FLJ32569
51	FLJ20308	AKNA	52	C14orf37	PCDHAC1
53	KIAA1001	TAS2R48	54	MGC21675	UBD
55	KIAA1001	APOBEC3B	56	MGC21675	TGM4
57	KIAA1001	PLA2R1	58	NOL4	SLC17A2
59	KIAA1001	RUNX3	60	PLD4	CAMKK1
61	GPR109B	SYNGAP1	62	AIF1	GREM1
63	FLJ10260	ATG10	64	ZNF297	FLJ34512
65	ATG10	KLHDC4	66	MGC16943	ZNF311
67	FLJ34512	FLJ21062	68	NRG1	PRR4
69	CILP	ARL11	70	NRG1	ZNF300
71	APOBEC3B	KIAA1102	72	SPG20	ZNF300
73	FLJ43752	LALBA	74	SPG20	ZNF300
75	HERC1	ZAN	76	MYO3A	DNAJC10
77	RUTBC1	KLHDC4	78	AKNA	ZNF311
79	CRNKL1	COMT	80	C14orf45	NR3C2
81	P2RX3	KIAA1102	82	ARL11	LOC285331
83	CXCL16	FAM55D	84	WFDC3	RAP140
85	MGC21675	SYNCRIP	86	FBXO30	ZNF300
87	MGC21675	FAM55D	88	MRPL52	RAP140
89	C6orf205	KLHDC4	90	MRPL52	RAP140
91	LRAP	COL6A2	92	COG6	OR52H1
93	GPA33	MTP	94	COG6	OR52H1
95	C2	KIAA1609	96	C14orf169	FLJ32569
97	SLAC2-B	BCAS1	98	HSPC065	WWOX
99	C3orf6	HLA-DQA1	100	TFRC	RPP21
101	PLEKHG2	PSMB7	102	RAI1	C6orf170
103	NRG1	FLJ32569	104	PLA2R1	KLRA1
105	NRG1	FLJ32569	106	TATDN2	SYNCRIP
107	NRG1	ASB10	108	TATDN2	ITM2C
109	SPG20	ZNF300	110	BCAS1	TOPBP1
111	DHTKD1	KIAA1609	112	BCAS1	C6orf205
113	KIAA1755	DDR1	114	KIAA1609	FLJ34512
115	AKNA	SYNCRIP	116	KIAA1609	RAP140
117	C14orf45	KLHDC4	118	AFG3L1	LALBA
119	ASB17	AVO3	120	C20orf161	ASCC3
121	ARL11	RAP140	122	HLA-DQA1	TEP1
123	ARL11	RAP140	124	HLA-DQA1	TEP1
125	TFRC	NEK9	126	HLA-DQA1	KLRA1
127	TATDN2	MYLIP	128	HLA-DQA1	PSF1
129	BCAS1	TOPBP1	130	ALDH1A2	AMACR
131	BCAS1	KLHDC4	132	ALDH1A2	FLJ32569
133	ALDH1A2	APOL4	134	ALDH1A2	COL11A2
135	ALDH1A2	C6orf205	136	ALDH1A2	KLRA1
137	ALDH1A2	MYLIP	138	ALDH1A2	APOL4
139	CASP5	COL11A2	140	FAM55D	ZNF300
141	RAB27A	SCN10A	142	FAM55D	GNL3
143	FAM55D	MGC16943	144	FAM55D	ZNF300
145	SMC1L2	MYLIP	146	FAM55D	ITIH3
147	FLJ41993	OR51B2	148	FAM55D	ZNF300
149	MOBKL2C	LALBA	150	HLA-DPA1	TFRC
151	TEDDM1	KIAA1001	152	CYP4F12	ZNF300
153	LDLRAP1	C14orf37	154	CYP4F12	ZNF300
155	HDCMA18P	WWOX	156	CYP4F12	ZNF300

157	HLA-DQA1	PSF1	158	LOC388915	KIAA1609
159	CACNA1E	RAP140	160	FSHR	MGC16943
161	PKD1L2	HAK	162	TYR	WDSUB1
163	LOC400499	OR51Q1	164	EPHA10	FLJ32569
165	LOC400499	FLJ32569	166	MOBKL2C	NR3C2
167	NOL4	ITM2C	168	MOBKL2C	FLJ10260
169	FBN3	ZNF239	170	MOBKL2C	ZNF311
171	EIF5B	CRNKL1	172	MOBKL2C	ZNF311
173	ZNF300	ARP10	174	MOBKL2C	KLHDC4
175	ZNF300	C6orf205	176	RPL11	ZNF300
177	LOC400499	CLCN2	178	RPL11	ZNF300
179	SPTBN4	CACNA2D1	180	RPL11	ZNF300
181	HKDC1	FLJ35429	182	GRM1	NYD-SP20
183	SCYL1BP1	KLRA1	184	GRM1	MYLIP
185	OR5U1	SYNGAP1	186	GC	WWOX
187	MYLIP	KIAA1102	188	PKD1L2	HAK
189	MYLIP	HIST1H2BI	190	LOC400499	ZNF300
191	LOC441459	FLJ32569	192	LOC400499	FLJ10260
193	C18orf26	RAP140	194	LOC400499	CRISP3
195	C14orf104	C6orf170	196	LOC400499	VPS52
197	UBD	COL11A2	198	LOC400499	HIST1H2BI
199	GMPR	ZNF300	200	FBN3	ZNF239
201	KIAA0143	ZNF300	202	LALBA	FLJ10260
203	USH1C	FLJ32569	204	LALBA	SERTAD1
205	PRR4	PDZK3	206	LALBA	C20orf26
207	PRR4	KLHDC4	208	ZNF300	ARP10
209	CD1E	WWOX	210	ZNF300	HTR3C
211	CD1E	ALDH1A2	212	C6orf170	ZC3HC1
213	OR13D1	FLJ32569	214	PTF1A	ZNF300
215	OR1N1	FLJ41993	216	PTF1A	WDR4
217	ZNF300	ARP10	218	LOC338773	FLJ41993
219	ZNF300	CRNKL1	220	DCT	PSF1
221	ZNF300	C6orf205	222	LOC400499	HIST1H2BI
223	PADI4	ZNF311	224	LOC400499	RAP140
225	PADI4	ZNF311	226	SPEN	CAMKK1
227	DKFZp564N2472	FLJ32569	228	HKDC1	KIF7
229	YTHDC2	WWOX	230	SCYL1BP1	WDSUB1
231	RFP	AKNA	232	MYH4	HLA-DQA1
233	ZNF300	ARP10	234	OR5U1	MOBKL2C
235	ZNF300	HTR3C	236	MYLIP	DMXL2
237	PKD1L3	SLC28A1	238	MYLIP	FLJ10260
239	WWOX	TFRC	240	MYLIP	HLA-C
241	WWOX	PLA2R1	242	MYLIP	COQ6
243	WWOX	LCP2	244	DNAH11	FLJ32569
245	TUBA2	APOBEC3B	246	SLC4A1AP	RAP140
247	FLJ32569	MICA	248	NAALADL2	TUBA2
249	FLJ32569	C1orf173			

Appendix C

SNPs and Their Corresponding Genes Involved in Three-way Interactions

SNP List

Index	SNP1	SNP2	SNP3
1	rs10446759	rs1050150	rs7203179
2	rs1063193	rs10446759	rs1050150
3	rs10985704	rs664226	rs10839659
4	rs11604671	rs4642516	rs9835332
5	rs1361754	rs2523898	rs7203179
6	rs1361754	rs4534095	rs1851859
7	rs1462983	rs2076109	rs11734372
8	rs1501940	rs1042391	rs12656542
9	rs1501940	rs1558876	rs6672420
10	rs1536036	rs6766334	rs9370867
11	rs1558876	rs6672420	rs7591849
12	rs1558878	rs2279281	rs1995641
13	rs1558878	rs2279281	rs2070317
14	rs2035639	rs3795251	rs2043112
15	rs2035639	rs9989177	rs2043112
16	rs2304053	rs267745	rs2306985
17	rs2341432	rs462769	rs3178794
18	rs293813	rs11604671	rs4642516
19	rs3734804	rs10920362	rs1799908
20	rs3735781	rs1361754	rs3784038
21	rs398607	rs12918952	rs4779794
22	rs507879	rs1361754	rs1799908
23	rs5771069	rs10837814	rs703143
24	rs6007808	rs4695918	rs1501940
25	rs6081901	rs2076109	rs11734372
26	rs7206698	rs2279281	rs4779794
27	rs740182	rs394732	rs4642516
28	rs740182	rs7203179	rs9289122
29	rs745143	rs2168749	rs11556563
30	rs7522061	rs1233627	rs9370867
31	rs8017377	rs9370867	rs6456880
32	rs8017377	rs9370867	rs6941946
33	rs9257694	rs6671527	rs211456
34	rs966447	rs664226	rs1056513
35	rs10277	rs7970885	rs3213646
36	rs10446759	rs2227264	rs1050150
37	rs10446759	rs2234002	rs1050150
38	rs10761073	rs3735781	rs6668857
39	rs10875561	rs1052576	rs2523898

40	rs1143672	rs10446759	rs7203179
41	rs11604671	rs4642516	rs10446759
42	rs11903403	rs213202	rs5771069
43	rs1233627	rs9370867	rs1849733
44	rs12918952	rs4645434	rs1059307
45	rs12918952	rs703143	rs1059307
46	rs13182512	rs1361754	rs1995641
47	rs1345658	rs394558	rs3748176
48	rs1361754	rs17229	rs2523898
49	rs1361754	rs4534095	rs4598671
50	rs1419640	rs3816800	rs11254413
51	rs1462983	rs1051221	rs10875561
52	rs1494961	rs394732	rs1361754
53	rs1498486	rs704326	rs11622969
54	rs1558876	rs12656542	rs9393691
55	rs1558876	rs2076109	rs11734372
56	rs1558876	rs3828323	rs2294478
57	rs1558876	rs6671527	rs7203179
58	rs17563	rs1801257	rs1799908
59	rs17704679	rs31726	rs2303690
60	rs1781935	rs4642516	rs310589
61	rs1799841	rs1990760	rs1361754
62	rs1851724	rs1361754	rs3748176
63	rs1990760	rs1233627	rs9370867
64	rs2070426	rs1368883	rs7203179
65	rs2073711	rs3803185	rs12120084
66	rs2228510	rs3740015	rs436278
67	rs2257212	rs10446759	rs7203179
68	rs2274064	rs3748176	rs6941946
69	rs2274064	rs7745023	rs3748176
70	rs2276038	rs3748176	rs11734372
71	rs2288675	rs6923492	rs6870166
72	rs2295005	rs4805162	rs1042391
73	rs2341432	rs462769	rs2072032
74	rs2341432	rs462769	rs7037849
75	rs2341432	rs917361	rs1064017
76	rs240780	rs380421	rs6870166
77	rs240780	rs380421	rs7637449
78	rs240780	rs6017667	rs6870166
79	rs240780	rs6017667	rs7637449
80	rs2523898	rs1003582	rs11556563
81	rs2734335	rs7643531	rs436278
82	rs293813	rs2279281	rs2171509
83	rs3117021	rs3784038	rs9370867
84	rs3117328	rs591120	rs211449
85	rs31726	rs1330811	rs4148959
86	rs3735782	rs1673607	rs1361754
87	rs3735782	rs3117328	rs1361754
88	rs3735782	rs935172	rs1361754
89	rs3740015	rs436278	rs2073717
90	rs3813563	rs10761073	rs6081901
91	rs3818717	rs8023214	rs7745023
92	rs383362	rs10277	rs11640138
93	rs436278	rs9370867	rs1592624
94	rs4408545	rs6671527	rs7312017
95	rs4642516	rs6076347	rs3803530
96	rs4645434	rs1805152	rs7745023
97	rs4646626	rs6923492	rs6451173
98	rs4969258	rs704326	rs1849733
99	rs507879	rs1361754	rs10751735

100	rs507879	rs1361754	rs3130685
101	rs532841	rs1781935	rs1361754
102	rs5771069	rs10837814	rs11210490
103	rs5771069	rs10875561	rs1050150
104	rs588997	rs2274064	rs8500
105	rs591120	rs3744108	rs1042391
106	rs6007808	rs978009	rs436278
107	rs6557634	rs1799841	rs1361754
108	rs6671527	rs7312017	rs2849233
109	rs6844637	rs9370867	rs12102203
110	rs7026705	rs1361754	rs3748176
111	rs7026705	rs8050530	rs211453
112	rs706761	rs3817672	rs507879
113	rs7106548	rs1995641	rs1108842
114	rs7206698	rs394558	rs1361754
115	rs7232237	rs383362	rs2289235
116	rs7246479	rs1378602	rs3748176
117	rs7312017	rs7990565	rs6076347
118	rs7312017	rs874556	rs10875561
119	rs7312017	rs874556	rs12656542
120	rs7312017	rs874556	rs7727919
121	rs7315731	rs293813	rs4642516
122	rs745142	rs2168749	rs11556563
123	rs7624750	rs1053966	rs42664
124	rs7918487	rs12918952	rs3828323
125	rs8050530	rs2279281	rs4779794
126	rs8050530	rs394558	rs1801058
127	rs814501	rs1361754	rs3748176
128	rs913257	rs1109278	rs3745298
129	rs913257	rs394732	rs2168749
130	rs9370867	rs2859071	rs1849733
131	rs9370867	rs9835332	rs1849733
132	rs9989177	rs9639393	rs10941112

Their Corresponding Genes

Index	**Gene1**	**Gene2**	**Gene3**
1	NR3C2	SLC22A4	KLHDC4
2	PRR4	NR3C2	SLC22A4
3	OR1L8	ARL2	OR2D3
4	ANKK1	HLA-DQA1	RAP140
5	FLJ32569	C6orf205	KLHDC4
6	FLJ32569	PTK2B	OR4C13
7	OR56B4	APOBEC3B	KIAA1102
8	PRDM9	GMPR	ZNF300
9	PRDM9	KIAA1001	RUNX3
10	ITPR3	ZNF167	MYLIP
11	KIAA1001	RUNX3	WDSUB1
12	KIAA1001	MGC21675	TGM4
13	KIAA1001	MGC21675	BPIL3
14	SYT9	ASB17	AVO3
15	SYT9	C14orf104	AVO3
16	FAT2	GPA33	MTP
17	OR52B6	MGC26885	FLJ21062
18	C3orf6	ANKK1	HLA-DQA1
19	C6orf97	LGR6	COL11A2
20	NRG1	FLJ32569	C14orf45
21	GALC	WWOX	KIAA1018
22	CASP5	FLJ32569	COL11A2

23	FLJ41993	OR51B2	OR6K3
24	LOC388915	KIAA1712	PRDM9
25	C20orf26	APOBEC3B	KIAA1102
26	LOC400499	MGC21675	KIAA1018
27	SCAP2	BCAS1	HLA-DQA1
28	SCAP2	KLHDC4	FLJ32859
29	HLC-8	KLRA1	NYD-SP20
30	FCRL3	RFP	MYLIP
31	KIAA0323	MYLIP	ZNF311
32	KIAA0323	MYLIP	ZNF311
33	OR5U1	MOBKL2C	SYNGAP1
34	LOC441459	ARL2	INADL
35	LOC51149	OR10P1	MGC16943
36	NR3C2	TAS2R5	SLC22A4
37	NR3C2	TAS2R4	SLC22A4
38	OR13D1	NRG1	RHBG
39	ZNF300	CASP9	C6orf205
40	SLC15A2	NR3C2	KLHDC4
41	ANKK1	HLA-DQA1	NR3C2
42	CPO	VPS52	FLJ41993
43	RFP	MYLIP	FLJ10260
44	WWOX	SYNE1	SYNCRIP
45	WWOX	OR6K3	SYNCRIP
46	JMY	FLJ32569	TGM4
47	ZNF30	TATDN2	AKNA
48	FLJ32569	PRSS1	C6orf205
49	FLJ32569	PTK2B	OR4C12
50	OR12D3	ATP10A	DNMT2
51	OR56B4	KIAA0143	ZNF300
52	HEL308	BCAS1	FLJ32569
53	OR51I1	CACNA1E	OR6S1
54	KIAA1001	ZNF300	HIST1H2BI
55	KIAA1001	APOBEC3B	KIAA1102
56	KIAA1001	PLA2R1	COL11A2
57	KIAA1001	MOBKL2C	KLHDC4
58	BMP4	GPR56	COL11A2
59	MGC33887	PLEKHG2	ELSPBP1
60	AKR1CL1	HLA-DQA1	ASB10
61	CST5	IFIH1	FLJ32569
62	OR13C5	FLJ32569	AKNA
63	IFIH1	RFP	MYLIP
64	PCNT2	C1orf179	KLHDC4
65	CILP	ARL11	KIF14
66	HERC1	DHTKD1	KIAA1609
67	SLC15A2	NR3C2	KLHDC4
68	NCF2	AKNA	ZNF311
69	NCF2	C6orf170	AKNA
70	P2RX3	AKNA	KIAA1102
71	ANP32C	GRM1	PDZK3
72	HINT3	ZNF565	GMPR
73	OR52B6	MGC26885	TNN
74	OR52B6	MGC26885	UAP1L1
75	OR52B6	MYH4	TCL1B
76	ASCC3	WFDC3	PDZK3
77	ASCC3	WFDC3	LOC285331
78	ASCC3	WFDC13	PDZK3
79	ASCC3	WFDC13	LOC285331
80	C6orf205	UBD	NYD-SP20
81	C2	MITF	KIAA1609
82	C3orf6	MGC21675	CD200R1

83	HLA-DPB1	C14orf45	MYLIP
84	OR5U1	LZTR2	KIFC1
85	PLEKHG2	PSMB7	NDUFV1
86	NRG1	Cep70	FLJ32569
87	NRG1	OR5U1	FLJ32569
88	NRG1	CIB4	FLJ32569
89	DHTKD1	KIAA1609	CCHCR1
90	C14orf169	OR13D1	C20orf26
91	RAI1	RAD51L1	C6orf170
92	WWOX	LOC51149	TNP2
93	KIAA1609	MYLIP	CSMD3
94	AFG3L1	MOBKL2C	LALBA
95	HLA-DQA1	PSF1	KIF7
96	SYNE1	CLCNKA	C6orf170
97	ALDH1A2	GRM1	FLJ25439
98	FLJ44861	CACNA1E	FLJ10260
99	CASP5	FLJ32569	AIM1L
100	CASP5	FLJ32569	HLA-C
101	DLC1	AKR1CL1	FLJ32569
102	FLJ41993	OR51B2	C1orf173
103	FLJ41993	ZNF300	SLC22A4
104	HLA-DPA1	NCF2	COQ6
105	LZTR2	MTMR4	GMPR
106	LOC388915	RPP21	KIAA1609
107	TNFRSF10A	CST5	FLJ32569
108	MOBKL2C	LALBA	MRO
109	HDCMA18P	MYLIP	DMXL2
110	OR13C8	FLJ32569	AKNA
111	OR13C8	LOC400499	KIFC1
112	CPAMD8	TFRC	CASP5
113	USH1C	TGM4	GNL3
114	LOC400499	TATDN2	FLJ32569
115	NOL4	WWOX	ITM2C
116	BRSK1	FLJ20308	AKNA
117	LALBA	DCT	PSF1
118	LALBA	HKDC1	ZNF300
119	LALBA	HKDC1	ZNF300
120	LALBA	HKDC1	ZNF300
121	SFRS2IP	C3orf6	HLA-DQA1
122	HLC-8	KLRA1	NYD-SP20
123	OPA1	TTC3	LOC85865
124	PTF1A	WWOX	PLA2R1
125	LOC400499	MGC21675	KIAA1018
126	LOC400499	TATDN2	GRK4
127	SPTBN4	FLJ32569	AKNA
128	SCYL1BP1	BIRC5	HRC
129	SCYL1BP1	BCAS1	KLRA1
130	MYLIP	HLA-DQA2	FLJ10260
131	MYLIP	RAP140	FLJ10260
132	C14orf104	DNAH11	AMACR

Appendix D

SNPs and Their Corresponding Genes Involved in Four-way Interactions

SNP List

Index	SNP1	SNP2	SNP3	SNP4
1	rs1053966	rs10875561	rs626251	rs7257948
2	rs1061472	rs4785751	rs1673607	rs7312017
3	rs10985704	rs4701997	rs664226	rs10839659
4	rs1330811	rs10818708	rs1042391	rs5771069
5	rs2070426	rs1558876	rs6672420	rs7591849
6	rs229527	rs3735781	rs1361754	rs1867503
7	rs3742302	rs310586	rs10277	rs7745023
8	rs4408545	rs2923006	rs1897820	rs1078211
9	rs550897	rs10772420	rs3213646	rs9908414
10	rs6844637	rs6687605	rs4642516	rs211456
11	rs6897513	rs507879	rs1361754	rs1799908
12	rs6968293	rs11604671	rs4642516	rs958755
13	rs8108738	rs8100718	rs436278	rs6076347
14	rs1015443	rs6699146	rs3735781	rs1361754
15	rs11071986	rs6902723	rs4504745	rs9370867
16	rs1558876	rs2279281	rs2070317	rs550897
17	rs2228059	rs2250242	rs7591849	rs6000172
18	rs4786026	rs6902723	rs462769	rs3178794
19	rs5764698	rs2296545	rs2230660	rs7637449
20	rs6761276	rs314298	rs1558876	rs2294478
21	rs6844637	rs11604671	rs869111	rs4642516
22	rs6920606	rs848209	rs10151658	rs2303690
23	rs7312017	rs874556	rs10875561	rs3800962

Their Corresponding Genes

Index	Gene1	Gene2	Gene3	Gene4
1	TTC3	ZNF300	C1orf87	FBN3
2	ATP7B	FLJ20186	Cep70	LALBA
3	OR1L8	DNAH5	ARL2	OR2D3
4	PSMB7	OR1N1	GMPR	FLJ41993
5	PCNT2	KIAA1001	RUNX3	WDSUB1
6	C1QTNF6	NRG1	FLJ32569	TOPBP1
7	C13orf22	ASB10	LOC51149	C6orf170
8	AFG3L1	HLA-B	ZNF180	C6orf208

9	FAM55D	TAS2R48	MGC16943	USP6
10	HDCMA18P	LDLRAP1	HLA-DQA1	SYNGAP1
11	FLJ23577	CASP5	FLJ32569	COL11A2
12	ABCA13	ANKK1	HLA-DQA1	RAP140
13	PIK3R2	CEACAM20	KIAA1609	PSF1
14	TAS2R13	RABGAP1L	NRG1	FLJ32569
15	CALML4	HLA-DQB2	ZNF79	MYLIP
16	KIAA1001	MGC21675	BPIL3	FAM55D
17	IL15RA	AKNA	WDSUB1	APOL4
18	SB153	HLA-DQB2	MGC26885	FLJ21062
19	SMC1L2	C10orf59	ZNF239	LOC285331
20	IL1F10	ZAN	KIAA1001	COL11A2
21	HDCMA18P	ANKK1	OR2G2	HLA-DQA1
22	HLA-DOA	SPEN	SYNE2	ELSPBP1
23	LALBA	HKDC1	ZNF300	PRPS1L1

Bibliography

Aamodt, A. & Plaza, E. (1994), 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AI communications* **7**(1), 39–59.

Altmüller, J., Palmer, L., Fischer, G., Scherb, H. & Wjst, M. (2001), 'Genome-wide scans of complex human diseases: true linkage is hard to find', *The American Journal of Human Genetics* **69**(5), 936–950.

Altshuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., Donnelly, P. et al. (2005), 'A haplotype map of the human genome', *Nature* **437**(7063), 1299–1320.

Altshuler, D., Hirschhorn, J., Klannemark, M., Lindgren, C., Vohl, M., Nemesh, J., Lane, C., Schaffner, S., Bolk, S., Brewer, C. et al. (2000), 'The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes', *Nature Genetics* **26**(1), 76–80.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J. et al. (2000), 'Gene ontology: tool for the unification of biology', *Nature Genetics* **25**(1), 25.

Bahlo, M., Stankovich, J., Speed, T., Rubio, J., Burfoot, R. & Foote, S. (2006), 'Detecting genome wide haplotype sharing using SNP or microsatellite haplotype data', *Human genetics* **119**(1), 38–50.

Bailey, J. & Eichler, E. (2006), 'Primate segmental duplications: crucibles of evolution, diversity and disease', *Nature Reviews Genetics* **7**(7), 552–564.

Balding, D. (2006), 'A tutorial on statistical methods for population association studies', *Nature Reviews Genetics* **7**(10), 781–792.

Barker, D., Hansen, M., Faruqi, A., Giannola, D., Irsula, O., Lasken, R., Latterich, M., Makarov, V., Oliphant, A., Pinter, J. et al. (2004), 'Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel', *Genome Research* **14**(5), 901.

Beckmann, L., Fischer, C., Obreiter, M., Rabes, M. & Chang-Claude, J. (2005), 'Haplotype-sharing analysis using Mantel statistics for combined genetic effects', *BMC Genetics* **6**(Suppl 1), S70.

Belkin, M. & Niyogi, P. (2002), 'Laplacian eigenmaps and spectral techniques for embedding and clustering', *Advances in Neural Information Processing Systems* **14**, 585–591.

Belkin, M. & Niyogi, P. (2004), 'Semi-Supervised Learning on Riemannian Manifolds', *Machine Learning* **56**(1), 209–239.

Bengio, Y., Paiement, J. & Vincent, P. (2004), 'Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering', *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* .

Bird, A. (2007), 'Perceptions of epigenetics', *Nature* **447**(7143), 396–398.

Botstein, D. & Risch, N. (2003), 'Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease', *Nature Genetics* **33**, 228–237.

Braga-Neto, U., Hashimoto, R., Dougherty, E., Nguyen, D. & Carroll, R. (2004), 'Is cross-validation better than resubstitution for ranking genes?', *Bioinformatics* **20**(2), 253.

Brassat, D., Motsinger, A., Caillier, S., Erlich, H., Walker, K., Steiner, L., Cree, B., Barcellos, L., Pericak-Vance, M., Schmidt, S. et al. (2006), 'Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans', *Genes and Immunity* **7**(4), 310–315.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Brookes, A. (1999), 'The essence of SNPs', *Gene* **234**(2), 177–186.

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

Buchala, S., Davey, N., Frank, R. & Gale, T. (2004), 'Dimensionality reduction of face images for gender classification', *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference* **1**.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T. & Van Eerdewegh, P. (2005), 'Identifying SNPs predictive of phenotype using random forests', *Genetic Epidemiology* **28**(2), 171–182.

Burges, C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Buyse, I., Fang, P., Hoon, K., Amir, R., Zoghbi, H. & Roa, B. (2000), 'Diagnostic testing for Rett syndrome by DHPLC and direct sequencing analysis of the MECP2 gene: identification of several novel mutations and polymorphisms', *The American Journal of Human Genetics* **67**(6), 1428–1436.

Cardon, L. & Bell, J. (2001), 'Association study designs for complex diseases', *Nature Reviews Genetics* **2**(2), 91–99.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C., Lim, E., Kalyanaraman, N., Nemesh, J. et al. (1999), 'Characterization of single-nucleotide polymorphisms in coding regions of human genes', *Nature Genetics* **22**(3), 231–238.

Carlson, C., Eberle, M., Kruglyak, L. & Nickerson, D. (2004), 'Mapping complex disease loci in whole-genome association studies', *Nature* **429**(6990), 446–452.

Carlson, C., Eberle, M., Rieder, M., Smith, J., Kruglyak, L. & Nickerson, D. (2003), 'Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans', *Nature Genetics* **33**(4), 518–521.

Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L. & Nickerson, D. (2004), 'Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium', *The American Journal of Human Genetics* **74**(1), 106–120.

Cartegni, L., Chew, S. & Krainer, A. (2002), 'Listening to silence and understanding nonsense: exonic mutations that affect splicing', *Nature Reviews Genetics* **3**(4), 285–298.

Cavalli-Sforza, L. (1974), 'The genetics of human populations.', *Sci Am* **231**(3), 80–9.

Chai, H. & Domeniconi, C. (2004), An evaluation of gene selection methods for multi-class microarray data classification, *in* 'Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics', pp. 3–10.

Chakravarti, A. (1998), 'It's raining SNPs, hallelujah?', *Nature Genetics* **19**(3), 216.

Chakravarti, A. (1999), 'Population genetics-making sense out of sequence', *Nature Genetics* **21**(Suppl 1), 56–60.

Chang, C. & Lin, C. (2011), 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 27.

Chen, S., Sun, J., Dimitrov, L., Turner, A., Adams, T., Meyers, D., Chang, B., Zheng, S., Grönberg, H., Xu, J. et al. (2008), 'A support vector machine approach for detecting gene-gene interaction', *Genetic Epidemiology* **32**(2), 152–167.

Chen, X., Liu, C., Zhang, M. & Zhang, H. (2007), 'A forest-based approach to identifying gene and gene–gene interactions', *Proceedings of the National Academy of Sciences* **104**(49), 19199.

Cheng, R., Ma, J., Elston, R. & Li, M. (2005), 'Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs', *Annals of Human Genetics* **69**(1), 102–112.

Cho, Y., Ritchie, M., Moore, J., Park, J., Lee, K., Shin, H., Lee, H. & Park, K. (2004), 'Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus', *Diabetologia* **47**(3), 549–554.

Chung, Y., Lee, S., Elston, R. & Park, T. (2007), 'Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions', *Bioinformatics* **23**(1), 71.

Clark, A. (2004), 'The role of haplotypes in candidate gene studies', *Genetic Epidemiology* **27**(4), 321–333.

Clark, A., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengård, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. et al. (1998), 'Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase', *The American Journal of Human Genetics* **63**(2), 595–612.

Clark, T., De Iorio, M. & Griffiths, R. (2007), 'Bayesian logistic regression using a perfect phylogeny', *Biostatistics* **8**(1), 32.

Clayton, D., Chapman, J. & Cooper, J. (2004), 'Use of unphased multilocus genotype data in indirect association studies', *Genetic Epidemiology* **27**(4), 415–428.

Coates, M. & Tracey, E. (2001), 'Cancer in New South Wales. Incidence and mortality 1998 and Incidence for Selected Cancers 1999', *NSW Central Cancer Registry, Cancer Research and Registers Division, NSW Cancer Council* pp. 42–43.

Cohen, J., Kiss, R., Pertsemlidis, A., Marcel, Y., McPherson, R. & Hobbs, H. (2004), 'Multiple rare alleles contribute to low plasma levels of HDL cholesterol', *Science* **305**(5685), 869.

Collins, F., Brooks, L. & Chakravarti, A. (1998), 'A DNA polymorphism discovery resource for research on human genetic variation', *Genome Research* **8**(12), 1229.

Collins, F., Guyer, M. & Chakravarti, A. (1997), 'Variations on a theme: cataloging human DNA sequence variation', *Science* **278**(5343), 1580.

Conrad, D., Andrews, T., Carter, N., Hurles, M. & Pritchard, J. (2005), 'A high-resolution survey of deletion polymorphism in the human genome', *Nature Genetics* **38**(1), 75–81.

Cook, N., Zee, R. & Ridker, P. (2004), 'Tree and spline based association analysis of gene-gene interaction models for ischemic stroke', *Statistics in Medicine* **23**(9), 1439–1453.

Cordell, H. (2009), 'Detecting gene–gene interactions that underlie human diseases', *Nature Reviews Genetics* **10**(6), 392–404.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Costa, J. & Hero, A. (2005), 'Classification Constrained Dimensionality Reduction', *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing* **5**.

Couch, F., Weber, B. et al. (1996), 'Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene', *Human Mutation* **8**(1), 8–18.

Cox, T. & Cox, M. (2001), *Multidimensional Scaling*, CRC Press.

Craddock, N., Hurles, M., Cardin, N., Pearson, R., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D., Giannoulatou, E. et al. (2010), 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', *Nature* **464**(7289), 713–720.

Crawford, D., Bhangale, T., Li, N., Hellenthal, G., Rieder, M., Nickerson, D. & Stephens, M. (2004), 'Evidence for substantial fine-scale variation in recombination rates across the human genome', *Nature Genetics* **36**(7), 700–706.

Crawford, D. & Nickerson, D. (2005), 'Definition and Clinical importance of Haplotypes', *Annual Review of Medicine* **56**(1), 303–320.

Cristóbal, I., Blanco, F., Garcia-Orti, L., Marcotegui, N., Vicente, C., Rifon, J., Novo, F., Bandres, E., Calasanz, M., Bernabeu, C. et al. (2010), 'SETBP1 overexpression is a novel leukemogenic mechanism that predicts adverse outcome in elderly patients with acute myeloid leukemia', *Blood* **115**(3), 615–625.

Culverhouse, R., Klein, T. & Shannon, W. (2004), 'Detecting epistatic interactions contributing to quantitative traits', *Genetic Epidemiology* **27**(2), 141–152.

Da Wei Huang, B., Lempicki, R. et al. (2008), 'Systematic and integrative analysis of large gene lists using david bioinformatics resources', *Nature protocols* **4**(1), 44–57.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J., Sempoux, C., Machiels, J., Haustermans, K. & De Moor, B. (2009), 'A kernel-based integration of genome-wide data for clinical decision support', *Genome medicine* **1**(4), 39.

Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. (2001), 'High-resolution haplotype structure in the human genome', *Nature Genetics* **29**(2), 229–232.

Dawson, E., Abecasis, G., Bumpstead, S., Chen, Y., Hunt, S., Beare, D., Pabial, J., Dibling, T., Tinsley, E., Kirby, S. et al. (2002), 'A first-generation linkage disequilibrium map of human chromosome 22', *Nature* **418**(6897), 544–548.

Demartines, P. & Herault, J. (1997), 'Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets', *Neural Networks, IEEE Transactions on* **8**(1), 148–154.

Díaz-Uriarte, R. & de Andrés, A. (2006), 'Gene selection and classification of microarray data using random forest', *BMC Bioinformatics* **7**(1), 3.

Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X. & Shen, Y. (2005), 'The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection', *Molecular biology and evolution* **22**(1), 148.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W. & Li, Y. (2007), 'Exploration of gene–gene interaction effects using entropy-based methods', *European Journal of Human Genetics* **16**(2), 229–235.

Donnelly, J. (2004), 'Pharmacogenetics in cancer chemotherapy: balancing toxicity and response', *Therapeutic drug monitoring* **26**(2), 231.

Douglas, J., Boehnke, M., Gillanders, E., Trent, J. & Gruber, S. (2001), 'Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies', *Nature Genetics* **28**(4), 361–364.

Dudbridge, F., Gusnanto, A. & Koeleman, B. (2006), 'Detecting multiple associations in genome-wide studies', *Human Genomics* **2**(5), 310–317.

Durrant, C., Zondervan, K., Cardon, L., Hunt, S., Deloukas, P. & Morris, A. (2004), 'Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes', *The American Journal of Human Genetics* **75**(1), 35–43.

Fallin, D. & Schork, N. (2000), 'Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data', *The American Journal of Human Genetics* **67**(4), 947–959.

Fan, R., Chen, P. & Lin, C. (2005), 'Working set selection using second order information for training support vector machines', *The Journal of Machine Learning Research* **6**, 1889–1918.

Feero, W., Guttmacher, A. & Manolio, T. (2010), 'Genome-wide association studies and assessment of the risk of disease', *New England Journal of Medicine* **363**(2), 166–176.

Frazer, K., Murray, S., Schork, N. & Topol, E. (2009), 'Human genetic variation and its contribution to complex traits', *Nature Reviews Genetics* **10**(4), 241–251.

Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. (2002), 'Trend tests for case-control studies of genetic markers: Power, sample size and robustness', *Human Heredity* **53**(3), 146–152.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer Series in Statistics.

Fu, Y. & Huang, T. (2010), 'Manifold and Subspace Learning for Pattern Recognition', *Pattern Recognition and Machine Vision* **6**, 215.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. et al. (2002), 'The structure of haplotype blocks in the human genome', *Science* **296**(5576), 2225.

Gibbs, R., Belmont, J., Hardenbol, P., Willis, T., Yu, F., Yang, H., Ch'ang, L., Huang, W., Liu, B., Shen, Y. et al. (2003), 'The international HapMap project', *Nature* **426**(6968), 789–796.

Glaser, B., Nikolov, I., Chubb, D., Hamshere, M., Segurado, R., Moskvina, V. & Holmans, P. (2007), Analyses of single marker and pairwise effects of candidate loci for rheumatoid arthritis using logistic regression and random forests, *in* 'BMC proceedings', Vol. 1, BioMed Central Ltd, p. S54.

Goldstein, D., Ahmadi, K., Weale, M. & Wood, N. (2003), 'Genome scans and candidate gene approaches in the study of common diseases and variable drug responses', *TRENDS in Genetics* **19**(11), 615–622.

Gower, J. (1966), 'Some distance properties of latent root and vector methods used in multivariate analysis', *Biometrika* **53**(3-4), 325–338.

Greaves, M. & Wiemels, J. (2003), 'Origins of chromosome translocations in childhood leukaemia', *Nature Reviews Cancer* **3**(9), 639–649.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1), 389–422.

Hahn, L., Ritchie, M. & Moore, J. (2003), 'Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions', *Bioinformatics* **19**(3), 376.

Halushka, M., Fan, J., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. & Chakravarti, A. (1999), 'Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis', *Nature Genetics* **22**(3), 239–247.

He, X., Yan, S., Hu, Y., Niyogi, P. & Zhang, H. (2005), 'Face Recognition Using Laplacianfaces', *IEEE Transactions on pattern analysis and machine intelligence* pp. 328–340.

Hinton, G. & Roweis, S. (2003), 'Stochastic neighbor embedding', *Advances in Neural Information Processing Systems* **15**, 833–840.

Hirschhorn, J. & Altshuler, D. (2002), 'Once and Again–Issues Surrounding Replication in Genetic Association Studies', *Journal of Clinical Endocrinology and Metabolism* **87**(10), 4438.

Hirschhorn, J. & Daly, M. (2005), 'Genome-wide association studies for common diseases and complex traits', *Nature Reviews Genetics* **6**(2), 95–108.

Hirschhorn, J., Lohmueller, K., Byrne, E. & Hirschhorn, K. (2002), 'A comprehensive review of genetic association studies', *Genetics in Medicine* **4**(2), 45.

Hoh, J. & Ott, J. (2003), 'Mathematical multi-locus approaches to localizing complex human trait genes', *Nature Reviews Genetics* **4**(9), 701–709.

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K. & Ott, J. (2001), 'Selecting SNPs in two-stage analysis of disease association data: a model-free approach', *Annals of human genetics* **64**(05), 413–417.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24**(6), 417–441.

Huang, L., Hsu, S., Lin, E. et al. (2009), 'A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data', *Journal of translational medicine* **7**(1), 81.

Hugot, J., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J., Belaiche, J., Almer, S., Tysk, C., O'Morain, C., Gassull, M. et al. (2001), 'Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease', *Nature* **411**(6837), 599–603.

Irving, J., Bloodworth, L., Bown, N., Case, M., Hogarth, L. & Hall, A. (2005), 'Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis', *Cancer research* **65**(8), 3053.

Jiang, H., Deng, Y., Chen, H., Tao, L., Sha, Q., Chen, J., Tsai, C. & Zhang, S. (2004), 'Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes', *BMC Bioinformatics* **5**(1), 81.

Johnson, G., Esposito, L., Barratt, B., Smith, A., Heward, J., Di Genova, G., Ueda, H., Cordell, H., Eaves, I., Dudbridge, F. et al. (2001), 'Haplotype tagging for the identification of common disease genes', *Nature Genetics* **29**(2), 233–237.

Jorgenson, E. & Witte, J. (2006*a*), 'A gene-centric approach to genome-wide association studies', *Nature Reviews Genetics* **7**(11), 885–891.

Jorgenson, E. & Witte, J. (2006*b*), 'Coverage and power in genome-wide association studies', *The American Journal of Human Genetics* **78**(5), 884–888.

Judson, R., Salisbury, B., Schneider, J., Windemuth, A. & Stephens, J. (2002), 'How many snps does a genome-wide haplotype map require?', *Pharmacogenomics* **3**(3), 379–391.

Kang, G., Yue, W., Zhang, J., Huebner, M., Zhang, H., Ruan, Y., Lu, T., Ling, Y., Zuo, Y. & Zhang, D. (2008), 'Two-stage designs to identify the effects of SNP combinations on complex diseases', *Journal of Human Genetics* **53**(8), 739–746.

Kasabov, N. (2009), 'Soft computing methods for global, local and personalized modeling and applications in bioinformatics', *Soft computing based modeling in intelligent systems* pp. 1–18.

Kasabov, N. & Hu, Y. (2010), 'Integrated optimisation method for personalised modelling and case studies for medical decision support', *International Journal of Functional Informatics and Personalised Medicine* **3**(3), 236–256.

Kasabov, N., Song, Q., Benuskova, L., Gottgtroy, P., Jain, V., Verma, A., Havukkala, I., Rush, E., Pears, R., Tjahjana, A. et al. (2008), 'Integrating local and personalised modelling with global ontology knowledge bases for biomedical and bioinformatics decision support', *Computational Intelligence in Biomedicine and Bioinformatics* pp. 93–116.

Kaski, S., Nikkila, J., Oja, M., Venna, J., Toronen, P. & Castren, E. (2003), 'Trustworthiness and metrics in visualizing similarity of gene expression', *BMC Bioinformatics* **4**(1), 48.

Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A., Bentley, D. et al. (2004), 'The impact of SNP density on fine-scale patterns of linkage disequilibrium', *Human Molecular Genetics* **13**(6), 577.

Kelly, E., Sievers, F. & McManus, R. (2004), 'Haplotype frequency estimation error analysis in the presence of missing genotype data', *BMC Bioinformatics* **5**(1), 188.

King, R., Rotter, J. & Motulsky, A. (2002), *The genetic basis of common diseases*, Vol. 44, Oxford University Press, USA.

Kokiopoulou, E. & Saad, Y. (2005), 'Orthogonal Neighborhood Preserving Projections', *IEEE Int. Conf. on Data Mining* pp. 1–8.

Kokiopoulou, E. & Saad, Y. (2007), 'Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique', *IEEE Transactions on pattern analysis and machine intelligence* pp. 2143–2156.

Kruglyak, L. (1997), 'The use of a genetic map of biallelic markers in linkage studies', *Nature Genetics* **17**(1), 21–24.

Kruglyak, L. (1999), 'Prospects for whole-genome linkage disequilibrium mapping of common disease genes', *Nature Genetics* **22**, 139–144.

Kruglyak, L. (2008), 'The road to genome-wide association studies', *Nature Reviews Genetics* **9**(4), 314–318.

Kruskal, J. (1964), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* **29**(1), 1–27.

Lalouel, J. & Rohrwasser, A. (2002), 'Power and replication in case-control studies', *American journal of hypertension* **15**(2), 201–205.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature* **409**(6822), 860–921.

Lee, J., Lee, J., Park, M. & Song, S. (2005), 'An extensive comparison of recent classification tools applied to microarray data', *Computational Statistics & Data Analysis* **48**(4), 869–885.

Lee, J., Lendasse, A. & Verleysen, M. (2004), 'Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis', *Neurocomputing* **57**, 49–76.

Lee, P. (2006), Computational Haplotype Analysis: An overview of computational methods in genetic variation study, PhD thesis, Queen's University.

Leykin, I., Hao, K., Cheng, J., Meyer, N., Pollak, M., Smith, R., Wong, W., Rosenow, C. & Li, C. (2005), 'Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data', *BMC Genetics* **6**, 7.

Li, J. (2008), 'Prioritize and select SNPs for association studies with multi-stage designs', *Journal of Computational Biology* **15**(3), 241–257.

Liang, Y., Vasilakos, A. & Kelemen, A. (2007), 'Computational intelligence for genetic association study in complex disease: Review of Theory and Applications', *Advances in Computational Sciences and Technology* **1**(1), 77–90.

Lin, D. & Zeng, D. (2006), 'Likelihood-based inference on haplotype effects in genetic association studies', *Journal of the American Statistical Asso.* **101**(473), 89–104.

Little, R. & Rubin, D. (1987), *Statistical analysis with missing data*, Wiley New York.

Liu, C., Ackerman, H. & Carulli, J. (2011), 'A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility', *Human genetics* **129**(5), 473–485.

Liu, Q., Yang, J., Chen, Z., Yang, M., Sung, A. & Huang, X. (2008), 'Supervised learning-based tagSNP selection for genome-wide disease classifications', *BMC Genomics* **9**(Suppl 1), S6.

Lohmueller, K., Pearce, C., Pike, M., Lander, E. & Hirschhorn, J. (2003), 'Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease', *Nature Genetics* **33**(2), 177–182.

Ionita, L. & Med, M. (2006), 'Optimal two-stage strategy for detecting interacting genes in complex diseases', *BMC Genetics* **7**, 39.

Lowe, C., Cooper, J., Chapman, J., Barratt, B., Twells, R., Green, E., Savage, D., Guja, C., Ionescu-Tîrgovişte, C., Tuomilehto-Wolf, E. et al. (2004), 'Cost-effective analysis of candidate genes using htSNPs: a staged approach.', *Genes and Immunity* **5**(4), 301.

Lunetta, K., Hayward, L., Segal, J. & Van Eerdewegh, P. (2004), 'Screening large-scale association study data: exploiting interactions using random forests', *BMC Genetics* **5**(1), 32.

Manolio, T., Brooks, L. & Collins, F. (2008), 'A HapMap harvest of insights into the genetics of common disease', *The Journal of clinical investigation* **118**(5), 1590.

Mao, W. & Kelly, S. (2007), 'An optimum random forest model for prediction of genetic susceptibility to complex diseases', *Advances in Knowledge Discovery and Data Mining* pp. 193–204.

Mao, W. & Mao, J. (2008), The application of random forest in genetic case-control studies, *in* 'International Conference on Information Technology and Applications in Biomedicine, 2008', IEEE, pp. 370–373.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z., Munro, M. & Abecasis, G. (2006), 'A comparison of phasing algorithms for trios and unrelated individuals', *The American Journal of Human Genetics* **78**(3), 437–450.

Marchini, J., Donnelly, P. & Cardon, L. (2005), 'Genome-wide strategies for detecting multiple loci that influence complex diseases', *Nature genetics* **37**(4), 413–417.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nature Genetics* **39**(7), 906–913.

Martin, E., Lai, E., Gilbert, J., Rogala, A., Afshari, A., Riley, J., Finch, K., Stevens, J., Livak, K., Slotterbeck, B. et al. (2000), 'SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease', *The American Journal of Human Genetics* **67**(2), 383–394.

Mas, A., Blanco, E., Monux, G., Urcelay, E., Serrano, F., De La Concha, E. & Martinez, A. (2005), 'DRB1-TNF-α-TNF-β haplotype is strongly associated with severe aortoiliac occlusive disease, a clinical form of atherosclerosis', *Human immunology* **66**(10), 1062–1067.

Maurer, H., Melchinger, A. & Frisch, M. (2007), 'An incomplete enumeration algorithm for an exact test of Hardy-Weinberg proportions with multiple alleles', *Theoretical and Applied Genetics* **115**(3), 393–398.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J. & Hirschhorn, J. (2008), 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nature Reviews Genetics* **9**(5), 356–369.

McKinney, B., Reif, D., Ritchie, M. & Moore, J. (2006), 'Machine learning for detecting gene-gene interactions: a review', *Applied Bioinformatics* **5**(2), 77–88.

Memisevic, R. & Hinton, G. (2005), 'Improving dimensionality reduction with spectral gradient descent', *Neural Networks* **18**(5-6), 702–710.

Meng, Y., Yang, Q., Cuenco, K., Cupples, L., DeStefano, A. & Lunetta, K. (2007), Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and bayesian networks, *in* 'BMC Proceedings', Vol. 1, p. S56.

Meng, Y., Yu, Y., Cupples, L., Farrer, L. & Lunetta, K. (2009), 'Performance of random forest when SNPs are in linkage disequilibrium', *BMC Bioinformatics* **10**(1), 78.

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

Molinaro, A., Carriero, N., Bjornson, R., Hartge, P., Rothman, N. & Chatterjee, N. (2010), 'Power of Data Mining Methods to Detect Genetic Associations and Interactions', *Human Heredity* **72**(2), 85–97.

Molitor, J., Marjoram, P. & Thomas, D. (2003), 'Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques', *The American Journal of Human Genetics* **73**(6), 1368–1384.

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

Moore, J., Gilbert, J., Tsai, C., Chiang, F., Holden, T., Barney, N. & White, B. (2006), 'A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility', *Journal of theoretical biology* **241**(2), 252–261.

Morris, A. (2005), 'Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes', *Genetic Epidemiology* **29**(2), 91–107.

Morris, R. & Kaplan, N. (2002), 'On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles', *Genetic Epid.* **23**(3), 221–233.

Motsinger, A., Lee, S., Mellick, G. & Ritchie, M. (2006), '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.

Motsinger, A., Ritchie, M. & Reif, D. (2007), 'Novel methods for detecting epistasis in pharmacogenomics studies', *Pharmacogenomics* **8**(9), 1229–1241.

Motsinger-Reif, A., Dudek, S., Hahn, L. & Ritchie, M. (2008), 'Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology', *Genetic Epidemiology* **32**(4), 325.

Mullighan, C., Goorha, S., Radtke, I., Miller, C., Coustan-Smith, E., Dalton, J., Girtman, K., Mathew, S., Ma, J., Pounds, S. et al. (2007), 'Genome-wide analysis of genetic alterations in ALL leukaemia', *Nature* **446**(7137), 758–764.

Musani, S., Shriner, D., Liu, N., Feng, R., Coffey, C., Yi, N., Tiwari, H. & Allison, D. (2007), 'Detection of gene x gene interactions in genome-wide association studies of human population data', *Human Heredity* **63**(2), 67–84.

NCI Dictionary of Cancer Terms (2011), *National Cancer Institute: Dictionary of Cancer Terms*. Online Resource `http://www.cancer.gov/dictionary`.

Nelson, M., Kardia, S., Ferrell, R. & Sing, C. (2001), 'A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation', *Genome Research* **11**(3), 458.

Ng, A., Jordan, M. & Weiss, Y. (2002), 'On spectral clustering: Analysis and an algorithm', *Advances in neural information processing systems* **2**, 849–856.

Nguyen, G. & Worring, M. (2004), 'Optimizing similarity based visualization in content based image retrieval', *ICME'04 IEEE International Conference* **2**, 759–762.

Nguyen, H., Vu, T., Ohn, S., Park, Y., Han, M. & Kim, C. (2006), 'Feature elimination approach based on random forest for cancer diagnosis', *MICAI 2006: Advances in Artificial Intelligence* pp. 532–542.

Nielsen, D., Ehm, M. & Weir, B. (1998), 'Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus', *The American Journal of Human Genetics* **63**(5), 1531–1540.

Niknian, M. (1995), 'Permutation tests: A practical guide to resampling methods for testing hypotheses', *Technometrics* **37**(3), 341–342.

Nistico, L., Buzzetti, R., Pritchard, L., Van der Auwera, B., Giovannini, C., Bosi, E., Larrad, M., Rios, M., Chow, C., Cockram, C. et al. (1996), 'The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry', *Human Molecular Genetics* **5**(7), 1075.

Nolte, I., McCaffery, J. & Snieder, H. (2010), 'Candidate gene and genome-wide association studies in behavioral medicine', *Handbook of Behavioral Medicine: Methods and Applications* **3**, 423.

Nonyane, B. & Foulkes, A. (2008), 'Application of two machine learning algorithms to genetic association studies in the presence of covariates', *BMC Genetics* **9**(1), 71.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999), 'KEGG: Kyoto encyclopedia of genes and genomes', *Nucleic acids research* **27**(1), 29–34.

Oh, S., Lee, J., Kwon, M., Kim, K. & Park, T. (2011), Efficient and Fast Analysis for Detecting High Order Gene-by-Gene Interactions in a Genome-Wide Association Study, *in* 'IEEE International Conference on Bioinformatics and Biomedicine', pp. 83–88.

Osuna, E., Freund, R. & Girosi, F. (1997), Training support vector machines: an application to face detection, *in* 'Computer Vision and Pattern Recognition', Published by the IEEE Comp. Soc., p. 130.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. et al. (2002), 'Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction', *Nature Genetics* **32**(4), 650–654.

Pääbo, S. (2003), 'The mosaic that is our genome', *Nature* **421**(6921), 409–412.

Pagano, M. & Gauvreau, K. (n.d.), 'Principles of biostatistics. 2000', *Australia: Duxbury Thompson Learning* .

Pagano, M., Gauvreau, K. & Pagano, M. (2000), *Principles of biostatistics*, Duxbury Press Belmont, CA.

Palmer, L. & Cardon, L. (2005), 'Shaking the tree: mapping complex disease genes with linkage disequilibrium', *The Lancet* **366**(9492), 1223–1234.

Panagopoulos, I., Kerndrup, G., Carlsen, N., Strömbeck, B., Isaksson, M. & Johansson, B. (2007), 'Fusion of NUP98 and the set binding protein 1 (SETBP1) gene in a paediatric acute T cell lymphoblastic leukaemia with t (11; 18)(p15; q12)', *British Journal of Haematology* **136**(2), 294–296.

Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D. et al. (2001), 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21', *Science* **294**(5547), 1719.

Pattin, K., White, B., Barney, N., Gui, J., Nelson, H., Kelsey, K., Andrew, A., Karagas, M. & Moore, J. (2008), 'A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction.', *Genetic Epidemiology* **33**(1), 87.

Pennacchio, L., Olivier, M., Hubacek, J., Cohen, J., Cox, D., Fruchart, J., Krauss, R. & Rubin, E. (2001), 'An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing', *Science* **294**(5540), 169.

Pirooznia, M. & Deng, Y. (2006), 'SVM Classifier: a comprehensive Java interface for support vector machine classification of microarray data', *BMC Bioinfo.* **7**, S25.

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

Qin, Z., Gopalakrishnan, S. & Abecasis, G. (2006), 'An efficient comprehensive search algorithm for tagSNP selection using LD criteria', *Bioinformatics* **22**(2).

Rampersaud, E., Damcott, C. M., Fu, M., Shen, H., McArdle, P., Shi, X., Shelton, J., Yin, J., Chang, Y. P. & Ott, S. H. (2007), 'Genome-wide association study of 14,000 cases of 7 common diseases and 3,000 shared controls', *Nature* **447**(7145), 661–78.

Reich, D., Gabriel, S. & Altshuler, D. (2003), 'Quality and completeness of SNP databases', *Nature Genetics* **33**(4), 457–458.

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

Reif, A., Herterich, S., Strobel, A., Ehlis, A., Saur, D., Jacob, C., Wienker, T., Töpner, T., Fritzen, S., Walter, U. et al. (2006), 'A neuronal nitric oxide synthase (NOS-I) haplotype associated with schizophrenia modifies prefrontal cortex function', *Molecular Psychiatry* **11**(3), 286–300.

Rifkin, R., Mukherjee, S., Tamayo, P., Ramaswamy, S., Yeang, C., Angelo, M., Reich, M., Poggio, T., Lander, E., Golub, T. et al. (2003), 'An analytical method for multiclass molecular cancer classification', *Siam Review* **45**(4), 706–723.

Rioux, J., Xavier, R., Taylor, K., Silverberg, M., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M., Datta, L. et al. (2007), 'Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis', *Nature Genetics* **39**(5), 596 – 604.

Risch, N. (2000), 'Searching for genetic determinants in the new millennium', *Nature* **405**(6788), 847–856.

Risch, N., Merikangas, K. et al. (1996), 'The future of genetic studies of complex human diseases', *Science* **273**(5281), 1516–1516.

Ritchie, M., Hahn, L. & Moore, J. (2003), 'Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity', *Genetic Epidemiology* **24**(2), 150–157.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F. & Moore, J. (2001), '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.

Robnik-Šikonja, M. & Kononenko, I. (2003), 'Theoretical and empirical analysis of ReliefF and RReliefF', *Machine Learning* **53**(1), 23–69.

Rosenberg, N., Huang, L., Jewett, E., Szpiech, Z., Jankovic, I. & Boehnke, M. (2010), 'Genome-wide association studies in diverse populations', *Nature Reviews Genetics* **11**(5), 356–366.

Roweis, S. & Saul, L. (2000), 'Nonlinear Dimensionality Reduction by Locally Linear Embedding', *Science* **290**, 2323–2326.

Sabbagh, A. & Darlu, P. (2006), 'Data-Mining Methods as Useful Tools for Predicting Individual Drug Response: Application to CYP2D6 Data', *Hum Hered* **62**, 119–134.

Sachidanandam, R., Weissman, D., Schmidt, S., Kakol, J., Stein, L., Marth, G., Sherry, S., Mullikin, J., Mortimore, B., Willey, D. et al. (2001), 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature* **409**(6822), 928–933.

Sadava, D., Heller, H., Orians, G., Purves, W. & Hillis, D. (2006), *Life: the science of biology*, Vol. 2, WH Freeman & Co.

Sasieni, P. (1997), 'From genotypes to genes: doubling the sample size', *Biometrics* pp. 1253–1261.

Satagopan, J. & Elston, R. (2003), 'Optimal two-stage genotyping in population-based association studies', *Genetic Epidemiology* **25**(2), 149–157.

Saul, L. & Roweis, S. (2003), 'Think globally, fit locally: unsupervised learning of low dimensional manifolds', *The Journal of Machine Learning Research* **4**, 119–155.

Saxena, R., Voight, B., Lyssenko, V., Burtt, N., de Bakker, P., Chen, H., Roix, J., Kathiresan, S., Hirschhorn, J., Daly, M. et al. (2007), 'Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels', *Science* **316**(5829), 1331–1336.

Schaid, D. (2004), 'Evaluating associations of haplotypes with traits', *Genetic Epidemiology* **27**(4), 348–364.

Schaid, D., Rowland, C., Tines, D., Jacobson, R. & Poland, G. (2002), 'Score tests for association between traits and haplotypes when linkage phase is ambiguous', *The American Journal of Human Genetics* **70**(2), 425–434.

Schölkopf, B. & Smola, A. (2002), *Learning with kernels*, The MIT Press.

Schulze, T., Zhang, K., Chen, Y., Akula, N., Sun, F. & McMahon, F. (2004), 'Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome', *Human Molecular Genetics* **13**(3), 335.

Schwender, H., Zucknick, M., Ickstadt, K. & Bolt, H. (2004), 'A pilot study on the application of statistical classification procedures to molecular epidemiological data', *Toxicology Letters* **151**(1), 291–299.

Sha, Q., Zhang, Z., Schymick, J., Traynor, B. & Zhang, S. (2009), 'Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis', *BMC Medical Genetics* **10**(1), 86.

Sha, Q., Zhu, X., Zuo, Y., Cooper, R. & Zhang, S. (2006), 'A combinatorial searching method for detecting a set of interacting loci associated with complex traits', *Annals of Human Genetics* **70**(5), 677–692.

Sham, P. (1998), *Statistics in human genetics*, A Hodder Arnold Publication.

Shannon, C. (1948), 'A mathematical theory of communication', *Bell Systems Tech J* **27**, 379–423.

Shastry, B. (2003), 'SNPs and haplotypes: genetic markers for disease and drug response', *International Journal of Molecular Medicine* **11**(3), 379–382.

Shendure, J., Mitra, R., Varma, C. & Church, G. (2004), 'Advanced sequencing technologies: methods and goals', *Nature Reviews Genetics* **5**(5), 335–344.

Sherman, B., Lempicki, R. et al. (2009), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic acids research* **37**(1), 1–13.

Shi, J. & Malik, J. (2000), 'Normalized Cuts and Image Segmentation', *IEEE Transaction on pattern analysis and machine intelligence* pp. 888–905.

Shim, J., Sohn, I., Kim, S., Lee, J., Green, P. & Hwang, C. (2009), 'Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine', *Comp. Stat. and Data Analysis* **53**(5), 1736–1742.

Sindhwani, V., Bhattacharya, P. & Rakshit, S. (2001), Information theoretic feature crediting in multiclass support vector machines, *in* 'Proceedings of the First SIAM International Conference on Data Mining'.

Smith, D. & Lusis, A. (2002), 'The allelic structure of common disease', *Human Molecular Genetics* **11**(20), 2455.

Souverein, O., Zwinderman, A. & Tanck, M. (2006), 'Multiple imputation of missing genotype data for unrelated individuals', *Annals of human genetics* **70**(3), 372–381.

Stallings, R., Ford, A., Nelson, D., Torney, D., Hildebrand, C. & Moyzis, R. (1991), 'Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes', *Genomics* **10**(3), 807–815.

Statnikov, A., Wang, L. & Aliferis, C. (2008), 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification', *BMC Bioinformatics* **9**(1), 319.

Stefansson, H., Petursson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O. et al.

(2002), 'Neuregulin 1 and susceptibility to schizophrenia', *The American Journal of Human Genetics* **71**(4), 877–892.

Stephens, J., Schneider, J., Tanguay, D., Choi, J., Acharya, T., Stanley, S., Jiang, R., Messer, C., Chew, A., Han, J. et al. (2001), 'Haplotype variation and linkage disequilibrium in 313 human genes', *Science* **293**(5529), 489.

Stephens, M., Smith, N. & Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *The American Journal of Human Genetics* **68**(4), 978–989.

Stoll, M., Corneliussen, B., Costello, C., Waetzig, G., Mellgard, B., Koch, W., Rosenstiel, P. et al. (2004), 'Genetic variation in DLG5 is associated with inflammatory bowel disease', *Nature Genetics* **36**(5), 476–480.

Storey, J. & Tibshirani, R. (2003), 'Statistical significance for genome-wide studies', *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9440.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T. & Zeileis, A. (2008), 'Conditional variable importance for random forests', *BMC Bioinformatics* **9**(1), 307.

Sun, Y., Cai, Z., Desai, K., Lawrance, R., Leff, R., Jawaid, A., Kardia, S. & Yang, H. (2007), Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests, *in* 'BMC proceedings', Vol. 1, BioMed Central Ltd, p. S62.

Suykens, J., De Brabanter, J., Lukas, L. & Vandewalle, J. (2002), 'Weighted least squares support vector machines: robustness and sparse approximation', *Neurocomputing* **48**(1-4), 85–105.

Svetnik, V., Liaw, A., Tong, C. & Wang, T. (2004), 'Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules', *Lecture Notes in Computer Science* **3077**, 334–343.

Tabor, H., Risch, N. & Myers, R. (2002), 'Candidate-gene approaches for studying complex genetic traits: practical considerations', *Nature Reviews Genetics* **3**(5), 391–397.

Tagaris, G., Richter, W., Kim, S., Pellizzer, G., Andersen, P., Ugurbil, K. & Georgopoulos, A. (1998), 'Functional magnetic resonance imaging of mental rotation and memory scanning: a multidimensional scaling analysis of brain activation patterns', *Brain Research Reviews* **26**(2-3), 106–112.

Tan, P., Steinbach, M. & Kumar, V. (2005), *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

The National Human Genome Research Institute (2003), *The National Human Genome Research Institute*. Online Resource `http://www.genome.gov/gwastudies/`.

Thomas, G., Jacobs, K., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N. et al. (2008), 'Multiple loci identified in a genome-wide association study of prostate cancer', *Nature Genetics* **40**(3), 310–315.

Thomas, J., Olson, J., Tapscott, S. & Zhao, L. (2001), 'An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles', *Genome Research* **11**(7), 1227–1236.

Tian, J., Wu, N., Guo, X., Guo, J., Zhang, J. & Fan, Y. (2007), 'Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines', *BMC Bioinformatics* **8**(1), 450.

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

Todd, J. (2006), 'Statistical false positive or true disease pathway?', *Nature Genetics* **38**(7), 731–734.

Toivonen, H., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., Herr, M. & Kere, J. (2000), 'Data mining applied to linkage disequilibrium mapping', *The American Journal of Human Genetics* **67**(1), 133–145.

Torgerson, W. (1952), 'Multidimensional scaling: I. Theory and method', *Psychometrika* **17**(4), 401–419.

Tzeng, J., Devlin, B., Wasserman, L. & Roeder, K. (2003), 'On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit', *The American Journal of Human Genetics* **72**(4), 891–902.

Tzeng, J., Wang, C., Kao, J. & Hsiao, C. (2006), 'Regression-based association analysis with clustered haplotypes through use of genotypes', *The American Journal of Human Genetics* **78**(2), 231–242.

van den Oord, E. & Sullivan, P. (2003), 'False discoveries and models for gene discovery', *TRENDS in Genetics* **19**(10), 537–542.

Van Der Maaten, L., Postma, E. & Van Den Herik, H. (2007), 'Dimensionality reduction: A comparative review', *Published online* **10**(February), 1–35.

Vapnik, V., Golowich, S. & Smola, A. (1996), Support vector method for function approximation, regression estimation, and signal processing, *in* 'Advances in Neural Information Processing Systems 9'.

Venkatarajan, M. & Braun, W. (2004), 'New quantitative descriptors of amino-acids based on multidimensional scaling of a large number of physical-chemical properties', *Journal Molecular Modeling* **7**(12), 445–453.

Venna, J. & Kaski, S. (2006), 'Local multidimensional scaling', *Neural Networks* **19**(6-7), 889–899.

Venna, J. & Kaski, S. (2007), 'Nonlinear Dimensionality Reduction as Information Retrieval', *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS* 07)* pp. 568–575.

Venna, J., Peltonen, J., Nybo, K., Aidos, H. & Kaski, S. (2010), 'Information retrieval perspective to nonlinear dimensionality reduction for data visualization', *The Journal of Machine Learning Research* **11**, 451–490.

Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R. et al. (2001), 'The sequence of the human genome', *Science* **291**(5507), 1304.

Waldron, E., Whittaker, J. & Balding, D. (2006), 'Fine mapping of disease genes via haplotype clustering', *Genetic Epidemiology* **30**(2), 170–179.

Wallenstein, S., Hodge, S. & Weston, A. (1998), 'Logistic regression model for analyzing extended haplotype data', *Genetic Epidemiology* **15**(2), 173–181.

Wang, H., Thomas, D., Pe'er, I. & Stram, D. (2006), 'Optimal two-stage genotyping designs for genome-wide association scans', *Genetic Epidemiology* **30**(4), 356–368.

Wang, W., Barratt, B., Clayton, D. & Todd, J. (2005), 'Genome-wide association studies: theoretical and practical concerns', *Nature Reviews Genetics* **6**(2), 109–118.

Wang, X., Wu, M., Li, Z. & Chan, C. (2008), 'Short time-series microarray analysis: Methods and challenges', *BMC Systems Biology* **2**(1), 58.

Wang, Y., Liu, G., Feng, M. & Wong, L. (2011), 'An empirical comparison of several recent epistatic interaction detection methods', *Bioinformatics* **27**(21), 2936–2943.

Weir, B., Hill, W. & Cardon, L. (2004), 'Allelic association patterns for a dense SNP map', *Genetic Epidemiology* **27**(4), 442–450.

Wellcome Trust Case Control Consortium (2011), *Wellcome Trust Case Control Consortium* . Online Resource `https://www.wtccc.org.uk/index.shtml`.

Wessel, J. & Schork, N. (2006), 'Generalized genomic distance-based regression methodology for multilocus association analysis', *The American Journal of Human Genetics* **79**(5), 792–806.

Wigginton, J., Cutler, D. & Abecasis, G. (2005), 'A note on exact tests of Hardy-Weinberg equilibrium', *The American Journal of Human Genetics* **76**(5), 887–893.

Williams, S., Ritchie, M., Phillips III, J., Dawson, E., Prince, M., Dzhura, E., Willis, A., Semenya, A., Summar, M., White, B. et al. (2004), 'Multilocus analysis of hypertension: a hierarchical approach', *Hum Hered* **57**(1), 28–38.

Wirth, R. (2000), Crisp-dm: Towards a standard process model for data mining, *in* 'Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining', pp. 29–39.

Wu, J., Devlin, B., Ringquist, S., Trucco, M. & Roeder, K. (2010), 'Screen and clean: a tool for identifying interactions in genome-wide association studies', *Genetic Epidemiology* **34**(3), 275–285.

Wu, X., Wu, Z. & Li, K. (2008), Classification and identification of differential gene expression for microarray data: improvement of the random forest method, *in* 'The 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008', IEEE, pp. 763–766.

Xu, J., Lowey, J., Wiklund, F., Sun, J., Lindmark, F., Hsu, F., Dimitrov, L., Chang, B., Turner, A., Liu, W. et al. (2005), 'The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk', *Cancer Epidemiology Biomarkers and Prevention* **14**(11), 2563.

Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S. et al. (2010), 'A genome-wide association study in the japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B', *Nature Genetics* **42**(10), 864–868.

Yan, H., Papadopoulos, N., Marra, G., Perrera, C., Jiricny, J., Boland, C., Lynch, H., Chadwick, R., de La Chapelle, A., Berg, K. et al. (2000), 'Conversion of diploidy to haploidy', *Nature* **403**(6771), 723–724.

Yeoh, E., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A. et al. (2002), 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling', *Cancer cell* **1**(2), 133–143.

Zeggini, E., Weedon, M., Lindgren, C., Frayling, T., Elliott, K., Lango, H., Timpson, N., Perry, J., Rayner, N., Freathy, R. et al. (2007), 'Wellcome Trust Case Control Consortium (WTCCC), McCarthy MI, Hattersley AT: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes', *Science* **316**(5829), 1336–1341.

Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. (2002), 'Haplotype block structure

and its applications to association studies: power and study designs', *The American Journal of Human Genetics* **71**(6), 1386–1394.

Zhang, Z., Zhang, S., Wong, M., Wareham, N. & Sha, Q. (2008), 'An ensemble learning approach jointly modeling main and interaction effects in genetic association studies', *Genetic Epidemiology* **32**(4), 285.

Zhao, H., Pfeiffer, R. & Gail, M. (2003), 'Haplotype analysis in population genetics and association studies', *Pharmacogenomics* **4**(2), 171–178.

Zhao, J., Boerwinkle, E. & Xiong, M. (2005), 'An entropy-based statistic for genome-wide association studies', *The American Journal of Human Genetics* **77**(1), 27–40.

Zhao, Z. & Liu, H. (2009), 'Searching for interacting features in subset selection', *Intelligent Data Analysis* **13**(2), 207–228.

Zhernakova, A., Stahl, E., Trynka, G., Raychaudhuri, S., Festen, E., Franke, L., Westra, H., Fehrmann, R., Kurreeman, F., Thomson, B. et al. (2011), 'Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-hla shared loci', *PLoS Genetics* **7**(2).

Zollner, S. & Pritchard, J. (2005), 'Coalescent-based association mapping and fine mapping of complex trait loci', *Genetics* **169**(2), 1071.