Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR GENOME WIDE ASSOCIATION STUDIES

by
Neha Singh

Variations present in human genome play a vital role in the emergence of genetic disorders and abnormal traits. Single Nucleotide Polymorphism (SNP) is considered as the most common source of genetic variations. Genome Wide Association Studies (GWAS) probe these variations present in human population and find their association with complex genetic disorders. Now these days, recent advances in technology and drastic reduction in costs of Genome Wide Association Studies provide the opportunity to have a plethora of genomic data that delivers huge information of these variations to analyze. In fact, there is significant difference in pace of data generation and analysis, which led to new statistical, computational and biological challenges. Scientists are using numerous approaches to solve the current problems in Genome Wide Association Studies.

In this thesis, a comparative analysis of three Machine learning algorithms is done on simulated GWAS datasets. The methods used for analysis are Recursive Partitioning, Logistic Regression and Naïve Bayes Classifier. The classification accuracy of these algorithms is calculated in terms of area under the receiver operating characteristic curve (AUC). Conclusively, the logistic regression model with binary classification seems to be the most promising one among the other four algorithms, as it outperformed the other tools in the AUC value.

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR GENOME WIDE ASSOCIATION STUDIES

by
Neha Singh

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

May 2012

APPROVAL PAGE

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR GENOME WIDE ASSOCIATION STUDIES

Neha Singh

Dr. Jason T. L. Wang, Thesis Advisor — Date
Professor of Computer Science, NJIT

Dr. James Geller, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Neha Singh

Degree: Master of Science

Date: May 2012

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2012

- Bachelor of Technology in Bioinformatics,
  Amity University, Noida, India, 2010

Major: Bioinformatics

I dedicate this thesis to my parents, Manju Singh and Vijayendra Pal Singh, for their unconditional support, love, encouragement and belief in me.

ACKNOWLEDGMENT

I would like to thank my thesis advisor and mentor, Dr. Jason T. L. Wang, for his belief in me and my ideas, and for his support and guidance that encouraged me at every step and showed me the path to proceed. I would also like to thank Dr. James Geller and Dr. Zhi Wei for agreeing to be on my thesis committee, for their valuable inputs and the moral support throughout.

I would also like to thank my family and friends for their unconditional support and love throughout my journey till now. I want to thank my parents for their patience and teachings which have given me the strong foundation and strength to follow my path with determination.

I would like to acknowledge and thank all the authors and researchers whose work has been a strong motivation and reference for my work.

Last, but not least, I want to thank Almighty for giving me the strength and confidence required.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AUC	Area Under Curve
CEPH	Centre d'Etude du Polymorphisme Humain
Df	Degrees of freedom
DNA	Deoxyribonucleic Acid
GWAS	Genome Wide Association Studies
LD	Linkage Disequilibrium
R	R Statistical Software
RNA	Ribonucleic Acid
ROC	Receiver Operating Curve
RPART	Recursive Partitioning
SNP	Single Nucleotide Polymorphism

LIST OF DEFINITIONS

Allele	Also called allelomorph is any one of two or more genes that may occur alternatively at a given site (locus) on a chromosome are called as alleles, also known as allelomorph.
Autosomes	Chromosomal pairs which are other than sex chromosome present in somatic cells.
Centromere	It is a central region where the two chromatids are held together to each other and attach to spindle fibers in mitosis and meiosis stage.
Chromatid	It is one of the two identical, threadlike filaments of a chromosome. Chromatids are produced by the self-replication of the chromosome during interphase and are held together by a common centromere.
Diploid stage	Chromosomes are present in duplication, which means two chromatids per chromosome.
Genetic Locus	The specific position on a chromosome where a gene is located.
Heterozygous	Non-identical copies of genes alleles at corresponding loci on homologous chromosomes. An individual inherits an allele for that trait from one parent and an alternative allele from the other parent.
Homozygous	Identical alleles are present at corresponding loci on homologous chromosomes. An individual inherits from each parent one allele for that trait.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The increasing power of Genome Wide Association Studies enables researchers to investigate the association of genomic variations with complex human genetic diseases such as Bipolar disorder (BD), Rheumatoid arthritis (RA), Type 1 diabetes (T1D), Type 2 diabetes (T2D), Coronary Artery Disease (CAD) etc. (WTCCC, 2007). Now these days, recent advancement in technology and drastic reduction in costs of Genome Wide Association Studies provide the opportunity to have plethora of genomic data that delivers huge human variation information to analyze. The amelioration of high throughput SNP genotyping technologies providing huge amount of Single Nucleotide Polymorphism data which fuels Genome Wide Association Studies, and which led us to new statistical, computational and biological challenges (Herbert, et al., 2006), (Ozaki & Ohnishi, 2002), (Roses, 2003). Every disease discovery project have aim to identify all genomic variation which leads to particular phenotype across the population which consists affected (Case) and unaffected (Controls). The result of these variations could be Disease Status, Drug Responder Status and Adverse Drug Reactions. GWAS raised the expectations of revealing the SNPs variations associations and their interactions involve in complex human genetic disorders, however the challenge is to deal with this huge amount of data and extract the underlying information. The considerable statistical and biological issues that are faced in the genomic datasets consists the dimensionality problem (Bellman, 1961) , Multiple Testing problem (Xie, Cai, Maris, & Li, 2010) and the presence of heterogeneity (Thornton-Wells, Moore, & Haines, 2006). But, this is

proved to be less fruitful than expected till this time as there are so many questions which need to be answer, as in a review study of 600 positive associations, some of which have been studied multiple times, only 6 association were consistent (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002), statistician and computational biologists need to apply some different methods and perspectives to reveal the underlying SNPs associations with genetic diseases.

To face the above mentioned issues there are many methods which have been applied to whole genome data like biological interpretation is incorporated into the statistical analysis to filter the data (Bush, Dudek, & Ritchie, 2009) and also statistical analysis results can be applied for further biological interpretation. To deal with above mentioned problems and to incorporate every technique one may follow multi- step approach (Kropff, 2008). Figure 1.1 describes the multi-step approach.

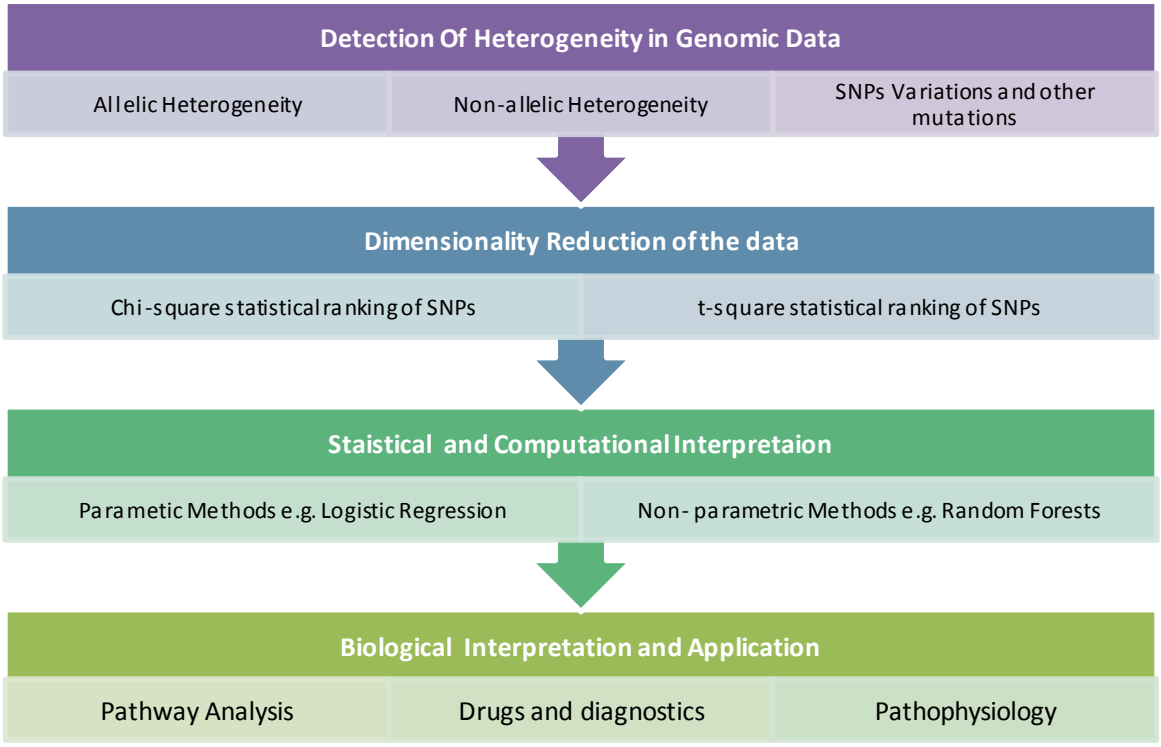

Figure 1.1 Multi-step approach towards genome wide association studies.

There are many Machine learning algorithms which have been already applied to Genome Wide Association Studies (Costello, Falk, & Ye, 2003) like classification and regression trees (CART) of (Breiman, 2001) (Uriarte & Andres, 2006), Support Vector Machines (SVM) (Vapnik, 1998) (Guyon, Weston, Barnhill, & Vapnik, 2002), Neural Networks (NN) (Bishop, 1995) and many more. At present, there is no single method which can be applied to all kind of datasets and deliver all the substantial information in Genome Wide Association Studies. This thesis work is basically focused over the application of some of the machine learning algorithms and their accuracy of classifying the data.

1.2 Objective

The objective of this thesis is to conduct the comparative analysis of the four Machine Learning (ML) Algorithms over simulated genomic data. The classifiers which are used for the study are Logistic Regression, Recursive Partitioning, and Naïve Bayes Classifier. These ML algorithms are implemented with the help of statistical software 'R'. The simulation program namely GWAsimulator (Li & Li, 2007) is used to simulate the whole genome data for this study.

The simulation is done five times on different control file for the program and these simulated datasets are divided into training and test datasets as per the Case- control study design. Then above mentioned classifiers ML algorithms are applied on each training dataset to create prediction models. Then these prediction models are applied to the training dataset for classification. The classification accuracy is predicted by means of Area under the Curve (AUC) of Receiver Operating Characteristic (ROC) graphs. The prediction methods are applied with the help of the ROCR package (Sing, Sander,

Beerenwinkel, & Lengauer, 2005) available in R which provides the standard methods for examining accuracy of the classifier by providing the specific performance measures.

1.3 Background

The whole stories of GWAS begins with the advent of the Human Genome Project in 2000, and also with this the SNP Consortium and first phase of the International Hap Map project (Gibbs, Belmont, Hardenbol, & Willis, 2003) put it forward. Then with the completion of second phase of the International Hap Map project in 2007 provided the strong foundation to this new era of whole genome studies. The International Hap Map project provided us with SNP frequencies, Genotypes and Haplotype structures which initiated the SNP genotyping and then eventually Genome Wide Association Studies. The Human Genome Project (International Human Genome Sequencing Consortium, 2001), the SNP consortium (The International SNP Map Working Group, 2001) and the International HapMap Project (The International HapMap consortium, 2007) collectively provided approximately 10 million DNA variants, mainly SNPs (The International HapMap 3 Consortium, 2010). The data generated in the above mentioned projects was available to public domain which proved to be the boost for genomic researches. Another main factor in increment of Genome Wide Association Studies was the evolution of Bio repositories. Bio repositories are bank of all the biological sequences which are potential research objects in Computational Biology, Genomics and so on. Essentially, in order to learn about the Genome Wide association studies it would be rational to have a brief look over the events which took place collectively to make platform for these studies.

Table 1.1 Time Line of Events Leads to Genome Wide Association Studies

Main Events	Years
Human Genome Project	2000-2004
The SNP Consortium	2000-2003
The International Hap Map Project	2002-2007
The SNP Genotyping	2005-Present
Genome Wide Association Studies	2007-Present

1.3.1 Human Genome Project

The whole approach of genome wide association studies is started after the completion of Human genome project (HGP) in 2003 (Ventor, Adams, Myers, Li, & Mural, 2001), which was a multi country 13 year program to genotype human genome and later the SNPs data coordinated by United States Department of Energy (DOE) and National Institute of Health (NIH). The main aim was to generate as much data as possible and store the data into databases for further studies. The pioneer contribution United States Department of Energy (Deegan, 1989) (Barnhart, 1989) ignited the fire of Human Genome project in the mind of scientists, and later, the efforts of Welcome Trust Case Control Consortium (WTCCC, 2007) and countries like UK, later Japan, France and more made this Human genome project a milestone in the field of Computational Genomics. The vast support achieved by Human Genome project tells the story of its critical importance and success achieved (Gert, 1996). The genetic information is then stored in open access sequence database GenBank database of National Centre for Biotechnology Information and related organizations of Europe and Japan. This made the

availability of human genome data to researchers which proved important in revealing the human variations responsible for common genetic diseases. This also helped in the understanding of complex human biology.

As the most important application of Human Genome Project, the Wellcome Trust Case-Control Consortium (WTCCC) undue approach towards the real SNPs data generation of the cases and controls of the seven complex diseases made the great contribution towards the analytical and computational solution of complex Human genetic diseases (WTCCC, 2007). Single Nucleotide Polymorphism is considered as the most common source of variations found in the human genome. As the result of Human Genome Project, it has been identified that Single Nucleotide Polymorphism occurs at approximately 1.4 million locations in humans (From genome to proteome., 2008). The results of HGP gave the platform which mobilized the investigations of locations and sequences of genes which are responsible complex human diseases.

Basically, with the results of Human Genome Project and advanced high-throughput technologies researchers could answer the complexity of human genome and complex diseases systematically and on a very big scale.

1.3.2 The SNP Consortium

The SNP Consortium (TSC) established in 1999 as the collaboration of major pharmaceutical companies, the WTCCC and academic centers (Holden, 2002). The main aim of the TSC was to identify more than 300000 SNPs up to 2001, which was resulted in exceeding of final results by release of approximately 1.4 million SNPs into the public domain (Sachidanandam, et al., 2001). The other objective of TSC was to manage the publications of the Haplotype Map (Holden, 2002) . The SNP Consortium is basically the

data repository which contains the initial data of SNP discovery process of that time and later on that SNP data is submitted to dbSNP (Thorisson & Stein, 2003). The Single Nucleotide Polymorphism data published as the result of the Human Genome Project was managed and analyzed by the SNP consortium and data management and analysis was conducted by Cold Spring Harbor Laboratory (SNP Fact Sheet, 2008).

1.3.3 The International HapMap Project

The International HapMap Project was initiated in 2002. This project was started with the collaboration among the researchers, laboratories, institutions and funding agencies form Japan, the United Kingdom, Canada, China, Nigeria and the United States (The International HapMap Project, 2002). This was the effort to investigate the genetic similarities and variations in human population (The International HapMap consortium, 2007). The main aim of the HapMap project was to describe the Haplotype map of the human genome to provide the solution to the problem of major genetic diseases. The Haplotype map includes the strongly associated SNPs and SNP tags in particular regions of chromosome which replicate together in diseased and healthy individuals. The huge data generated in all the three phases of the International HapMap project resulted in the substantial cost reduction of genotyping the SNP data which led to the increment in pace of Genome Wide association studies. Almost all parts of human genome are similar to each other, but they have differences in some common haplotypes. Therefore, to found the differences in haplotype frequency data is collected from four different regions namely Nigeria (Yoruba), Japan, China and U.S. residents with northern and western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH).

1.3.4 Genome Wide Association Studies

In defiance of the biological, statistical and computational intricacies related to discovering process of genomic variations in complex genetic disorders, the classic study design and analysis have not worked up to the mark. In 2001, linkage analysis has been done for T1D (European Consortium for IDDM genome Studies, 2001) and for many more diseases which produced some convincing results in diseases which have high sibling ratio (Altmuller, Palmer, Fischer, Scherb, & Wjst, 2001). But linkage studies could not find the genetic risk factors for familial Alzheimer's disease, Multiple Sclerosis and Autism, which are the very prominent candidates for linkage analysis, even after the number of studies. On the other hand, Linkage analysis was able to produce some significant results for rare forms of other familial phenotype, such as familial hypercholesterolaemia2 (Ott, Schrott, & Goldstein, 1974) (Ott, Kamatani, & Lathrop, Family based designs for genome-wide association studies., 2011) and familial breast cancer (Wooster & Weber, 2003). Similarly, genetic association studies proved less substantial when they are tested multiple times; therefore it's not wise to make conclusion over the association between genetic variant and susceptibility of disease from only one testing (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). Over the last two decades, the advancement in technology and drastic reduction in costs of Genome Wide Association Studies provided us the opportunity to investigate the intricacies of human genome variations which are responsible for complex diseases. The Genome Wide Association Studies aim to find out the difference in allelic frequencies in SNP haplotypes between healthy and diseased individuals. In these association studies about 1 million of SNPs, which responsible for maximum variations, are captured to find the

causal variations across the human genome (Barrett & Cardon, 2006). The basic principle of the Genome wide association studies is to follow the path of contiguous stretch of tagged SNPs or haplotypes which transmit from generation to generation through recombination. And, by further analysis the association between these markers and disease phenotype can be detected. This idea follows the Common Disease (CD) - Common Variant (CV) hypothesis, that onset of common genetic diseases relies on the common variations present in human genome (Shields, 2011).

The identification of Complement Factor H (CFH) as causal variant in Age-related Macular Degeneration was the inaugural success of in the field of GWAS (Klein, et al., 2005). Since, then it has been seen the regular increment in the acceptance of GWAS. Genome wide association studies statistically investigated about and over 200 disease traits in 700 genome wide association studies (Baker, 2010) which involves over 1200 human genome till December 2009 (Johnson & O'Donnell, 2009). These studies identify the association of causal SNPs with the complex diseases but cannot fully identify the cause of disease. The journey of Genome Wide Association Studies is well described by the review analysis of GWAS by (Manolio, Brooks, & Collins, 2008). The following figure is the extension of the work of (Manolio, Brooks, & Collins, 2008) which is regularly updated as the catalog of published Genome-Wide Association Studies by (Hindorff, et al., 2011).

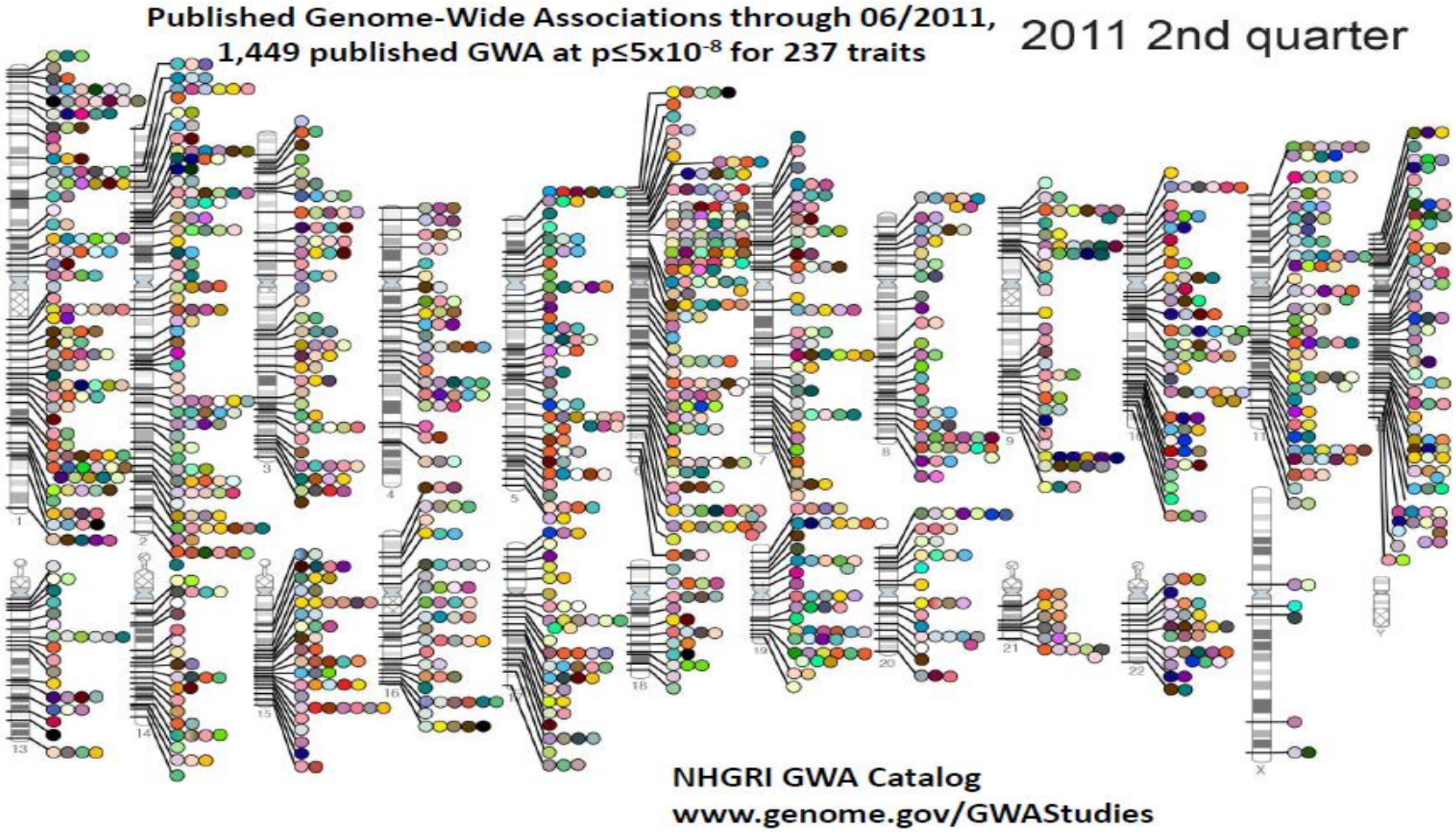

Figure 1.2 Karyogram of SNP- Trait association investigated in GWAS

Source: www.genome.gov/GWAStudies

Figure 1.3 Explanation of traits present in the GWA catalog.

Source: www.genome.gov/GWAStudies

CHAPTER 2

BIOLOGICAL OVERVIEW OF GWAS

2.1 Overview

Genome wide association studies are basically the answer of the ever existing question that why some people are predisposed towards a certain trait or disease while others lives a healthy life. At the time of the birth of Genomic era, it was there in in every conscience that this will improve the understanding of the hidden aspects of biology and human genetics. In the field of Genomics; science, technology and medicine developed and progressed at very high pace in last two decades (Guttmacher & Collins, 2003). It is believed that human genome contains about 20,000-25,000 genes which encodes proteins (Stein, 2004), which transcribes into Ribonucleic acid (RNA) and then direct the translation of RNA into proteins (Lander, 2011). Every mere functional, developmental and organizational phenomena of human body depends on the Central Dogma; the informational flow in biological systems shown in Figure 2.1.

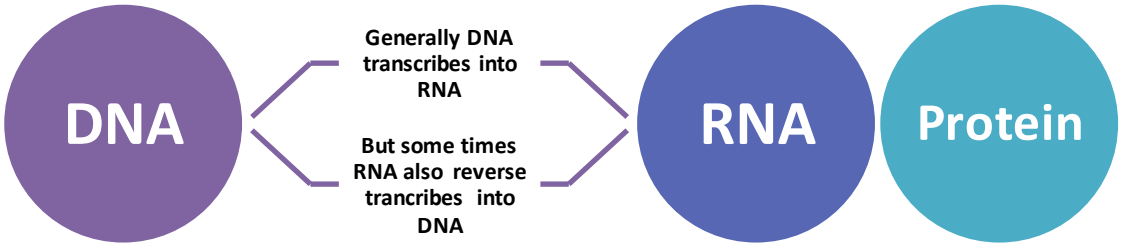

Figure 2.1 Central dogma: flow of information in biological systems.

Now, the question arises that, what varies a person from another person? Where these variations came from? Which part of the human genome they affect? How they are

associated with diseases and traits and so on. These answers can be found by looking into the biology behind this. This can be understood by consider Genomes as book of the life which contains 23 chapters called Chromosomes (Barlow-Stewart, 2004). The Genes are the sections of each chapter which are the functioning part of the book and these genes are comprised of collection of words called Deoxyribonucleic Acid (DNA). And, these words which are called as DNA are comprised of only four letters A, T, G and C. The following diagram is the illustration of the packaging of the whole genomic information.

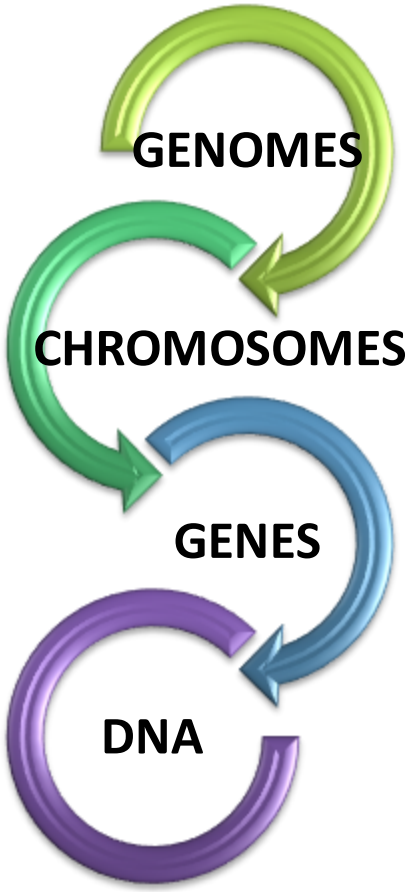

Figure 2.2 The packaging of genetic information in humans.

Genomes comprised of Chromosomes, Chromosomes comprised of Genes, and Genes are of DNA.

2.1.1 Genome

The Genome is the entity that carries the whole genetic information of organisms in encrypted format. It has the complete set of genetic instructions which guides the functioning of the cells of organism, and passes hereditary information to next generation. It contains all the coding and non-coding DNA and RNA (Ridley, 2006). Apparently, every cell of an organism contains the whole copy of its genome. As it is illustrated in Figure 2.2, Genome is made up set of Chromosomes, Chromosomes are made up Genes and Genes are made of DNA. Every organism has a particular number of chromosomes copies like some are diploid as humans, triploid, or haploid only one copy of all chromosomes. Therefore, when it is said that an organism's genome is sequenced, it implies that a haploid or single copy of chromosomes or single set of autosomes (Chromosomal set without sex determination chromosome) is sequenced and store in database. Humans have 3.2 billion base pairs and approximately 20,000 to 22,000 genes on 23 pairs of Chromosomes in all cells of human body and decide their structure and functions.

As we see, almost all the individuals of human population have same basic characteristics but yet different from each other. Consider these variations among human species; one cannot say a particular human genome a standard or normal. Everybody is abnormal in their own way; every genome is mutant (Feero, Guttmacher, & Collins, 2010). To study these differences, Genome wide association studies can be an answer (Guttmacher & Collins, 2003). There are mainly three basic types of variations: First is Single-base-pair changes, second is insertion and deletion of nucleotide, and third is frame-shift mutation. The single-base-pair mutation is also known as SNPs. Over the past

several years, the association, candidate gene and linkage studies have made it possible to quantify the association of these SNPs with diseases (Baker, 2010).

2.1.2 Chromosomes

The Chromosomes are compact organization of DNA and proteins (which are used in packaging of DNA in compact form) as single unit which have genes (coding), non-coding sequences and stackable proteins. In other terms it is a long chain of nucleotides which is compactly arranged in the form of chromatin which allows huge DNA molecules to fit into eukaryotic cells. Chromosomes are mostly found in pairs in human species and this is called diploid state. The diploid behavior of human chromosomes was observed about 50 years ago (Painter, 1924) (Jio & Levan, 1956). Chromosomes can be of different shapes and sizes, but humans and most of the eukaryotic organisms have linear shape. Chromosomes must be replicated and divided into single chromosomes and pass on to daughter cells to their later progeny. At this point both the sister chromatids are attached to each other. There is a constriction point which divides the chromosomes into two parts, called as Centromere. This constriction divides the chromatids into two parts; the shorter arm is called as p arm and longer one is called as q arm.

The genome of every organism is divided into Chromosomes. Human have 23 pairs of linear chromosomes which comprises of 22 pairs of autosomes and one pair of sex chromosomes. These vary slightly in shape, size and appearance. The chromosome contains a full stretch of a single DNA molecule. The number of chromosomes is nothing to do with the complexity of organisms, it's completely depend on nature as very small Goldfish has 94 pairs of chromosomes, on the other hand Cat has only 38 pairs of

Chromosomes. But on the contrary both the species have huge difference in their metabolic, structural and functional complexity.

2.1.3 Genes

The singular coding hereditary unit is called gene. Genes are the stretches of DNA which encodes proteins which are further responsible for specific traits and functions in the organisms. Genes are responsible for similarities and differences in the species, similarities like every human has "hair color gene" which codes for hair color but differences lies in which color like people have different color of hairs such as black, brown, grey, white, golden and many more (Davenport & Davenport, 1908). Mostly, all people have similar genes for each and every trait but these are alleles, the single variants of genes, which are responsible for variation in phenotype or physical appearance of people.

Then during the course of period the molecular biological definition of gene changed which says genes are the stretches of DNA which has definite end and beginning (Noble, 2008). The biochemical explanation of gene defines the ultimate process of transformation of Gene to physical form of trait expression. The gene is coding DNA which codes for protein and RNA, and this coding depends on Promoter and Enhancers. Here, promoters and enhancers decide which part of DNA will transcribe into pre-mRNA. The pre-mRNA is composed of Exon and Introns, where Exon is coding part of pre-mRNA which later encodes for proteins, and Introns are spliced during the transformation from pre-mRNA to mRNA. And, later this mRNA translated to resultant proteins. According to classic genetics, the definition and functioning of gene was simpler but it is becoming complex day by day with the fact of overlapping genes

sequences (Pearson, 2006). The more comprehensive study of gene functionality will open path for better understanding of both rare and common diseases (Feero, Guttmacher, & Collins, 2010).

2.1.4 DNA

Deoxyribonucleic acid (DNA) is the basic biochemical entity of the gene and genome. DNA is written in language of four bases: adenine (A), thymine (T), cytosine (C) and guanine (G). And these bases with sugar and the phosphate group make nucleotides which are the chemical units of DNA. It is long double helical chain like structure which consist repeated units of nucleotides. The order of these nucleotides determines the biological instructions on genes (National Human Genome Reseach Institute, 2011). DNA is get transmitted to generations to generations and in its coded language it guides cell about its function and organization (Hershey & Crick, 1952). There are about 3 billion bases in humans and these are almost similar up to 99% in all humans (Kidd, et al., 2008).

The very first time DNA was characterized by Friedrich Miescher in 1869 during the analysis of constituents of the cell (Dahm, 2004). And then in 1915 Phoebus Levene described the structure of the fundamental unit of DNA, called nucleotide (Levene, 1915). In 1953, James Watson and Francis Crick discover the double helical structure of DNA and in this study they explained the probable pairing of adenine (A) with thymine (T) and cytosine (C) with guanine (G) (Watson & Crick, 1953). This discovery was the extension of Erwin Chargaff assumption of that DNA has approximately equal amounts of the adenine (A)-thymine (T) and cytosine (C)-guanine (G).

2.2 Genetic Variation

Despite of all the similarities in the book of genome among human species, every individual's genome is slightly different from each other. Although all rules would still apply e.g. E. Chargaff's rule, but the two genome sequences would not match exactly base to base. Inheritance of variations in genome leads to difference in phenotypes which can increase the risk of disease and may environmental behavior. The common types of genetic variations are: Mutations, Genetic rearrangements and Polymorphisms. Mutations are the variations which present at the level of DNA in which random changes could happen to one or more base pairs. Genetics rearrangements happen at chromosome level in which deletions and insertions of DNA sequences take place in chromosomes. Polymorphisms are variations which present in each individual DNA but these are not mutations. These single base variations or differences are referred as alleles. This is mostly present in two forms Single Nucleotide Polymorphisms and Copy number variations (Rotimi & Jorde, 2010). Even after the rigorous studies of almost a decade the compendium of causal variants or SNPs is not complete, and this proves the need of introduction of new aspects of Genome Wide Association Studies over the wide range of populations (Rotimi & Jorde, 2010).

2.2.1 Allele

An Allele is one of the two or more variants of the gene. The entire genome of humans has two copies of it in each cell, which is called as the diploid state. One copy genome comes from mother and one comes from father. Therefore, an individual inherits two copies of each gene, which may have different phenotypic effects, called alleles. This inheritance is explained in Figure 2.3. There are two possibilities; if alleles are the same

then this is called as "homozygous" condition and if alleles are different then it is called as "heterozygous" condition. This can be explained by the condition that the same base pair position can be acquired by Cytosine in one individual and the same position can be acquired by Guanine in another individual. In this condition, the presence of two different nucleotides represents two alleles of same gene.

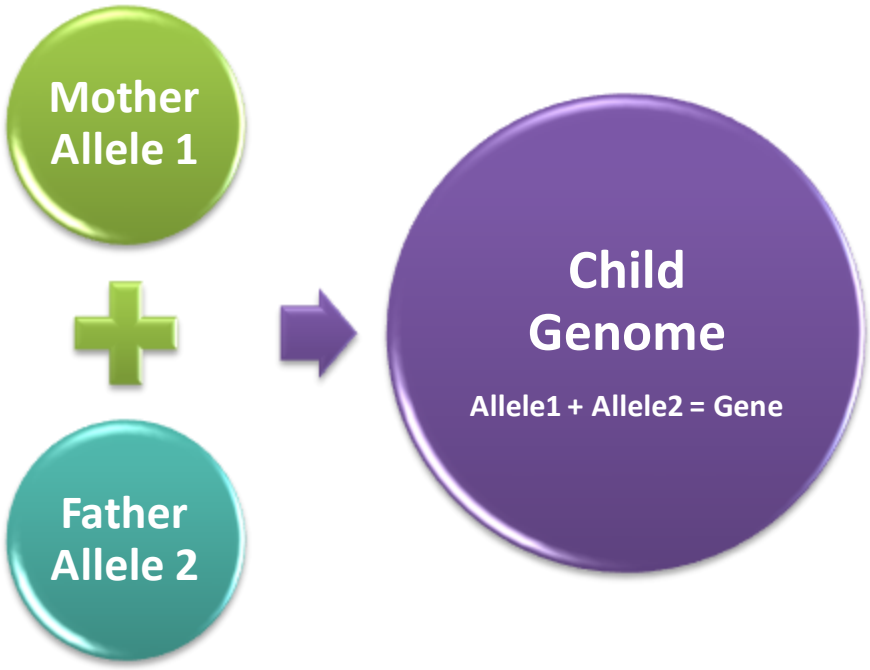

Figure 2.3 Each individual inherits two copies of a gene called alleles from each of his/her parents.

Out of the two alleles, one allele is always prevalent to another one in a particular population. The more frequent allele is often called as wild type and other allele is considered as mutation. Nevertheless, "mutation" is not the appropriate term for the less frequent allele because wild type or ancestral allele is not always the most frequent one. Therefore, "variation" will be the appropriate term should be used to describe the presence of alleles in genetics.

2.2.2 Single Nucleotide Polymorphism

Two or more than two variation of single DNA nucleotide at specific position among individuals is called SNPs. This can be explained as at a specific position one individual may have "A" in contrary of another individual who has "C".

SNP Position/ Person	*1*	*2*	*3*	*4*	*5*	6	*7*	*8*	9	*10*	*11*	*12*
John	A	T	G	A	C	***G***	C	C	***C***	T	G	A
Joseph	A	T	G	A	C	***G***	C	C	A	T	G	A
Thomas	A	T	G	A	C	T	C	C	A	T	G	A
Michelle	A	T	G	A	C	***G***	C	C	***C***	T	G	A
Acsede	A	T	G	A	C	T	C	C	***C***	T	G	A

Figure 2.4 The two SNP positions 6th and 9th in different individuals have difference in nucleotides. At 6th position G/T is SNP and at 9th position A/C is the SNP.

This type of variation is considered as the most common form of variation in human genome as this contributes about 80% of the total variations (Levy, et al., 2007). Any two individuals may differ in their genomes at the frequency of approximately 1 single nucleotide polymorphism in 1.9 kilobases (Sachidanandam, et al., 2001). SNPs are present throughout the genome, irrespective of coding and non-coding DNA (Musunuru, et al., 2010). In the matter of fact that SNPs are also present in non-coding DNA, the further study of SNPs association will be more complex.

CHAPTER 3

DATA SIMULATION AND METHODOLOGY

3.1 Data Simulation

Whole genome case-control study datasets for this work are simulated by GWAsimulator (Li & Li, 2007). GWAsimulator is a based on C++. This program uses user specified disease model to produce whole genome case-control SNPs data. It simulates one causal SNP at each disease locus of the described disease model genotyped Single Nucleotide Polymorphisms chips data on the basis of rapid moving-window algorithm (Durrant, Zondervan, Cardon, Hunt, Deloukas, & Morris, 2004). This program takes phased genotypes as input and the output is based on local linkage disequilibrium (LD) patterns of the input data. For this study we used HapMap project (International Human Genome Sequencing Consortium, 2001) phased genotype of HapMap CEU population sample (Utah residents with Northern and Western European ancestry from the CEPH) which consists 120 phased autosomes for 90 individuals.

The simulation program precisely follows the LD pattern of the input data. For the data generation of 2000 cases and 2000 controls, window size 5 is selected. Seven disease locus are specified, one causal SNP per chromosome, with disease prevalence of 0.1 to 0.01. The information of disease loci like chromosome number, SNP position, disease variant allele, genotypic relative risks and start and end positions is given in Table 3.1. The multiplicative genetic model is used with the relative risk of 1.5. Approximately 1000 to 2000 SNPs are simulated around the causal SNP, which gives the total simulation of around 19000 SNPs.

Table 3.1 Description of Disease Model used for Simulation

Locus	Chromosome Number	SNP Position	Disease Variant Allele	Genotype Relative Risk	Start Position	End Position
1	2	10714	0	1.5	10000	12000
2	6	4322	1	1.5	3000	5000
3	11	9067	1	1.5	8000	10000
4	18	9659	1	1.5	6000	10000
9	19	2885	1	1.5	1000	4000
6	20	3357	0	1.5	1000	5000
7	23	7607	0	1.5	7000	9000

For this study we simulated five training datasets with disease prevalence of 0.1, 0.075, 0.05, 0.025 and 0.01 respectively, with all the parameters same as above specified. Five test datasets are simulated to calculate the disease risk prediction accuracy with all parameters same and respective values of disease prevalence, as of training dataset simulation are used, except that of number of subjects, i.e. 200 cases and 200 controls. The GWAsimulator can provide the data output in three formats namely linkage, genotype and phased data. For this work genotype output format is selected, in which the datasets are kind of matrix where each column represents SNPs and each row represents an individual with genotype 0, 1 and 2, which tells the number of copies of allele 1 (as alleles have two copies per SNP position, “1” = allele 0 and “2” = allele 1). The whole representation of genotypic data is explained in Appendix A.

3.2 Methodology

The two important and challenging problems in Genome wide association studies are prediction accuracy and interpretation. This work is basically focused over the prediction of classification accuracy of the statistical models created by four machine learning algorithms. In this work two-stage testing is applied which was proposed by Van Steen (Steen, et al., 2005) is used. The two-stage testing approach is basically have two statistically independent steps, first is the screening or filtering step and the second is testing or prediction step (Murphy, Weiss, & Lange, 2010). Previous studies shows that the application of two-stage analysis by using Chi-square statistics for SNP ranking i.e. for screening step and then application of other testing methods over highly ranked SNPs, improves the ranking and stability of SNP (Roshan, Chikkagoudar, Wei, Wang, & Hakonarson, 2011). Chi-square statistics is the most commonly applied method over the Genome Wide Association data till yet (Wang, Chen, & Zhang, 2010) (Jewell, 2003). There are lots of other machine learning approaches which have also been applied on the case-control study of Genome Wide Association Studies like classification and regression trees (CART) of (Breiman, 2001) (Uriarte & Andres, 2006) (Roshan, Chikkagoudar, Wei, Wang, & Hakonarson, 2011), Support Vector Machines (SVM) (Vapnik, 1998) (Guyon, Weston, Barnhill, & Vapnik, 2002), Neural Networks (NN) (Bishop, 1995) and many more. Another quality control issue is to control Type 1 error or family-wise error rate in these studies, which occurs due to increment in chance of false discoveries in multiple testing scenarios. There are many methods which have been used to control family-wise error rate in previous studies (Duggal, Gillanders, Holmes, & Bailey-Wilson, 2008) like permutation testing (Dudbridge, 2006), false discovery rate (Benjamini & Hochberg,

1993), Bayesian factors (Marchini, Howie, Myers, Myers, & Donnelly, 2007) and Bonferroni correction (Duggal, Gillanders, Holmes, & Bailey-Wilson, 2008). Among these Bonferroni correction is the most applied method but this has some limitations by considering all the SNPs independent.

In this study 2-df chi-square statistics and holm's procedure is used for the screening step of the two-stage process. The SNPs are ranked with 2-df chi-square statistics with the help of GWAsimulator incorporation of user specific "dataanalysis" function. And then according to results of the application of Holm's procedure top ranked SNPs are screened from each dataset for further statistical analysis. Further, Logistic Regression, Recursive Partitioning and Naïve Bayes Classifier are applied on the screened dataset for the prediction of classification accuracy, at testing step of the study.

3.2.1 Logistic Regression

Logistic regression is the parametric form of statistical methods which has been extensively applied in the field of Genome Wide Association studies (Albert & Zhang, 1984) (Park & Hastie, 2007). Despite of the presence of many methods which can be used as test for association studies, logistic regression proved to be the consistent and reliable method to predict the association of causal variants and phenotype in case-control studies (Nagelkerke, Smits, Cessie, & Houwelingen, 1997). Basically, logistic regression is often used in the presence of dichotomous response variable. The logit function can be described as follows:

$$logit\, p = \log\,(odds) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

In logistic regression structure, binary trees represent prediction models, where leaves signifies the variables used in prediction and nodes of the tree are binary expressions. Logistic regression frames the classifiers by simulating the prognostic combinations of dichotomous variables. Its primary aim is to predict the non-linear and additive interactions among the binary features for prediction (Ruczinski, Kooperberg, & LeBlanc, 2003).

3.2.2 Recursive Partitioning (rpart)/CART

Recursive partitioning is the technique which is adopted by many of the classification algorithms. The Classification and Regression trees method, which are popularly called as CART, is one of the most important among them (Breiman, Friedman, Stone, & Olshen, 1984). Each tree in CART method is based on recursive partitioning principle. Classification and regression trees have been applied to wide range of data mining problems (Hastie, Tibshirani, & Friedman, 2001).

In this thesis work recursive partitioning is applied with the help of routine rpart in R (Therneau & Atkinson, 2011). The rpart routine uses a two stage procedure to structure general classification and regression models. Classification models which are generated by rpart represented as binary tree. In the first stage of the application; the algorithm adopts stepwise procedure to build the complex tree. In the process of building a tree, the splitting criterion is to decrease the risk. Let's say if a node A is split into two nodes B and C, then criteria is described as follows,

$$P(B)r(B) + P(C)r(C) \leq P(A)r(A)$$

The correctness and accuracy of the first stage is predicted by node impurity like Gini index or entropy. The splitting process continues till the daughter nodes cannot be

dividing further (Zhang, Wang, & Chen, 2009). And in the second stage of the application; the algorithm trims back the whole tree by using cross-validation techniques. This is done by sequential regression and stop at when F-test cannot achieve a particular level of significance (α). The best value for α is chose by cross-validation technique.

3.2.3 Naïve Bayes Classifier

Naïve Bayes classifier (Tan, Steinbach, & Kumar, 2006)works on the assumption that feature vector is independent of the class. Even though it has been always observed that assumption of independence proved inefficient, but Naïve Bayes Classifier has given remarkable accuracy in many prognostic applications, like in the field of classification of text, diagnostics in medical field and performance management of systems (Domingos & Pazzani, 1997), (Hellerstein, Thathachar, & Rish, 2000), (Mitchell, 1997). Basically it estimates the conditional probability of class by assuming that features are conditionally independent:

$$P(X/C) = \prod_{i=1}^{n} P(X_i/C)$$

Where, X = $(X_1, \ldots\ldots\ldots, X_n)$, represents the feature vector, n is the number of feature variables in the model and C represents the class (Positive and Negative in binary classification problem). The probability of the feature class is predicted by $P(X/C)$, in this thesis work it is determined by the training datasets of cases and controls. This algorithm assumes that the distribution of variables is normal. The Naïve Bayes classifier calculates the posterior probability of each class, symbolizes as ω_i:

$$p(\omega_i/x) = P(\omega_i) \prod_{j=1,d} p(x_j/\omega_i)/P(x)$$

Where $p(\omega_i/x)$ represents the posterior probability of class ω_i, $p(x_j/\omega_i)$ represents the class-conditional probability of feature j, $P(\omega_i)$ represents the prior probability of the class ω_i, and P(x) represents the prior probability of x. As the prior probability of x is fixed for all the value of ω, the classifier chooses that particular value of class or ω that maximizes the numerator.

Naïve Bayes classifier has its own simple approach to compute the classification, robust to background noise and good in feature selection by disregarding the irrelevant features (Tan, Steinbach, & Kumar, 2006). At the same place, its assumption that each feature set has normal distribution and those features are independent of each other are the disadvantage of Naïve Bayes Classifier.

CHAPTER 4

RESULTS

4.1 Datasets

As described in Chapter 3, in this thesis five training datasets are simulated to train the models of the four classification algorithms. The classification models of the following algorithms, Logistic Regression, Recursive Partitioning and Naïve Bayes are trained on each dataset separately. Each training dataset has 2000 cases and 2000 controls, so total 4000 individuals. From now on these five datasets will be referred as Dataset1, Dataset2, Dataset3, Dataset4 and Dataset5, respectively. Also, five test dataset containing 200 cases and 200 controls are simulated respective to the training dataset. After the first screening stage, the resultant datasets have the following number of SNPs, shown in Table 4.1.

Table 4.1 Resultant Number of SNPs Dataset After the Application Chi-square Statistics and Holms Procedure at the Screening Level

Name of Dataset	*Number of SNPs*
Dataset 1	149
Dataset 2	164
Dataset 3	152
Dataset 4	155
Dataset 5	171

In this thesis work, the classification accuracy is calculated in terms of the area under the ROC (Receiver Operating Curve) on the five datasets separately. The AUC (Area

Under the Curve) values of the test data prediction of classification algorithms at 100%, 75%, 50% and 25 % of top ranked SNPs also compared.

4.1.1 Receiver Operating Curve

Receiver Operating Curve methodology has been applied to many practical problems of classification since 1950 (Green & Swets, 1966) (Metz, 1986). The ROC curve has been proved the best tool to measure the discriminative and classification ability of the algorithm. The ROC curve is a curve between the classification's true positive rate (Sensitivity) and false positive rate (1- Specificity). The ROC accumulates all possible combination of Sensitivity and Specificity, and hence it gives a comprehensive review of a classifier's discriminative accuracy over the whole possibilities of the scenario.

4.1.2 Area under the ROC Curve (AUC)

AUC is most promising indexes among the other summary indexes of the Receiver Operating Curve. The AUC is connected to two most important statistics: Mann-Whitney statistic and P $(x_1 > x_0)$. Where, Mann-Whitney statistic gives a non-parametric way to estimate the area under the ROC curve with the standard error. And, P $(x_1 > x_0)$ defines for AUC as the probability of randomly chosen cases ranked higher than the randomly chosen control subject (Hanley & McNeil, 1982). The most important thing here which makes AUC as most reliable statistic is that it considers average value of True Positive Rate (Sensitivity) over the complete range of False Positive Range (1-Specificity).

4.2 Individual Application of Algorithms

The four machine learning algorithms are applied on the five datasets individually and the prediction accuracy as the AUC value is calculated on the 100%, 75%, 50% and 25% of SNPs. The average AUC value is calculated for all the five datasets at 100 % of SNPs, 75% of SNPs, 50% of SNPs and 25% of SNPs separately to estimates the classification accuracy of algorithms at different number of SNPs. The application of the following tools Logistic Regression, Recursive Partitioning and Naïve Bayes Classifier is done in R. The source code for R is provided in Appendix B. The functions which are used for creating the models are described in the Table 4.2. Table 4.3 lists the R packages required for the each tool and also the common packages for other estimations.

Table 4.2 R Functions used to Create Model on the Training Data

Machine Learning Algorithm	*Function of R*
Logistic Regression	**glm**
Recursive Partitioning	**rpart**
Naïve Bayes Classifier	**naiveBayes**

Table 4.3 R Packages used in the Study

R Packages	*Description*
DESIGN	**Regression Modeling**
e1071	**Misc. Functions of Department of Statistics for Naïve Bayes Classification**
gllm	**Generalized log-linear model**
glm2	**Fitting Generalized Linear Models**
gplots	**Plotting of Data**
gtools	**Basic functionality tools**
MASS	**Support functions and dataset**
ROCR	**Visualizing performance of scoring classifiers**
rpart	**Recursive Partitioning**

4.2.1 Logistic Regression Results

Logistic Regression model classifies the test dataset with fairly high AUC values. Almost all the dataset are following the same pattern in the AUC values for different number of SNPs. It is observed that Logistic Regression gives the highest AUC values at 75% of SNPs. It gives average AUC value of 0.729632 at 100% of the SNPs, 0.738643 which is highest at 75% of SNPs, 0.733541 at 50% of SNPs and 0.706394 at 25% of SNPs. Table

4.4 list the AUC values for each dataset at different number of SNPs with average AUC values. And, figure 4.1 shows the difference in AUC values at different number of SNPs.

Table 4.4 List of all the AUC Values for Logistic Regression

Percentage of SNPs	*DATASET1 AUC VALUES*	*DATASET2 AUC VALUES*	*DATASET3 AUC VALUES*	*DATASET4 AUC VALUES*	*DATASET5 AUC VALUES*	*AVERAGE AUC VALUES*
100	0.7189464	0.7202455	0.7185465	0.7555652	0.7348556	0.729632
75	0.7198683	0.7298855	0.7581559	0.7588455	0.7264598	0.738643
50	0.7184816	0.7250758	0.7411554	0.7593560	0.7236341	0.733541
25	0.6784452	0.7104564	0.7324498	0.7330510	0.6775686	0.706394

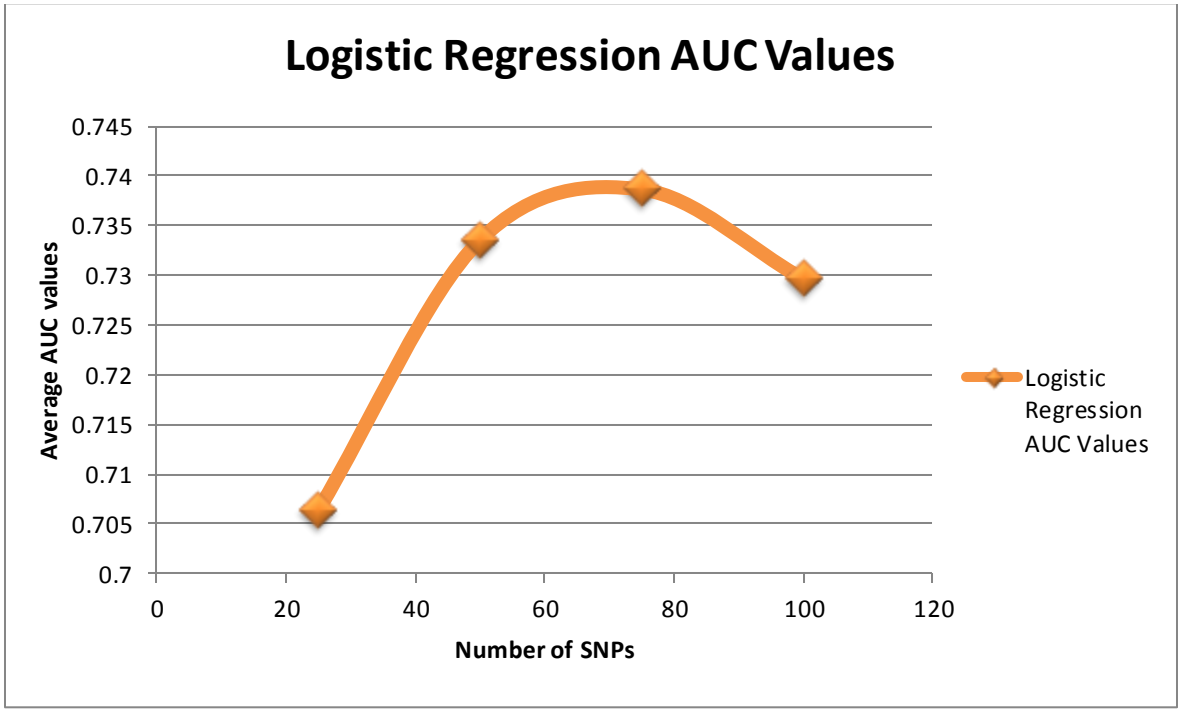

Figure 4.1 Graphical representation of Average AUC values at different number of SNPs for Logistic Regression.

4.2.2 Recursive Partitioning

Recursive Partitioning model classifies the test dataset with fairly high AUC values but comparatively lower than logistic regression. Almost all the dataset are following the same pattern in the AUC values for different number of SNPs. It is observed that Recursive Partitioning gives the highest AUC values at 75% of SNPs. It gives average AUC value of 0.698692 at 100% of the SNPs, 0.711312 which is highest at 75% of SNPs, 0.709135 at 50% of SNPs and 0.689681 at 25% of SNPs. Table 4.5 list the AUC values for each dataset at different number of SNPs with average AUC values. And, figure 4.2 shows the difference in AUC values at different number of SNPs.

Table 4.5 List of all the AUC Values for Recursive Partitioning

Percentage of SNPs	***DATASET1 AUC VALUES***	***DATASET2 AUC VALUES***	***DATASET3 AUC VALUES***	***DATASET4 AUC VALUES***	***DATASET5 AUC VALUES***	***AVERAGE AUC VALUES***
100	0.6262709	0.6836918	0.7187203	0.7085345	0.7562431	0.698692
75	0.6347955	0.7022565	0.7298123	0.7198512	0.7698454	0.711312
50	0.6300086	0.7156654	0.7199965	0.7124651	0.7675412	0.709135
25	0.6095652	0.6802354	0.7100245	0.6984552	0.7501248	0.689681

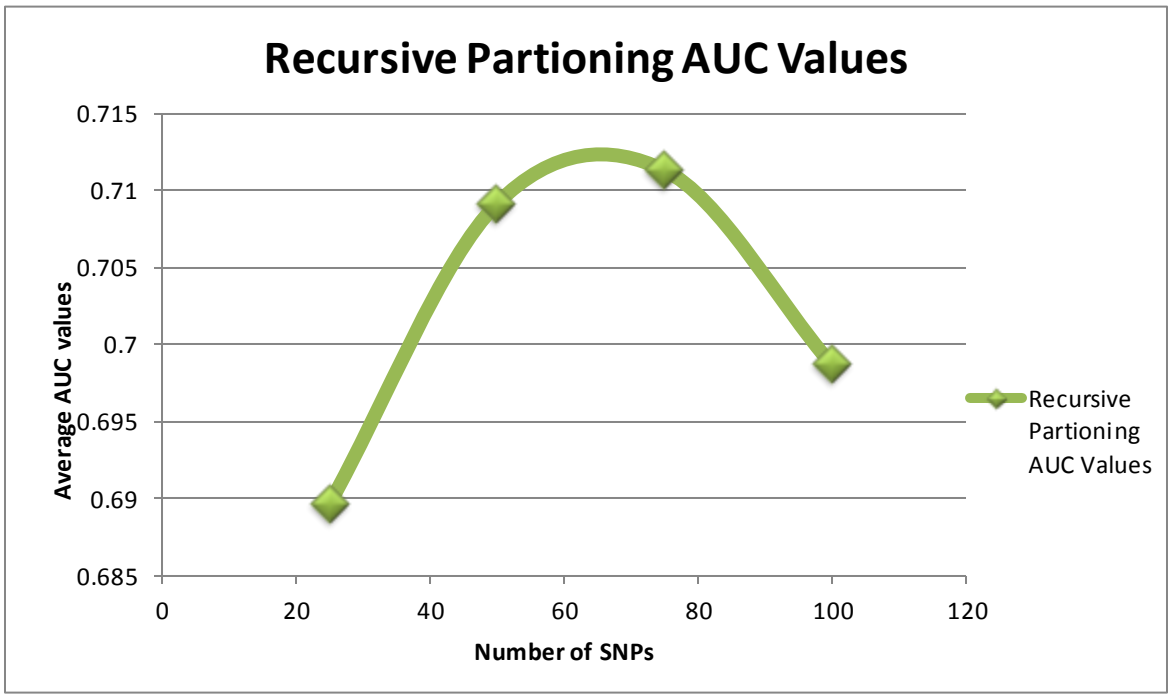

Figure 4.2 Graphical representation of Average AUC values at different number of SNPs for Recursive Partitioning.

4.2.3 Naïve Bayes Classifier

Naïve Bayes Classifier model classifies the test dataset with moderate AUC values which are comparatively lower than Logistic Regression and Recursive Partitioning classification algorithm. Almost all the dataset are following the same pattern in the AUC values for different number of SNPs. It is observed that Naïve Bayes Classifier gives the highest AUC values at 75% of SNPs. It gives average AUC value of 0.53753 at 100% of the SNPs, 0.542118 which is highest at 75% of SNPs, 0.538542 at 50% of SNPs and 0.536684 at 25% of SNPs. Table 4.7 list the AUC values for each dataset at different number of SNPs with average AUC values. And, figure 4.4 shows the difference in AUC values at different number of SNPs.

Table 4.6 List of all the AUC Values for Naïve Bayes Classifier

Percentage of SNPs	*DATASET1 AUC VALUES*	*DATASET2 AUC VALUES*	*DATASET3 AUC VALUES*	*DATASET4 AUC VALUES*	*DATASET5 AUC VALUES*	*AVERAGE AUC VALUES*
100	0.5386914	0.5124555	0.5189625	0.5555958	0.5619432	0.53753
75	0.5421558	0.5247845	0.5581712	0.5588509	0.5266294	0.542118
50	0.5401578	0.5283762	0.5411754	0.5593721	0.5236303	0.538542
25	0.5298722	0.5104656	0.5324214	0.5330846	0.5775748	0.536684

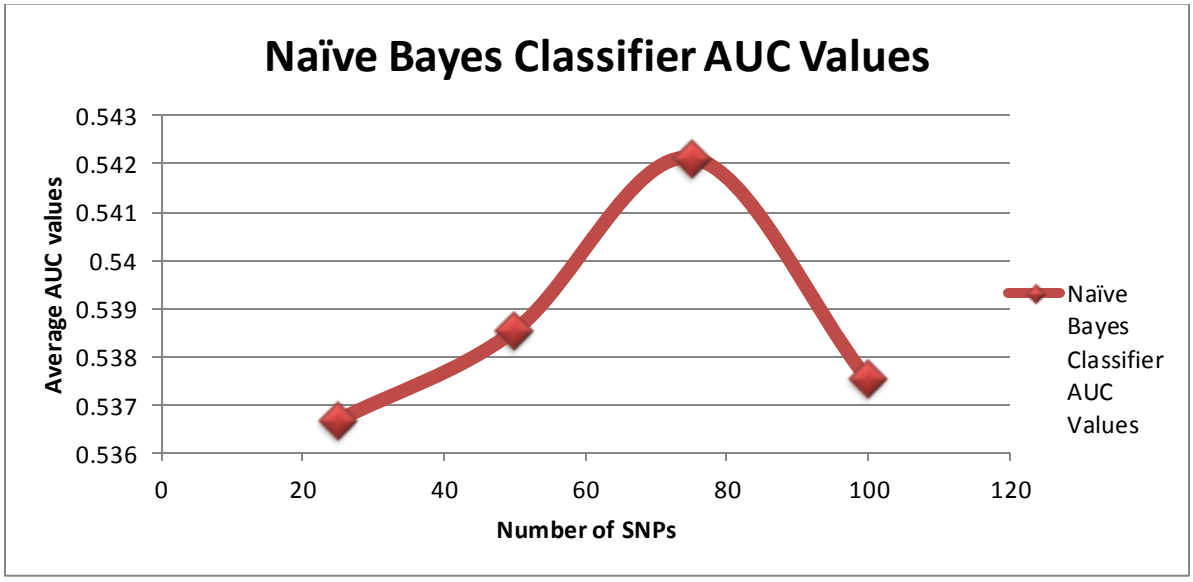

Figure 4.3 Graphical representation of Average AUC values at different number of SNPs for Naïve Bayes Classifier.

4.3 Comparative Analysis of all the four machine learning algorithms

Logistic Regression algorithm performed best among the other four tools which have been used in this thesis work over the simulated dataset. The Recursive Partitioning algorithm is also performed somewhere equivalent to the Logistic Regression. The Logistic Regression got highest value of overall AUC value that is 0.727052; overall AUC value for Recursive Partitioning is 0.702205; and overall AUC value for Naïve Bayes Algorithm is 0.53871853. Table 4.8 lists the values of overall AUC values of the four Machine Learning Algorithms used in this thesis work.

As described earlier and also we can observe it from the figure 4.1, 4.2, 4.3 and 4.4 that average value of AUC for all the four machine learning algorithms peaked at 75% of SNPs. Therefore it shows that all the four classifiers are performing better with a particular number of SNPs. Figure 4.5 shows the overall performance of all the tools over the simulated datasets.

Table 4.7 Overall AUC Values of Four Machine Learning Algorithms

Machine Learning Algorithms	*Overall AUC values*
Logistic Regression	0.727052
Recursive Partitioning	0.702205
Naïve Bayes Algorithm	0.538718

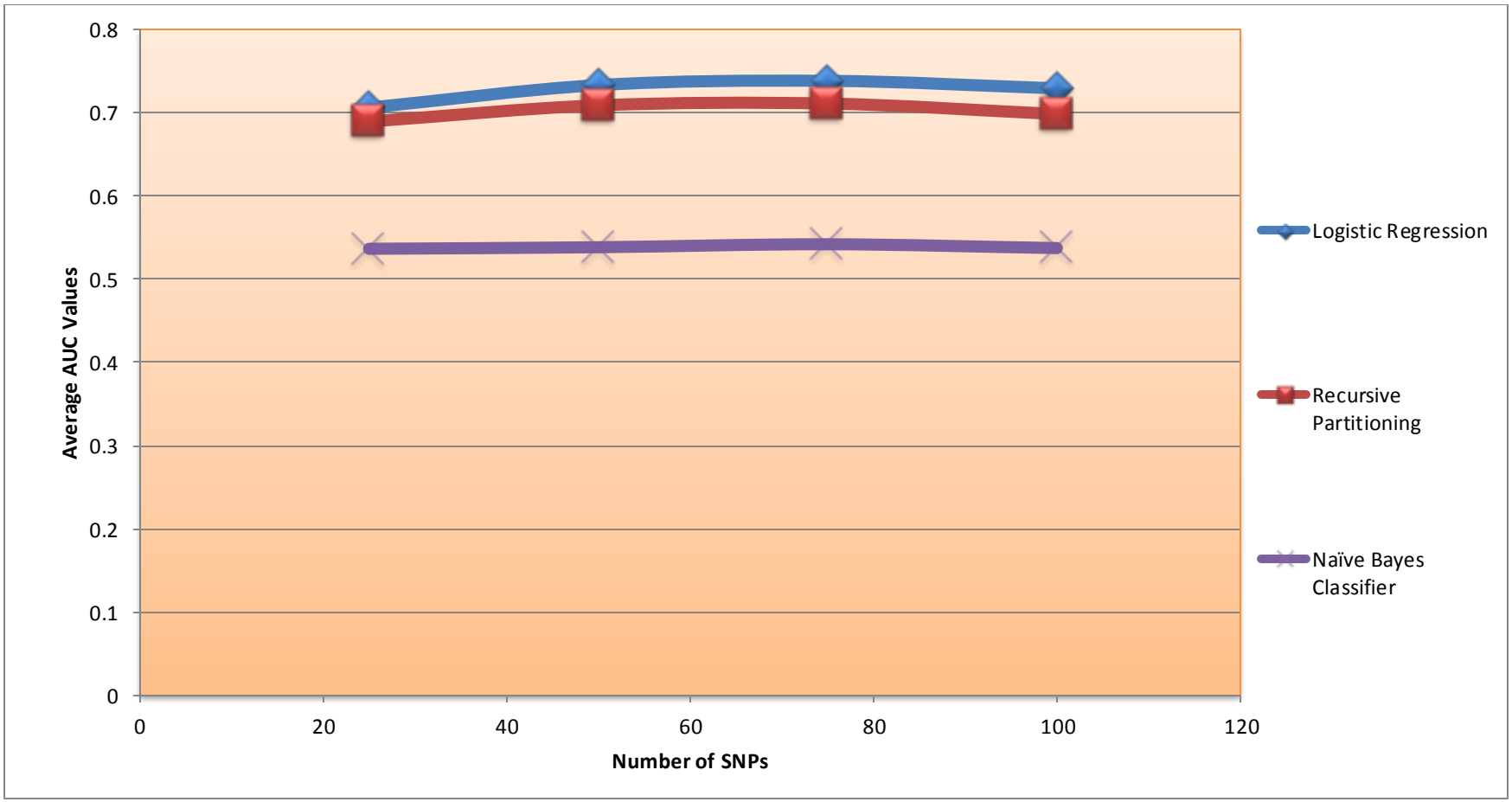

Figure 4.4 Graphical representation and comparison of Average AUC values at different number of SNPs for Logistic Regression, Recursive Partitioning and Naïve Bayes Classifier. This shows that Logistic Regression performed best among the all, and Recursive Partitioning performed almost similar to it. Naïve Bayes Classifier is comparatively lower than the above mentioned two tools.

CHAPTER 5

CONCLUSION

It has been shown that the Logistic Regression using binary model with classification function with a target variable SNPs set is a superior predictor of cases and controls in test dataset as compared with other classification models under study. Logistic regression is more sensitive over the whole range of specificity which is clearly shown by the area under the receiver operating curve. The two-stage testing which is used in this work can be compared to other testing criteria and can be refined by implementation of other features.

This Classification strategy can be tested on the real data to see the classification accuracy in it. And also can be applied to other case-control studies in genetics and medical field to see its performance on class prediction. Also the performance can be elevated by using some better screening techniques and other quality control measures, as we observed the better values of AUC for particular set ranked SNPs. The combination of other screening and testing strategies can be used to improve the classification accuracy.

APPENDIX A

BASICS OF STANDARD NUMERICAL ENCODING OF SNPS AND SNP GENOTYPING

The GWAsimulator uses standard method of encoding of SNP genotypic data is based on the method of Principal Component Analysis (Edwarde, 2003), which is initially applied to genetic data for population stratification (Price, Patterson, Plenge, Weinblatt, Shadick, & Reich, 2006). Genotyping of Single nucleotide polymorphism is the procedure to transform the SNPs alphabetical data to numerical data for statistical, mathematical and computational applications (Gunderson, et al., 2006). The SNP genotypic data is a matrix in which each column is SNP and each is an individual. Each SNP has two copies of alleles represents as first copy is "allele 0" and second copy is "allele 1". Let's assume "allele 0" as "A" and "allele 1" as "B". So, the total possibilities of the combination of alleles at one position are AA, AB and BB.

The main idea behind this conversion of data is that we have to consider SNPs in alphabetical order. Let say if A/B is the SNP name which is in alphabetical order then to change it in numerical data we have to count the number of time B appears in a SNP. Suppose we have several SNPs positions for different subjects in our data for consideration and also we have the SNP name according to their real and replaced nucleotides. This alphabetical name is then transformed to numerical data by counting the number of allele 1 i.e. "B" (which comes later in alphabetical order). The final conversion of the data for all three possibilities is given in the following Table A.1 and Table A.2.

Table A.1 Numerical Encoding of Genomes

Allele Combination	**Numeric Genotype**	**Reason**
AA	**0**	**Number of "allele 1" or B is 0**
AB	**1**	**Number of "allele 1" or B is 1**
BB	**2**	**Number of "allele 1" or B is 2**

Table A.2 Real Time Scenario of SNP Genotyping

SNP Name	A/T	C/T	G/T	…	A/T	C/T	G/T...
Individual 1	AA	TT	GG	…	0	2	0 …
Individual 2	AT	CC	GT	…	1	0	1 …
Individual 3	AA	CT	GT	…	0	1	1 ...

APPENDIX B

SOURCE CODE FOR IMPLEMENTATION OF CLASSIFICATION ALGORITHMS IN R

The following code is the implementation of the Logistic Regression, Recursive Partitioning and Naïve Bayes Classifier in R. Here it is provided for Dataset1 for 100% of SNPs.

B.1 Logistic Regression

DATASET1

```
> train01<-read.table("train1")
>y<-c(rep(0,2000),rep(1,2000))
>names(train01)
>attach(train01)
>train01.logr<-glm(y ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 +
V12 +  V13 + V14 + V15 + V16 + V17 + V18 + V19 + V20 + V21 + V22 + V23 + V24
+ V25 + V26 + V27 +  V28 + V29 + V30 + V31 + V32 + V33 + V34 + V35 + V36 +
V37 + V38 + V39  +V40 + V41 + V42 + V43 + V44 + V45 + V46 + V47 + V48 + V49 +
V50 + V51 + V52 + V53 + V54 + V55 + V56 + V57 + V58 + V59 + V60 + V61 + V62 +
V63 + V64 + V65 + V66 + V67 + V68 + V69 + V70 + V71 + V72 + V73 + V74 + V75 +
V76 + V77 + V78 + V79 + V80 + V81 + V82 + V83 + V84 + V85 + V86 + V87 + V88 +
V89 +  V90 + V91 + V92 + V93 + V94 + V95 + V96 + V97 + V98 + V99 + V100 +
V101 + V102 + V103 + V104 + V105 + V106 + V107 + V108 + V109 + V110 + V111 +
V112 + V113 + V114 + V115 + V116 + V117 + V118 + V119 + V120 + V121 + V122 +
```

```
V123 + V124 + V125 + V126 + V127 + V128 + V129 + V130 + V131 + V132 + V133 +
V134  + V135  + V136  + V137 + V138 + V139 + V140 + V141 + V142 + V143 + V144
+ V145 + V146 + V147 + V148 + V149, family=binomial("logit"))
> test01<-read.table("test1")
 >predictiontest01=predict(train01.logr,test01)
> prediction01<-inv.logit(predictiontest01)
 >ytest01<-c(rep(0,200),rep(1,200))
 >pred01<-prediction(prediction01,ytest01)
 >auc01<-performance(pred01,measure="auc")
 >auc01.75<-performance(pred01.75,measure="auc")
 >auc01.50<-performance(pred01.50,measure="auc")
 >auc01.25<-performance(pred01.25,measure="auc")
```

B.2 Recursive Partitioning

DATASET1

```
> train01<-read.table("train1")
 >y<-c(rep(0,2000),rep(1,2000))
 >names(train01)
 >attach(train01)
>train01.rpart<-rpart(y ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11
+ V12 + V13 + V14 + V15 + V16 + V17 + V18 + V19 + V20 + V21 + V22 + V23 + V24
+ V25 + V26 + V27 +  V28 + V29 + V30 + V31 + V32 + V33 + V34 + V35 + V36 +
V37 + V38 + V39  +V40 + V41 + V42 + V43 + V44 + V45 + V46 + V47 + V48 +
```

V49 + V50 + V51 + V52 + V53 + V54 + V55 + V56 + V57 + V58 + V59 + V60 + V61 + V62 + V63 + V64 + V65 + V66 + V67 + V68 + V69 + V70 + V71 + V72 + V73 + V74 + V75 + V76 + V77 + V78 + V79 + V80 + V81 + V82 + V83 + V84 + V85 + V86 + V87 + V88 + V89 + V90 + V91 + V92 + V93 + V94 + V95 + V96 + V97 + V98 + V99 + V100 + V101 + V102 + V103 + V104 + V105 + V106 + V107 + V108 + V109 + V110 + V111 + V112 + V113 + V114 + V115 + V116 + V117 + V118 + V119 + V120 + V121 + V122 + V123 + V124 + V125 + V126 + V127 + V128 + V129 + V130 + V131 + V132 + V133 + V134 + V135 + V136 + V137 + V138 + V139 + V140 + V141 + V142 + V143 + V144 + V145 + V146 + V147 + V148 + V149, >train01,method="anova")

```
>test01<-read.table("test1")
>pred01rpart<-predict(train01.rpart,test01)
>ytest01<-c(rep(0,200),rep(1,200))
>predict01rpart<-prediction(pred01rpart,ytest01)
>auc01<-performance(predict01rpart,measure="auc")
```

B.3 Naïve Bayes Classifier

DATASET1

```
> train01<-read.table("train1")
 >y<-c(rep(0,2000),rep(1,2000))
>test01<-read.table("test1")
>pred01rpart<-predict(train01.rpart,test01)
 >names(train01)
 >attach(train01)
```

```
>train00.nb<-naiveBayes(y ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 +
V11 + V12 + V13 + V14 + V15 + V16 + V17 + V18 + V19 + V20 + V21 + V22 + V23 +
V24 + V25 + V26 + V27 +  V28 + V29 + V30 + V31 + V32 + V33 + V34 + V35 + V36
+ V37 + V38 + V39  +V40 + V41 + V42 + V43 + V44 + V45 + V46 + V47 + V48 + V49
+ V50 + V51 + V52 + V53 + V54 + V55 + V56 + V57 + V58 + V59 + V60 + V61 + V62
+ V63 + V64 + V65 + V66 + V67 + V68 + V69 + V70 + V71 + V72 + V73 + V74 + V75
+ V76 + V77 + V78 + V79 + V80 + V81 + V82 + V83 + V84 + V85 + V86 + V87 + V88
+ V89 +  V90 + V91 + V92 + V93 + V94 + V95 + V96 + V97 + V98 + V99 + V100 +
V101 + V102 + V103 + V104 + V105 + V106 + V107 + V108 + V109 + V110 + V111 +
V112 + V113 + V114 + V115 + V116 + V117 + V118 + V119 + V120 + V121 + V122 +
V123 + V124 + V125 + V126 + V127 + V128 + V129 + V130 + V131 + V132 + V133 +
V134  + V135  + V136  + V137 + V138 + V139 + V140 + V141 + V142 + V143 + V144
+ V145 + V146 + V147 + V148 + V149 + V150 + V151 + V152 + V153 + V154 + V155
+ V156 + V157 + V158 + V159 + V160 + V161 + V162 + V163,train00)
> predict01nb<-predict(train01.nb,test00,type="raw")
> predictiontest01nb<-predict01nb[,1]
> prediction01nb<-inv.logit(predictiontest01nb)
>ytest01<-c(rep(0,200),rep(1,200))
> pred01nb<-prediction(prediction01nb,ytest01)
> auc01nb<-performance(pred01nb,measure="auc")
> auc01nb
```

REFERENCES

From genome to proteome. (2008, March 26). Retrieved February 12, 2012, from **U.S. DOE Human Genome Project**, http://www.ornl.gov/hgmis/home.shtml

SNP Fact Sheet. (2008, Sepetember 19). Retrieved February 14, 2012, from **U.S. DOE Human Genome Project**: http://www.ornl.gov/hgmis/home.shtml

Albert, A., & Zhang, L. (1984). **On the existence of maximum likelihood estimates in logistic regression models**. *Biometrika*, 1-10.

Altmuller, J., Palmer, J. L., Fischer, G., Scherb, H., & Wjst, M. (2001). **Genomewide scans of complex human diseases: true linkage is hard to find**. *American Journal of Human Genetics*, 936-950.

Baker, M. (2010). **Genomics: The search for association**. *Nature*, 1135-1138.

Balakrishnama, S., & Ganapathiraju, A. (1998). ***Linear discriminant analysis- A brief tutorial.*** Retrieved 2012, from http://www.isip.msstate.edu/publications/reports/isip_internal/1998/linear_discrim_analysis/lda_theory.pdf

Barlow-Stewart, K. (2004). ***Genes and Chromosomes***. Retrieved 2012, from The Centre for Genetics Education.: www.genetics.edu.au

Barnhart, B. J. (1989). **DOE Human Genome Program**. *Human Genome Program, U.S. Department of Energy, Human Genome News (v1n1).*

Barrett, J. C., & Cardon, L. R. (2006).**Evaluating coverage of genome-wide association studies**. *Nature Genetics*, 659-662.

Bellman, R. (1961). ***Adaptive control processes: A guided tour***. Princeton, N.J.: Princeton University Press.

Benjamini, Y., & Hochberg, Y. (1993). **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *JSTOR*, 289-300.

Bishop, C. M. (1995). **Neural networks for pattern recognition**. *Clarendon Press-Oxford*.

Breiman, L. (2001). **Random Forests**. In *Machine Learning* (pp. 5-32).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). ***Classification and regression trees.***

Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). **Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies**. *Pacific Symposium On Biocomputing*, 368-379.

Costello, T. J., Falk, C. T., & Ye, K. Q. (2003). **Dataminig and computationally intensive methods: Summary of group 7 contributions to genetic analysis workshop 13**. *Human Genetics*, 57-63.

Dahm, R. (2004). **Friedrich Miescher and the discovery of DNA**. *Developmental Biology*, 274-288.

Davenport, G. C., & Davenport, C. B. (1908). **Heredity of hair-form in man**. *The American Naturalist*, 341-349.

Deegan, R. C. (1989). **The alta summit, December 198**4. *Genomics*, 661-663.

Domingos, P., & Pazzani, M. (1997). **On the optimality of the simple bayesian classifier under zero-one loss**. In *Machine Learning* (pp. 103-130). Netherland: Kluwer Academic Publishers.

Dudbridge, F. (2006). **A note on permutation tests in multistage association scans**. *The American Journal of Human Genetics*, 1094-1095.

Duggal, P., Gillanders, E. M., Holmes, T. N., & Bailey-Wilson, J. E. (2008). **Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies**. *BMC Genomics*, 516.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., & Morris, A. P. (2004). **Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes**. *American Journal of Human Genetics*, 35-43.

Edwarde, J. J. (2003). ***A user's guide to principal components***. New York: John Wiley & Sons.

European Consortium for IDDM genome Studies. (2001). **A genome-wide scan for type 1 diabetes susceptibility genes in Scandinavian families. Identification of new loci with evidence of interaction**. *American Journal of Human Genetics*, 1301-1313.

Feero, G. W., Guttmacher, A. E., & Collins, F. S. (2010). **Genomic medicine — An updated primer.** *The New England Journal of Medicine*, 2001-2010.

George, A. (2008). **Multi-modal biometrics human verification using LDA and DFB**. *International Journal of Biometric and Bioinformatics*.

Gert, B. (1996). **Morality and the new genetics: A guide for students and health care providers**. Boston, MA: Jones and Bartlett, Publishers.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., & Willis, T. D. (2003). **The International HapMap projec**t. *Nature*, 789-796.

Green, D. M., & Swets, J. A. (1966). ***Signal detection theory and psychophysics***. New York: Wiley.

Gunderson, K. L., Steemers, F. J., Ren, H., Ng, P., Zhou, L., Tsan, C., et al. (2006). **Whole-genome genotyping.** *Methods Enzymol*, 359-376.

Guttmacher, A. E., & Collins, F. S. (2003). **Welcome to the genomic era.** *The New England journal of Medicine*, 996-998.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). **Gene selection for cancer classifiaction using support vector machines.** *Machine learning,1-3.*

Hanley, J. A., & McNeil, B. J. (1982). **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology*, 29-36.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). ***The elements of sStatistical learning: data mining, inference, and prediction***. New York: Springer.

Hellerstein, J., Thathachar, J., & Rish, I. (2000). **Recognizing end-user transactions in performance management.** *Proceedings of AAAI-2000*, (pp. 596-602). Austin, Texas.

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T. H., et al. (2006). **A common genetic variant is associated with adult and childhood obesity.** *Science*, 279-283.

Hershey, A., & Crick, F. (1952). **Independent functions of viral protein and nucleic.** *General Physiology*, 39-56.

Hindorff, L. A., MacArthur, J., Wise, A., Junkins, H. A., Hall, P. N., Klemm, A. K., et al. (2011). ***A catalog of published genome-wide association studies***. Retrieved March 13, 2012, from National Human Genome Research Institute: http://www.genome.gov/gwastudies/

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). **A comprehensive review of genetic association studies**. *Genetics in Medicine*, 45-61.

Holden, A. L. (2002). **The SNP consortium: summary of a private**. *BioTechniques*, 22-26.

International Human Genome Sequencing Consortium. (2001). **Initial Sequencing and analysis of the human genome.** *Nature*, 860-921.

Jewell, N. P. (2003). ***Statistics for epidemiology***. New York: Chapman & Hall.

Jio, J. H., & Levan, A. (1956). **The chromosome number of man.** *Hereditas*, 1-6.

Johnson, A. D., & O'Donnell, C. J. (2009). **An open access database of genome-wide association results**. *BMC Medical Genetics*, 1471-2350-10-6.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). **Mapping and sequencing of structural variation from eight human genomes.** *Nature*, 56-64.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). **Complement factor- H polymorphism in age-related macular degeneration.** *Science*, 385-389.

Kropff, M. J. (2008). ***Statistical applications in nutrigenomics***. Wageningen: Wageningen University, Netherlands.

Lander, E. S. (2011). **Initial impact of the sequencing of the human genome**. *Nature*, 187-197.

Levene, P. (1915). **On Chondrosamine**. *Proceedings of the National Academy of Sciences of USA*, 190-191.

Levy, S., Sutton, G., Ng, P. C., Halpern, A. L., Walenz, B. P., Axelrod, N., et al. (2007). **The diploid genome sequence of an individual human**. *PLoS Biology*, 254.

Li, C., & Li, M. (2007). **GWAsimulator: a rapid whole-genome simulation program.** *Bioinformatics*, 140-142.

Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). **A HapMap harvest of insights into the genetics of common disease**. *The Journal of Clinical Investigation*, 1590-1605.

Marchini, J., Howie, B., Myers, S., Myers, G., & Donnelly, P. (2007). **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nature Genetics*, 906-913.

Metz, C. E. (1986). **ROC methodology in radiologic imaging**. *Invest Radiology*, 720-733.

Mitchell, T. M. (1997). ***Machine Learning.*** McGraw-Hill.

Murphy, A., Weiss, S. T., & Lange, C. (2010). **Two-stage testing strategies for genome-wide association studies in family-based designs.** In *Statistical Methods in Molecular Biology* (pp. 485-496). Clifton, NJ: Springer Protocols.

Musunuru, K., Strong, A., Kamenetsky, M. F., Lee, N. E., Ahfeldt, T., Sachs, K. V., et al. (2010). **From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus**. *Nature*, 714-719.

Nagelkerke, N., Smits, J., Cessie, S. L., & Houwelingen, H. V. (1997). **Testing goodness-of-fit of the logistic regression model in case–control studies using sample reweighting**. *Statistics in Medicine*, 121-130.

National Human Genome Reseach Institute. (2011). ***Deoxyribonucleic Acid.*** Retrieved 2012, from National Human Genome Reseach Institute: http://www.genome.gov/25520880

Noble, D. (2008). **Genes and causation**. *Philosofical transactions of The Royal Society A: Mathematical, Physical & Engineering Sciences*, 3001-3015.

Ott, J., Kamatani, Y., & Lathrop, M. (2011). **Family based designs for genome-wide association studies**. *Nature Reviews Genetics*, 465-474.

Ott, J., Schrott, H. G., & Goldstein, J. L. (1974). **Linkage Studies in a large kindred with familial hypercholesterolemia.** *American Journal of Human Genetics*, 598-603.

Ozaki, K., & Ohnishi, Y. (2002). **Functional SNPs in the lymphotoxin-alpha gene that are associated with suceptibility to myocardial infarction**. *Nature Genetics*, 650-654.

Painter, T. E. (1924). **The sex chromosomes of man**. *The American Naturalist*, 506-524.

Park, M. Y., & Hastie, T. (2007). **Penalized logistic regression for detecting gene interactions.** *Biostatistics*, 30-50.

Pearson, H. (2006). **Genetics: What is a gene?** *Nature*, 398-401.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). **Principal components analysis corrects for stratification**. *Nature Genetics*, 904-909.

Ridley, M. (2006). ***Genome***. New York: Harper Perennial.

Roses, A. (2003). **The genome era begins**. *Nature Genetics*, 217.

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., & Hakonarson, H. (2011). **Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest**. *Nucleic Acids Research*, 1-8.

Rotimi, C. N., & Jorde, L. B. (2010). **Ancestry and disease in the age of genomic medicine**. *The New England Journal of Medicine*, 1551-1558.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). **Logic Regression**. *Journal of computaional and graphical statistics.*

Sachidanandam, R., Weissman, D., Schmidt, S., Kakol, J., Stein, L., Marth, G., et al. (2001). **The IntA map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature*, 928-933.

Shields, R. (2011). **Common Disease: Are Causative Alleles Common or Rare?** *PLOS Biology*,1.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). **ROCR: visualizing classifier performance in R**. *Bioinformatics*, 3940-3941.

Steen, V. K., McQueen, M. B., Herbert, A., Rosenow, C., Silverman, E. K., Laird, N. M., et al. (2005). **Genomic screening and replication using the same data set in family-based association testing**. *Nature Genetics*, 683-691.

Stein, L. D. (2004). **Human genome: End of the beginning**. *Nature*, 915-916.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). ***Introduction to Data Mining.*** Addison-Wesley Press, Minneapolis, MN.

The Internation SNP Map Working Group. (2001). **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature*, 928-933.

The International HapMap 3 Consortium. (2010). **Integrating common and rare genetic variations in diverse human populations.** *Nature*, 52-58.

The International HapMap consortium. (2007). **A second genration human Haplotype map over 3.1 million SNPs.** *Nature*, 851-861.

The International HapMap Project. (2002, October). ***About the Hap Map***. Retrieved January 2012, from International HapMap Project: http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en

Therneau, T. M., & Atkinson, E. J. (2011). ***An introduction to recursive partitioning using the RPART routines.*** Mayo Foundation, Rochester, MN.

Thorisson, G. A., & Stein, L. D. (2003). **The SNP Consortium website: past, present and future**. *Nucleic Acids Research*, 124-127.

Thornton-Wells, T. A., Moore, J. H., & Haines, J. L. (2006). **Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data.** *BMC Boiinformatics*,204.

Uriarte, R. D., & Andres, A. D. (2006). **Gene selection and classification of microarray data using random forest**. *BMC Bioinformatics*, 7-11.

Vapnik, V. N. (1998). **Statistical learning theory**. *Wiley, New York,NY.*

Ventor, C. J., Adams, M. D., Myers, E. W., Li, P. W., & Mural, R. J. (2001). **The Sequence of the human genome**. *Science*, 1304-1351.

Wang, M., Chen, X., & Zhang, H. (2010). **Maximal conditional chi-square importance in random forests**. *Bioinformatics*, 831-837.

Watson, J. D., & Crick, F. H. (1953). **Genetical implications of the structure of deoxyribonucleic acid**. *Nature*, 964-967.

Wooster, R., & Weber, B. L. (2003). **Breast and Ovarian Cancer**. *The New England Journal Of Medicine*, 2339-2347.

WTCCC. (2007). **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature*, 661-678.

Xie, J., Cai, T. T., Maris, J., & Li, H. (2010). **False Discovery rate control for high dimensional dependent data with an application to large-scale genetic associations.** *Annals of applied Statistics*.417-430.

Zhang, H., Wang, M., & Chen, X. (2009). **Willows: a memory efficient tree and forest construction package**. *BMC Bioinformatics*, 130-131.