AN INTEGRATIVE APPROACH TO STRUCTURED SNP PRIORITIZATION AND REPRESENTATIVE SNP SELECTION FOR GENOME-WIDE ASSOCIATION STUDIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÜRKAN ÜSTÜNKAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
INFORMATION SYSTEMS

JANUARY 2011

Approval of the Graduate School of Informatics

_____

Prof. Dr. Nazife Baykal
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

_____

Prof. Dr. Yasemin Yardımcı Çetin
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

_____

Prof. Dr. Gerhard-Wilhelm Weber
Co-Supervisor

_____

Assist. Prof. Dr. Yeşim Aydın Son
Supervisor

Examining Committee Members

Prof. Dr. Yasemin Yardımcı Çetin (METU, II) _____

Assist. Prof. Dr. Yeşim Aydın Son (METU, II) _____

Assist. Prof. Dr. Tolga Can (METU, CENG) _____

Assist. Prof. Dr. Zeynep Kalaylıoğlu (METU, STAT) _____

Assist. Prof. Dr. Süreyya Özöğür Akyüz (BAHÇEŞEHİR, MATH) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Gürkan Üstünkar

Signature:

ABSTRACT

AN INTEGRATIVE APPROACH TO STRUCTURED SNP PRIORITIZATION AND REPRESENTATIVE SNP SELECTION FOR GENOME-WIDE ASSOCIATION STUDIES

Üstünkar, Gürkan

Ph.D., Department of Information Systems

Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

Co-Supervisor: Prof. Dr. Gerhard-Wilhelm Weber

January 2011, 149 pages

Single Nucleotide Polymorphisms (SNPs) are the most frequent genomic variations and the main basis for genetic differences among individuals and many diseases. As genotyping millions of SNPs at once is now possible with the microarrays and advanced sequencing technologies, SNPs are becoming more popular as genomic biomarkers. Like other high-throughput research techniques, genome wide association studies (GWAS) of SNPs usually hit a bottleneck after statistical analysis of significantly associated SNPs, as there is no standardized approach to prioritize SNPs or to select representative SNPs that show association with the conditions under study. In this study, a java based integrated system that makes use of major public databases to prioritize SNPs according to their biological relevance and statistical significance has been constructed. The Analytic Hierarchy Process, has been utilized for objective prioritization of SNPs and a new emerging methodology for second-wave analysis of genes and pathways related to disease associated SNPs based on a combined p-value approach is applied into the prioritization scheme. Using the subset of SNPs that is most representative of all SNPs associated with the diseases reduces the required computational power for analysis and decreases cost of

following association and biomarker discovery studies. In addition to the proposed prioritization system, we have developed a novel feature selection method based on Simulated Annealing (SA) for representative SNP selection. The validity and accuracy of developed model has been tested on real life case control data set and produced biologically meaningful results. The integrated desktop application developed in our study will facilitate reliable identification of SNPs that are involved in the etiology of complex diseases, ultimately supporting timely identification of genomic disease biomarkers, and development of personalized medicine approaches and targeted drug discoveries.

Keywords: Biomarkers Discovery, SNP-Complex Disease Association, Representative SNP Selection, SNP Prioritization, Pathway Discovery

ÖZ

GENOM BOYUTUNDA İLİŞKİLENDİRME ÇALIŞMALARINDA YAPILANDIRILMIŞ SNP ÖNCELİKLENDİRMESİ VE TEMSİLCİ SNP SEÇİMİ İÇİN BÜTÜNLEŞİK BİR YAKLAŞIM

Üstünkar, Gürkan

Doktora, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Yard. Doç. Dr. Yeşim Aydın Son

Ortak Tez Yöneticisi: Prof. Dr. Gerhard-Wilhelm Weber

Ocak 2011, 149 sayfa

Tek Nükleotit Polymorfizmi (SNP) en yaygın olarak görülen genom çeşitliliği ve kişiler arasındaki genetik farklılıkların ve birçok hastalığın temel nedenidir. Günümüzde milyonlarca SNP genotipinin bir seferde belirlenmesi mikro dizilim ve ileri sekanslama teknolojileri sayesinde mümkündür. Yüksek verimli teknolojilerin kullanıma girmesi ile SNPler gözde biyolojik göstergeler arasına girmiştir. SNPlerin genom boyutunda ilişkilendirme çalışmalarında istatistiksel analiz safhasından sonra gerek SNPlerin önceliklendirilmesinde gerekse temsilci bir SNP kümesinin seçilmesinde mevcut standart bir yöntem olmadığı için bir darboğazla karşılaşılmaktadır. Bu çalışmada, SNPleri biyolojik ve istatistiksel önemlerine göre önceliklendirilmesinde kullanılabilecek java tabanlı bütünleşik bir sistem oluşturulmuştur. SNPlerin objektif olarak önceliklendirilebilmesi için Analitik Hiyerarşi Prosesinden yararlanılmıştır. Ek olarak birleşik p-değeri yöntemine dayanan ve hastalıkla ilgili SNPlerle ilişkili gen ve biyolojik yolakların ikincil analizinde yeni kullanılmaya başlanan bir yöntem de önceliklendirme esnasında kullanılmıştır. Hastalıkla ilişkili bütün SNPler için temsilci bir SNP alt kümesinin seçilmesi daha az kapasiteli bilgisayarlarla ve daha ucuza analiz ve ardıl ilişkilendirme ve biyolojik gösterge bulma işlemlerinin yapılabilmesini sağlar. Bu amaçla temsilci SNP seçimi için Benzetilmiş Tavlama algoritmasına dayalı yenilikçi bir algoritma

geliştirilmiştir. Geliştirilen yöntemlerin geçerliliği ve doğruluğu gerçek bir vaka-kontrol çalışmasından alınan verilerle kontrol edilmiş, biyolojik olarak anlamlı sonuçlara ulaşılmıştır. Bu çalışmada geliştirilen bütünleşik masaüstü uygulamasının karmaşık hastalıklarla ilişkili SNPlerin güvenilir bir biçimde belirlenmesinde büyük rolü olacaktır ve ileride hastalıkla ilişkili biyolojik göstergelerin ortaya çıkarılmasını ve kişiselleştirilmiş ve hedefe yönelik ilaçların keşfedilmesini destekleyeceğini öngörmekteyiz.

Anahtar Kelimeler: Biyolojik Gösterge Bulma, SNP-Karmaşık Hastalık İlişkilendirmesi, Temsilci SNP Seçimi, SNP Önceliklendirmesi, Biyolojik Yolak Bulma

To My Family

ACKNOWLEDGMENTS

During the course of this study, I changed my research topic three times, read dozens of articles that did not have any contribution to this thesis work, completed my military service, got divorced after seven years, changed the city I lived in and started over at the age of 30. As a result, I thought of quitting many times but at the end, being one of the most important achievements in my life, I completed this research work. It means a lot to me, much more than a degree and I am grateful to many people for making this possible.

First of all, I am thankful to Prof. Dr. Nazife Baykal. She has supported me from the very beginning of this study and she never lost her confidence in me "no matter what". She is one of the reasons I changed my mind on quitting, because I was indebted to her. I am also very much indebted to my supervisor Assist. Prof. Dr. Yeşim Aydın Son. Without her support, interest and guidance I would have been lost in this research topic, about which I have no previous relevant background, from day one. I will never forget our Google Talk sessions on the research, which took place as late as 2 a.m. many nights. I admire her dedication to work and she has been and she will be a role model for me as a business professional. Next, I thank Assist. Prof. Dr. Süreyya Özöğür-Akyüz. It has been one of the luckiest coincidences in my life to get the chance to meet her. Her motivation for success has been an inspiration for me and she put me back in track and helped me to regain my motivation for research during possibly the worst times of my life. I am also grateful to my co-supervisor Prof. Dr. Gerhard-Wilhelm Weber, dear Willi, for his continuous support and guidance, which helped me a lot to get the work done. I, many times, found myself thinking that he cared possibly more than me about my success and I will never ever forget this. Next, I need to express my gratitude to my examining committee members Assist. Prof. Dr. Tolga Can and Assist. Prof. Dr. Zeynep Kalaylıoğlu. Their positive approach and feedbacks have encouraged me a lot to continue on the thesis work and shaped the structure of my thesis. Inputs from H. Alper Döm, Burak Demiralay, Onat Kadıoğlu and Yener Tuncel made the AHP based analysis more reliable. I thank them and wish them success during their academic endeavors. I also thank Prof. Dr. Uğur Sezerman for introducing the SNPs as a research topic.

I am also grateful to my family, for being there for me every time I need. I thank my parents for devoting their life on me and and my sister. I thank Aslı, my little sister, for bringing joy to my life and for making me laugh even if I am in the worst mood possible. I am also deeply grateful to Deniz, my love, who has given me the motivation to start over to, literally, everything.

Last, I thank my managers at Vestel White, especially Mr. Gökmen Çeşmeci, who supported me for getting this degree and showed tolerance every time I need. I also gratefully acknowledge the support of Genetic Analysis Workshops organization committee and ADNI investigators for sharing the data sets used in this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

1.1 Motivation

Bioinformatics is the science of analyzing, extracting, and interpreting information from biological sequences and molecules. Upon completion of human genome project in April 2003, this area has drawn even more attention. With more projects undertaken, such as sequencing projects of other genomes and mapping of human haplotype of genomic variations, data in the field is exponentially growing. Facing this huge amount of data, the biologist cannot simply use the traditional analysis techniques in biology. In order to extract useful and meaningful information from this precious data, instead, information technologies are needed.

One of the recent areas that draw attention in bioinformatics field is to try to find a solution to the problem of identifying genetic variations that are the major reasons of complex diseases. Our understanding of the genetic etiology of human disease is limited because of the massive number of genetic variations on the human genome, as well as the complex relationships between multiple genes and environmental factors underlying disease. An association study is a widely used and potentially very powerful strategy for finding genetic and environmental risk factors causing human diseases. The underlying principle is a very simple one: if a DNA variant or an environmental factor increases disease susceptibility, it is expected to be observed more frequently among those who are affected than those who are not. An association is said to be found between two variables if it can be shown that their values are correlated, i.e. they are dependent on each other.

With the completion of Human Genome Project in 2003, it is now possible to convey research studies to associate genetic variations in the human genome with common and complex diseases. The approach in these studies is to scan markers within DNA, or genomes, of many people for finding genetic and environmental risk factors underlying human diseases and quantitative traits. In an association analysis, the aim is to identify a particular variant whose alleles or haplotypes are observed significantly more frequent in cases than controls therefore

signalling an association with disease phenotype. A large percentage of association studies concentrate on candidate genes that are thought to be biologically relevant to a disease through linkage analysis. Main downside of candidate gene approach is that its success mainly depends on proper selection of candidate genes by biological judgment. Apart from these, somewhat biased ways of associating genetic markers with complex diseases, another approach is to use "most of the genome" to investigate variants. These studies are called **Genome-wide Association Studies** (GWAS).

SNPs are the most frequently observed genetic variations and they are heavily used in GWAS alongside with microsatellites, copy number variation (CNV) and other markers. The studies related to other markers than SNPs are beyond the scope of this work. Recent developments in genotyping technologies, ability to access genetic information held within public databases and the start of the International HapMap Project [1, 2] have facilitated the implementation of SNP based genome-wide association studies. The main objective of the HapMap was to genotype SNPs with sufficient density across the human genome, eventually achieving a resolution of one SNP in every one to two kilobases. With the help of this, the scientific community is now able to collect valuable genome sequence and SNP data and gain insights on etiology and pathogenesis of various complex genetic diseases, which would in turn have great effects through drug development.

Number of SNPs and required sample size of individuals are relatively larger for a typical research study, in which genetic factors associated with complex diseases are investigated because of the contribution of multiple genes on the disease phenotype. This is not unusual as the number of SNPs under study and genotyped individuals have a significant effect on the statistical power of the study. However this is not an easy task because genotyping that much SNPs and individuals is not only very expensive but also suffers from various problems such as missingness, accuracy etc. It is therefore desirable to select a SNP subset, which is free from biological and statistical redundancy.

In this research, we have developed two novel methods for prioritizing and selecting SNPs consecutively for stronger association with complex disease after following an integrative biological scoring and filtering approach for supporting Genome-Wide Association Studies. A java based integrated system, **Most Efficient Tagging Utility for SNPs** (METU-SNP), which makes use of major public databases to prioritize SNPs according to their biological and statistical relevance has been constructed. The Analytic Hierarchy Process (AHP), a well-known Multi-Criteria Decision Making Method, has been utilized for objective prioritization of SNPs. Along with the integrated prioritization system; a novel feature selection method based on Simulated Annealing (SA) has been developed for informative SNP selection. In this chapter a brief overview of the dissertation work will be presented.

1.2 Prioritization and Representative SNP Selection

The human genome can be represented as an array of 3.3 billion letters and each letter is from the set of {A, C, G, T} representing nucleotides Adenine, Cytosine, Guanine and Thymine. The nucleotide sequence does not differ across the population in more than 99% of the positions on the genome. However, people possess a unique genetic composition in about 1% of their genome [3]. Those genetic variations include different nucleotide occurrences, called (if occurred at least 1% of a given population) Single Nucleotide Polymorphisms (SNPs – pronounced 'snips'), deletion/insertion of one or more nucleotides, or variations in the number of multiple nucleotide repetitions.

Recent research on molecular epidemiology has focused on finding genetic biomarkers that contribute to disease and act as decent predictors of the disease phenotypes. Usually multiple genetic loci are involved in complex disease development as well as environmental factors. Therefore, one needs to search through all polymorphisms present in the biologically more relevant (functionally important) regions of candidate genes as well as utilizing the information about biological pathways (such as cellular adhesion, inflammation, lipid metabolism, etc.), which can be regarded as gene networks, constituting biological systems of great importance for thorough analysis of these kinds of diseases [4-6].

Association studies are among promising ways of dealing with the problem of finding disease causing variants and such association studies typically make use of SNPs as they are the most common form of genetic variations and they can represent an individual's genetic variability in greatest detail [7]. However, the enormous number of SNPs (estimated more than 11 million) makes it infeasible to gather information and perform analysis on all the SNPs in the human genome. Therefore in order to save resources and make problem computationally feasible while keeping the power of the statistical tests at acceptable levels genetics researcher would prefer working on a "considerably smaller" subset of the entire SNP set for performing an association study. In order to come up with such a SNP subset it is required to first prioritize the SNPs according to well defined criteria so that informative SNPs are not overlooked. An intelligent way of performing prioritization task is needed to calculate a score for each SNP under study that would reflect SNP's biological and statistical relevance and rank SNPs accordingly. Following prioritization task, feature selection mechanisms should be applied to select the SNP subset with best prediction performance. In summary, finding a subset of SNPs that is informative enough to perform association studies but still small enough to reduce the analysis workload, to which we refer as **prioritization and representative SNP selection**, has become an important step for disease-gene association studies.

1.3 Thesis Statement and Contributions of the Thesis

The ultimate goal of this research is to advance the state-of-the-art in SNP-complex disease association studies by introducing an integrated system that would be used for Genome-Wide Association Studies (GWAS), case-control studies in particular. We aim to provide researches with a tool that would enable them to prioritize and select representative SNPs among massive SNP sets for subsequent GWAS. By integrating pathway based analysis and feature selection schemes we believe that we will enable the comprehensive analysis of SNP genotype data. We expect the new methods, introduced throughout this thesis, to provide an advanced SNP selection framework for facilitating disease-gene association studies. We hope that results of this research will support timely diagnosis, personalized treatments, and targeted drug design, through facilitating reliable identification of SNPs that are involved in the etiology of common and complex diseases. In the following sections we briefly introduce key contributions of this research study.

1.3.1 AHP Based Prioritization Scheme for Multi-Hierarchical Filtering of Informative SNPs after Genome-Wide Association Study

A typical approach in a Genome Wide Association Study (GWAS) is to repeat the study on an independent sample by genotyping the top signals. It is evident that this approach would be facilitated by incorporating information from biological databases so as to give SNPs that has the potential to possess biological relevance higher priorities. We introduce a method that would be used to achieve this strategy in a structured way by combining information on genomic location, biological significance and evolutionary conservation alongside with statistical evidence of genotype–phenotype correlation. Our approach is based on the well known multi-criteria decision making method called Analytic Hierarchy Process (AHP) [8], which is a structured technique for dealing with complex decisions.

Selecting functionally and biologically important SNPs associated with a particular complex disease is set as a goal and we create a hierarchy tree to compare SNPs accordingly. Our AHP study is combined with a Delphi technique [9] as we perform the study with inputs from five different biologists. We have created a hiearchy tree that structurized SNP prioritization process. We calculate scores for each SNP depending on *p*-values of association and we combine this information with SNP's biological importance depending on three different criteria: (1) Evolutionary Conservation, (2) Gene Association and (3) Genomic Location. In order to calculate scores for SNPs we integrate information from major databases listed in Section 1.2.1. Our method is a novel one as it integrates pathway based information in SNP prioritization process and makes use of a very well known technique, AHP, to the best of our knowledge for the first time in this domain.

1.3.2 Simulated Annealing Based Feature Selection Scheme for Representative SNP Selection

In a disease association study, the goal is to identify genetic factors that contribute to disease. The common and basic approach to finding genotype-phenotype associations is to apply statistical hypothesis-testing procedures. Another, somewhat more advanced, approach is making use of classification models, in which the different genotypes for a SNP can be regarded as inputs (attributes) and the phenotype (disease under study) can be regarded as the class to be predicted. Various machine learning techniques, such as logistic regression or support vector machines can be used for this purpose. Feature selection procedures are thought to be useful to enhance the prediction performance of multi-SNP models and are suggested for reducing the dimension of highly-correlated SNPs [10].

One can define the problem of feature selection in Representative SNP selection context as to find a minimum set of SNPs $S = \{ SNP_1 , ..., SNP_k \}$ out of n SNPs in a particular genomic location (a chromosome or haplotype) such that $k < n$ and S can predict the remaining unselected SNPs with minimum error. To tackle this problem, we introduce a Simulated Annealing (SA) based Informative SNP selection algorithm in which we create a random binary coding of size n as an initial solution and test the accuracy of the solution using Naive Bayes by calculating the mean error for $n - k$ supervised learning iterations. We use a trade of between accuracy and the number of SNPs in the SNP set. Therefore, we also try to minimize the number of chosen SNPs k.

1.3.3 METU-SNP: An Integrated Software System for SNP-Complex Disease Association Analysis

There are various tools for use in GWAS and analysis of SNP genotype data for Case-Control studies such as PLINK [11], BEAGLE [12], SNPTEST [13], GENABEL [14] etc. Among those tools, although being really capable, some of them are far from being user-friendly and based on Command Line Interface (CLI). Others are introduced as desktop applications or web based applets but usually lacks the functionality for a comprehensive analysis or presented as R packages, which are far from being efficient. Having analyzed the available tools, we decided to come up with an Integrated Software System, which is specifically designed for use in GWAS and equipped with only the necessary analysis schemes. We call this tool METU-SNP and we believe that as the name implies it will be regarded as **Most Effective Tagging Utility** among researchers of molecular epidemiology.

METU-SNP is a java based desktop application: it can be used in various platforms. It make use of data from major public databases such as dbSNP [15], Entrez Gene [16], KEGG [17], Gene Ontology [18] etc. It is equipped with a state-of-the-art novel SNP prioritization and Gene Set Enrichment Analysis frameworks. It also has the ability to select informative SNPs by making use of machine learning algorithms. Ability of comprehensive SNP-complex disease

association analysis, modularity and extendibility gives METU-SNP a great advantage over existing platforms.

1.4 Experiment Data Sets

During this research study, one of the major drawbacks has been finding proper data sets for evaluating the performance of the suggested analysis methods and algorithms. As the proper data for analysis consists of DNA data collected from individuals, it has usually been regarded as classified and therefore related parties have shown reluctance to share the data. One of major milestones for this study, therefore, has been achieved when we get a hand on first data DVD, which has been shared with us with the courtesy Dr. Jean W. Maccluer, head of Genetics Department of Southwest Foundation for Biomedical Research.

We used two data sets for evaluation purposes. The first data set is whole genome association Rheumatoid Arthritis (RA) data from the North American Rheumatoid Arthritis Consortium (NARAC) including 868 cases and 1,194 controls. The data was used in Genetic Analysis Workshop 16 (GAW 16). It consists of 501,463 SNP-genotype fields from the Illumina 550K chip. The second data set, a relatively smaller one compared to the first one, is whole genome association data for Alzheimer's disease (AD). The data was obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database[1]. The used subset of the ADNI data includes 149 AD cases and 182 controls. It consists of 555,850 SNP-genotype fields from the Illumina 610Quad chip. We used the former data set to test the performance of METU-SNP in Chapter 6. The latter one has been used in Chapter 3 and 4 to test the performance of the algorithms developed.

1.5 Thesis Organization

This thesis is composed of seven main chapters and an Appendix. Brief contents are given below:

- **Chapter 1** introduces the problem we aim to tackle during this research study and outlinesmajor contributions of this dissertation.
- **Chapter 2** provides biological background relevant to disease-association studies.
- **Chapter 3** depicts the fundementals of AHP and presents our SNP prioritization process.
- **Chapter 4** starts with an extensive literature review on Representative SNP Selection methods and introduces our Simulated Annealing based selection scheme.
- **Chapter 5** explains our integrated software application METU-SNP in detail.
- **Chapter 6** provides experimental results and throws performance of our contributions into sharp relief.

[1] http://www.loni.ucla.edu/ADNI

- **Chapter 7** concludes the dissertation work and outlines possible directions for future research.

CHAPTER 2

BIOLOGICAL BACKGROUND

In this chapter a brief biological background on SNPs and why they are associated with complex diseases will be discussed. In particular the basic concepts in genetics and molecular epidemiology that are relevant to SNP-Complex Disease Association studies will be explained. First a brief overview on human genome will be presented. Then, variations on human genome and SNPs will be introduced and how SNPs influence a person's health by introducing the types of SNPs and their functional effects will be explained. Following that, we will explore an important feature of human genome called Linkage Disequilibrium (LD). Next, complex diseases and association studies that map SNPs and those diseases will be addressed. Lastly, the introduction to Biological Pathways will be given and how this concept is related to complex disease association studies will be explained.

2.1 Human Genome: Set of Instructions for Life

Human genome can be regarded as the set of instructions for life. It consists of chemical molecules (nucleotides) acting as building blocks. A nucleotide is composed of a nucleobase from the set {Adenine, Cytosine, Thymine, Guanine} (See Figure 2.1- used with the permission of [19]), phosphate groups (number changes from one to three) and a five-carbon sugar (deoxyribose). Together, nucleotides form a linear chain, a long continuous molecule, containing roughly 3.2 billion chemical bases, which is known as deoxyribonucleic acid (DNA). Nucleotides contain two types of nucleobases: purines (Adenine (A), Guanine (G)) and pyrimidines (Cytosine (C), Thymine (T)). Specific hydrogen bonds are formed between the bases (A pairs with T and C pairs with G) so as to stabilize the structure of the DNA, which is called "the double helix". Since hydrogen bonds are not covalent they can be broken and rejoined easily hence giving DNA a unique dichotomious feature.

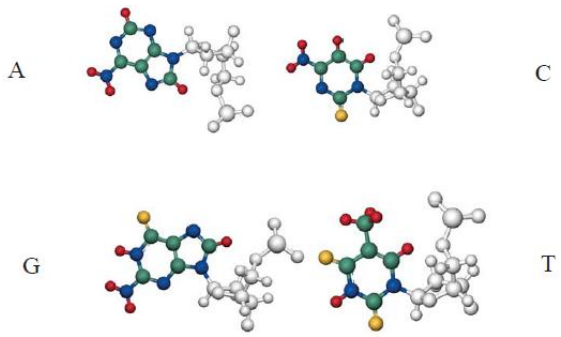


Figure 2. 1 Structures of DNA building blocks A, C, G, and T.

Double helix structure consists of two strands arranged so that paired bases meet in the middle and backbones run up the outside edges. This special structure shown in Figure 2.2 (used with the permission of [20]) is the key to pairing of DNA, which forms the basis for protein encoding and gene functioning. The DNA encodes the proteins and defines who we are biologically. Except for red blood cells, human genome is located in nucleus of cells in the body in which it is organized into 46 chromosomes.

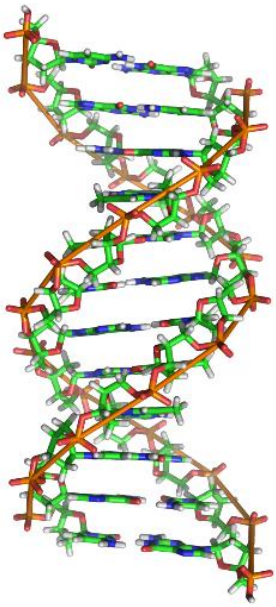


Figure 2. 2 Double helix structure of DNA.

The genetic information in a genome is held within order of nucleotides, which constitutes the genomic sequence, and the complete set of this information in an organism is called its genotype. Only about 3 percent of the information human genome carries is actually translated into biologically active molecules. These regions are called coding regions, or in other words genes, and they are scattered throughout the chromosomes. A gene is a unique DNA sequence within a chromosome that ultimately directs the building of a specific protein with a specific function that influences a particular characteristic in an organism. Close to each gene is a regulatory sequence of DNA, which is able to turn the gene on or off. Remaining 97 percent of

the genome, areas which are known as noncoding regions or junk DNA, has no defined function yet.

Transition from DNA to proteins, also known as central dogma of molecular biology, follows a specific path of molecular processes: Transcription – Splicing – Translation (See Figure 2.3 - used with the permission of [21]). Transcription is the first step leading to gene expression. During transcription, RNA complementary to DNA sequence of the coding region is created. Only one strand of the DNA corresponding to that gene (known as the template strand) is copied into an RNA molecule. During the second step, which is called splicing, a modification of an RNA after transcription, in which introns (non-coding sections within a gene that is not translated into protein) are removed and exons (contains part of the open reading frame that codes for a specific portion of the complete protein) are joined. During the last step messenger RNA (mRNA) produced in transcription is translated into a specific amino acid chain, called polypeptide that will later form an active protein.

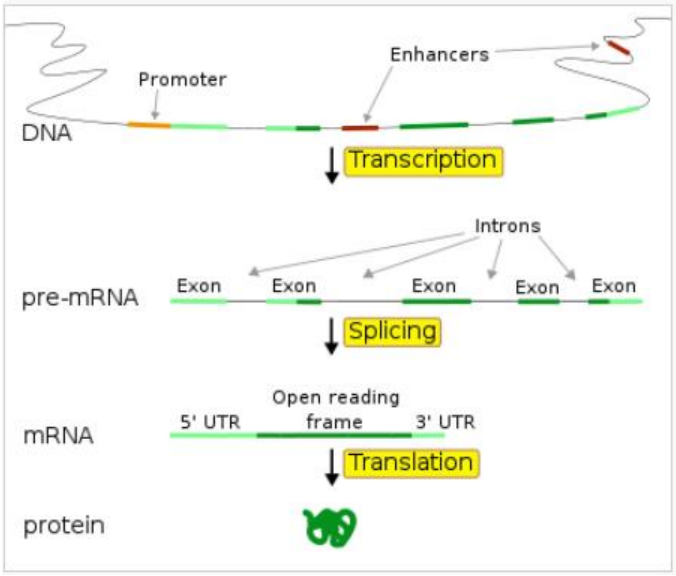


Figure 2. 3 Typical eukaryotic protein-coding gene.

There are twenty different amino acid building blocks used to make proteins in human cell. Three bases in a row (called codon) are “read” to specify a single amino acid. Each codon specifies the incorporation of one and only one amino acid. There are a total of 64 codon combinations but the total number of amino acids is only 20. So we can incur that some amino acids such as methionine can be encoded by only one codon, AUG, but some amino acids such as leucine can be designated by as many as six different codons [19]. The codon AUG is the start codon and codes for an amino acid called methionine that always signals the start of protein building. UAG, UGA, and UAA are stop codons and do not code for any amino acid. Instead, they act as a signal to stop the transcription of DNA into RNA.

Whenever the order of the bases in the DNA sequence changes a variation occurs. Variations can involve only one base or many bases. How they affect the biological functioning can vary from no alteration in biological activity to serious malfunctioning that can be the

molecular basis of a disease. Combined variations in particular sections of the human genome alongside with environmental effects can be regarded as the reason for complex diseases as they may trigger folding of different proteins than expected or proteins functionality may be affected.

2.2 Mutations and Variations in the Human Genome

Compared to the huge size of the genome, variation in the DNA sequence only accounts for a small portion. When the genome of one person is compared to that of another, of the 3.2 billion bases, roughly 99.9 percent are the same. It is the variation in the remaining 0.1 percent of the genome that makes a person unique. This small amount of variation determines attributes such as how a person looks or tendency to develop complex diseases.

A mutation is a stable and heritable change in the genome of an organism, which refers to a change in the sequence of the DNA. Mutations include changes from one base pair to another (for example, A-T to G-C), deletion of one or more base pairs, or insertion of one or more bases. Depending on where mutations occur, they may affect if and when, where, how much a gene will be transcriped and translated into a protein and if it will code for an altered product.

There are many mutations that accumulate on genome, changing the DNA sequence. These mutations can be classified into 5 types depending on their effect on protein sequence [22]:

- **A frameshift mutation** is caused by insertion or deletion of a number of nucleotides that disturbs the triplet nature of a codon, therefore causing mistranslation of the reading frame that code for the protein sequence. Frameshift mutations result in a completely different translation from the original protein sequence.
- **A nonsense mutation** occurs when a premature stop codon or a nonsense codon is coded as a result of an exchange of a single nucleotide for another in the transcribed mRNA. These mutations often result in truncated and nonfunctional protein products.
- **A missense mutation** (nonsynonymous mutation) is a type of mutation caused by change of a single nucleotide, which results in substitution of a different amino acid in the protein sequence. This would result in protein malfunction.
- **A neutral mutation** occurs in the coding region of a gene, which results in the use of a different, but chemically similar, amino acid in the protein sequence. The similarity between the two is enough that little or no change occurs in the protein structure or function.
- **A silent mutation** does not result in a change of the amino acid sequence of a protein. They may occur in a non-coding region or within a codon but they do not alter the final amino acid sequence.

2.3 Single Nucleotide Polymorphism

Among 3.2 billion nucleotides that build our genome, there is only a 0.1% difference or variation between two randomly selected individuals. These variations arise from mutations and are called **polymorphisms** if observed in more than 1% of a given population. The simplest form of DNA variation among individuals is the substitution of one single nucleotide for another at a homologous site in a population. This type of change is called **single nucleotide polymorphism** (SNP-pronounced snip) and they classify 90% of all variations seen between individuals (See Figure 2.4 - used with the permission of [23]). The nucleotide at a position in which a SNP occurs is called an **allele**. The allele with the dominant occurrence within a population is called the major allele, while those occurring less frequently are called the **minor alleles** [24].

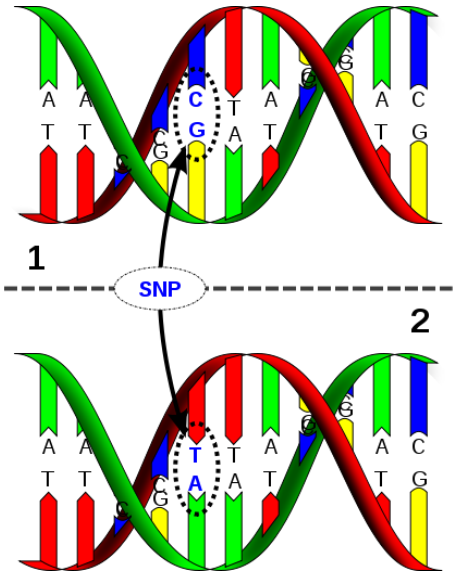


Figure 2. 4 A C/T polymorphism between two DNA molecules.

One may expect to observe a single-nucleotide difference between two haploid genomes in the range of one in every 100–300 base pairs. This means that entire human genome hosts approximately 10 to 30 million potential SNPs. National Center for Biotechnology Information's (NCBI) current SNP variation database (dbSNP build 131) holds more than 14 million validated SNPs out of estimated 30 million potential SNPs within the human genome [25]. As of September 2010 over 1.6 million common SNPs between 11 different populations have been genotyped within the scope of International HapMap Project (HapMap3) [26]. Although the vast majority of SNPs are shared between populations [27], it is evident that many are specific to populations or continental grouping of populations that share recent history.

SNPs exist both within coding regions and non-coding regions of the entire human genome. SNPs present in non-coding regions can occur in intergenic regions, intronic regions, 5'-UTR (untranslated) or 3'-UTR regions, and associated non-coding regions such as transcription factor binding sites, promoters, upstream and downstream sequences. Those falling within

coding regions can be further categorized into two groups: synonymous (silent) and nonsynonymous (see Table 2.1 for various types) [28]. Both coding and non-coding SNPs can be utilized as genomic biomarkers. If a particular SNP is located near a gene (linked to the gene), every time that gene is inherited, the SNPs are passed along with the linked allele of the gene. This enables researchers to assume that when they find the same SNP profiles in a group of individual genomes, the associated allele for the gene is also present.

A synonymous SNP's impact on gene function is mostly via effect on mRNA splicing. This is well documented in the literature [29, 30] and can result in distinct phenotypes, underlying many diseases [31]. The structure of mRNA is crucial for its functioning and it is mainly determined by its primary nucleotide sequence. SNPs that perturbates mRNA stability and translation also affects gene expression [32]. Additionally, it has been shown that SNPs have an influence on protein folding and ultimate protein function [33].

Table 2. 1 Frequency of various types of single nucleotide polymorphisms (NCBI Database, Human Genome).

Nucleotide change	Coding synonymous	Coding nonsynonymous	Intron	Splice site	mRNA UTR	Flanking 2000 up/500 down	Outside coding regions	Total
A-C	3,701	8,855	438,464	62	13,992	26,104	810,028	**1,301,206**
A-G	29,049	30,783	1,694,999	176	48,035	91,977	2,928,878	**4,823,897**
A-T	1594	4,924	360,280	40	9,361	18,367	695,762	**1,090,328**
C-G	4,416	11,620	447,12	68	14,732	30,048	753,613	**1,261,617**
C-T	31,706	28,366	1,692,412	326	48,426	93,000	2,929,112	**4,823,348**
G-T	3,329	9,018	440,620	98	13,170	25,745	810,560	**1,302,540**
A/C/G	75	220	2485	4	130	247	5,885	**9,046**
A/C/T	60	114	2166	0	108	191	5,570	**8,209**
A/G/T	65	146	2047	3	86	193	5,556	**8,096**
C/G/T	78	160	2444	0	135	255	5,806	**8,878**
All others (including deletions)	1,735	13,542	1,328,009	300	34,160	74,582	1,806,707	**3,259,035**
Total	**75,808**	**107,748**	**6,411,225**	**1,077**	**182,361**	**360,737**	**10,757,498**	**17,896,454**

The complete list of functional types for SNPs that has potential for deleterious effects are listed below [34].

SNPS on coding region:

- **Non-synonymous SNPs (also known as missense mutation)** - SNPs that are located in protein coding regions and lead to an amino acid change in an encoded protein sequence.
- **Synonymous SNPs (also known as silent mutation)** - SNPs that are located in protein coding regions, but do not result in a change of an amino acid sequence.
- **Frameshift variations (also known as nonsense mutation)** - SNPs that are located in protein coding regions, and result in a frameshift.
- **Stop lost (also known as nonsense mutation)** - SNPs that are located in protein coding regions, and result in the loss of a stop codon.

- **Stop gained (also known as nonsense mutation)** - SNPs that are located in protein coding regions, and result in the gain of a stop codon. As a result, these SNPs lead to a curtailed protein sequence.

  SNPS on non-coding region:
- **Essential splice site** - SNPs that are located in the first two or the last two base pairs of an intron.
- **Splice site** - SNPs that are located in 1-3 base pairs into an exon or 3-8 base pairs into an intron.
- **Upstream variations** - SNPs that are located within a 5 kb (kilo base) upstream region of the 5-prime end of a transcript.
- **Regulatory region variations** - SNPs that are located in regulatory regions, annotated by Ensembl or dbSNP.
- **5-prime UTR variations** - SNPs that are located in the 5-prime untranslated region (UTR).
- **Intronic variations** - SNPs that are located in an intron.
- **3-prime UTR variations** - SNPs that are located in the 3-prime UTR.
- **Downstream variations** - SNPs that are located within a 5 kb downstream region of the 3-prime end of a transcript.
- **Intergenic variations** - SNPs that are located more than 5 kb either upstream or downstream of a transcript.

2.4 Block Structure of Human Genome

For each of the SNPs on a chromosome, every individual has two alleles, one on the paternal chromosome and the other on the maternal chromosome. Each of consecutive SNPs present on the same chromosome for a particular individual is referred to as a haplotype. It is a set of closely linked alleles (SNPs) inherited as a unit. If we assume that there is a gene with three SNPs and represent two alleles of each SNP with A and B, the number of possible combinations of the three SNPs is eight (as listed here: A-A-A, A-A-B, A-B-A, A-B-B, B-A-A, B-A-B, B-B-A, B-B-B). However, the total number of common (i.e., frequency > 5%) haplotypes in the population usually much less than eight, for example, it can be three (A-A-A, A-B-B, B-B-A). The number of common haplotypes varies depending on the chromosome regions. Due to high cost and lengthy processing time it is difficult to distinguish the paternal and maternal origin of each allele (therefore haplotypes) for long DNA sequences (Even most efficient methods are limited to 10 to 20 kilobase pairs of DNA) [35]. Instead, the usual procedure is to simply associate the two alleles with the SNP position without determining which one of the two chromosomes carries which allele. The combined allele information for a particular locus is called a genotype, and the experimental procedure used for extracting the genotype information is called genotyping. Figure 2.5 (used with the permission of [36]) represents the difference

between haplotypes and genotypes, where gray boxes indicate major allele and black boxes indicates minor allele for haplotypes.

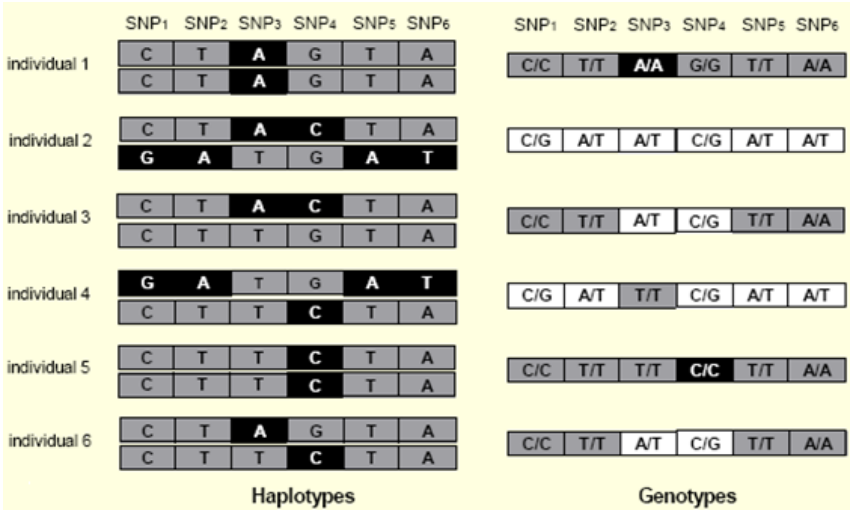


Figure 2. 5 Haplotypes and genotypes.

In a population, certain combinations of genomic variations, like SNPs, may occur with a different frequency than would be expected from a random distribution of haplotypes based on their allelic frequencies. This observation has been described as Linkage Disequilibrium (LD), which is the result of the non-random association of alleles at two or more loci, on the same or different chromosomes. Basically LD measures the difference between observed and expected random distributions for allelic frequencies, thus non-random associations between polymorphisms for different loci is determined.. To give a mathematical definition of the degree of LD, let SNP_1 and SNP_2 be two SNPs at under study. Let $|SNP_1|$ and $|SNP_2|$ represent the cardinality of allele set for corresponding SNPs. Let s_{1i} denote the i^{th} allele of SNP_1, and s_{2j} denote the j^{th} allele of SNP_2. If SNP_1 and SNP_2 were to be independent then we would say the joint probability of s_{1i} and s_{2j} to occur for all i, j is $Pr(s_{1i}, s_{2j}) = Pr(s_{1i}).Pr(s_{2j})$. When this equation does not hold, say if $Pr(s_{1i}, s_{2j}) = Pr(s_{1i}).Pr(s_{2j}) + D$, we conclude that two alleles are not independent and consequently they are in a state of LD and degree of LD is denoted by D[37]. Several measures of pairwise LD are routinely used when describing marker–marker correlation and are central to SNP tagging. The two most commonly used are D' (standardized LD coefficient, D) and r^2 (correlation coefficient) [38].

In general, SNPs physically close to each other are hypothesized to be in high LD, as the recombination probability is higher where two SNPs are distant to each other [39]. Therefore, SNPs close to each other tend to be inherited together to descendants from their ancestor. This phenomenon results in their alleles often being highly correlated with each other, and therefore number of unique haplotypes containing these SNPs is considerably smaller than one may expect under the independence assumption as explained above.

2.5 Complex Diseases

Genetic disorders are caused by the unusual characteristics of genes or chromosomes within certain individuals. Those disorders, which can be attributed to observable defects in chromosomes and mutations that negatively affect specific function of a single gene are classified as Mendelian diseases (over 1,500 detected). These are fairly rare diseases and can be usually identified by specific patterns of transmission: dominant, recessive, sex linkage [40]. Various other chronic diseases usually show different characteristics than Mendelian family patterns and they are thought to be caused by interplay of multiple genes, as well as involvement of environmental factors. Among such diseases are hypertension, coronary heart disease, obesity, diabetes, Parkinson's disease, Alzheimer's disease, Epilepsy, various cancers, and others [41]. These diseases are called complex diseases and occur commonly in the population and are a major source of disability and death worldwide. Complex diseases are characterized by the following features [42]:

- Multiple genes are thought to be involved,
- Caused also by both environmental factors and behaviors that elevate the risk of disease,
- Susceptibility alleles have a high population frequency,
- Non-Mendelian.

As genetics studies fastly approach to completing the human genome project, an exciting era with a great potential to associate DNA sequence variation with disease susceptibility comes into play. SNPs are the most common of all DNA sequence variations. SNPs may occur within coding sequences of genes, noncoding regions of genes, or in the regions between genes. Although the vast majority of the SNPs are found in noncoding regions of the genome and most of the SNPs found in coding regions are silent or have subtle functional effects, SNPs are thought to be the basis for much of the genetic variation found in humans from physical appearance to susceptibility to disease.

In the following subsections we will explain two complex diseases that are studied for GWAS in this thesis work: Alzheimer's Disease (AD) and Rheumatoid Arthritis (RA).

2.5.1 Alzheimer's Disease

AD is an incurable, degenerative and terminal disease (fewer than three percent of individuals live more than fourteen years after diagnosis) characterized by the deposition into the brain of amyloid peptide, which originates a cascade of inflammatory events leading to neuronal death (see Figure 2.6 - used with the permission of [43]). It is expected that in year 2050 1 of every 85 individuals will suffer from this disease. AD usually slowly progresses and results in memory loss, alterations of emotional stability, higher intellectual function, and cognitive abilities [44].

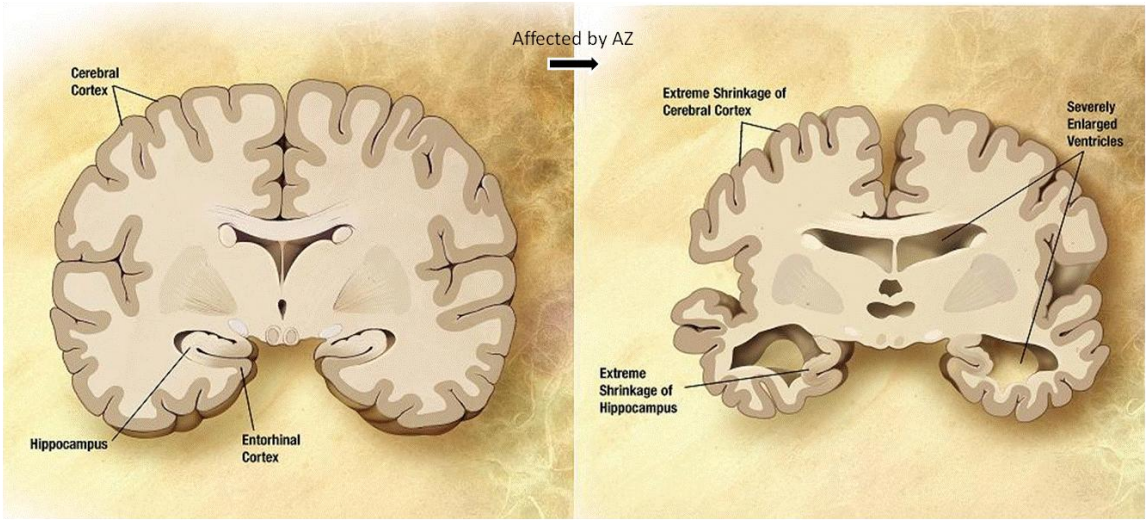


Figure 2. 6 Comparison of a normal aged brain and an Alzheimer's patient's brain.

AD has very complex genetic roots involving both gene–gene and gene–environment interactions. As for the environmental risk factors that would be regarded as major causes of AD, we can count injuries related with head. Serious head injuries increases future risk of Alzheimer's, especially when trauma occurs repeatedly or involves loss of consciousness. Additionally, the risk of developing AD is thought to be increased by many conditions that damage the heart or blood vessels. One can count diabete**s**, heart disease, high blood pressure, stroke, and high cholesterol among such conditions.

Among genetic causes, scientists have discovered variations that directly cause AD in the genes coding three proteins: amyloid precursor protein (APP), presenilin-1 (PS-1) and presenilin-2 (PS-2). Also, several risk genes implicated in Alzheimer's disease have been identified in past years. The risk gene with the strongest influence is called apolipoprotein E-e4 (APOE-e4) [45].

2.5.2 Rheumatoid Arthritis

RA is a chronic and systemic inflammatory disorder. It is a major cause of disability affecting about 1 per cent of world's population (women three times more often than men). This disability results from changes in the anatomy and functioning of the synovial joints (See Figure 2.6 - used with the permission of [46]) caused by autoreactive tissue-destructive immune response(s). The main symptoms individuals with RA report are pain, stiffness, joint swelling and fatigue. As the disease progresses joint tissue may become permanently damaged. The combined effects of inflammation and joint damage result in progressive disability.

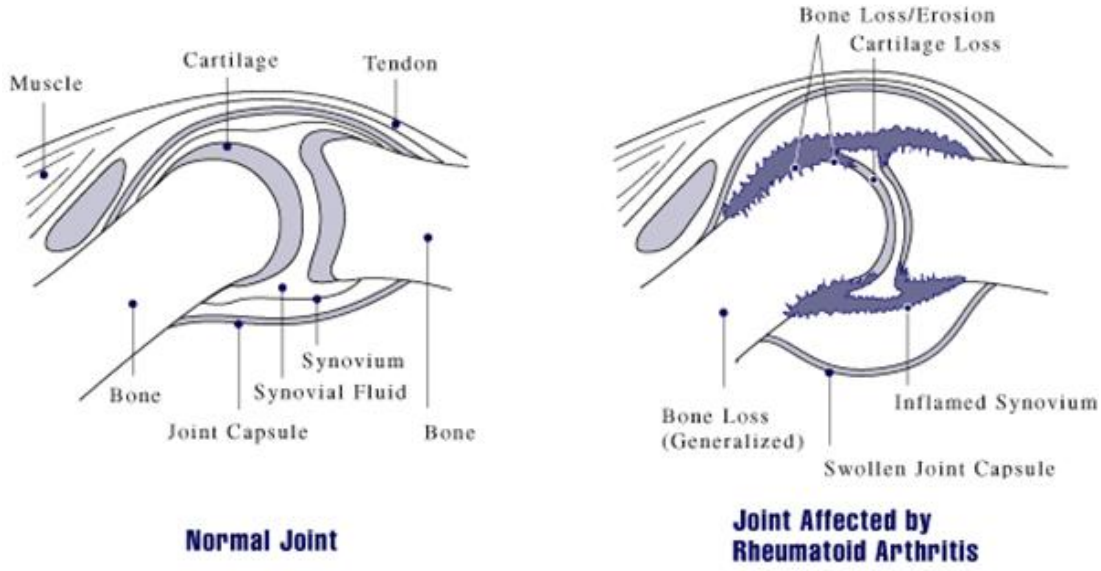


Figure 2. 7 Effect of RA on a joint.

The exact causes of RA are, as yet, unknown. It is generally considered an auto-immune disease, but factors that trigger this destructive response of the body's immune system are not clear. Some research evidence supports exposure to infection as a trigger for RA (e.g. Epstein–Barr virus, parvovirus and some bacteria such as Proteus and Mycoplasma). In [47], it is suggested that conditions that effects level of female hormones such as using oral contraceptive pill or pregnancy may be somewhat protective for developing RA; for example, the use of the and pregnancy are both associated with a decreased risk. However, it is also possible that such phenomena are secondary to the exposure with RA [48]. It is however clear that triggered T-cells, which are normally part of the body's adaptive immune system and which normally respond against foreign bodies, are induced to attack the individual's joints in the case of RA [49].

Apart from mentioned environmental factors, genetic basis of RA has been subject of extensive research in the past years [47, 50, 51]. Majority of the performed analysis focus genes within the major histocompatibility complex (MHC) class II (HLA-DR - human leukocyte antigen) chromosomal region (located on the short arm of chromosome 6 - 6p21.3) [52]. Other studies identified non-MHC genes such as corticotrophin releasing hormone [53], CYP19 (estrogen synthase) [54], IFN-γ [55] and other cytokines [56] linked to RA. Possibility that association with RA for other genetic regions linked to other autoimmune diseases, such as insulin-dependent diabetes [57], was also investigated.

Although the precise etiology of RA is not yet known, it is clearly multifactorial and complex, with contributions from both genetic and non-genetic factors.

2.6 SNP - Complex Disease Association: Genome-wide Association Studies

SNPs are by far the most common of genetic variations and they are promising genomic biomarkers[2] for population genetic studies and for spotting genetic variations responsible for complex diseases. Most of the complex diseases are assumed to be attributable to deleterious effects of SNPs on protein functionality and/or change in the regulation of genes. A number of instances are known for where a particular nucleotide change at a SNP locus (i.e., a particular SNP allele) is associated with an individual's propensity to develop a disease [58]. There have also been a number of reports that show some SNPs in certain genes can determine whether a drug can treat a disease more effectively in individual with certain genotypes compared to those who do not carry such SNPs [59].

The completion of human genome project as well as introduction of the high throughput genotyping technologies allow researchers to analyze the DNA sequences to select a set of single-nucleotide polymorphisms (SNPs), which represents the entire haplotype, and utilize these SNPs to assess genetic variation between individuals. Two types of approaches have been commonly utilized for locating genetic factors that are responsible for or associated with the disease**: Candidate-gene-based approaches** and **Genome-wide Association Studies (GWAS)**. In the formation of Candidate-gene-based approach a hypothesis is involved so as to associate genes with a disease. In order to do that, gene should be sequenced in case-control groups and the variant that differentiates cases and controls are investigated. Genome-wide association studies on the other hand are unbiased in their attempt to identify disease causing variants as search is done across most of the genome. They have the capability to test few thousands to more than million markers at a time. They are hypothesis-free, and aimed at the discovery of novel disease-casual variants. In this research study we will refer to GWAS when we mention genotype-phenotype association studies (See Appendix A for glossary of terms frequently used in GWAS).

National Institutes of Health (NIH) defines GWAS as a study of common genetic variations across the entire human genome, designed to identify genetic associations with observable traits [60]. In a typical GWAS, researcher follows a common 5 step process: (1) a large number of individuals with disease or another trait of interest alongside with a suitable comparison group is selected; (2) in order to assure high genotyping quality, DNA isolation, genotyping and data review is performed; (3) statistical tests are applied to observe the association between SNPs and disease/trait; (4) identified associations are replicated in an independent population sample, (5) Fine mapping to DNA locus and biological interpretation is applied. Figure 2.8 summarizes the process (used with the permission of [61]).

[2] DNA sequence with a known location on a chromosome and associated with a particular gene or trait.

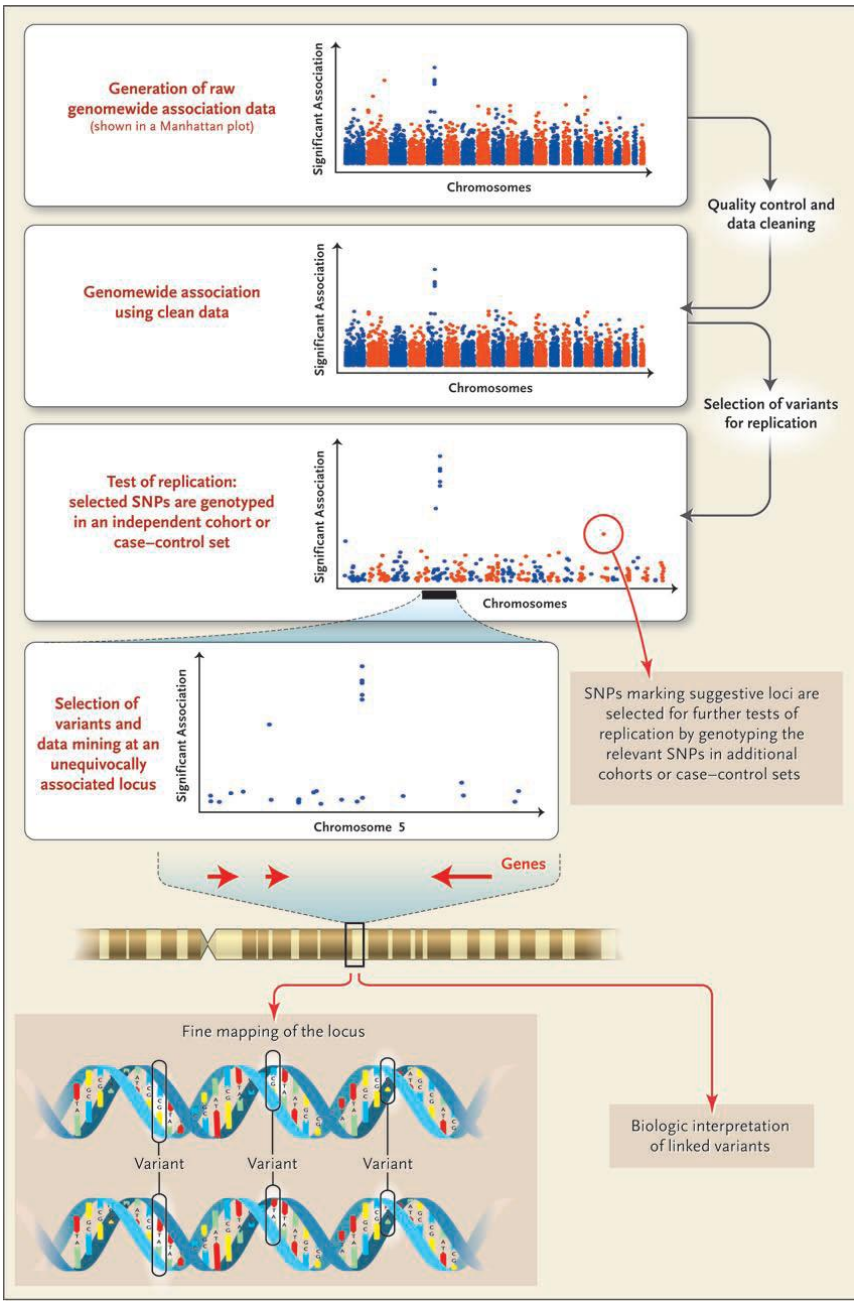


Figure 2. 8 Process steps for a GWAS.

The major study design used in a typical GWAS is the "case-control" study approach in which allele frequencies in patients with the disease of interest are compared to those in a disease-free comparison group. The other design types are "cohort", in which baseline information in a large number of individuals is collected and these individuals are then monitored to identify disease in subgroups defined by and "trio", which includes the affected case participant and both of his/her parents. Assumptions, advantages and disadvantages in using different design patterns are depicted in Table 2.2 [60]. In this research study, we are only interested in case-control design and our analysis is developed for this type of design. Figure 2.9 depicts a case-control GWAS (used with the permission of [62]).

Table 2. 2 Commonly used design types for a typical GWAS.

	Assumptions	Advantages	Disadvantages
Case-Control	Case and control participants are drawn from the same population	Short time frame	Prone to a number of biases including population stratification
	Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified	Large numbers of case and control participants can be assembled	Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases
	Genomic and epidemiologic data are collected similarly in cases and controls	Optimal epidemiologic design for studying rare diseases	Overestimate relative risk for common diseases
	Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls		
Cohort	Participants under study are more representative of the population from which they are drawn	Cases are incident (developing during observation) and free of survival bias	Large sample size needed for genotyping if incidence is low
	Diseases and traits are ascertained similarly in individuals with and without the gene variant	Direct measure of risk	Expensive and lengthy follow-up
		Fewer biases than case-control studies	Existing consent may be insufficient for GWA genotyping or data sharing
		Continuum of health-related measures available in population samples not selected for presence of disease	Requires variation in trait being studied
			Poorly suited for studying rare diseases
Trio	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents	Controls for population structure; immune to population stratification	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset
		Allows checks for Mendelian inheritance patterns in genotyping quality control	Highly sensitive to genotyping error
		Logistically simpler for studies of children's conditions	
		Does not require phenotyping of parents	

In Figure 2.9, in Panel A genotyping process is depicted. Two SNPs from chromosome 9 is genotyped. In Panel B association between SNPs genotyped throughout the entire genome (at least 100,000 SNPs should be used) and the observed disease/trait is calculated. In this particular example SNP_1 and SNP_2 is highly associated with the phenotype with *p*-values 10^{-12} and 10^{-8} respectively. Panel C shows the *p*-values for al genotyped SNPs that have survived a quality control screen, where each chromosome is shown in a different color. The results implicate a highly associated locus on chromosome 9.

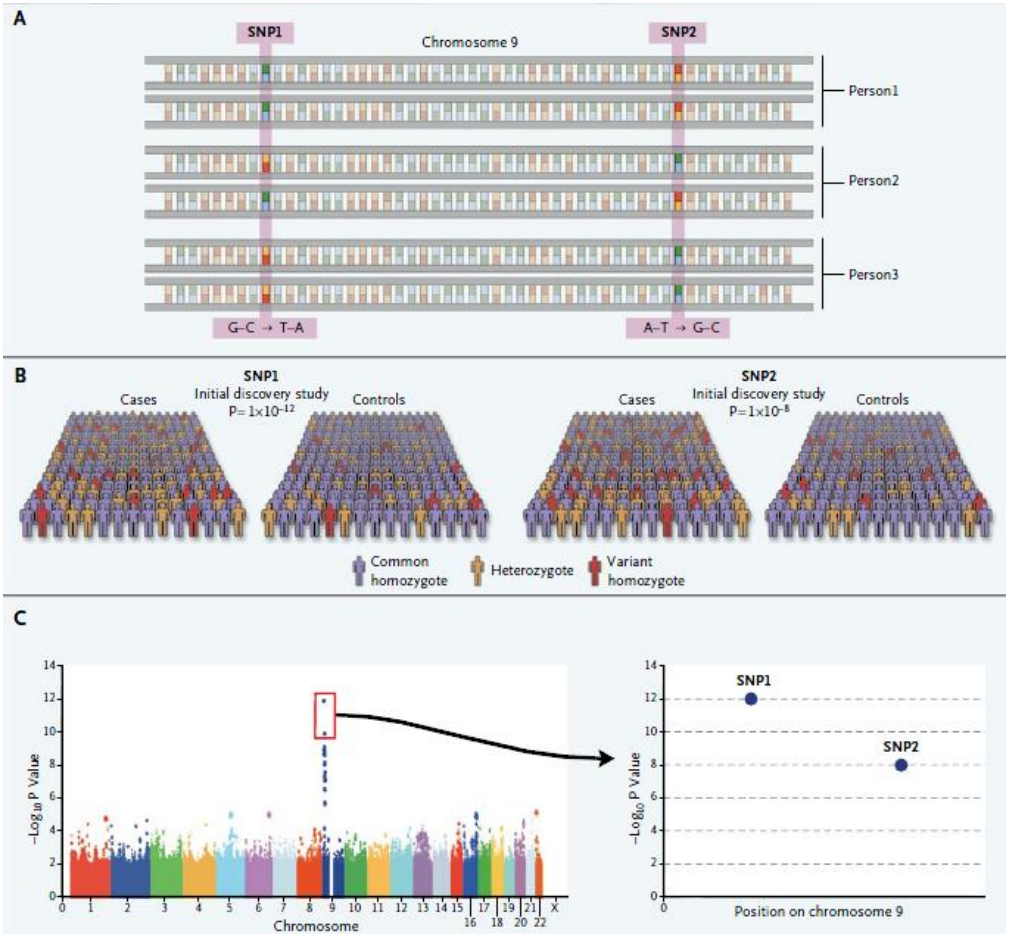


Figure 2. 9 An example case-control GWAS

As it will be covered in Chapter 4 in greater detail, one of the major problems related with a GWAS is the enormous number of simultaneous statistical tests that needs to be applied and the boosted false positive results following the tests. To account for this problem, usual approach is to perform multi-stage designs in which after performing an initial genome-wide scan on case-control participants, a smaller number of associated SNPs is replicated in a second or third group of participants. Initially, a small number of participants with a large number of SNPs are used in some studies to minimize the false negative results. Other studies begin with more participants and carry forward a smaller proportion of associated SNPs. Table 2.3 shows examples of multi-stage designs.

Table 2. 3 Example of multi-stage designs in GWAS

	3-Stage Study		4-Stage Study	
Stage	Case Participants/ Control Participants	SNPs Analyzed	Case Participants/ Control Participants	SNPs Analyzed
1	400/400	500 000	2000/2000	100 000
2	4000/4000	25 000	2000/2000	1000
3	20 000/20 000	25	2000/2000	20
4			2000/2000	5

GWAS has been a popular way of identifying disease causing variants but it comes with certain challenges in terms of logistical and technical requirements. Firstly, a suitable disease should be selected for the analysis and this is not an easy task as the selection should favor specifically diagnosable and measurable phenotype for a successful analysis. Additionally for such studies, extremely large sample sizes (in thousands) are required for cases and controls, which make it hard to collect and manage genetic data. Table 2.4 presents benefits, misconceptions and limitations regarding GWAS [61].

Table 2. 4 Benefits, misconceptions and limitations of GWAS

Benefits
Does not require an initial hypothesis
Uses digital and additive data that can be mined and augmented without data degradation
Encourages the formation of collaborative consortia, which tend to continue their collaboration for subsequent analyses
Rules out specific genetic associations (e.g., by showing that no common alleles, other than APOE, are associated with Alzheimer's disease with a relative risk of more than 2)
Provides data on the ancestry of each subject, which assists in matching case subjects with control subjects
Provides data on both sequence and copy-number variations
Misconceptions
Thought to provide data on all genetic variability associated with disease, when in reality only common alleles with large effects are identified
Thought to screen out alleles with a small effect size, when in reality such findings may still be very useful in determining pathogenic biochemical pathways, even though low-risk alleles may be of little predictive value
Limitations
Requires samples from a large number of case subjects and control subjects and therefore can be challenging to organize
Finds loci, not genes, which can complicate the identification of pathogenic changes on an associated haplotype
Detects only alleles that are common (>5%) in a population
Requires replication in a similarly large number of samples

2.7 Biological Pathways

Biological pathways are networks of complex reactions at the molecular level. They model how biological molecules interact to perform a biological function in response to each other and their environment. One can represent pathway as a collection of biochemical entities serving as the nodes of a network. The edges, in this context, then can be thought as interaction between these entities in the form of protein – protein interactions or protein – DNA binding interactions, regulation events or modifications. In this representation, the global bimolecular network acts as a graph containing all cellular molecular entities and all possible molecular events linking them and a pathway can be regarded as a subgraph. Importantly, not every particular subgraph can be a pathway. There needs to be an actual flow of information (signal) and/or chemical reaction flow that would turn part of the network into a pathway. Signals can be

involved in chemical reactions, turning genes on or off, or regulating the cell function and determine cell's faith [63]. Figure 2.10 (excerpted from KEGG website) depicts the mitogen-activated protein kinase (MAPK) pathway, which is involved in various cellular functions, including cell proliferation, differentiation and migration.

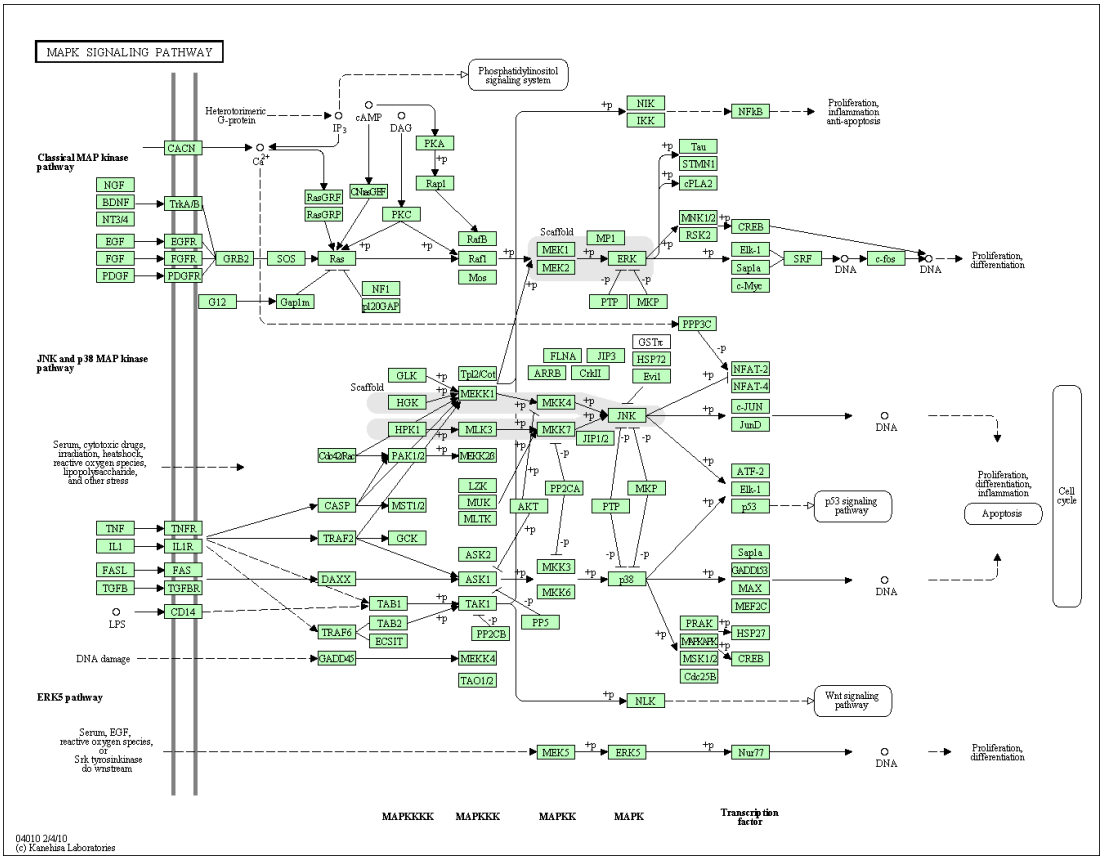


Figure 2. 10 MAPK Signaling Pathway

Recently, there has been a focus on investigating pathogenetic effects of pathways and identifying involvement of certain functional and mechanistic pathways in a variety of disease processes [61, 64]. MAPK Signaling Pathway shown in Figure 2.10, for example, is found out to be significantly overrepresented for RA [65]. During our analysis, we regard a pathway as **a collection of genes**. In our SNP prioritization algorithm, we have integrated pathway based data into GWAS. The process is described in greater detail in the next chapter.

2.8 Conclusion

In this chapter, we aimed to introduce the biological concepts that are important for the development of the ideas that will shape the rest of this research work. We carefully tried to avoid getting into deep towards biological terms to a level that would distract the reader. Instead, we introduced and mentioned the most relevant terms such as mutations, SNPs, genes, complex diseases, GWAS and biological pathways. More specific terms and concepts are explained in subsequent chapters when needed.

CHAPTER 3

ANALYTIC HIERARCHY PROCESS BASED PRIORITIZATION SCHEME FOR MULTI-HIERARCHICAL FILTERING OF INFORMATIVE SNPS AFTER GENOME-WIDE ASSOCIATION STUDY

A typical approach in a Genome Wide Association Study (GWAS) is to repeat the study on an independent sample by genotyping the top signals. It is evident that this approach would be facilitated by incorporating information from biological databases to determine and rank the SNPs that has the highest potential to effect biological functions. In this chapter, we introduce a method that would be used to achieve this strategy in an objective and structured way by combining information on genomic location, biological significance and evolutionary conservation alongside with statistical evidence of genotype–phenotype correlation. Our approach is based on the well-known multi-criteria decision making method called Analytic Hierarchy Process (AHP), which is a structured technique for dealing with complex decisions.

3.1 Introduction

In a biomarker discovery study, the major aim is to identify potentially significant differential variants, such as gene expressions or protein production, in a large set of variables (i.e. potential biomarkers) across control and case groups. The number of variables changes from hundreds to thousands in a typical biomarker discovery study. If we consider that thousands of SNPs may be tested in a study scanning hundreds of genes in a candidate pathway or that hundreds of thousands may be tested in a GWAS, it quickly becomes clear that high numbers of falsified associations would be detected at common thresholds of statistical significance (such as $p = 0.05$) because the simultaneous testing of hundreds of thousands of SNP markers means that a high number of hits would occur by chance [66]. This type of analysis is called as multiple-hypotheses testing and it hinders the replicability of findings in follow-up association studies. Standard procedures that are used to cope with the boost of false positives as a result of multiple-

hypothesis testing (such as Bonferroni correction) are considered too conservative for GWAS, as they tend to ignore the dependency between markers such as those in LD with each other [67].

On the other hand, even if the disadvantages of multiple hypothesis testing are taken care of, the resulting *p*-values alone should be interpreted with utmost caution. For example, it is a common practice for many researchers to make excessive use of the term "significant" without taking neither statistical nor research domain contexts into consideration. Usually, any findings with *p*-values below 0.05 are automatically defined and reported as significant. It may often be the case for many researchers to ignore those findings falling above this threshold, because they are simply not significant. A better way would be to understand the meaning of the parameters and the outcomes of the hypothesis testing process and to interpret them considering the goals, limitations, assumptions and characteristics of the investigation. Another common mistake is to believe that a *p*-value solely represents the probability of not significance, or the probability that the null hypothesis is true. On the contrary, it should be remembered that *p*-value is a measure for the "strength of the evidence found against the null hypothesis". Moreover, hypothesis testing results with *p*-values below 0.05 may not necessarily represent strong evidence against the null hypothesis [10].

The usual approach after an initial GWAS run is to replicate the findings in an independent sample to prove that the association observed was not random and it is reproducible. However performing a second full GWAS is very costly with limited genotyping resources. Even if unlimited resources are provided, the prioritization of SNPs is still required in order to select a subset of SNPs for cost reduction in downstream functional experiments, which aims to extract the exact molecular genetic mechanism responsible for the disease, such as the effect of the genetic variant on transcription, structure of the protein product, or regulatory mechanism [68, 69].

Additionally, focusing on the statistical evidence alone is not a valid approach for SNP biomarker discovery in a GWAS setting because of Linkage Disequibrium (LD). As a result of the strong physical association between certain SNPs (LD proxies), it is highly unlikely to spot the true causal polymorphisms by looking at only the *p*-values of association as the genotypes will be very similar. Then, the necessity of prioritizing and selecting a subset of SNPs for subsequent GWAS studies and/or functional experiments brings forward the question of what is the best approach to prioritize SNPs. In the literature, alongside with statistical evidence, use of biological information such as functional effects of SNPs is favored for this purpose [6-12]. In fact many software tools have been proposed recently to systematically prioritize SNPs based on the wealth of biological information available in public databases such as SPOT [69], FunctSNP [70], FASTSNP [71], SNPLogic [72], SNPinfo [73] and SNPit [74].

In this chapter a novel method for SNP biomarker scoring and prioritizing based on the well-known multi-criteria decision making method Analytic Hierarchy Process (AHP) will be introduced. The proposed hierarchy tree and the scoring system to structurize SNPs for objective

prioritization will be presented. SNP biomarker scoring system is based on both *p*-values of association and the biological importance of SNPs depending on three different criteria: (1) Evolutionary Conservation, (2) Gene Association and (3) Genomic Location. In order to calculate scores for SNPs the information from major public databases are integrated. In the subsequent sections we will review the prioritization approaches in the literature, explain the major public databases on which our scoring mechanism is constructed, discuss available tools for prioritization and give details of our novel prioritization scheme proposed for the first time in this study.

3.2 Prioritizing SNPs after GWAS

Individual SNP based association analysis should only be regarded as a means to an end for revealing the etiology of complex diseases. Recently Hardy et al. [75] state that the genetics of complex diseases can be visualized as a jigsaw puzzle and by performing an individual SNP based analysis one only constructs the frame of the puzzle by putting the edges and corners. To complete the puzzle the preliminary genetic information available for additional analysis by statistical procedures that accumulate evidence should be used effectively. After the preliminary step, secondary analyses should be performed to prioritize SNPs so as to guide subsequent studies and experiments. Cantor et al. [76] present a very detailed review on the techniques that can be used for prioritizing SNPs after a GWAS. Meta-analysis and pathway analysis are suggested as the available approaches.

Meta-analysis is a statistical technique that can be used to combine the results of independent studies, which examine the same research hypothesis. As it does not require use of original data, it introduces considerable advantage and is advocated as a proper method to increase power in genetic analysis [77-79]. In recent studies, use of meta-analysis is also favored for prioritizing genes and SNPs for subsequent studies [80-82]. Combining test statistics from comparable studies have been proposed for application of meta-analysis and among proper statistics, Fisher's method of combining *p*-values [83] is widely used [84-86]. Converting test statistics into z scores [84, 87-89] and using odds ratios and regression coefficients is another approach favored by many researchers. There are various methods for use in SNP prioritization following a meta-analysis. Some of them are based solely on meta-analysis *p*-values [82, 88] whereas others use predetermined decision rules as prior evidence in addition to meta-analysis results to guide the decision [84, 89].

Second method available for prioritization based on GWAS pathway analysis (GWASPA), which integrates results of a GWAS and the genes in a known biological pathway to test whether the pathway is associated with the disorder or not. This approach is promising because it enables combination of statistical results via use of pathways and suggests a biological interpretation. Various analytical methods have been proposed to calculate measures that would be used to rank SNPs in a GWASPA. Subramanian et al. [90] propose gene-set enrichment

analysis algorithm for combining *p*-values in which genes in the pathway are ranked according to the strongest empirical *p*-value statistic. A running sum similar to Kolmogorov-Smirnov is used to calculate test statistics for genes within pathway. Yu et al. [91] introduce an analysis based on adaptive rank truncated product statistic *p* value, which aims to select *n* most significant genes or genes satisfying a user defined threshold. Baranzini et al. [92] define a test statistic calculated over a two-step procedure. First, *p*-value of each gene is converted into a *z* score and then *z* scores are accumulated over the genes in the pathway contributing to the overall score. Once the significant genes and pathways are found through calculation of test statistics and measures, most frequent approach is to select a single SNP with the strongest association signal from each gene.

During our analysis phase in order to find statistically significant (enriched) genes and pathways we have followed the approach stated in [85]. In order to combine *p*-values of all SNPs within a gene we used Fisher's combination test, in which the statistic for combining *K* SNPs is given by

$$Z_F = -2\sum_{i=1}^{K} lnP_i, \quad \textbf{(3.1)}$$

which follows χ^2_{2K} distribution. Following that to search for overrepresentation of significantly associated genes among all genes in a pathway we used Hypergeometric test (Fisher's exact test). Assuming that total number of genes is *N*, number of genes that are significantly associated with the disease is *S* and the number of genes in the pathway is *m; p*-value of observing *k*-significant genes in the pathway is calculated by:

$$p = 1 - \sum_{i=0}^{k} \frac{\binom{S}{i}\binom{N-S}{m-i}}{\binom{N}{m}}. \quad \textbf{(3.2)}$$

In order to calculate the aforementioned statistics, one has to have access to the relevant biological information. Fortunately, recent studies enabled collection of such information within web-based databases. Most of these databases are publicly accessible. Our AHP based prioritization scheme requires integration of primary public databases because of the need of mappings for:

- SNP to Gene,
- Gene to Disease,
- Gene to Biological Pathway.

3.2.1 SNP and Gene Repositories

The International HapMap Project [1, 2] has been identifying patterns of DNA sequence variation (linkage disequilibrium (LD)) in human, since it was initiated in 2002. Generated data has been used to assist researchers in the mapping of loci associated with disease, drug response and other human features. Genotype data, allele and genotype frequencies, LD data, phase information, SNP assay details, protocols and sample documentations are available for download from the HapMap website. Novel SNPs identified by the HapMap Project have been submitted to

the public databases such as dbSNP[3] [15] and JSNP [93], and the data incorporated into Ensembl [94] and the UCSC Generic Genome Browser [95]. See Table 3.1 for HapMap based databases.

Table 3. 1 Publicly available databases incorporating HapMap data.

Database	Description	URL
HapMap	Primary data source	http://www.hapmap.org
dbSNP	Individual genotypes, allele and genotype frequencies, LD plots	http://www.ncbi.nlm.nih.gov
JSNP	Genotype frequencies	http://snp.ims.u-tokyo.ac.jp
Ensembl	Individual genotypes, allele and genotype frequencies, LD plots, tag SNP identification	http://www.ensembl.org
UCSC Genome Browser	Recombination rates, hotspots, sequencing coverage and allele frequencies in encode regions	http://genome.ucsc.edu

Among other heavily used data repositories are SNP500 [96], Environmental Genome Project (EGP) [97], Seattle SNPs [98] and Human Gene Mutation Database (HGMD) [99]. The SNP500 database is almost entirely a gene-centric database and is home to the resequencing data from genes thought to be of importance in cancer. It aims to provide a framework to the molecular epidemiologists for the design of cancer-based association studies. The premise of the EGP is to identify polymorphic variations in candidate genes that are believed to be at the interface between genetics and response to environmental stimulus. Seattle SNPs concentrates on genes with relevance to inflammation, but also clotting and heart, lung and blood-related phenotypes. It provides assay conditions and resources for assay design. HGMD aims to collate data on genetic variation pertaining to human disease.

As for the gene information the most relied source is NCBI Entrez Gene [16]. It is NCBI's repository for gene-specific information such as gene symbol, alias names, chromosomal location, and gene type. The gene centric information is provided through the NCBI ENTREZ and the information about a single gene, associated sequences, structures, variations, and more for around 38,550 human genes are also visualized through its genome browsers. The gene record offers a summary information about the genomic region/transcripts/products, genomic context, a bibliography, interaction data, general gene and protein information, references sequences and more. See Appendix B for Entity-Relationship Diagram of Entrez Gene.

3.2.2 Gene and Disease Repositories

Genetic disorders are result of the unusual characteristics of genes or chromosomes within certain individuals. Genetic disorders can be classified into Mendelian diseases in which chromosomal defects interfere with functions of a single gene. A catalog of Mendelian diseases can be found at the Online Mendelian Inheritance in Man (OMIM) database. In case of complex diseases, the interplay of multiple genes alongside with environmental factors triggers disease

[3] The most recent build for human SNPs is build 131 and it is dated 25.03.2010. It provides annotation for 12,017,369 validated human SNPs.

status. Coronary heart disease, hypertension, diabetes, obesity, various cancers, Alzheimer's disease, Parkinson's disease, Epilepsy are among complex diseases [41].

NCBI's Genes and Disease[4] is a rich and comprehensive resource providing information on genes and the diseases that they cause. The genetic disorders are organized, so that they are grouped by the parts of the body that they affect. A list of 85 human genetic disorders, categorized by the 16 body parts that they affect is presented. Figure 3.1(excerpted from KEGG website) provides an example in which a couple of genetic disorders are mapped to chromosome 4. A much more detailed (gene – disease) mapping is provided for each chromosome.

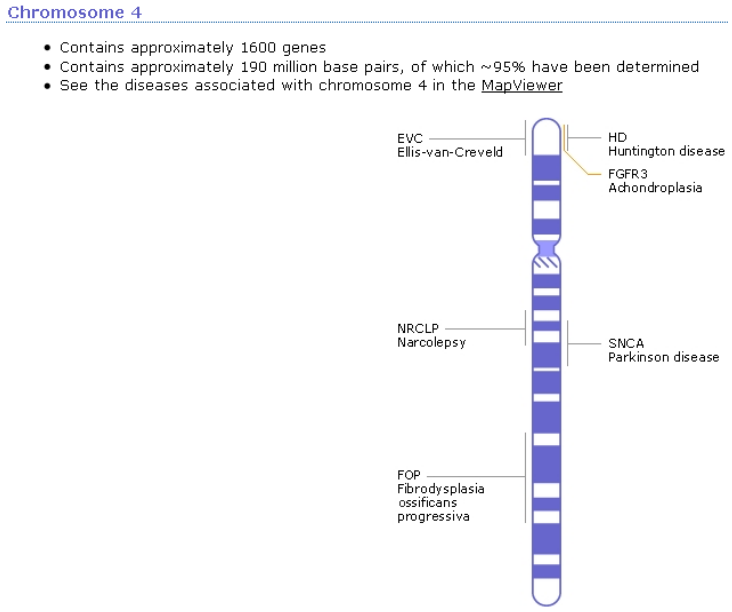


Figure 3. 1 A sample mapping from NCBI's genes and disease.

NCBI's OMIM database[5] [100] provides users with very detailed information of human genes and genetic phenotypes. OMIM contains information on all known Mendelian disorders and over 12,000 genes. It can be regarded as a great source to map genes to diseases. Additionally, the Genetic Association Database[6] can be used to exploit human genetic association studies of complex diseases and disorders. The data is presented in a gene centric fashion and is extracted from published scientific papers. Data fields common to genetic association studies such as disease phenotypes, sample sizes, significance values, population information and allele descriptions are identified. Recently disease annotations using GeneRIF is suggested as a promising approach for increased accuracy in disease identification [101]. A GeneRIF (Gene Reference into Function) is a brief (up to 255 characters) annotation to a gene in the NCBI database. It contains gene specific information including disease associations. To provide disease to gene annotation, the Disease Ontology (DO) is used to identify the relevant diseases in

[4] http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gnd
[5]http:// www.ncbi.nih.gov/omim/
[6] http://geneticassociationdb.nih.gov/

GeneRIFs. Mappings are stored in the Open Biomedical Ontologies[7]. Table 3.2 shows a sample mapping.

Table 3. 2 A sample Disease Ontology - GeneRIF mapping.

Do ID	Disease Name	Gene ID	Gene Name	GeneRIF
7148	Rheumatoid arthritis	11278	Kruppel-like factor 12	KLF12 as a new susceptibility gene for rheumatoid arthritis.
7148	Rheumatoid arthritis	10563	Chemokine (C-X-C motif) ligand 13	CCL23, M-CSF, TNFRSF9, TNF-alpha, and CXCL13 are predictive of rheumatoid arthritis disease activity and may be useful in the definition of disease subphenotypes and in the measurement of response to therapy in clinical studies.

3.2.3 Biological Pathway Databases

Biological pathways are networks of complex molecular level interactions of living cells. They model how biological molecules interact to perform specific functions in respond to the biological and environmental signals. A pathway therefore can be formally thought of as a collection of "nodes" representing biochemical entities, connected by edges that represent interactions, regulation events or modifications. A comprehensive list of recently identified pathway resources can be found in the Pathguide[8].

Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] is a widely utilized database providing gene lists, diagrams, and pathway classification tools for many aspects of biology. It is an integrated database resource consisting of 16 main databases, categorized into systems information, genomic information, and chemical information. An alternative source is Gene Ontology (GO) [18], which contains ontology, gene product, sequence, and manual annotation data. Mainly three domains are covered by GO:

- **Cellular component**, the parts of a cell or its environment,
- **Molecular function**, the elemental activities of a gene product at the molecular level,
- **Biological process**, operations or sets of molecular events with a defined beginning and end.

BioCyc [102], BioCarta[9] and Wikipathways [103] are among other relevant pathway resources. The BioCyc collection of Pathway/Genome Databases (PGDBs) provides information on the genomes and metabolic pathways of sequenced organisms. BioCyc databases are integrated with other biological databases containing protein and nucleic-acid sequence data, bibliographic data, protein structures, and descriptions of different strains. BioCarta provides graphical tools for observing how genes interact in metabolic pathways and signaling pathways.

[7] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology
[8] Pathguide contains information **about 325** biological pathway related resources and molecular interaction related resources: http://www.pathguide.org/
[9] http://www.biocarta.com/

It contains important resources providing information for over 120,000 genes from multiple species. WikiPathways is another online resource for biological pathway information. It acts as a platform for community-based curation. As of December 2010, WikiPathways contains more than 600 pathways, representing species including bacteria, fungi, plants and animals.

3.2.4 Available Tools for SNP Prioritization

There have been extensive number studies on exploring and analyzing the data gathered in database repositories listed in previous sections. This resulted in various software tools and integrated systems that can be used for prioritizing SNPs before or after GWAS using statistical data alongside with biological data. Among those tools are SPOT [69], SNPLogic [72], FunctSNP [70], FastSNP [71] and Pupasuite [104].

SPOT recently introduced the genomic information network (GIN) method, which is a directed graph whose nodes are features from a biological database. The GIN process begins with a SNP and ends in the terminal node with the calculation of its overall prioritization score S. The overall score is formed by biological relevance obtained by combining information from multiple databases. For example, if a SNP is in a gene this will increase the overall score. Once the overall score S determined, they rank SNPs from a GWAS for further study by $\frac{p}{10^S}$.

SNPLogic is a web service, which can be used for SNP selection, annotation and prioritization. The integrated information sheds light upon the genetic context of SNPs, genotypic data, functional predictions, biological pathways and gene - disease associations. The interface facilitates construction of SNP lists and user defined scoring rules would be used to rank those SNP lists. The lists can be stored and accessed indefinitely. One major problem related to the approach of "user defined scoring" is the lack of objective evaluation criteria for standardization of SNP scoring and prioritization. For most users, biologists or bioinformaticians, the process of defining which criteria should outweigh the other is a complicated decision and assigning a biological relevance score between the criteria is a much more confusing task.

FunctSNP is an R-Package backed up with integrated databases for five different species including Homo sapiens. Its aim is to efficiently screen and select the GWAS significant SNPs which are more likely to be reliable as DNA markers and link GWAS derived SNPs to biological pathways and gene networks. Screening of SNPs is made possible by determining physical location of SNPs with respect to genes and finding evidence for a functional role of the significant SNPs.

FastSNP is a web server that allows identification and prioritization of high-risk SNPs according to their phenotypic risks and deleterious functional effects. It extracts the information from 11 external web servers, which makes always up-to-date querying of relevant biological information possible. It provides reporting facilities for genomic information, SNP functional effects, transcription regulatory, alternative splicing regulatory, mRNA/protein domain effects and protein structure effects. It utilizes a decision tree approach to assign risk rankings for SNP

prioritization. The decision tree utilized by the system is depicted in Figure 3.2 (used with the permission of [71]). The tree structure used for the prioritization only depends on genomic location of SNPs and lacks the essential information such as evolutionary conservation, gene association and biological pathways. Depending on the functional effects, each SNP is assigned a risk factor and ranking is done accordingly. Therefore compared to hierarchy tree proposed in our AHP study, the decision tree utilized by FastSNP is fairly simple and not suitable for comprehensive analysis.

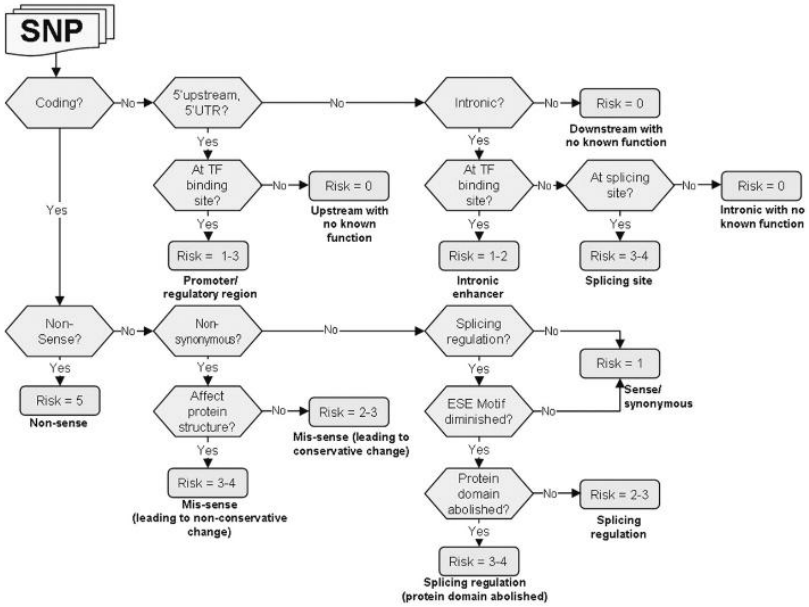


Figure 3. 2 FastSNP decision tree.

Pupasuite is a web tool that can be used to select SNPs with possible effect on different phenotypes. It aims to facilitate large scale genotyping studies. It integrates various heterogeneous resources and provides pre-calculated predictions for selection of optimal SNP sets. Selection is based on different characteristics of SNPs such as type, frequency, validation status, deleterious functional properties, LD parameters etc.

Table 3.3 summarizes features of these tools:

Table 3. 3 SNP prioritization tools comparison.

Function/ Application	SPOT	SNPLOGIC	FunctSNP	FASTSNP	Pupasuite
SNP Prioritization					
Genomic Location	NO	YES	YES	YES	YES
SNP effect miRNA target sequences	NO	NO	NO	NO	NO
SNP effect on promoter CpG Islands	NO	NO	NO	NO	NO
Biological Associations					
via Gene	YES	LIMITED	YES	LIMITED	YES
via Pathway	NO	NO	YES	NO	NO
via Linkage Disequilibrium	YES	NO	YES	NO	YES
Evolutionary Conservation	YES	YES	NO	NO	YES
Pathway Mapping of Prioritized SNPs	NO	NO	NO	NO	NO
Application Features					
Desktop Application	NO	NO	YES	NO	NO
Web based server	YES	NOT RELIABLE	NO	YES	NO
Frequently updated , maintained databases	YES	NO	PARTLY	YES	NO

3.3 AHP Based Prioritization Scheme

3.3.1 AHP Method Basics

AHP first proposed in 1980 for addressing complex decision making problems with multiple criteria [8]. It is very well suited to cases in which evaluation criteria can be organized into sub-criteria in a hierarchical way. In AHP studies, the decision making problem is divided into a four stage process [105]:

1. Hierarchical structuring of the problem,
2. Data input,
3. Relative weights estimation for evaluation criteria,
4. An overall evaluation of the alternatives through combination of the relative weights.

In the first stage, the problem at hand is represented as a multi-hierarchy structure. Figure 3.3 depicts a general form of such a structure. The top level of the hierarchy represents the general objective or goal of the problem. The second level includes evaluation criteria. Each criterion is analyzed in the subsequent levels into sub-criteria. Finally, the last level of the hierarchy contains the objects to be evaluated.

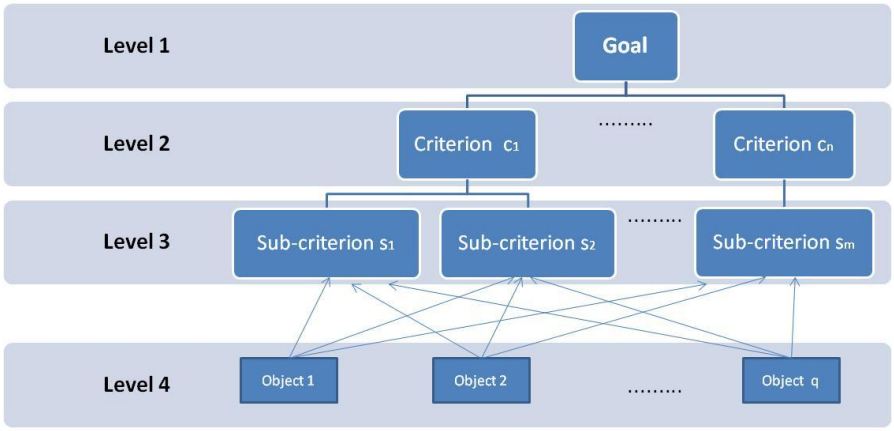


Figure 3. 3 The hierarchical structure of a decision making problem in AHP context.

Once the hierarchy is formed, next step is to perform a pairwise comparison of all elements at each level of the hierarchy. Each of these comparisons is based on the elements of the upper level of the hierarchy. For instance, as the first level involves only one element, considering the general hierarchy of Figure 2, no comparisons are required. In the second level, representing the evaluation criteria, all elements are compared in a pairwise fashion based on the objective of the problem stated at the first level of the hierarchy. Then, similarly the sub-criteria of the third level are compared each time from a different point of view depending on each criterion of the second level of the hierarchy. It is easy to see that if *n* sub-criterion is present under a particular criterion, there will be $\binom{n}{2}$ comparisons. This process is continued until all elements of the hierarchy are compared. The objective of all these comparisons is to find out the relative significance of all elements of the hierarchy in a structured way for making the final

decision according to the initial objective. The usual way of making comparisons is to use a 9-point scale depicted in Table 3.4.

Table 3. 4 Pairwise comparison scale for element i vs j in an AHP hierarchy.

Intensity	Definition	Explanation
1	i and j has equal importance	Two elements contribute equally to the objective
3	i has modarate importance over j	Experience and judgement slightly favor i over j
5	i has strong importance over j	Experience and judgement strongly favor i over j
7	i has very strong importance over j	i is favored very strongly over j; its dominance is demonstrated in practice
9	i has extreme importance over j	The evidence favoring i over j is of the highest possible order of affirmatio

Intensities of 2, 4, 6 and 8 can be used to express immediate values. Intensities 1.1, 1.2, 1.3, etc. can be used for elements that are very close in importance.

The results of the comparisons are used to form a comparison matrix of size *n*x*n* for every comparison performed in each level. Let $\boldsymbol{C_k}$ denotes the comparison matrix. Then $\boldsymbol{C_k}$ is defined as:

$$\boldsymbol{C_k} = \begin{vmatrix} 1 & p_{k1}/p_{k2} & \ldots\ldots\ldots & p_{k1}/p_{kn} \\ p_{k2}/p_{k1} & 1 & \ldots\ldots\ldots & p_{k2}/p_{kn} \\ . & . & \ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots & . \\ p_{kn}/p_{k1} & p_{kn}/p_{k2} & \ldots\ldots\ldots & 1 \end{vmatrix}$$

where $p_k = (p_{k1}, p_{k2}, \ldots, p_{kn})$ denotes the actual weights assigned to each element included at level k of the hierarchy as opposed to a specific element of level (k-1). Under the assumption of consistent comparisons one can estimate the weights through the solution of following system of linear equalities:

$$C_k . p_k = n . p_k . \quad \textbf{(3.3)}$$

This relation can be used to solve for p_k if C_k is known. Providing a non-zero solution to this problem is known as the *Eigen value problem* and it is represented as:

$$C_k^{'} . p_k^{'} = \lambda_{\max} . p_k^{'}, \quad \textbf{(3.4)}$$

where $C_k^{'}$ is the matrix formed by making use of the comparisons, $\lambda_{\max}$ is the largest Eigen value of $C_k^{'}$ ($\lambda_{\max} \geq n$), which is also known as priority vector and $p_k^{'}$ is the vector of the estimates of the actual weights.

As a last step, weights defined in the previous stage are combined to account for an overall evaluation of the elements belonging in the final level of the hierarchy (level k). The combination is based on the initial objective of the analysis (first level of the hierarchy) and represented as:

$$S_{1k} = \prod_{j=2}^{k} W_j, \quad \textbf{(3.5)}$$

where S_{1k} is a vector of the global evaluations for the elements of level k and W_j is a matrix of the weights of the elements in level j as opposed to the elements of level $(j-1)$ [105].

3.3.2 AHP Based SNP Prioritization Scheme

In this study we have proposed a novel AHP based prioritization scheme for objective and structured prioritization of SNP biomarkers. We first constructed the hierarchy tree by setting *Functionally and Biologically Important SNPs Associated with the Condition* as our objective. In addition to the biological information we have incorporated GWAS results into prioritization mechanism unlike another recently proposed decision tree based approach [74]. The tree structure is depicted in Appendix C and described in Table 3.5 in detail. Formation of this tree structure has been directed by the type of information we can access via biological databases, relevance of the criteria for achieving the objective and ability to make meaningful and smooth pairwise comparisons for the criteria on the same level. The current tree structure has been optimized after comprehensive evaluation on six different versions. To the best of our knowledge, our tree structure has been the most detailed one as, apart from functional effects of SNPs: we have integrated pathway based data, disease annotation data and statistical information simultaneously for the first time in this domain. SNPs that are overlapping with 5'UTR CpG islands or 3'UTR miRNA binding sites are also integrated in the AHP structure for scoring, as recent studies have shown the effect of the regulatory structures on gene function.

Table 3. 5 AHP tree details.

Node	Description	Explanation	Source
0	GWAS Results	Statistical relevance of the SNPs are compared	GWAS
1	Biological Facts	Biological significance of the SNPs are compared	Biological Databases
0.1	Individual SNP	Represents GWAS p-values for individual SNPs	GWAS
0.2	Significant Gene	SNPs related with significant genes according to combined p-value	GWAS, Pearson's Chi-Square
0.2.1	Significant Gene - Via LD	SNP-to-Significant Gene association through another SNP in high LD	Entrez Gene, dbSNP, HapMap
0.2.2	Significant Gene - Via Direct	Direct SNP-to-Significant Gene association	Entrez Gene, dbSNP
0.2.3	Significant Gene - Via Pathway	SNPs related with significant genes present in significant pathways according to combined p-value	KEGG, GO, Biocyc, Biocarta, Wikipathways
0.3	Significant Pathway Gene	SNPs related with significant pathways according to combined p-value	KEGG, GO, Biocyc, Biocarta, Wikipathways
0.3.1	Significant Pathway Gene - Via LD	SNP-to-Significant Pathway Gene association through another SNP in high LD	KEGG, GO, Biocyc, Biocarta, Wikipathways, HapMap
0.3.2	Significant Pathway Gene - Via Direct	Direct SNP-to-Significant Pathway Gene association	KEGG, GO, Biocyc, Biocarta, Wikipathways
0.3.3	Significant Pathway Gene - Via Pathway	SNPs related with significant pathway genes present in significant pathways according to combined p-value	KEGG, GO, Biocyc, Biocarta, Wikipathways
1.1	Evolutionary Conservation	Favor SNPs in evolutionary conserved regions	ECRBase

Table 3.5 (cont.)

1.1.1	Vertebrate	Evolutionary conserved region in vertebrate	ECRBase
1.1.2	Mammalian	Evolutionary conserved region in mammalian	ECRBase
1.1.2.1	Mammalian - Significant Mouse ECR	Evolutionary conserved region in mammalian, ECR > 0.8 in mouse	ECRBase
1.1.2.2	Mammalian - Other Mammalian	Evolutionary conserved region in mammalian, ECR < 0.8 in mouse	ECRBase
1.2	Gene Association	Favor SNPs associated with genes	Entrez Gene, dbSNP
1.2.1	Disease Gene	SNPs associated with genes proven to be associated with complex disease under consideration	GeneRIF, DO
1.2.1.1	Disease Gene - Via LD	SNP-to-Disease Gene association through another SNP in high LD	GeneRIF, DO, HapMap, Entrez Gene
1.2.1.2	Disease Gene - Via Direct	Direct SNP-to-Disease Gene association	GeneRIF, DO, Entrez Gene
1.2.1.3	Disease Gene - Via Pathway	SNPs related with genes which are in the same pathway with a disease gene	GeneRIF, DO, Pathway databases
1.2.2	Other Gene	SNPs associated with genes not proven to be associated with complex disease under study	Entrez Gene, dbSNP
1.2.2.1	Other Gene - Other Disease	SNPs associated with genes proven to be associated with another complex disease	Entrez Gene, dbSNP, GeneRIF, DO
1.2.2.1.1	Other Gene - Other Disease - Via LD	SNP-to-Disease Gene association through another SNP in high LD	Entrez Gene, dbSNP, GeneRIF, DO, HapMap
1.2.2.1.2	Other Gene - Other Disease - Via Direct	Direct SNP-to-Disease Gene association	Entrez Gene, dbSNP, GeneRIF, DO
1.2.2.1.3	Other Gene - Other Disease - Via Pathway	SNPs related with genes which are in the same pathway with a disease gene	Entrez Gene, dbSNP, GeneRIF, DO, Pathway databases
1.2.2.2	Other Gene - Neutral	SNPs associated with genes not proven to be associated with any disease	Entrez Gene, dbSNP
1.2.2.2.1	Other Gene - Neutral - Via LD	SNP-to-Gene association through another SNP in high LD	Entrez Gene, dbSNP, HapMap
1.2.2.2.2	Other Gene - Neutral - Via Direct	Direct SNP-to-Gene association	Entrez Gene, dbSNP
1.2.2.2.3	Other Gene - Neutral - Via Pathway	SNPs related with genes which are in the same pathway with a neutral gene	Entrez Gene, dbSNP, Pathway databases
1.3	Genomic Location	Favor SNPs with functional effects	dbSNP, PolyPhen
1.3.1	Non-Coding		dbSNP
1.3.1.1	Non-Coding- UTR-3	3 prime untranslated region	dbSNP
1.3.1.1.1	Non-Coding- UTR-3 - MiRNA Prediction	3 prime untranslated region with MiRNA Prediction	dbSNP
1.3.1.1.2	Non-Coding- UTR-3 - No MiRNA Prediction	3 prime untranslated region and no MiRNA Prediction	dbSNP
1.3.1.2	Non-Coding- UTR-5	5 prime untranslated region	dbSNP
1.3.1.2.1	Non-Coding- UTR-5 - CpG Island	5 prime untranslated region near CpG Island	dbSNP
1.3.1.2.2	Non-Coding- UTR-5 - No CpG Island	5 prime untranslated region not near CpG Island	dbSNP
1.3.1.3	Non-Coding - Intronic	The variation is in the intron of a gene but not in the first two or last two bases of the intron	dbSNP
1.3.1.4	Non-Coding - Near Gene 3	Within 3' 0.5kb to a gene.	dbSNP
1.3.1.5	Non-Coding - Near Gene 5	Within 5' 2kb to a gene	dbSNP

Table 3.5 (cont.)

1.3.1.5.1	Non-Coding - Near Gene 5 - CpG Island	Within 5' 2kb to a gene and near CpG Island	dbSNP
1.3.1.5.2	Non-Coding - Near Gene 5 - No CpG Island	Within 5' 2kb to a gene and not near CpG Island	dbSNP
1.3.1.6	Non-Coding - Splice3	The variation is in the first two bases of the intron	dbSNP
1.3.1.7	Non-Coding - Splice 5	The variation is in the last two bases of the intron	dbSNP
1.3.2	Coding	Variation is in the coding region of the gene	dbSNP
1.3.2.1	Coding - Frameshift	Indel SNP causing frameshift	dbSNP
1.3.2.2	Coding - CDS Non Syn	Nonsynonymous change	dbSNP
1.3.2.2.1	Coding - CDS Non Syn - Polyphen Benign	Nonsynonymous change and polyphen benign	dbSNP, PolyPhen
1.3.2.2.2	Coding - CDS Non Syn - Possibly Damaging	Nonsynonymous change and possibly damaging	dbSNP, PolyPhen
1.3.2.2.3	Coding - CDS Non Syn - Probably Damaging	Nonsynonymous change and probably damaging	dbSNP, PolyPhen
1.3.2.2.4	Coding - CDS Non Syn - Completely Determine	Nonsynonymous change and complete deleterious effect	dbSNP, PolyPhen

After the formation of the tree structure, pairwise associations for each level of the tree are performed. Inspired by Kelder et al. [106], we have performed an AHP-Delphi study with inputs from 5 biologists trained in bioinformatics to improve the reliability. Pairwise comparison tables and related summary statistics for each node in the tree are presented in Appendix D. The priority vectors for each evaluation for different experts are presented in Table 3.6.

Table 3. 6 Priority vectors for each node of the AHP tree.

Node	Description	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
0	Gwas Results	0.33	0.25	0.83	0.14	0.36
1	Biological Facts	0.67	0.75	0.17	0.86	0.64
0.1	Individual SNP	0.07	0.07	0.08	0.16	0.06
0.2	Significant Gene	0.64	0.64	0.19	0.75	0.27
0.2.1	Significant Gene - Via LD	0.06	0.11	0.07	0.07	0.11
0.2.2	Significant Gene - Via Direct	0.66	0.63	0.75	0.81	0.33
0.2.3	Significant Gene - Via Pathway	0.28	0.26	0.18	0.12	0.56
0.3	Significant Pathway Gene	0.28	0.28	0.72	0.09	0.67
0.3.1	Significant Pathway Gene - Via LD	0.06	0.11	0.08	0.11	0.11
0.3.2	Significant Pathway Gene - Via Direct	0.66	0.63	0.69	0.7	0.33
0.3.3	Significant Pathway Gene - Via Pathway	0.28	0.26	0.23	0.19	0.56
1.1	Evolutionary Conservation	0.06	0.12	0.26	0.31	0.11
1.1.1	Vertebrate	0.33	0.17	0.9	0.13	0.25
1.1.2	Mammalian	0.67	0.83	0.1	0.88	0.75
1.1.2.1	Mammalian - Significant Mouse ECR	0.67	0.83	0.25	0.75	0.8
1.1.2.2	Mammalian - Other Mammalian	0.33	0.17	0.75	0.25	0.2
1.2	Gene Association	0.66	0.32	0.63	0.62	0.58
1.2.1	Disease Gene	0.9	0.83	0.88	0.9	0.88
1.2.1.1	Disease Gene - Via LD	0.06	0.11	0.09	0.23	0.11

Table 3.6 (cont.)

1.2.1.2	Disease Gene - Via Direct	0.66	0.63	0.75	0.68	0.33
1.2.1.3	Disease Gene - Via Pathway	0.28	0.26	0.16	0.08	0.56
1.2.2	Other Gene	0.1	0.17	0.13	0.1	0.13
1.2.2.1	Other Gene - Other Disease	0.75	0.88	0.75	0.75	0.88
1.2.2.1.1	Other Gene - Other Disease - Via LD	0.06	0.11	0.08	0.46	0.11
1.2.2.1.2	Other Gene - Other Disease - Via Direct	0.66	0.63	0.69	0.28	0.33
1.2.2.1.3	Other Gene - Other Disease - Via Pathway	0.28	0.26	0.23	0.26	0.56
1.2.2.2	Other Gene - Neutral	0.25	0.13	0.25	0.25	0.13
1.2.2.2.1	Other Gene - Neutral - Via LD	0.06	0.11	0.08	0.46	0.11
1.2.2.2.2	Other Gene - Neutral - Via Direct	0.66	0.63	0.69	0.28	0.33
1.2.2.2.3	Other Gene - Neutral - Via Pathway	0.28	0.26	0.23	0.26	0.56
1.3	Genomic Location	0.27	0.56	0.11	0.08	0.31
1.3.1	Non-Coding	0.13	0.13	0.17	0.1	0.13
1.3.1.1	Non-Coding- UTR-3	0.07	0.05	0.09	0.09	0.1
1.3.1.1.1	Non-Coding- UTR-3 - MiRNA Prediction	0.88	0.88	0.13	0.5	0.9
1.3.1.1.2	Non-Coding- UTR-3 - No MiRNA Prediction	0.13	0.13	0.88	0.5	0.1
1.3.1.2	Non-Coding- UTR-5	0.1	0.11	0.14	0.06	0.2
1.3.1.2.1	Non-Coding- UTR-5 - CpG Island	0.83	0.83	0.83	0.42	0.9
1.3.1.2.2	Non-Coding- UTR-5 - No CpG Island	0.17	0.17	0.17	0.58	0.1
1.3.1.3	Non-Coding - Intronic	0.03	0.03	0.03	0.07	0.03
1.3.1.4	Non-Coding - Near Gene 3	0.04	0.08	0.06	0.13	0.03
1.3.1.5	Non-Coding - Near Gene 5	0.15	0.12	0.12	0.16	0.15
1.3.1.5.1	Non-Coding - Near Gene 5 - CpG Island	0.75	0.83	0.75	0.83	0.9
1.3.1.5.2	Non-Coding - Near Gene 5 - No CpG Island	0.25	0.17	0.25	0.17	0.1
1.3.1.6	Non-Coding - Splice3	0.22	0.3	0.27	0.24	0.19
1.3.1.7	Non-Coding - Splice 5	0.39	0.3	0.29	0.24	0.3
1.3.2	Coding	0.88	0.88	0.83	0.9	0.88
1.3.2.1	Coding - Frameshift	0.74	0.6	0.8	0.74	0.73
1.3.2.3	Coding - CDS Non Syn	0.26	0.4	0.2	0.26	0.27
1.3.2.3.1	Coding - CDS Non Syn - Polyphen Benign	0.05	0.04	0.06	0.05	0.05
1.3.2.3.2	Coding - CDS Non Syn - Possibly Damaging	0.15	0.14	0.09	0.09	0.12
1.3.2.3.3	Coding - CDS Non Syn - Probably Damaging	0.3	0.27	0.19	0.17	0.22
1.3.2.3.4	Coding - CDS Non Syn - Completely Determine	0.5	0.55	0.66	0.69	0.61

Relative weights for each leaf node after taking the arithmetic averages is calculated by using Equation 3.5 and presented in Table 3.7. For example, the weight for Intronic ($WL_{1.3.1.3}$) is calculated by $WL_{1.3.1.3} = W_1 \, x \, W_{1.3} \, x \, W_{1.3.1} \, x \, W_{1.3.1.3} = 0.618 \, x \, 0.266 \, x \, 0.132 \, x \, 0.038 = 0.000825$.

Table 3. 7 Weights for the leaf nodes of the hierarchy tree.

	Leaf	Description	Score
GWAS RELATED	0.1	Individual SNP	0.033616
	0.2.1	Significant Gene - Via LD	0.01598
	0.2.2	Significant Gene - Via Direct	0.12099

Table 3.7 (cont.)

0.2.3	Significant Gene - Via Pathway	0.053266
0.3.1	Significant Pathway Gene - Via LD	0.01465
0.3.2	Significant Pathway Gene - Via Direct	0.093825
0.3.3	Significant Pathway Gene - Via Pathway	0.04738
1.2.1.1	Disease Gene - Via LD	0.036593
1.2.1.2	Disease Gene - Via Direct	0.186016
1.2.1.3	Disease Gene - Via Pathway	0.081725
1.2.2.1.1	Other Gene - Other Disease - Via LD	0.005756
1.2.2.1.2	Other Gene - Other Disease - Via Direct	0.01818
1.2.2.1.3	Other Gene - Other Disease - Via Pathway	0.011161
1.2.2.2.1	Other Gene - Neutral - Via LD	0.00145
1.2.2.2.2	Other Gene - Neutral - Via Direct	0.004579
1.2.2.2.3	Other Gene - Neutral - Via Pathway	0.002811
1.1.1	Vertebrate	0.037841
1.1.2.1	Mammalian - Significant Mouse ECR	0.04532
1.1.2.2	Mammalian - Other Mammalian	0.023347
1.3.1.1.1	Non-Coding- UTR-3 - MiRNA Prediction	0.001142
1.3.1.1.2	Non-Coding- UTR-3 - No MiRNA Prediction	0.000604
1.3.1.2.1	Non-Coding- UTR-5 - CpG Island	0.002017
1.3.1.2.2	Non-Coding- UTR-5 - No CpG Island	0.00063
1.3.1.3	Non-Coding - Intronic	0.000825
1.3.1.4	Non-Coding - Near Gene 3	0.001476
1.3.1.5.1	Non-Coding - Near Gene 5 - CpG Island	0.002467
1.3.1.5.2	Non-Coding - Near Gene 5 - No CpG Island	0.000571
1.3.1.6	Non-Coding - Splice3	0.005295
1.3.1.7	Non-Coding - Splice 5	0.006597
1.3.2.1	Coding - Frameshift	0.103733
1.3.2.3.1	Coding - CDS Non Syn - Polyphen Benign	0.001997
1.3.2.3.2	Coding - CDS Non Syn - Possibly Damaging	0.004713
1.3.2.3.3	Coding - CDS Non Syn - Probably Damaging	0.009187
1.3.2.3.4	Coding - CDS Non Syn - Completely Determine	0.024045

GENETIC

In order to calculate the final score, which will guide the ranking of SNPs in the prioritization process, we first define an indicator function $I_k(SNP_i)$ as follows:

$$I_k(SNP_i) = \left\{ \begin{array}{l} 1, \ \text{if } SNP_i \text{ is relevant for comparison for leaf node } k \\ 0, \ otherwise. \end{array} \right\}.$$

For example, if a SNP is found out to cause frameshift mutation, then $I_{1.3.2.1}(SNP_i) = 1$ for this particular SNP. Here, we slightly shift from AHP methodology, for which a pairwise comparison using a 9-point scale is also suggested for the leaf nodes in the tree. However, considering the binary nature of the SNPs (either relevant or not relevant) even if one does a pairwise comparison, the relevant SNPs should be favored to the extremes by using 9 as intensity of importance over not relevant ones. This would lead a negligible priority vector value for those SNPs, which are not relevant for comparison. Therefore we chose to approximate using the

specified indicator function. Since the number of SNPs that will be prioritized after a GWAS is in the scale of hundreds of thousands, without this approximation a scoring would be computationally highly complex, if not uncomputable, as one would need almost 5 billion comparisons for a single node even if 100,000 SNPs are being studied. This being said, the final score $S(\text{SNP}_i)$ can be calculated for a particular SNP by using:

$$S(\text{SNP}_\text{i}) = \sum_{k=1}^{n} I_k(SNP_i)W_k \text{ for i = 1,..., } m, \quad \textbf{(3.6)}$$

where n is the number of leaf nodes, m is the total number of SNPs and W_k is the weights specified in Table 3.7.

3.4 Experimental Study

In order to evaluate the performance of our AHP based prioritization algorithm we performed a comparative study against SPOT. We have used Alzheimer Disease data introduced in Chapter 1 and separated 15% of the overall data for testing purposes (another 15% is separated for selection process explained in Chapter 4). We performed an initial quality control based filtering by using following thresholds: Minor Allele Frequency = 0.05, SNP Missingness Rate = 0.1, Individual Missingness Rate = 0.1, Hardy Weinberg Equilibrium = 0.001.

We apply BEAGLE based imputation (details of which are presented in Chapter 5) by using 0.95 as allelic r^2 threshold. By doing that the number of SNPs is decreassed from 555,850 to 517,003. In this training data set, there were 112 cases and 121 controls whereas test data set for prioritization consisted of 22 cases and 27 controls.

Following initial pass of quality control based on filtering and imputation, the GWAS is performed. We have chosen to use 0.05 as *p*-value threshold. Uncorrected *p*-values (not corrected for multiple testing) are preferred as correction via Bonferroni [107] or False Discovery Rate (FDR) [108] left no statistically relevant SNPs even with higher *p*-value thresholds. We used 0.05 thresholds for combined *p*-values for both Fisher's combination test (genes) and Fisher's exact test (pathways) for finding significant genes and pathways. Table 3.8 depicts first 20 SNPs with smallest *p*-value of association. It is evident from the *p*-values that correction for multiple testing has been too conservative for this data set. Table 3.9 presents top 10 significant genes and Table 3.10 presents top 10 significant pathways according to the calculated combined *p*-values.

Table 3. 8 Individual SNP p-values of association of GWAS for Alzheimer's Disease.

CHR	SNP	UNADJ	BONF	FDR
17	rs4795895	8.27E-07	0.4277	0.2186
17	rs1233651	1.27E-06	0.6558	0.2186
17	rs885691	1.27E-06	0.6558	0.2186
18	rs12457258	1.84E-06	0.9487	0.2372
17	rs6505403	4.91E-06	1	0.4329
16	rs7191801	6.55E-06	1	0.4329
18	rs12605132	6.71E-06	1	0.4329

Table 3.8 (cont.)

5	rs10941091	7.43E-06	1	0.4329
10	rs7911085	7.54E-06	1	0.4329
18	rs11660401	1.19E-05	1	0.6151
2	rs6729218	1.60E-05	1	0.657
2	rs13006848	1.79E-05	1	0.657
17	rs3138039	2.03E-05	1	0.657
22	rs17365991	2.21E-05	1	0.657
19	rs2075650	2.32E-05	1	0.657
10	rs2394109	2.45E-05	1	0.657
10	rs1915633	2.56E-05	1	0.657
5	rs6859143	2.57E-05	1	0.657
5	rs2548032	2.68E-05	1	0.657
5	rs7443549	2.73E-05	1	0.657

Table 3. 9 Top 10 significant genes according to combined p-value of GWAS for AD data.

Entrez Gene ID	Full Name	Location	P-Value
220963	Solute carrier family 16, member 9 (monocarboxylic acid transporter 9)	10q21.2	~0.0
10665	Chromosome 6 open reading frame 10	6p21.3	~0.0
84679	Solute carrier family 9 (sodium/hydrogen exchanger), member 7	Xp11.3-p11.23	~0.0
83891	Sorting nexin 25	4q35.1	~0.0
84182	Family with sequence similarity 188, member B	7p14.3	~0.0
9951	Heparan sulfate (glucosamine) 3-O-sulfotransferase 4	16p11.2	~0.0
654463	Fer-1-like 6 (C. elegans), null	8q24.1	~0.0
8854	Aldehyde dehydrogenase 1 family, member A2	15q21.3	~0.0
202559	KH domain containing, RNA binding, signal transduction associated 2	6q11.1	~0.0
81792	ADAM metallopeptidase with thrombospondin type 1 motif, 12	5q35	~0.0

Table 3. 10 Top 10 significant pathways according to combined p-value of GWAS for AD data.

Pathway System	Pathway Title	Gene Count	P-Value
KEGG	3-Chloroacrylic acid degradation	15	5.09E-06
WikiPathways	IL-1 NetPath 13	11	4.19E-04
WikiPathways	Ribosomal Proteins	88	4.3E-04
GO Component	Golgi apparatus	187	4.39E-04
GO Component	Cell junction	199	6.75E-04
KEGG	Metabolism of xenobiotics by cytochrome P450	70	6.87E-04
GO Function	ATP binding	229	7.19E-04
GO Process	Regulation of cell shape	41	7.31E-04
GO Process	Organic anion transport	13	0.0124
GO Process	Regulation of Rho protein signal transduction	68	0.0133

Next step of our analysis involved running AHP based prioritization algorithm. To perform a comparative study we used individual *p*-values of association and calculated combined *p*-values for genes and pathways and ran our prioritization algorithm for those SNPs with *p*-value less than 0.05. This way we calculated AHP scores for 26,545 SNPs. We selected first 10,000 SNPs for ranking for computational limitations. Then, by using individual *p*-values of association for SNPs we ran SPOT's prioritization algorithm. To compare the performance we have again selected first 10,000 SNPs from SPOT's list. Table 3.11 lists first 20 SNPs for AHP based prioritization and SPOT based prioritization. Appendix F presents the scoring details of AHP based analysis for SNPs.

Table 3. 11 Top 20 SNPs according to SPOT and AHP prioritization of GWAS for AD data.

SPOT Ranking					AHP Ranking			
RANK	**CHR**	**SNP**	**P-value**	**P-Value Rank**	**CHR**	**SNP**	**P-value**	**P-Value Rank**
1	17	rs4795895	8.3E-07	1	1	rs4651138	0.03702	19,320
2	22	rs17365991	2.2E-05	14	11	rs2070045	0.02771	14,429
3	1	rs3795263	0.00125	690	1	rs4652769	0.02968	15,460
4	2	rs4426564	0.00228	1,213	8	rs3779870	0.04915	25,918
5	19	rs2075650	2.3E-05	15	8	rs10808738	0.02714	14,117
6	18	rs12605132	6.7E-06	7	8	rs4395923	0.02714	14,119
7	6	rs9268368	0.00235	1,273	11	rs4936637	0.02771	14,428
8	5	rs10941091	7.4E-06	8	1	rs6424883	0.04604	23,915
9	11	rs667782	0.00508	2,644	1	rs10752893	0.04604	23,920
10	17	rs885691	1.3E-06	3	X	rs1800464	0.03104	16,393
11	17	rs1233651	1.3E-06	2	15	rs1606659	0.00546	2,853
12	12	rs5442	0.00533	2,763	5	rs2966952	0.00215	1,141
13	3	rs12489170	0.00539	2,797	2	rs17561	0.00276	1,472
14	2	rs6729218	1.6E-05	11	5	rs1532268	0.0381	19,848
15	2	rs13006848	1.8E-05	12	3	rs2280294	0.00581	3,039
16	18	rs12457258	1.8E-06	4	8	rs1986181	0.00909	4,726
17	20	rs6020624	0.00071	386	3	rs9881879	0.0097	5,009
18	11	rs4935801	7.2E-05	47	11	rs1010158	0.0156	8,164
19	7	rs3735080	0.00792	4,114	21	rs2830052	0.02168	11,293
20	18	rs3862683	8E-05	52	8	rs4486246	0.03844	20,080

Table 3.12 presents prediction performance of AHP based list and SPOT based list via 5-fold Cross Validation (CV) run using Naive Bayes classifier as the supervised learning scheme. In order to evaluate the prediction performance of two prioritization algorithms we used following measures:

- Accuracy: (TP + TN) / (P + N),
- Recall: TP / (TP + FN),
- Negative Predictive Value (NPV): TN / (TN + FN),
- Precision: TP / (TP + FP),
- Specificity: TN / (FP + TN),

where TP denotes True Positive, TN denotes True Negative, FP denotes False Positive and FN denotes False Negative for a $2x2$ confusion matrix.

Table 3. 12 5-fold Cross Validation results for AHP and SPOT based list of SNPs over disease trait for AD data.

	Accuracy	**Recall**	**NPV**	**Precision**	**Specificity**
AHP	**0,571**	**0,636**	**0,636**	0,519	0,519
SPOT	0,490	0,444	0,444	0,545	0,545

From Table 3.12 it is evident that AHP based prioritization outperform SPOT in most of the classification measures even though we used the same number of SNPs in analysis. From the selected list of SNPs it can also be seen that SPOT list highly depends on *p*-values of association although SPOT also makes use of various biological databases. AHP method on the other hand makes use of *p*-values in a much more effective way by integrating combined *p*-values for genes and pathways and enhancing this with data from biological databases. To compare the biological relevance of the selected SNPs for both methods, we have checked whether top ranking SNPs have been identified as significantly associated with an OMIM gene or not. Next, the SNP's association with previously described AD loci is investigated through the OMIM genes they are linked. We have found that only 50% of the SPOT list of SNPs are associated with an OMIM gene, whereas this proportion is 95% for AHP based list. It has also been verified that of the 20 SNPs in AHP list, 6 of them are experimentally associated with AD. None of the SNPs in SPOT list has been identified as associated with AD in the literature. Therefore we can conclude that our algorithm offers a much more reliable list of SNPs prioritized based on biological factors and therefore we suggest it should be preferred over SPOT for subsequent GWAS analysis. Table 3.13 summarizes our discussion.

Table 3. 13 Comparison of biological relevance of SPOT and AHP lists for AD.

	SPOT List			AHP List		
RANK	**SNP**	**OMIM Gene**	**AD Association**	**SNP**	**OMIM Gene**	**AD Association**
1	rs4795895			rs4651138	LAMC1	
2	rs17365991	TEF		rs2070045	SORL2	Yes
3	rs3795263	ACTRT2		rs4652769	LAMC1	
4	rs4426564	ADRA2B		rs3779870	CYP7B3	
5	rs2075650	TOMM40		rs10808738	CYP7B1	
6	rs12605132			rs4395923	CYP7B4	
7	rs9268368			rs4936637	SORL2	Yes
8	rs10941091	ADAMTS12		rs6424883	LAMC1	
9	rs667782	HYLS1		rs10752893	LAMC1	
10	rs885691			rs1800464	MAOA	
11	rs1233651			rs1606659	CHRNA7	
12	rs5442	GNB3		rs2966952		
13	rs12489170	PARP15		rs17561	IL1A	Yes
14	rs6729218			rs1532268	MTRR	
15	rs13006848			rs2280294	MFI2	
16	rs12457258			rs1986181	CYP7B2	
17	rs6020624			rs9881879	MME	Yes
18	rs4935801	UBASH3B		rs1010158	SORL1	Yes
19	rs3735080			rs2830052	APP	Yes
20	rs3862683	DCC		rs4486246	CYP7B5	

3.5 Conclusion

In this chapter we have presented a novel prioritization algorithm that can be used to find a biologically and statistically relevant set of SNPs for use in the subsequent GWAS. The backbone of the algorithm is based on well-known multi-criteria decision making method Analytic Hierarchy Process. We have developed a hierarchy tree, whose nodes represent biological and statistical criteria that would be used to evaluate the SNPs. In each node we have specified weights by following AHP methodology and this allowed us to calculate a final score for each SNP depending on *p*-values of association and functional information gathered from major biological databases. Then, we sorted SNPs according to the prioritization scores. The prediction performance of the prioritized SNP set is an important measure for utilization of the proposed approach in subsequent studies. Therefore, the performance of AHP is measured and compared to an alternative application, SPOT, by applying 5-fold Cross Validation using Naive Bayes as supervised learning scheme. When the classification performance measures are compared we have shown that AHP based SNPs outperformed SPOT's prediction performance. Following that, we compared the biological relevance of top ranking SNPs in SPOT based and AHP based lists. We have found out that AHP offers a much more reliable prioritization as it points us to SNPs, which are highly associated with OMIM genes and loci proven to be associated with AD in the literature. Therefore we are suggesting that the novel AHP based prioritization scheme steps forward as a reliable and effective method for use in GWAS bringing a new, objective and structured approach to SNP biomarker scoring and ranking considering overall functional effects of the SNP in a biological system.

CHAPTER 4

SELECTION OF REPRESENTATIVE SNP SETS FOR GENOME-WIDE ASSOCIATION STUDIES: A METAHEURISTIC APPROACH

After the completion of Human Genome Project in 2003, it is now possible to convey the research studies to associate genetic variations in the human genome with common and complex diseases. The Single Nucleotide Polymorphism (SNP) biomarkers across the complete sets of DNA, or genomes, of 11 different populations are scanned for revealing genetic risk factors and quantitative traits associated with human diseases. The current challenge is to utilize the genome data efficiently and to develop tools that improve our understanding of etiology of complex diseases. Many of the algorithms needed to solve these problems are strongly supported by management science and operations research (OR) methods. One application is to select a subset of SNPs from the whole SNP set that is informative and small enough to convey subsequent association studies. In this chapter, we present an OR application for representative SNP selection that makes use of our novel Simulated Annealing (SA) based feature selection algorithm.

4.1 Introduction

Recently, the research focus in molecular epidemiology is to find genetic markers, haplotypes[10], and potentially other variables that together contribute to a disease and serve as good predictors of the disease phenotypes. Complex diseases are typically associated with multiple genetic loci and several environmental factors. Therefore, it is essential to investigate all polymorphisms located in the functional regions of candidate genes [4, 5] and integrate the information about the network of genes involved in biological systems of major physiological importance [6] for thorough analysis of these biologically complex diseases.

[10] A *haplotype* is a combination of alleles (DNA sequences) at different places (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome.

Association studies are among the promising ways of dealing with the problem of finding disease causing variants and such association studies typically make use of SNPs as they are the most common form of genetic variations and they can represent an individual's genetic variability in greatest detail [7]. However, the enormous number of SNPs (estimated up to 30 million) makes it infeasible to gather information and perform analysis on all of the SNPs in the human genome. Thus, while performing a disease association study, the geneticist would prefer to experimentally test for association by only considering a subset of the entire SNP set instead of all of the SNPs, thereby considerably saving resources (alternatively, increasing the power of the statistical tests by increasing the number of individuals) as well as making the problem computationally feasible. Therefore, selecting a subset of SNPs that is informative enough to perform association studies but still small enough to reduce the analysis workload, to which we refer as ***representative SNP selection***, has become an important step for disease-gene association studies.

Reducing biological and statistical redundancy from hundreds of thousands of SNPs is the key for representative SNP selection. Dealing with many dependent association tests is one of the emerging issues on the statistical and computational side. SNP vs. disease data, in addition to being large, redundant, diverse and distributed, has three important characteristics posing challenges for the data analysis and modeling: (1) heterogeneity, (2) a constantly evolving biological nature and (3) complexity. Therefore intelligent methods are needed to find SNPs associated with the disease and extract biologically relevant subsets. The problem of SNP selection has been proven to be NP-hard (i.e., non-deterministic polynomial-time hard, harder than those problems that a non-deterministic turing machine can solve in polynamial time) in general [109] and current selection methods possess certain restrictions and require use of heuristics for reducing the complexity of the problem. OR methods have been used recently to the problem of representative SNP selection [110-114].

In this chapter, we will present a method for selecting representative SNP subset for stronger association with complex disease after following an integrative biological scoring and filtering approach. An OR class novel feature selection method based on Simulated Annealing (SA) has been developed for representative SNP selection, in which we try to maximize tagged SNP prediction while minimizing cardinality of the selected SNP subset.

4.2 Problem Definition

The aim of the representative SNP selection approach is to find a minimal subset of SNPs, whose allele[11] information can explain the whole set of SNPs in the candidate region under study (a whole chromosome, or a target region) to the greatest detail. A formal definition of the problem can be stated as follows: Let $S = \{ SNP_1 , ..., SNP_n \}$ be a set of n SNPs in a candidate region and $G = \{g_1, ..., g_m\}$ be a data set of m genotypes, where each genotype g_i

[11] An *allele* is defined as one of two or more forms of the DNA sequence of a particular gene.

consists of the consecutive allele information of the n SNPs: SNP_1 , ... , SNP_n. For simplicity we represent $g_i \in G$ be a vector of size n whose vector element is 0 when the allele of a SNP is homozygous dominant[12], 1 when it is heterozygote[13] and 2 when it is homozygous recessive[14]. Alternatively, data can be gathered from a case-control study. In this case, researcher would also have a hand on the phenotype data for the particular patient and this information can be matched with genotype information. Phenotype variable will be represented by P and it takes the value -1 if it belongs to case group and 1 if it belongs to control group. Matrix A in Figure 4.1 represents such data.

Suppose that the maximum number of SNPs that can be selected is k (which can be alternatively be a variable for the problem), and a function $f(R|G,P)$ evaluates how well the allele information of SNPs in subset $R \subset S$ retains the allele information of all SNPs in S based on the genotype data G and classification performance of selected set R on disease phenotype are.

	SNP_1	SNP_2	SNP_3	.....	SNP_n	P
g_1	0	0	0	.....	1	1
g_2	2	1	1	.....	1	1
g_3	2	0	0	.....	2	-1
.	.	.	.	.....	.	.
.	.	.	.	.....	.	.
.	.	.	.	.....	.	.
.	.	.	.	.....	.	.
.	.	.	.	.....	.	.
g_m	1	0	2	.....	1	1

Figure 4. 1 SNP-Genotype Matrix A.

Given S, G, P and k the Representative SNP Selection problem is the following optimization problem:

max $F(R|G,P)$

subject to: $R \subset S$,

$|R| \leq k$,

$k > 0$,

k integer.

To solve *Representative SNP Selection problem*, one needs to find an optimal subset of SNPs, R, of size less than or equal to k, based on the given evaluation function $F(R|G,P)$. From a set theoretic point of view, it is computationally intractable to examine all possible subsets of the given set of SNPs to select a set of representative markers, except for very small data sets. The problem is proven to be NP-hard [110]. To cope with this difficulty, it is possible to divide

[12] An individual that is **homozygous dominant** for a particular trait carries two copies of the allele that codes for the dominant trait.
[13] A person possessing two different forms of a particular gene, one inherited from each parent.
[14] An individual that is **homozygous recessive** for a particular trait carries two copies of the allele that codes for the recessive trait.

Representative SNP selection into three largely independent steps: (1) identifying genomic segments where the selection will be performed, (2) defining a measure to quantify how well a set of SNPs can predict all observed and/or unobserved SNPs and (3) searching a minimum set of Representative SNPs that meets a desired threshold.

4.3 Related Work

Application of statistical hypothesis-testing procedures is the basic approach for finding genotype-phenotype associations. The null hypothesis to be tested is that there is no difference between two study groups with respect to the genotype frequencies (i.e., genotype proportions) observed in each group. The chi-square and Fisher's exact tests may be applied in this task [115]. Odds-ratios are also commonly used to indicate differences between groups on the basis of their genotype frequencies. Methods for multiple testing (such as Bonferroni or False Discovery Rate) in high-dimensional settings can be applied when many SNPs are considered simultaneously.

In addition to statistical hypothesis testing in which causative SNPs are identified, one may choose to use classification models for genotype-phenotype association modeling. This can be done by representing different genotypes for a particular SNP as inputs and phenotype as label. Different statistical and machine learning techniques, such as logistic regression and support vector machines, can be applied for this purpose. Not only the genotype information extracted from multiple SNPs but also information related to environmental exposure factors and other biomarkers can be incorporated by introducing multivariable statistical and machine learning models in this context. Tagging and different feature selection procedures are useful to improve the prediction performance of multiple-SNPs models. The former can be applied to problems with a large number of SNPs in which haplotype data is present. Feature selection is recommended to reduce the number of highly-correlated SNPs, in which high Linkage Disequibrium[15] (LD) makes it difficult to select true disease causing variant. These methods are presented in the subsequent sections.

4.3.1 Statistical Methods

In order to select a subset of SNPs in genome-wide complex disease association studies, various statistical measures and testing based approaches have been introduced specific to the problem domain. The paper [116] proposes a sliding window approach, which made use of combination of p-values from multiple independent tests by making use of

$$X^2 = -2 \sum_{i=1}^{m} log(p_i) \sim X^2_{2m} . \quad \textbf{(4.1)}$$

Here, p_i denotes *p*-value of association between SNP_i and disease presence and *m* is the number of SNPs in the sliding window. It is shown that test statistic X^2 follows a Chi-square distribution with *2m* degrees of freedom. The basic advantage of this approach is that it takes into

[15] *Linkage disequilibrium* is the non-random association of alleles at two or more loci.

account the ordering of SNPs on the chromosome and allows detection of chromosome regions with significant associations by merging adjacent windows [117, 118]. However an implicit assumption is made that the distance between any two adjacent SNPs is constant.

Other scan statistics have been developed that also considers the ordering and spacing of SNPs on the chromosome [119-122]. For example in [119] a two-step procedure was presented for calculating chromosomal scan statistic: (1) identify SNP clusters and (2) extract clusters with significant disease association.

It is assumed here that position of each SNP follows a Poisson distribution. Therefore length between two adjacent SNPs is assumed to have exponential distribution and distance between two particular SNP is assumed to follow a Gamma distribution. Using these assumptions one can identify the clusters of SNPs by testing the hypothesis that whether the observed length between a set of SNPs is equal or less than the expected length. If the hypothesis is rejected then this group of SNPs is identified as a cluster. Then to test the significance of disease association for a particular cluster Pearson Chi-square p-values are calculated. However this type of scan approaches has the disadvantage that they do no incorporate gene-gene interactions.

4.3.2 Tagging and Machine Learning Methods

One obvious observation from the formal definition of representative SNP selection problem is the selected subset's dependence on the function *F*. In the literature, various objective functions have been defined to represent the allele information of genotypes in *G* using SNPs in *S* and solve the problem accordingly. One can classify the proposed approaches into three categories according to how they try to measure the allele information of genotypes: (1) Haplotype Diversity based approaches, (2) Pairwise Association based approaches and (3) Predicting Tagged SNPs.

Haplotype Diversity based approaches are inspired by the fact that DNA can be partitioned into discrete blocks such that within each block high LD is observed and between blocks low LD is observed [123, 124]. As a result of this feature, number of distinct haplotypes consisting of the SNPs is very small across a population. Hence, one would try to find the subset of SNPs, which are responsible for the "limited haplotype diversity" in order to find the representative SNP set. Different studies have been conveyed to see how well diverse haplotypes can be distinguished depending on a selected "diversity measure" and chose the best one. A detailed explanation on the different types of measures used in the literature has been provided in [125-128]. The usual approach among these methods is to exhaustively list and search SNPs through every subsets of the set of haplotypes. Therefore, only a small number of SNPs can be analyzed. To cope with this problem, efficient heuristics have been proposed using Dynamic Programming [111-113], Principal Component Analysis [129-131] and Greedy Algorithm [132]. Although haplotype diversity based methods are simple to implement they depend on the block

partitioning method used for a target locus. In addition, the union of the candidate SNP sets for each block may not be an optimal set for the overall locus.

Pairwise Association based approaches are based on the principle that all the SNPs in the target locus are highly associated with at least one of the SNPs in the selected SNP subset. This way, although a SNP that can be used to predict a disease causing variant may not be selected as a representative SNP, the association may be indirectly assumed from the selected SNP that is highly associated with it. The associations between SNPs can be estimated using LD. The common solution approach for these methods is to cluster the SNPs into different subsets and choose a representative SNP (or SNPs) from each cluster [133-135]. Although with their $O(cgs^2)$ complexity (c being number of clusters, g being number of genotypes and s being number of SNPs) pairwise association methods are so much faster than haplotype diversity based methods, they have a major shortcoming as they cannot explain multi-SNP dependencies [109] and they tend to select more tag SNPs [136].

Predicting Tagged SNPs is motivated by the idea of reconstructing the genotype data from an initial set of selected SNPs in order to minimize the error of prediction for unselected SNPs. Those prediction methods have a certain advantage over Pairwise Association methods as they would take multi-SNP dependencies into consideration. Bafna et al. [109] proposes a measure called "Informativeness" and used dynamic programming to solve the problem of finding the optimal subset of SNPs that can best predict the remaining (tagged) SNPs. Let $E^s_{i,j}$ be the event that genotypes g_i and g_j have a different allele at SNP s, and $E^S_{i,j}$ be the event that genotypes g_i and g_j have a different allele at some SNP in S. To measure how well a set of SNPs, $S = \{SNP_1, ..., SNP_k\}$, can predict the SNP, s, the used measure is as follows:

$$I(S,s) = P_{i \neq j}(E^S_{i,j} \mid E^s_{i,j}). \quad \textbf{(4.2)}$$

A more recent approach with using dynamic programming is proposed by Halperin [110] through fixing the number of representative SNPs for each tagged SNP to 2. Lee et al. [137] proposed a heuristic algorithm that uses the probabilistic framework of Bayesian networks to effectively identify a set of predictive SNPs. The heuristic approach improves the tag SNP selection compared to the current predicting based methods by allowing multi-allelic prediction (instead of bi-allelic) and not restricting the number of representative SNPs.

Our proposed algorithm is among the methods that utilizes a heuristic approach as in the last example and explained in detail in Section 4.4.

4.4 Proposed Methodology

The proposed representative SNP set selection methodology in this research can be divided into 4 consecutive steps based on the working SNP set (1) Initial Set, (2) Filtered Set, (3) Biologically and Statistically Relevant Set (4) Representative SNP set as presented in Figure 4.2. At first step the initial filtering based on quality control measures Minor Allele Frequency, missingness and Hardy-Weinberg equilibrium is applied. Next, the calculations for multiple

testing adjusted p-values of association are performed and then biologically most relevant SNPs using the AHP procedure is selected as explained in Chapter 3. Finally, the proposed SA feature selection algorithm is applied for each individual chromosome and the selected SNP subsets are merged to reveal the final representative SNP set.

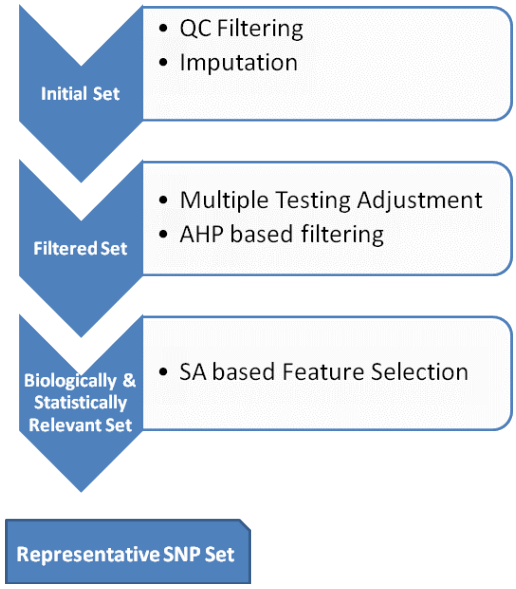


Figure 4. 2 Process steps for finding representative SNP set.

In Section 4.2, it has been stated that the optimal SNP subset depends on the selected evaluation function. It is also pointed out that maximizing the prediction accuracy of selected SNPs over unselected SNPs is an approach used in the literature for representative SNP selection. We set our goal as to find a minimum size set of representative SNPs and a prediction algorithm, such that the prediction error is minimized. Then our objective function becomes:

$$\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{T_i}) + NaiveBayes(G_R, P), \quad \textbf{(4.3)}$$

where G_R denotes genotype data related with representative SNP set R, G_{T_i} denotes genotype data related with $SNP_i \in S \backslash R$ and

$$NaiveBayes(F,L) = argmax_L \; P(L = l) \prod_{i=1}^{n} p(F_i = f_i | L = l), \quad \textbf{(4.4)}$$

denotes a Naive Bayes classifier where F is the feature set (SNP set in our context) and L is the label. We calculate 5-fold Cross Validation (CV) based classification and find classification accuracy.

In order to solve our problem, we used Simulated Annealing (SA) [138], which is a local search algorithm that strives for the best solution starting from a randomly created solution. Each step of the SA algorithm replaces the current solution by a "nearby" solution. The new solutions are chosen depending on an evaluation function and a global parameter T (temperature). T value is gradually decreased during the process. We have proposed to utilize the SA, for the first time, as a feature selection approach for selecting representative SNP set. Fundamental to the SA structure is the binary coding scheme. Let C_i represents i^{th} coding where each code containing n

SNPs (dimension). Each code of the length n is a sequence over $\{0, 1\}^n$ (0 represents a non-selected SNP and 1 represents a selected SNP). For example, assume there is a code represented by C_i = {1, 0, 1, 0, 0, 1, 0}. In this encoding scheme SNP_1, SNP_3 and SNP_6 are selected SNPs. A neighbor for a coding scheme C_i is another coding scheme, which is one bit different than $C_{i.}$

A random binary coding of size n as an initial solution is created and the accuracy of the solution using Naive Bayes by calculating the mean classification error for (n-k) supervised learning iterations, where k is the number of selected SNPs in a particular iteration have been tested. The proposed SA algorithm runs for a certain amount of steps (user defined), with a tradeoff between accuracy and the number of SNPs in the representative SNP set. Therefore we have also tried to minimize the number of chosen SNPs (k). The pseudocode of the proposed algorithm is given in Algorithm 4.1.

Algorithm 4. 1 Simulated Annealing based representative SNP selection.

Input:

s_0 initial randomly selected SNP set

t simulated annealing parameter temperature.

d simulated annealing parameter decreasing factor.

c_{max} number of iterations.

Output:

s_{best} representative SNP set

1. $s \leftarrow s_0; e \leftarrow E(s)$
2. $s_{best} \leftarrow s; e_{\text{best}} \leftarrow e$
3. **For** $c = 1$ to $c_{\max}$
   $s_{new} \leftarrow \text{neighbor}(s)$
   $e_{new} \leftarrow E(s_{new})$
4. **if** $e_{new} > e_{best}$ **then**
5. $s_{best} \leftarrow s_{new}; e_{best} \leftarrow e_{new}$
6. **if** $P(e_{new}, t) <$ **random() then**
7. $s \leftarrow s_{new}; e \leftarrow e_{new}; t = t * d$
8. **Next** c
9. **return** s_{best}

Here, s_0 is the initial randomly selected SNP set (R) and $E(s)$ is an evaluation function denoted by:

$$\boldsymbol{E(s) = w(\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{T_i}) + NaiveBayes(G_R, P)) + (1 - w)k}\,, \quad \textbf{(4.5)}$$

presenting our objective function. Cardinality of the representative SNP set is denoted by k. We have used two user specified arguments for the algorithm: c_{max} denotes the number of iterations and w ($\boldsymbol{0 \leq w \leq 1}$) denotes weight that specifies tradeoff. If a w value smaller than 0.5 is chosen, decreasing the cardinality of representative SNP set is favored over classification performance of the SNP set. One need to note that, while other parameters are the same, use of a smaller w does not necessarily mean to have a set with smaller cardinality as a random set of SNPs at initialization step is chosen. The simulated annealing parameters are represented as t, d and $P(E(S), t)$ (energy), where temperature is denoted by t and decreasing factor is denoted by d. $P(E(S),t)$ is calculated by:

$$P(E(s), t) = \exp\,(-\frac{E(s)}{t}). \quad \textbf{(4.6)}$$

4.5 Experimental Study

In order to test the prediction performance of the representative SNP set, we first applied an initial split (70% training, 15% test for AHP based prioritization and 15% test for SA selection) on the experimental genotyping data as described in Chapter 3. Therefore 112 cases and 121 controls for training data set and 15 cases and 34 controls for test data set for representative SNP selection are denoted. AHP based prioritization is performed where first 10,000 SNPs are selected for subsequent analysis of representative SNP selection. Following that, we ran our simulated annealing based representative SNP selection algorithm on the training set. As the algorithm is based on the idea of selecting the subset of SNPs, which best predicts the genotypic profile of the remaining SNPs for a particular genomic region; we ran the algorithm for each chromosome individually and merged the selected representative SNP sets as the final representative SNP set. For computational reasons we ended the algorithm if evaluation function value does not change in 40 SA step in a row. The prediction performance (average accuracy) of the selected SNPs on unselected SNPs is presented in Tables 4.1 for each chromosome and for different w values (0.3, 0.5 and 0.7, respectively).

Table 4. 1 Prediction performance of representative SNP selection algorithm: t = 10, d = 0.1, c_{max} = 1,000.

	w = 0.3			w = 0.5			w = 0.7		
CHR	**INI**	**SEL**	**ACA**	**INI**	**SEL**	**ACA**	**INI**	**SEL**	**ACA**
1	773	65	0.607	773	61	0.627	773	106	0.665
2	699	46	0.593	699	56	0.638	699	109	0.675

3	786	53	0.597	786	70	0.624	786	115	0.657
4	521	32	0.601	521	50	0.644	521	78	0.689

Table 4.1 (cont.)

5	592	45	0.627	592	46	0.668	592	70	0.683
6	688	47	0.639	688	45	0.663	688	83	0.682
7	614	30	0.608	614	46	0.628	614	91	0.669
8	571	25	0.621	571	56	0.648	571	76	0.666
9	413	19	0.609	413	24	0.641	413	60	0.674
10	473	24	0.621	473	52	0.646	473	61	0.67
11	504	37	0.6	504	43	0.639	504	66	0.667
12	492	21	0.583	492	34	0.619	492	69	0.658
13	250	9	0.632	250	13	0.653	250	41	0.709
14	328	23	0.618	328	20	0.649	328	43	0.682
15	297	15	0.622	297	15	0.648	297	37	0.681
16	332	23	0.616	332	24	0.656	332	50	0.671
17	373	22	0.588	373	29	0.606	373	50	0.649
18	300	22	0.621	300	23	0.651	300	54	0.703
19	232	5	0.609	232	19	0.627	232	23	0.633
20	286	17	0.614	286	29	0.654	286	57	0.71
21	150	2	0.614	150	15	0.68	150	24	0.72
22	148	8	0.627	148	6	0.67	148	28	0.733
X	178	6	0.633	178	11	0.649	179	12	0.641
TOTAL/AVERAGE		**596**	**0.613**		**787**	**0.645**		**1403**	**0.678**

CHR: Chromosome number, **SEL:** Selected number, **ACA:** Average classification accuracy.

Using representative SNP selection algorithm we have managed to decrease the dimensions considerably. For example for $w = 0.5$ the number of SNPs is decreased from 10,000 to 787 for AD data. Average classification accuracy for the representative SNP set over unselected for each chromosome is 0.645. This means that although the dimension is decreasing more than 90%, we are not introducing a significant information loss (in terms of classification accuracy over unselected SNPs). To observe the prediction performance of the selected set over the **disease phenotype**, we have compared the performance against two filtering based attribute selection scheme from the WEKA tool set (Relief-F and Chi-Square). In order to achieve that, we have selected the same set of SNPs used as the training set for the test sets and applied a 10-fold Cross Validation run using Naive Bayes classifier as the supervised learning scheme. Classification measures are explained in Chapter 3 in greater detail. Results are as presented in Table 4.2.

Table 4. 2 Prediction performance comparison for SA algorithm and WEKA based algorithms.

	w = 0.3, 596 SNPs			w = 0.5, 787 SNPs			w = 0.7, 1403 SNPs		
Measure	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F
Accuracy	**0.5306**	0.6327	0.4898	**0.4898**	0.6327	0.4898	0.4694	0.4898	0.4898
Recall	0.5333	0.6667	0.6000	0.3333	0.6000	0.6000	0.4000	0.4000	0.4667
NPV	**0.7200**	0.8077	0.7143	0.6552	0.7857	0.7143	0.6538	0.6667	0.6800
Precision	**0.3333**	0.4348	0.3214	0.2500	0.4286	0.3214	0.2609	0.2727	0.2917
Specificity	**0.5294**	0.6176	0.4412	**0.5588**	0.6471	0.4412	**0.5000**	0.5294	0.5000

The comparison between the SA and WEKA algorithms revealed that the proposed SA based algorithm outperforms Relief-F for w = 0.3 and it shows comparative performance to WEKA based attribute selection schemes for high dimensional data. For greater values of c_{max} we expect to have better results in terms of classification performance, especially if the premature endings, which are applied to gain time, are omitted. Still it is important to note that even though SA algorithm performs better with high dimensional data, it is also very demanding and time consuming.

4.6 Conclusion

In this chapter, we have presented a novel representative SNP Selection algorithm based on the idea of maximizing prediction accuracy of selected SNP set over non-selected. We have developed a methodology based on SA in order to select a representative set among the top 10,000 prioritized SNPs after GWAS. We have performed biological prioritization and SNP selection on real life data belonging to Alzheimer's disease and showed that the proposed SA based algorithm is capable of reducing the dimension of the data without much information loss. We have performed a comparative study with two well-known attribute selection schemes. Our algorithm performed reasonably well against filtering based approaches.

We have showed the benefits of exploring alternative analysis methods in the representative SNP selection while introducing a new research and application are for operational research (OR) in field of applied mathematics. We hope that our work will encourage OR community to expand their studies to representative SNP selection and other topics in biomarker research and suggest other advanced OR methods such as conic programming, multi-objective optimization, stochastic programming, and optimization in data mining and in computational statistics for research topic in bioinformatics.

CHAPTER 5

METU-SNP: AN INTEGRATED SOFTWARE SYSTEM FOR SNP-COMPLEX DISEASE ASSOCIATION ANALYSIS

We have implemented the ideas stated in the previous chapters to build a java based Integrated Software System (ISS), which is specifically designed as an all-in-one GWAS application. We call this tool METU-SNP and we believe that as the name implies it will be regarded as *Most Effective Tagging Utility* among researchers of molecular epidemiology. It makes use of data from major public databases such as dbSNP, Entrez Gene, KEGG, Gene Ontology etc. It is equipped with a state-of-the-art AHP based SNP prioritization and Gene Set Enrichment Analysis frameworks. It also has the ability to select representative SNPs by making use of SA based machine learning algorithms. Modularity and extendibility give METU-SNP a great advantage over existing platforms. In this chapter we will introduce the software and details of the functionality.

5.1 System Architecture

METU-SNP is a desktop application written in Java. The program was written with Java Swing GUI (Graphical User Interface) architecture using JDBC to interact with the database (See Figure 5.1). Java Swing is routed in a slightly modified version of classical model-view-controller (MVC) design: **model-delegate** pattern. The model-delegate pattern combines the view and the controller into a single object that presents information to and interacts with the user and that object delegates to its model, which holds data specific to the application. This architecture allows:

- Cross-platform consistency and easy maintenance,
- Plugging of various custom implementations for extendibility,
- Customization through fine control over the details of rendering of a component,
- Changing look and feel at runtime and therefore configurability.

In the delegator object, there exist algorithms related with data imputation, association analysis, SNP prioritization and selection written with Java via partly involvement of 3rd party

tools alongside with the User Interface (UI). Model object holds biological data incorporated from major repositories resides in a relational database. The application can be installed and run on a standalone computer in which Java Run Time Environment and MySQL database is previously installed.

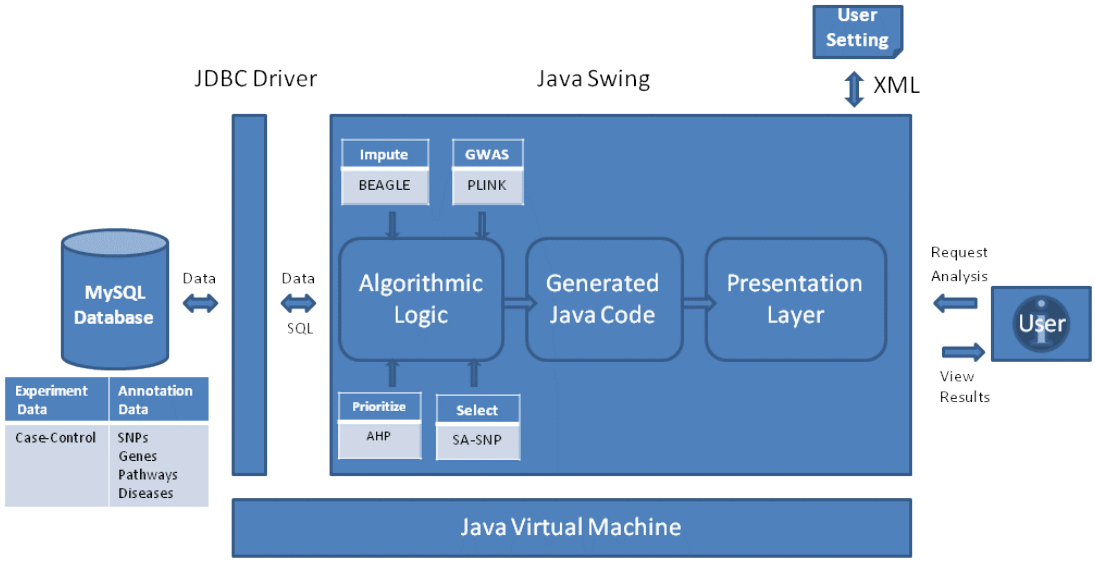


Figure 5. 1 System architecture for METU-SNP.

5.2 Third Party Tools

METU-SNP utilizes functionality offered by 3rd party tools: PLINK [11], BEAGLE [139] and WEKA [140]. The logical flow of the METU-SNP system that also shows involvement of third party tools is presented in Figure 5.2.

5.2.1 PLINK

Developed by Shaun Purcell at the Center for Human Genetic Research, PLINK is an open-source whole genome association analysis toolset that offers immense set of functions for GWAS including: data management, summary statistics for quality control, population stratification detection, basic association testing, copy number variant analysis, meta-analysis, result annotation and reporting. PLINK can be regarded as the Swiss-army-knife of the genetic epidemiologist as there is little that cannot be done with it. However sadly, PLINK is offered as single executable program, that needs to be run from command line. This definitely isn't the most user friendly approach for the researcher. The functionality is offered in a set of operations, which can be cumbersome to excel to the fullest if the user is not computer-savvy. In fact it is quite easy to get lost within the set of commands and sometimes finding the right command to perform the required function takes time. Therefore we used PLINK as a callable executable file to perform the required operations:

1. To divide the overall data set into chromosomes to be used as input to BEAGLE (for imputation).

2. To perform quality control based filtering including the removal of those SNPs or individuals that does not comply with user defined thresholds.

3. To perform association analysis

4. To perform certain data management tasks (extract set of SNPs, individuals etc.) that would be otherwise hard because of the size of the data sets (in gigabytes).

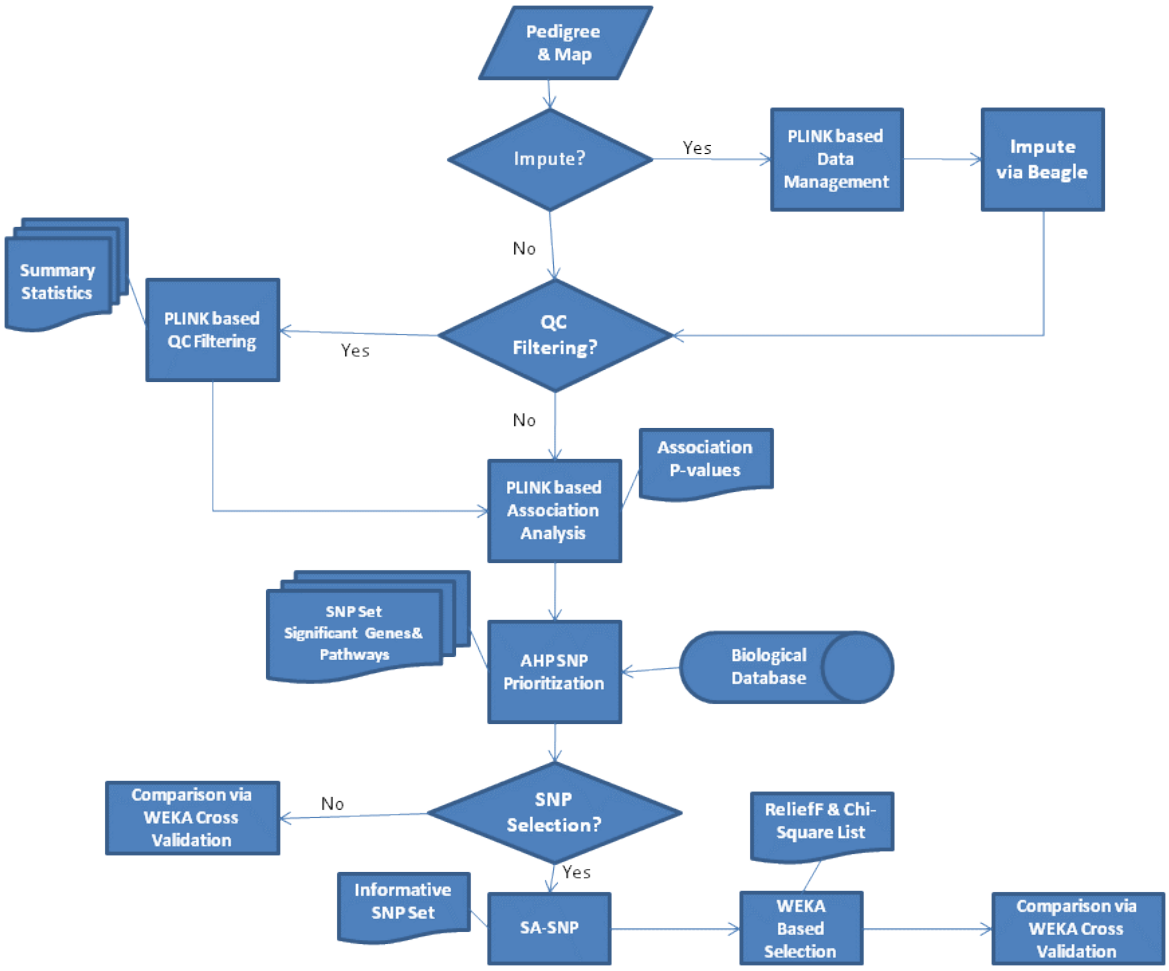


Figure 5. 2 Logic flow of METU-SNP software system.

PLINK is crucial within our framework as we rely on it for two major steps of the analysis process: (1) Quality control based filtering and (2) association analysis. In the former step the user is allowed to filter out those SNPs and individuals that do not conform to the user defined thresholds. Related thresholds are for missingness, minimum allele frequency and Hardy-Weinberg equilibrium (See Appendix E for mathematical background of the relevant processes). We have also allowed the user to perform a multiple adjusted (corrected) association analysis using functionality offered by PLINK. User is given the option to select (1) uncorrected *p*-value, (2) Bonferroni single-step adjusted *p*-value [107] or (3) step up False Discovery Rate (FDR) control *p*-value [108] with appropriate thresholds for further analysis.

5.2.2 BEAGLE

Imputation is a statistical method to substitute a calculated value for a missing data point. In order to increase the power of existing marker sets in GWAS, various methods have been proposed to predict sporadic missing data by imputation [141]. Imputation can be used to replace missing/un-genotyped data when genotyping percentage has failed (by exceeding a certain threshold specified by the researcher) in certain number of the typed SNPs in order to expand the coverage of SNPs in case control data sets beyond what has been genotyped. Lately, imputation is becoming a part of the GWAS and it has been used in the meta-analysis of different diseases and traits [142, 143]. Several software programs have been developed so far to account for the missing data in genetic data sets. Among those are BEAGLE, IMPUTE [13], MACH [144], fastPHASE [145] and PLINK. Recently, it has been stated that BEAGLE, IMPUTE and MACH step forward in terms of accuracy and execution time [141, 146]. We have chosen to integrate BEAGLE as it is presented as a standalone executable java archive (jar) file and it uses a very similar data format as input to that of PLINK's, which has been used for association studies within the METU-SNP framework.

BEAGLE is a software program that can be used for genotype imputation, haplotype phase inferring, and genetic association analysis. It can be effectively and efficiently used for data sets with the order of hundreds of thousands markers and thousands of samples. BEAGLE is written in Java and runs on most computing platforms (e.g., Windows, Unix, Linux, Solaris, and Mac). We offer imputation functionality via integration of executable jar file of the BEAGLE within our framework. We used scripts that allowed us to convert PLINK outputs to BEAGLE inputs and vice and versa. BEAGLE assumes that the supplied genotype data is on the same strand for each individual and crashes otherwise; therefore the same assumption is used for METU-SNP. A python script that allows strand-checking and switching is offered in BEAGLE website[16].

In the METU-SNP framework the whole genome genotyping data is split into individual chromosome files to be used as input files for BEAGLE. After imputation process, BEAGLE provides three output files for each chromosome: (1) phased file (.phased.gz), (2) genotype probabilities file (.gprobs.gz) and (3) allelic r^2 file (.r2). Phased file gives imputed missing data, whereas the other two files indicate how accurate the imputation process has been for a particular marker. Allelic r^2 file contains two columns: the marker identifier and estimated squared correlation ($0 \leq r^2 \leq 1$) between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker. Larger values of allelic r^2 indicate more accurate genotype imputation. METU-SNP allows users to specify a threshold (default being 0.95), providing a flexibility to include only "well" imputed markers in subsequent analysis.

[16] http://www.stat.auckland.ac.nz/~browning/beagle/strand_switching/strand_switching.html

5.2.3 WEKA

WEKA is an open source machine learning and data mining tool developed and maintained by University of Waikato, New Zealand. An extensive set of algorithms for pre-processing, classification, regression, clustering and association is included within the WEKA collection and can be directly applied to a data set using the GUI offered by the software or calling from an independent java code. We used the latter approach and used WEKA's executable jar file for evaluating the prediction accuracy of the selected SNP sets after AHP based prioritization and simulated annealing based informative SNP selection steps via Cross Validation. WEKA is also used for comparison purposes via filtering based attribute selection schemes Relief-F and Chi-square as presented in detail in Chapter 4.

5.3 The METU-SNP Database

METU-SNP database is MySQL 5 based relational database that incorporates data from major biological databases. Entity-Relationship diagram of the database that presents table details can be found in Figure 5.3. Sources of data for individual SNPs, genes, pathways and diseases is discussed in detail under the corresponding subtitles.

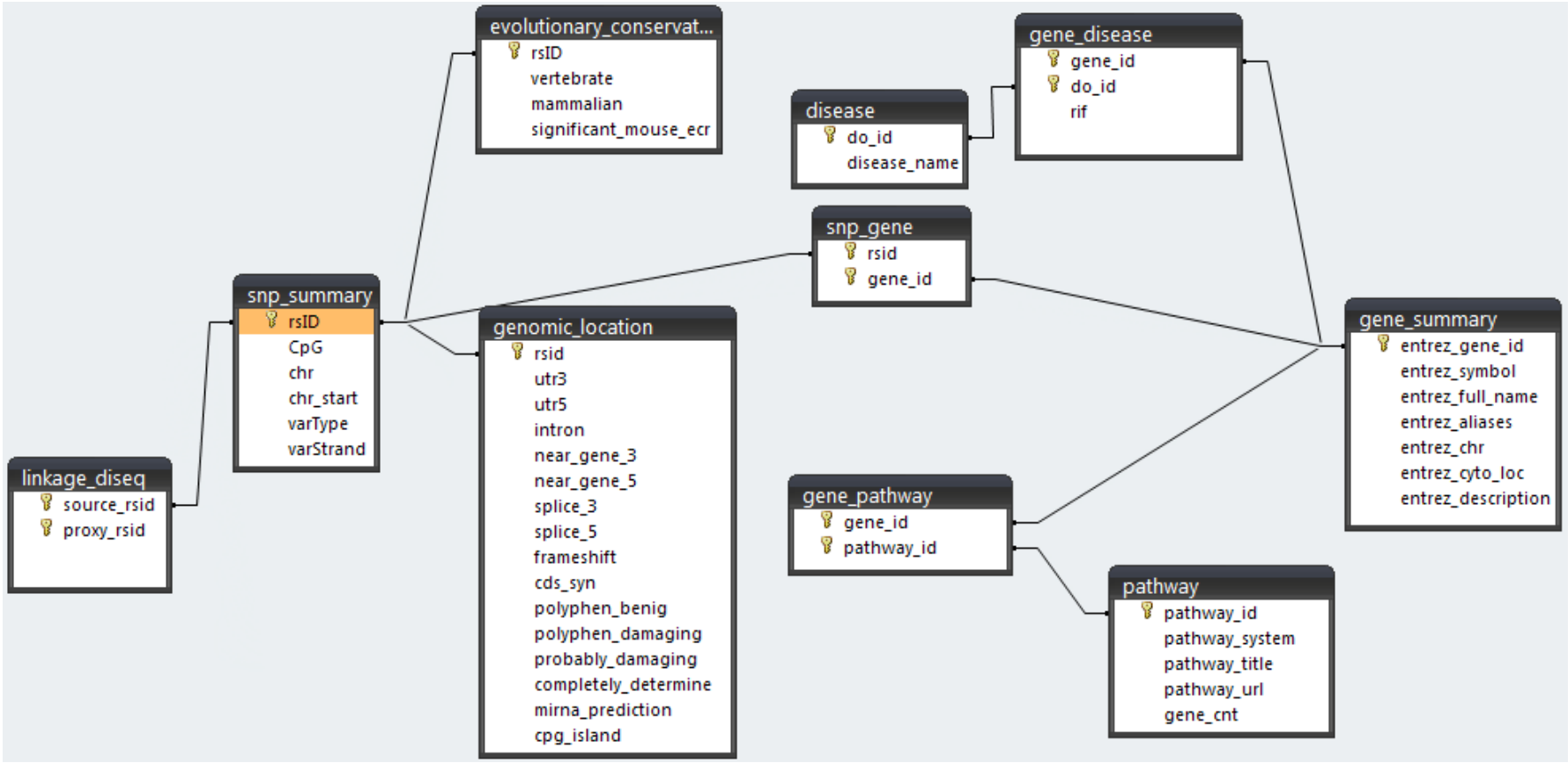


Figure 5. 3 ER diagram of METU-SNP relational database.

5.3.1 SNP Data

Our SNP related information is mainly based on SNPLogic's integrated database [72]. Annotation information is presented in detail in Table 5.1.

Table 5. 1 SNP Annotation sources integrated into METU-SNP.

Resource	Description of extracted annotations	SNPs annotated
dbSNP, build 128	SNPs rs IDs and basic annotations	11,833,664
HapMap	Allele frequencies from HapMap project	3,967,349
Tagger (HapMap)	Haplotype tags in CEU (0.8 r2 cutoff)	tags 695,153,tagged 2,009,725
Tagger (HapMap)	Haplotype tags in CHB (0.8 r2 cutoff)	tags 580,509,tagged 1,908,721
Tagger (HapMap)	Haplotype tags in JPT (0.8 r2 cutoff)	tags 562,741,tagged 1,883,580
Tagger (HapMap)	Haplotype tags in YRI (0.8 r2 cutoff)	tags 1,282,451,tagged 1,571,139
UCSC	PhastCons conserved elements, 28-way vertebrate	434,235
UCSC	PhastCons conserved elements, 28-way mammal	322,704
Delta-MATCH	Transcription factor binding sites, scored by ΔZ	2,456,473
PupaSuite	Transcription factor binding sites (Transfac)	81,293
PupaSuite	Transcription factor binding sites (JASPER)	60,082
PupaSuite	DNA triplex sequences	439,350
PupaSuite	Exonic splicing enhancers (ESE)	153,523
PupaSuite	Exonic splicing silencers (ESS)	22,926
PupaSuite	miRNA sequences	20,716
PupaSuite	New splice site formation	13,415
PupaSuite	Splice site disruption	1574
Affymetrix	Genome-Wide Human SNP Array 6.0 (+11 others)	924,216
Illumina	Human 1M BeadChip (+7 others)	1,126,075
Polyphen	Structure-based predictions	53,720
SNP3D	Structure-based predictions	4792
SNP3D	Sequence-based predictions	28,136

Reference SNP IDs (rsID) assigned by dbSNP is used as a primary key for SNP related tables of our database as rsID uniquely identifies particular SNPs. Basic annotations for all human SNPs were extracted from dbSNP [15] and among these annotations most important ones to our AHP based scoring scheme were function class, associated gene ID and symbol. Additionally, PupaSuite [104] annotations allowed scoring of SNPs based on their overlap with

splicing regulatory elements, miRNA and conserved regions in the genome. UCSC [147] provides genomic coordinates of highly conserved elements across multiple species, allowing for identification of SNPs overlapping with evolutionarily conserved regions of the human genome[17]. Additionally database[18] tables of SPOT [69] online prioritization tool related with linkage disequilibrium correlation values (r^2) (originally extracted from HapMap) and SNPs with significant (greater than 0.8) mouse ECR (Evolutionary ConservedRegions) values (originally extracted from ECRBase) are integrated into METU-SNP database. SPOT offers the option to specify a certain population for LD correlations. Instead, in order to identify the SNPs with high LD we have used SNP Source-Proxy pairs with $r^2 \geq 0.8$ in each population.

5.3.2 Gene Data

Entrez Gene ID is used as the primary key for identifying a particular gene. Most gene related information is integrated from SPOT database, which is originally extracted from NCBI Entrez Gene. Also SNP-Gene associations are extracted from NCBI and dbSNP. 45,379 genes are annotated and relevant information is organized in the database as presented in detail in Table 5.2.

Table 5. 2 Gene based annotation from NCBI Entrez Gene.

Field	Description
Entrez Symbol	NCBI Entrez Gene official gene symbol
Entrez Gene ID	NCBI Entrez Gene ID
Gene type	Gene type: protein-coding, tRNA, etc.
Entrez full name	Full name from NCBI Entrez Gene
Chr	Chromosome
Start Pos (bp)	Start Position in base pairs (NCBI Mapview)
Stop Pos (bp)	Stop Position in base pairs (NCBI Mapview)
Size (kb)	Size of transcript in kb (NCBI Mapview)
Cytogenetic Pos.	Cytogenetic Position

5.3.3 Pathway Data

Pathway-based analysis of GWA data is emerging as a useful tool for discovery of underlying molecular mechanisms of diseases associated with particular SNP biomarkers. In second wave GWAS studies where combined *p*-value approach is used to identify associated genes and pathways to a phenotype through SNP genotyping, it is assumed that markers underlying a disease or phenotype are enriched in genes belonging to the same pathway. Thus, the gene and pathway information in METU-SNP database is integrated data from major

[17] The data from the major databases are integrated into a single SQL dump file and we acknowledge the help of SNPLogic developers for sharing this file with us.
[18] https://spot.cgsmd.isi.edu/doc/gin_primary_2.sql.gz

biological repositories following the same assumption. A summary of the integrated resources is presented in Table 5.3.

Table 5. 3 Biological pathway resources used for annotation.

Resource	Description of extracted annotations	Number of Pathways	Total Number of Distinct Genes
Gene Ontology	Molecular Function term associations via genes	2,479	10,644
Gene Ontology	Biological Process term associations via genes	3,066	10,793
Gene Ontology	Cellular Component term associations via genes	636	6,236
KEGG	Pathway associations via genes	177	3,901
WikiPathways	Pathway associations via genes	106	3,089
BioCarta	Pathway associations via genes	314	1,375
BioCyc	Pathway associations via genes	179	452

For each particular pathway residing in a pathway system METU-SNP provides the gene IDs within the pathway and an URL link to the listed pathway that would help researcher to browse the page and visualize the pathway in better detail.

5.3.4 Disease Data

We have utilized a recently suggested GeneRIF-Disease Ontology (DO) mapping approach [101, 148] to construct our gene-disease association tables. It is suggested that this approach performs better when the prediction performances are compared with OMIM. Mapping process is illustrated in Figure 5.4 (used with the permission of [101]). In the figure, (A) the use of MetaMap Transfer tool to annotate GeneRIFs with DO and (B) association between Gene ID: 7040 and DO ID: 2585 are shown. We have incorporated the same mapping that has been provided in a relational database format at DO-RIF project page of Northwestern University[19]. The summary statistics of the annotation data residing in our database is presented in Table 5.4

[19] http://projects.bioinformatics.northwestern.edu/do_rif/

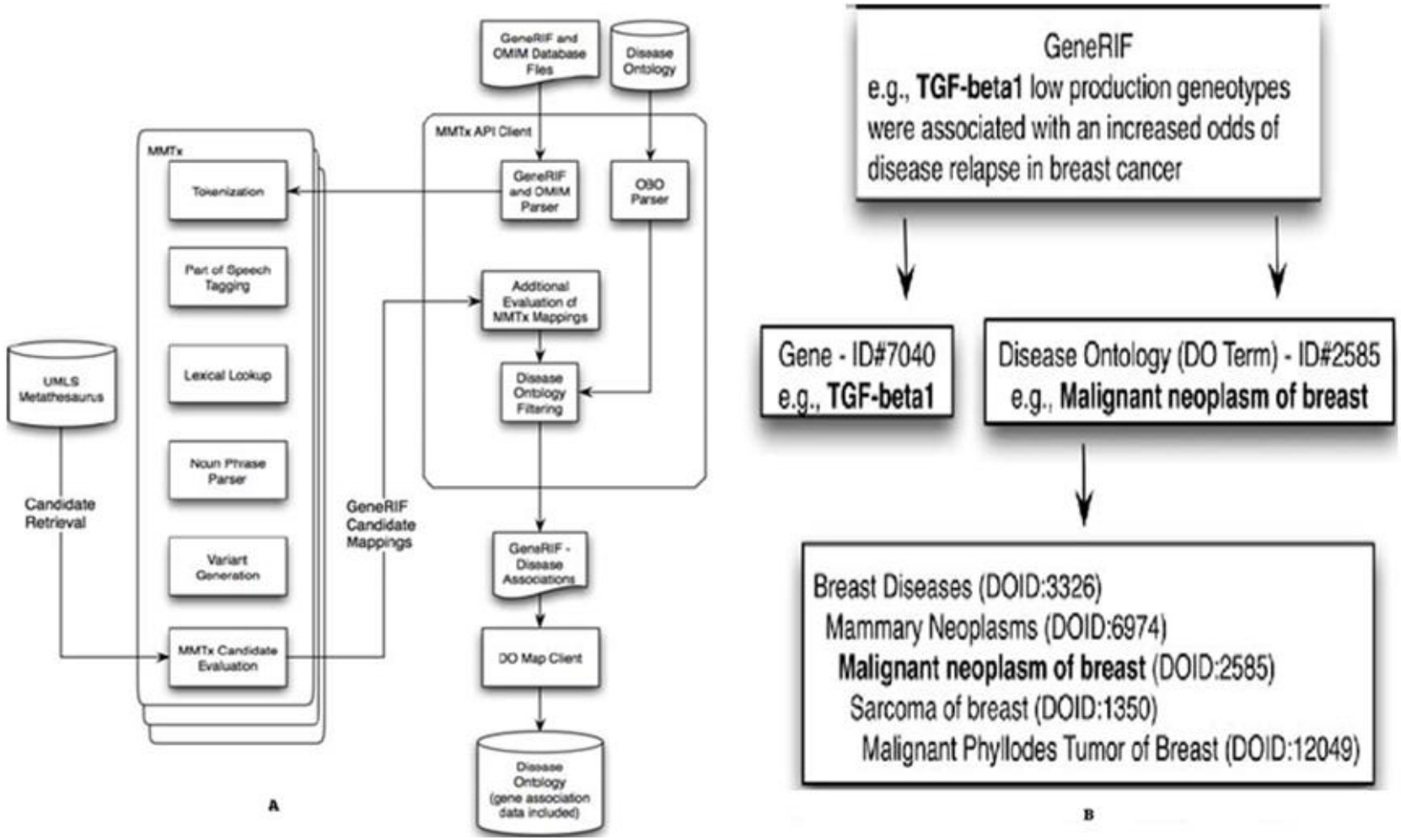


Figure 5. 4 Diagram of disease ontology annotation of the human genome.

Table 5. 4 Annotation statistics related with disease data.

Annotations	METU-SNP
# of Disease	14,889
# of Diseases with at least one mapping	1,851
# Genes with at least one mapping	4,070
Average mappings per disease	14.64

5.4 Utilized Algorithms

METU-SNP's logical flow depicted in Figure 5.2 also gives some hints on the utilized algorithms within SNP prioritization and representative SNP selection framework. One can classify the functionality offered by METU-SNP into four: (1) preprocessing, (2) association, (3) prioritization and (4) selection.

5.4.1 Preprocessing

In the first step of the analysis the required input files for PLINK, pedigree file and map file, are uploaded and imputation and quality control based filtering are applied. Output is provided in the form of PLINK based binary data file, which holds the imputed genotype data that satisfies quality thresholds. Preprocessing algorithm is presented in Algorithm 5.1.

Algorithm 5. 1 Preprocessing genotype data - METU-SNP

Input:

P Genotype data in pedigree format, all data in the same strand.

M Map data.

maf Minor allele frequency threshold.

S_m SNP missingness rate threshold.

I_m Individual missingness rate threshold.

H Hardy Weinberg equilibrium threshold.

rsq Allelic r^2 threshold.

Output:

Cleaned Quality controlling applied SNP set.

1. $S \leftarrow P$
2. **If** impute = true **Then**
3. **While** Chromosome i in *M*
4. $S_i \leftarrow SplitChromosome(S)$
5. $Si \leftarrow BEAGLE(Si, rsq)$ //imputation performed
6. count++
7. **Next** i
8 $Merged = \emptyset$

8. **For** j = 1 to count

Algorithm 5.1 (cont.)

9. $Merged = Merged \cup S_i$
10. **Next** j
11. $Cleaned \leftarrow PLINK(Merged, maf, H)$
12. **Else** $Cleaned \leftarrow PLINK(S, Sm, Im, maf, H)$
13. **Return** Cleaned

5.4.2 Association

Second step of the analysis involves a genome-wide association run that identifies the SNP biomarkers that are "significantly" associated with the disease phenotype. Using the calculated *p*-values of the SNPs found to be associated in the initial step, the significantly associated gene and pathways are investigated based on the combined *p*-value approach during the second-wave analysis [85]. Algorithm 5.2 presents details of this step.

Algorithm 5. 2 Two-wave genome-wide association run - METU-SNP

Input:

Cleaned Quality controlling applied SNP set.

Type Individual SNP *p*-value type: Bonferroni, FDR or uncorrected.

p_{SNP} *p*-value threshold for significance for SNPs.

p_{GENE} *p*-value threshold for significance for genes.

p_{PATH} *p*-value threshold for significance for pathways.

Output:

$sign_{SNP}$ Statistically significant SNP set.

$sign_{GENE}$ Statistically significant gene set.

$sign_{PATH}$ Statistically significant pathway set.

S List of SNPs and *p*-values.

1 $sign_{SNP} = \emptyset,\ sign_{GENE} = \emptyset,\ sign_{PATH} = \emptyset,\ GeneList = \emptyset,\ PathwayList = \emptyset$

2. $S \leftarrow PLINK(Cleaned, Type)$

3 *countS* = | *S* |

4. **For** *i* = 1 to countS

5. $GeneList = GeneList \cup AssociatedGene(S_i.rsID)$ // Database integration SNP-Gene

6. **If** $S_i.p_{value} < p_{SNP}$ **Then** $sign_{SNP} = sign_{SNP} \cup S_i.rsID$

7. **Next** *i*

8. *countG* = |*GeneList*|

9. For *j* = 1 to *countG*

10. $PathwayList = PathwayList \cup AssociatedPathway(GeneList_j.GeneID)$ // Database integration Gene-Pathway

Algorithm 5.2 (cont.)

11 . **If** $CHIDIST(-2\sum_{k=1}^{|GeneList_j.SNPset|}\log(GeneList_j.SNPset_k.p_{value}), 2\,|GeneList_j.SNPset|) < p_{GENE}$ **Then** $sign_{GENE} = sign_{GENE} \cup GeneList_i$ // One-tail probability of Chi-square distribution

12. **Next** *j*

13. *countP* = |*PathwayList*|

14. For *l* = 1 to *countP*

15. $m = |PathwayList_l.GeneSet|$

16. $n = |PathwayList_l.GeneSet \cap sign_{GENE}|$

17. **If** $(1 - \sum_{p=0}^{n}\frac{\binom{sign_{GENE}}{p}\binom{countG-sign_{GENE}}{m-p}}{\binom{countG}{m}}) < p_{PATH})$ **Then** $sign_{PATH} = sign_{PATH} \cup PathwayList_l$)

18. **Next** *l*

19. **Return** [$sign_{SNP}$, $sign_{GENE}$, $sign_{PATH}$, *S*]

5.4.3 Prioritization

In the third step of our analysis we utilize the information of statistically significant SNPs, genes and pathways to determine the prioritized set of SNPs as explained in detail in Chapter 3. Shortly, the AHP methodology is followed for scoring of genotyped SNPs and the SNPs are ranked according to their AHP scores. User is given the option to filter relevant SNPs either depending on the calculated *p*-values or AHP scores or both. The algorithm related to this step is presented in Algorithm 5.3.

Algorithm 5. 3 AHP based prioritization - METU-SNP

Input:

$sign_{SNP}$	Statistically significant SNP set.
$sign_{GENE}$	Statistically significant gene set.
$sign_{PATH}$	Statistically significant pathway set.
disease	Complex disease under study
p_{SNP}	*p*-value threshold for prioritization for SNPs.
rank	Number of SNPs to select
S	List of SNPs and *p*-values of selected type in previous step.
AHPscore	Table that holds AHP scores for each SNP (rsID) in {0, 1} format based on biological properties (independent of GWAS) of the particular SNP as well as calculated scores on run time.
LD	Table that holds SNP pairs with LD r^2 of 0.8 in each HapMap population.
GenePwayGene	Table that holds gene pairs sharing (reside in) at least one pathway together.
GeneDisease	Table that holds gene-disease association.

Genes Table that holds gene set.

AHPweights Table 3.7- Weights for leaf nodes.

Algorithm 5.3 (cont.)

Output:

$prioritized_{SNP}$ Prioritized SNP set.

Totalscores Table that holds total AHP scores for each SNP (rsID). Score is between 0 and 1.

1. $prioritizedSNP = \emptyset, AHPscore_{0.1} = AHPscore_{0.2.1} = AHPscore_{0.2.2} = AHPscore_{0.2.3} = AHPscore_{0.2.2} = AHPscore_{0.3.1} = AHPscore_{0.3.2} = AHPscore_{0.3.3} = AHPscore_{1.2.1.1} = AHPscore_{1.2.1.2} = AHPscore_{1.2.1.3} = AHPscore_{1.2.2.1.1} = AHPscore_{1.2.2.1.2} = AHPscore_{1.2.2.1.3} = AHPscore_{1.2.2.2.1} = AHPscore_{1.2.2.2.2} = AHPscore_{1.2.2.2.3} = 0$
2. *countS* = | *S* |
3. **For** *i* = 1 to *countS*
4. **If** $S_i.p_{value} > p_{SNP}$ **Then Break**
5. **If** $S_i.rsID \in signSNP$ **Then** $AHPscore_{0.1}(S_i.rsID) = 1$
6. **If** $AssociatedGene(LD(S_i.rsID)) \in signGENE$ **Then** $AHPscore_{0.2.1}(S_i.rsID) = 1$
7. **If** $AssociatedGene(S_i.rsID) \in signGENE$ **Then** $AHPscore_{0.2.2}(S_i.rsID) = 1$
8. **If** $GenePwayGene(AssociatedGene(S_i.rsID)) \cap signGENE \neq \emptyset$ **Then** $AHPscore_{0.2.3}(S_i.rsID) = 1$
9. $|sign_{PATH}| = countP$
10. **For** *j* = 1 to *countP*
11. **If** $sign_{PATH_j}.GeneSet \cap AssociatedGene(LD(S_i.rsID)) \neq \emptyset$ **Then** $AHPscore_{0.3.1}(S_i.rsID) = 1$
12. **Next** *j*
13. **For** *j* = 1 to *countP*
14. **If** $sign_{PATH_j}.GeneSet \cap AssociatedGene(S_i.rsID) \neq \emptyset$ **Then** $AHPscore_{0.3.2}(S_i.rsID) = 1$
15. **Next** *j*
16. **For** *j* = 1 to *countP*
17. **If** $sign_{PATH_j}.GeneSet \cap GenePwayGene(AssociatedGene(S_i.rsID)) \neq \emptyset$ **Then** $AHPscore_{0.3.3}(S_i.rsID) = 1$
18. **Next** *j*
19. **If** $GeneDisease(disease) \cap AssociatedGene(LD(S_i.rsID)) \neq \emptyset$ **Then** $AHPscore_{1.2.1.1}(S_i.rsID) = 1$
20. **If** $GeneDisease(disease) \cap AssociatedGene(S_i.rsID) \neq \emptyset$ **Then** $AHPscore_{1.2.1.2} = 1$
21. **If** $GenePwayGene(AssociatedGene(S_i.rsID)) \cap GeneDisease(disease) \neq \emptyset$ **Then** $AHPscore_{1.2.1.3}(S_i.rsID) = 1$

22. **If** $GeneDisease(!\,disease) \cap AssociatedGene(LD(S_i.rsID)) \neq \emptyset$ **Then**
$AHPscore_{1.2.2.1.1}(S_i.rsID) = 1$

Algorithm 5.3 (cont.)

23. **If** $GeneDisease(!\,disease) \cap AssociatedGene(S_i.rsID) \neq \emptyset$ **Then**
$AHPscore_{1.2.2.1.2}(S_i.rsID) = 1$
24. **If** $GenePwayGene(AssociatedGene(S_i.rsID)) \cap GeneDisease(!\,disease) \neq \emptyset$ **Then**
$AHPscore_{1.2.2.1.3}(S_i.rsID) = 1$
25. $NeutralGene \leftarrow Genes \,/\, GeneDisease.GeneSet$
26. **If** $NeutralGene \cap AssociatedGene(LD(S_i.rsID)) \neq \emptyset$ **Then**
$AHPscore_{1.2.2.2.1}(S_i.rsID) = 1$
27. **If** $NeutralGene \cap AssociatedGene(S_i.rsID) \neq \emptyset$ **Then**
$AHPscore_{1.2.2.2.2}(S_i.rsID) = 1$
28. **If** $GenePwayGene(AssociatedGene(S_i.rsID)) \cap NeutralGene \neq \emptyset$ **Then**
$AHPscore_{1.2.2.2.3}(S_i.rsID) = 1$
29. $Totalscores(S_i.rsID) = \sum_j AHPscore_j\,(S_i.rsID)\; x\; AHPweights_j(S_i.rsID)$
30. **Next** *i*
31. *AHPscore.sortdescending(scores)*
32. **While** *k* < *rank* $prioritized_{SNP} = prioritized_{SNP} \cup Totalscores_k(rsID)$
33. **Return** [$prioritized_{SNP}$, *Totalscores*]

5.4.4 Selection

In the last step of our METU-SNP application, Simulated Annealing based SNP selection algorithm, presented in great detail in Chapter 4, is applied to the AHP based prioritized set of SNPs to select a representative SNP set. The algorithm related to this step is based on Algorithm 4.1 and presented in detail in Algorithm 5.4.

Algorithm 5. 4 Simulated Annealing based selection - METU-SNP

Input:

$prioritized_{SNP}$	Prioritized SNP set.
t	Simulated annealing parameter temperature.
d	Simulated annealing parameter decreasing factor.
c_{max}	Number of iterations.
w	Weight for specifying tradeoff between SNP number and accuracy
P	Phenotype vector

Output:

$selected_{SNP}$	Representative SNP set.

1. $s_0 \in prioritized_{SNP}$

Algorithm 5.4 (cont.)

2. $s_{best} \leftarrow s_o, e_{best} \leftarrow e$
3. **For** c = 1 to c_{max}

$s_{new} \leftarrow neighbor(s_0)$

$G_T = prioritized_{SNP} - s_{new}$

$k = |s_{new}|$

$e_{new} \leftarrow w(\sum_{i=1}^{n-k} NaiveBayes(s_{new}, G_{T_i}) + NaiveBayes(s_{new}, P)) + (1-w)\boldsymbol{k}$

if $e_{new} > e_{best}$ **then**

$\boldsymbol{s_{best} \leftarrow s_{new}; e_{best} \leftarrow e_{new}}$

if exp $(-\frac{e_{new}}{t})$. < random() **then**

$\boldsymbol{s \leftarrow s_{new}; e \leftarrow e_{new}; t = txd}$

Next c

4. $selected_{SNP} = \boldsymbol{s_{best}}$
5. **return** $selected_{SNP}$

5.5 User Interface

METU-SNP User Interface (UI) is designed so as to guide the user through analysis process. UI consists of 6 tabs corresponding to 6 steps of the analysis: (1) configuration, (2) preprocess, (3) genome-wide association, (4) SNP prioritization, (5) SNP selection and (6) performance.

5.5.1 Configuration

Configuration tab is depicted in Figure 5.5. This tab allows the user to enter database properties such as database name, user name and password that will allow METU-SNP to connect with MySQL database. User may also specify maximum number of connections that will be used for threading for database read/write. Once the database properties are set, user can connect to the database and select the disease under study from drop down menu.

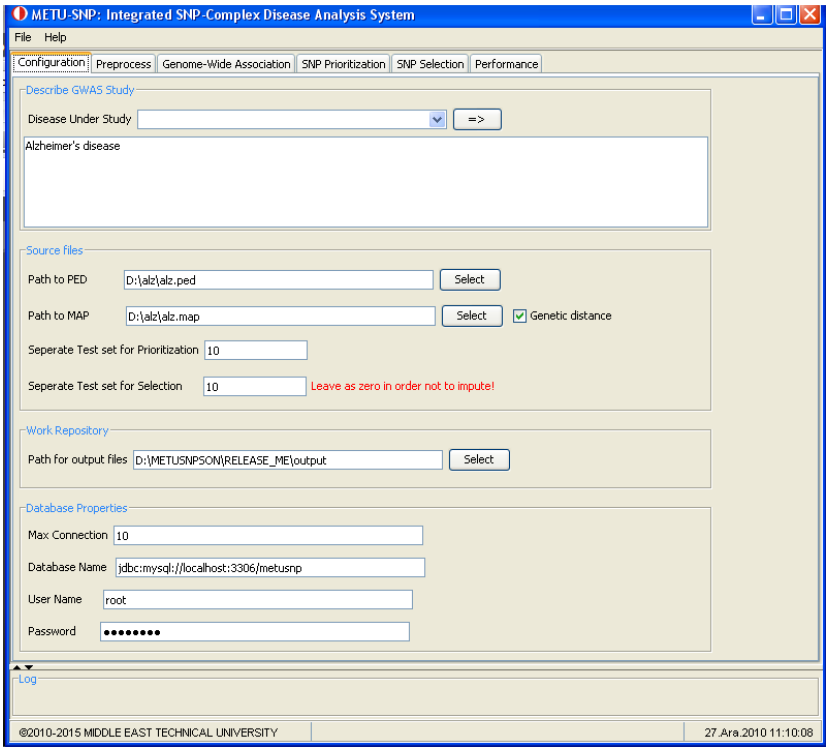


Figure 5. 5 Configuration tab - METU-SNP.

Current version of METU-SNP only accepts data files in Pedigree and Map format[20] compatible with PLINK. Map format is either 3 column (chromosome number, reference sequence ID for SNP and base pair location) or 4 column (additionally genetic distance in centimorgans) and user is expected to specify this using Genetic Distance check box. User is also given the option to separate test data for prioritization and selection steps. "Percentage" of the overall data that is required to be separated for testing can be entered into appropriate text box. Since METU-SNP makes use of 3rd party software, lot of files are created during algorithm runs, therefore it is suggested to create a work repository for output files.

5.5.2 Preprocess

Preprocess tab is depicted in Figure 5.6. After configuration step, user can start on the analysis phase and the first step of the analysis is the preprocessing of the data at hand. This involves quality control (QC) based filtering (offered by PLINK) and imputation (offered by BEAGLE). For QC thresholds user is given the option to filter out those SNPs/individuals according to minor allele frequency, missingness and Hardy Weinberg equilibrium. A comprehensive guide containing further information on the thresholds is provided on PLINK website[21]. Mathematical background of the process is explained in greater detail in Appendix E. Default values for the analysis (0.05 for Minor Allele Frequency, 0.1 for SNP Missingness Rate,

[20] http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml
[21] http://pngu.mgh.harvard.edu/~purcell/plink/thresh.shtml

0.1 for Individual Missingness Rate and 0.001 for Hardy Weinberg equilibrium) that are commonly used in various GWAS studies are set.

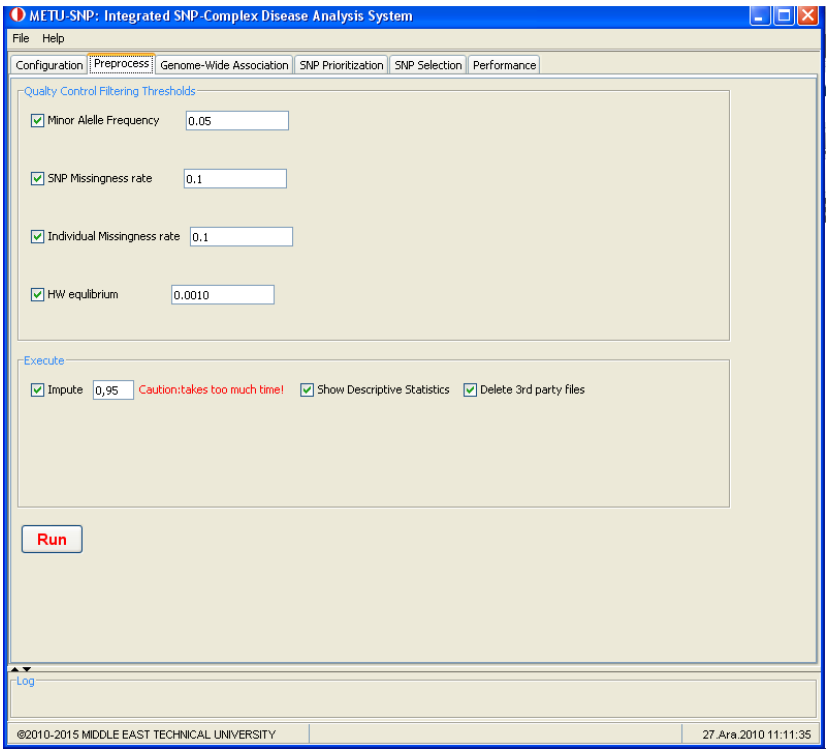


Figure 5. 6 Preprocess tab - METU-SNP.

After specifying QC thresholds, user is given the option to impute the data for missingness. Imputation must be applied to perform SNP selection phase, as the algorithm crashes in the presence of missing values. However, if only prioritization is to be applied, user may choose not to impute as the process takes considerably long time. As stated in section 5.2.2 user can specify a threshold for allelic r^2 (default being 0.95) in order to include only "well" imputed markers in subsequent analysis. To get a basic idea on the data at hand, user may choose to view the descriptive statistics files (freq.frq, missing.imiss, missing.lmiss, hardy.hwe) created during PLINK based QC analysis. Table 5.5 explains descriptive statistics files.

Table 5. 5 Descriptive statistics files.

File	Field Name	Description
freq.frq	CHR	Chromosome
	SNP	SNP identifier
	A1	Allele 1 code (minor allele)
	A2	Allele 2 code (major allele)
	MAF	Minor allele frequency
	NCHROBS	Non-missing allele count
missing.imiss	FID	Family ID
	IID	Individual ID
	MISS_PHENO	Missing phenotype? (Y/N)
	N_MISS	Number of missing SNPs
	N_GENO	Number of non-obligatory missing genotypes
	F_MISS	Proportion of missing SNPs
missing.lmiss	SNP	SNP identifier
	CHR	Chromosome
	N_MISS	Number of individuals missing this SNP
	N_GENO	Number of non-obligatory missing genotypes
	F_MISS	Proportion of sample missing for this SNP
hardy.hwe	SNP	SNP identifier
	TEST	Code indicating sample
	A1	Minor allele code
	A2	Major allele code
	GENO	Genotype counts: 11/12/22
	O(HET)	Observed heterozygosity
	E(HET)	Expected heterozygosity
	P	H-W p-value

Depending on the user specifications, initial pedigree and map files are preprocessed and a cleaned and imputed binary file is created at the end of this step.

5.5.3 Genome-wide Association

Genome-wide association tab is depicted in Figure 5.7. At this step of the analysis user runs an association analysis to find SNPs significantly associated with disease/trait under study. Depending on user's choice, three different methods can be used to calculate *p*-values: (1) uncorrected, (2) Bonferroni and (3) False Discovery Rate. The latter two approaches include adjusting for multiple testing, which is explained in greater detail in Chapter 3. Depending on the threshold set by the user, SNPs are labeled as significant or not. Most widely accepted threshold for *p*-value is 0.05, however according to the requirements of the analysis, it is possible to specify any threshold value using the text boxes provided.

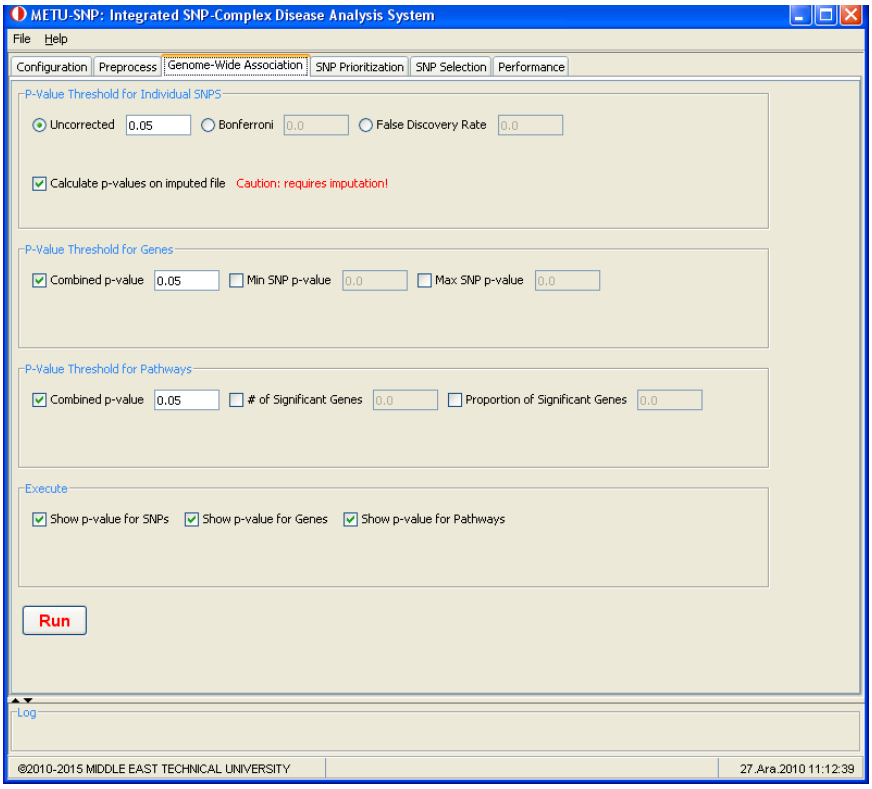


Figure 5. 7 Genome-wide association tab - METU-SNP.

This step of the analysis also includes identifying "significant" genes and pathways, which will be used in SNP prioritization phase. To label a gene as significant user can use three thresholds: (1) combined *p*-value, (2) min SNP *p*-value and (3) max SNP *p*-value. Combined p-value for a gene is calculated using Fisher's combination test, which is explained in Section 3.2 and used in Algorithm 5.2. If combined value for the gene is less than the threshold, gene is labeled as significant. As the gene is regarded as a combination of SNPs in our analysis, one may choose to specify thresholds regarding the *p*-values of individual SNPs associated with the gene. It is therefore possible to limit how big the minimum *p*-value of the SNPs or max p-value of the SNPs associated with the gene. In order to determine if a pathway is significant user also has three thresholds: (1) combined *p*-value, (2) number of significant genes and (3) proportion of significant genes. Combined p-value for a gene is calculated using Fisher's exact test (hypergeometric test), which is explained in Section 3.2 and used in Algorithm 5.2. If combined value for the pathway is less than the threshold, pathway is labeled as significant. We regard pathways as a combination of genes. Therefore, it is possible to value the significance of the pathway by evaluating how many significant genes there are within the pathway (2) or proportion of significant genes over all the genes associated with the pathway (3). Following this step, three files are created: (1) snp.txt, (2) gene.txt and (3) pathway.txt. Table 5.6 explains contents of these files.

Table 5. 6 Files created during GWAS.

File	Field #	Description
snp.txt	1	SNP rs ID (as in dbSNP)
	2	p-value (according to the specified type of test)
	3	Significance (0 = not significant, 1 = significant)
gene.txt	1	Entrez gene ID
	2	p-value (according to the specified threshold)
	3	Significance (0 = not significant, 1 = significant)
pathway.txt	1	pathway ID (as in MySQL database)
	2	p-value / significant gene info (according to the specified threshold)
	3	Significance (0 = not significant, 1 = significant)

5.5.4 SNP Prioritization

SNP prioritization tab is depicted in Figure 5.8. The functionality depends on Algorithm 5.3. AHP scores are created for those SNPs meeting the individual SNP *p*-value threshold (less than or equal to the specified *p*-value) specified by the user. User may also set a limit on the number of SNPs to be used for subsequent analysis after prioritization. This can be done by entering a value (default is 10,000) in the SNP ranking text box. This way if there are more than 10,000 SNPs, which satisfy *p*-value threshold, only first 10,000 of them are listed according to the ranking done with respect to AHP scores and *p*-values.

Most significant SNPs, genes and pathways can be viewed in results panel after the prioritization run. In SNPs panel, SNP rsID, p-value and significance is written for the first *n* SNPs (*n* is user defined via text box) as in snp.txt file. In Genes panel Entrez gene ID, full name and cytolocation are listed for the most significant genes. In Pathways panel, pathway ID, pathway system (database), full pathway title, pathway URL (web site for the particular pathway if exists) and gene count for the pathway are listed.

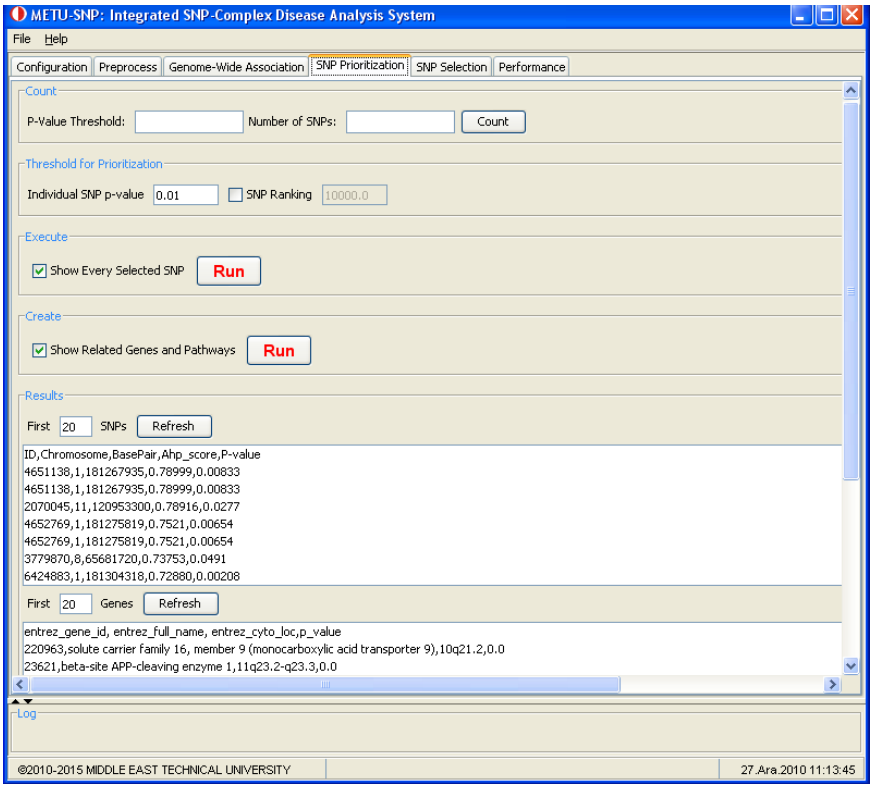


Figure 5. 8 SNP prioritization tab - METU-SNP.

5.5.5 SNP Selection

SNP selection tab is depicted in Figure 5.9. The functionality depends on the Algorithm 5.4. User enters Simulated Annealing algorithm parameters and s/he can specify the classifier that will be used for supervised learning step. Default classifier is Naive Bayes. However, user is given the option to use different classifiers offered by WEKA. For a complete list of classifiers, please refer to WEKA website[22]. Full weka path should be specified, therefore user should enter full path specified as in weka.classifiers package (for example **functions.SMO** should be entered for sequential minimal optimization (SMO) classifier implemented in package weka.classifiers.**functions.SMO**). One needs to note that only those WEKA classifiers, which support multi-valued nominal attributes, are supported for METU-SNP.

User may choose to include (classification accuracy) or not to include (tagged prediction) phenotype data for the algorithm. If "Show Resulting SNP set" option is selected user can view the list of rsIDs for representative SNP set (featuresSelected.txt).

[22] http://weka.sourceforge.net/doc/weka/classifiers/Classifier.html

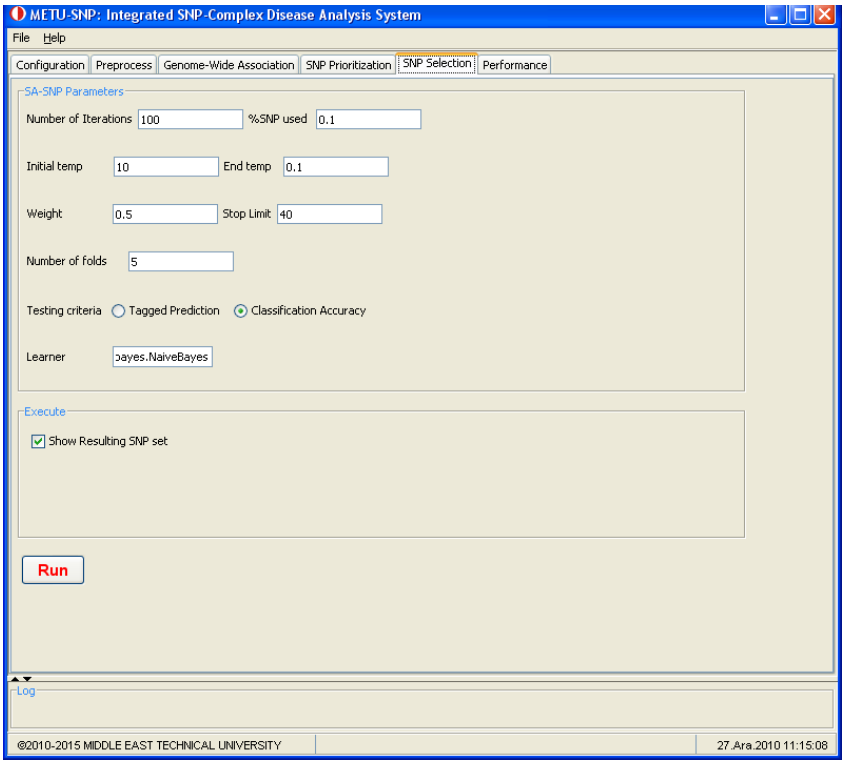


Figure 5. 9 SNP selection tab - METU-SNP.

5.5.6 Performance

Last tab in UI of METU-SNP is the performance tab, which aims to measure the prediction performance of the prioritization and selection step. Performance tab is depicted in Figure 5.10. In order to evaluate the performance of the algorithms, user has to separate a test set at the configuration step. User is given the option to perform a *k*-fold cross validation using classifiers offered by WEKA. As in the SNP selection step, full WEKA path for the classifier should be entered in appropriate text box.

If user chooses to measure the prediction performance of SNP set created at prioritization step, a text file (prioritization_performance.txt) is created that contains certain classification measures such as correctly classified instances, incorrectly classified instances, Kappa statistic, mean absolute error, root mean squared error, root relative square error and confusion matrix. Similarly, if selection performance is to be measured, three text files are created (selection_performance.txt, selection_performance_relieff.txt and selection_performance _chisquare.txt). The content of these files are similar to prioritization_performance.txt and allow user to compare the performance of Simulated Annealing based selection with WEKA based attribute selection schemes Chi-square and Relief F.

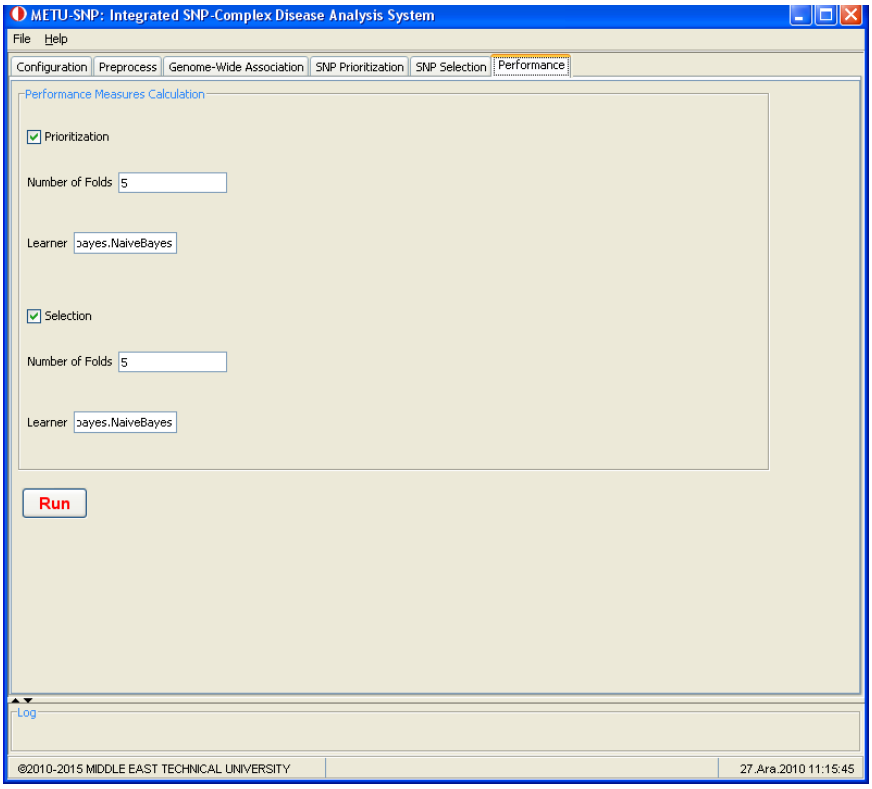


Figure 5. 10 Performance tab - METU-SNP.

5.6 Conclusion

In this chapter we have introduced a java based integrated software system, METU-SNP, which can be effectively used for GWAS and post-GWAS analysis of case-control based data for SNP-complex disease association. It makes use of data from major public databases such as dbSNP, Entrez Gene, KEGG, Gene Ontology etc. It offers state-of-the-art AHP based SNP prioritization and Gene Set Enrichment Analysis frameworks. METU-SNP integrates well known 3rd party tools such as PLINK and BEAGLE, which are well-known tools for GWAS. However, these are command line tools requiring a deeper knowledge and experience on various commands to get benefit from the functionality offered. METU-SNP, as an all-in-one GWAS application, offers a user friendly Graphical User Interface to manage 3rd party tools and it is equipped with necessary functionality to prioritize and filter the most relevant SNPs from a massive initial SNP set. We believe that METU-SNP will facilitate reliable identification of SNPs that are involved in the etiology of complex diseases and ultimately supporting timely identification of genomic disease biomarkers, and development of personalized medicine approaches and targeted drug discoveries.

CHAPTER 6

A CASE STUDY: USE OF METU-SNP TO ANALYZE GWAS CASE-CONTROL DATA FOR RHEUMATOID ARTHRITIS

In this chapter we will demonstrate the functionality offered by METU-SNP with the analysis of the case control data for Rheumatoid Arthritis (RA) disease. The details of the features of the experimental data have been explained in Chapter 1. The quality control based filtering and imputation has been performed to increase data quality. Then the overall data has been split into three: training set (%60) and 2 different test sets for AHP based prioritization (%20) and SA based representative SNP selection (%20). The comparisons of the analysis results with SPOT for prioritization and with WEKA based Relief and Chi-square attribute selection schemes for the representative SNP selection have been presented. Additionally, the SNPs, genes and pathways associated with RA have been identified. The genes and pathways revealed by METU-SNP analysis results were in confirmation with previous molecular genetics research described in the literature on the molecular basis of RA. The overall analysis of the results confirms that the integrated all-in-one METU-SNP application is an effective and user friendly tool for use in GWAS.

6.1 Data Preprocessing and Cleaning

First step of our analysis involves performing PLINK based quality control filtering and imputation via BEAGLE. Details of the process are explained in greater detail in Chapter 5. Before frequency and genotyping pruning, there were 501,463 SNPs studied on 868 cases and 1,194 controls of which 569 were males and 1,493 were females. Table 6.1-4 represents summary statistics of the data at hand.

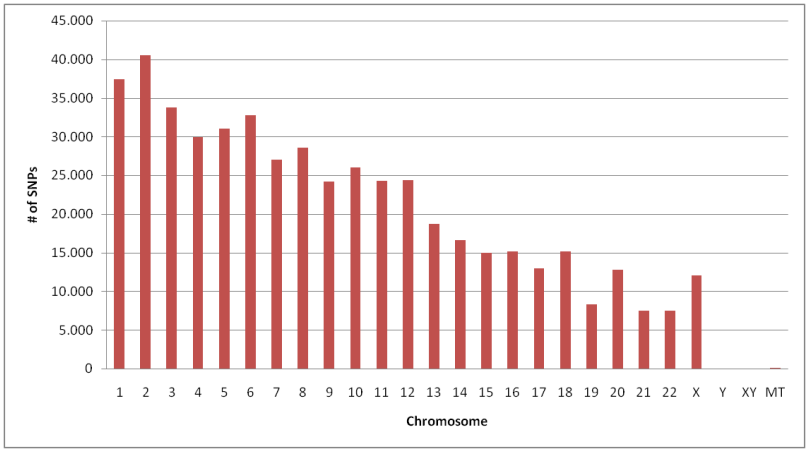


Figure 6. 1 Number of SNPs per chromosome for RA data.

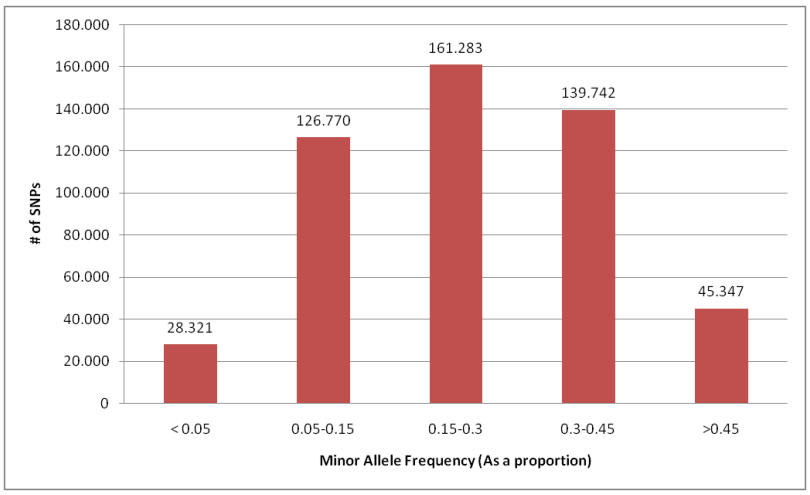


Figure 6. 2 Minor Allele Frequency distribution for SNPs for RA data.

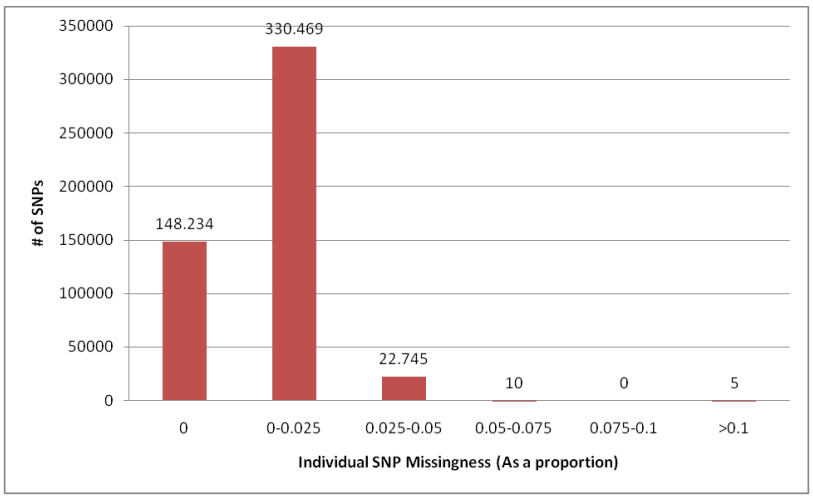


Figure 6. 3 SNP missingness rate for RA data.

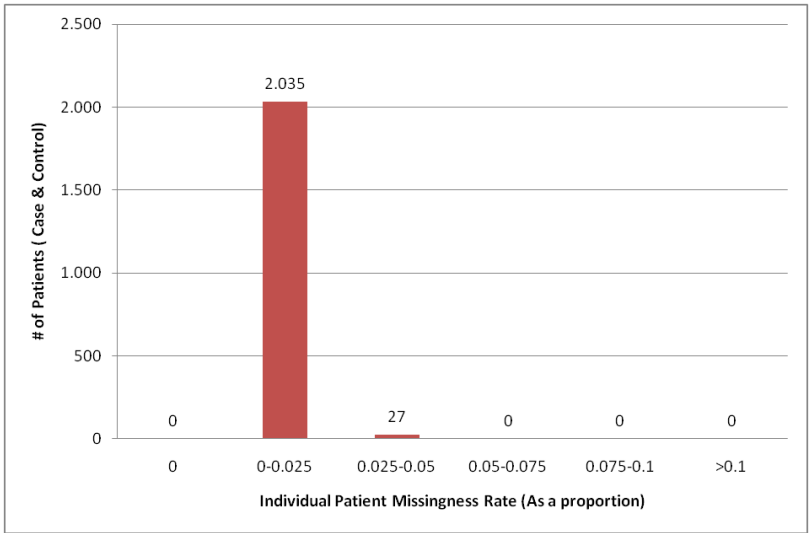


Figure 6. 4 Individual missingness rate for RA data.

In order to handle with the missingness we applied imputation via BEAGLE. During the imputation process 910 SNPs with allelic r^2 frequency less than 0.95 have been removed and 500,553 SNPs remained available for the downstream applications. Among these SNPs, 36 SNPs are excluded based on Hardy Weinberg Equilibrium threshold ($p \leq 0.001$). Additionally 26,904 SNPs are removed due to failure to comply with frequency test (MAF $\leq$ 0.05). At the end of the preprocessing step, 473,613 SNPs from the RA genotyping data are selected for subsequent analysis. The overall data is split into three: training set, test set for prioritization and test set for selection. For training data set we had 522 cases and 716 controls (325 males, 913 females). For test set for prioritization we had 180 cases and 232 controls (129 males, 283 females). 166 cases and 246 controls (115 males, 297 females) are reserved for test set for selection.

6.2 GWAS for RA Data

In the second step of METU-SNP based analysis, the preprocessed RA data has been used for GWAS. The multiple testing adjusted *p*-values (False Discovery Rate) is selected for the statistical analysis. The threshold *p*-value is specified as 0.05 for measuring statistical significance of individual SNPs. Same threshold was also used for combined p-value threshold for genes (Fisher's combination) and pathways (Hypergeometric test).

The overall results of the association analysis have been presented in the Figure 6.5, where the negative logs of *p*-values versus chromosome distributions are plotted. It is clearly observed that chromosome 6 is highly associated with the RA disease phenotype chromosome (which is also pointed out in Chapter 2) as a large portion of SNPs with lowest *p*-value association are located on to this chromosome. Additionally chromosomes 1, 9 and 10 host SNPs that are significantly associated with RA.

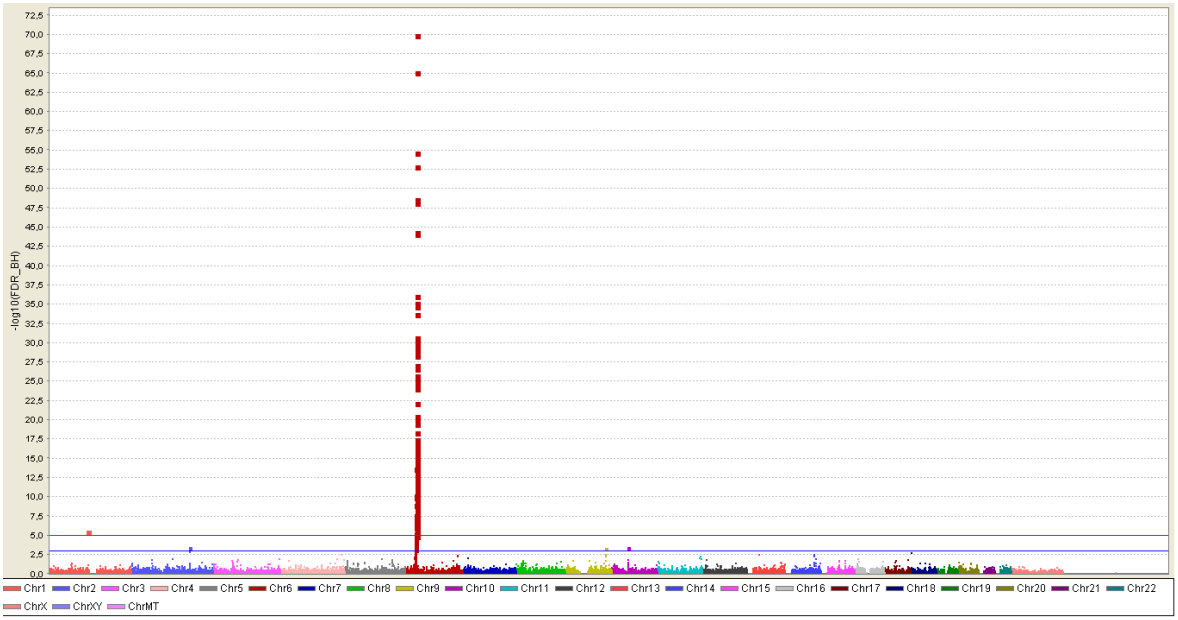


Figure 6. 5 Plot of negative logarithm p-value of association for individual SNPs and their distribution on individual chromosomes.

Table 6.1 depicts the top 20 SNPs found to be significantly associated with RA. The SNPs with smallest *p*-value are rs2395175 (*p*-value = 1.02E-67), rs660895 (*p*-value = 7.10E-66), rs2395163 (*p*-value = 2.04E-52) and rs6910071 (*p*-value = 1.34E-50), which are also shown to be in association with RA in previous studies in the literature [149-152]. All of these SNPs are located around the HLA-DRB1 gene in chromosome 6. As explained in Chapter 2, this locus is suspected to constitute the genetic basis for the etiology of RA. So, the integrated analysis approach we have applied through METU-SNP was able to reveal the SNPs with previously established association to the RA disease phenotype, supporting the literature.

Table 6. 1 Individual SNP p-values of association of GWAS for RA data.

RANK	CHR	SNP	FDR
1	6	rs2395175	1.02E-67
2	6	rs660895	7.10E-66
3	6	rs2395163	2.04E-52
4	6	rs6910071	1.34E-50
5	6	rs3763309	1.96E-46
6	6	rs3763312	6.22E-46
7	6	rs9275224	3.45E-45
8	6	rs6457617	7.41E-42
9	6	rs2395185	6.96E-34
10	6	rs9275555	6.60E-33
11	6	rs2516049	9.23E-33
12	6	rs477515	1.56E-32
13	6	rs9275595	1.53E-31
14	6	rs3817973	1.99E-28
15	6	rs4424066	3.27E-28

Table 6.1 (cont.)

16	6	rs9275406	3.27E-28
17	6	rs9275374	3.27E-28
18	6	rs9275407	3.27E-28
19	6	rs9275393	3.27E-28
20	6	rs9275418	3.27E-28

Following this first step of GWAS run, using the calculated p-values for individual SNPs, as a second-wave analysis we calculated combined p-values for genes and pathways as explained in Chapter 3 in greater detail. Table 6.2 and 6.3 presents top 20 statistically significant genes and pathways found out to be in association with RA as well as calculated *p*-values using Fisher's combination test (for genes) and Hypergeometric test (for pathways). As in individual SNP based analysis, those genes and pathways with calculated p-value less than 0.05 are labeled as "significant". For practical reasons those pathways consisting of a single gene are removed from the list.

Table 6. 2 Top 20 significant genes according to combined p-value of GWAS for RA data.

Entrez Gene ID	Full Name	Location	P-Value
177	Advanced glycosylation end product-specific receptor	6p21.3	~0.0
7916	HLA-B associated transcript 2	6p21.3	~0.0
3122	Major histocompatibility complex\, class II\, DR alpha	6p21.3	~0.0
6891	Transporter 2\, ATP-binding cassette\, sub-family B (MDR/TAP)	6p21.3	~0.0
7148	Tenascin XB	6p21.3	~0.0
731881	Hypothetical protein LOC731881	6p21	~0.0
6048	Ring finger protein 5	6p21.3	~0.0
4855	Notch homolog 4 (Drosophila)	6p21.3	~0.0
629	Complement factor B	6p21.3	~0.0
1388	Activating transcription factor 6 beta	6p21.3	6.21E-13
3118	Major histocompatibility complex\, class II\, DQ alpha 2	6p21.3	1.74E-10
3113	Major histocompatibility complex\, class II\, DP alpha 1	6p21.3	4.09E-10
534	ATPase\, H+ transporting\, lysosomal 13kDa\, V1 subunit G2	6p21.3	5.75E-10
135644	Tripartite motif-containing 40	6p22.1	8.94E-10
199	Allograft inflammatory factor 1	6p21.3	1.01E-09
10107	Tripartite motif-containing 10	6p21.3	1.12E-09
10919	Euchromatic histone-lysine N-methyltransferase 2	6p21.31	1.15E-09
8859	Serine/threonine kinase 19	6p21.3	1.39E-09
80741	Lymphocyte antigen 6 complex\, locus G5C	6p21.33	1.46E-09
259215	Lymphocyte antigen 6 complex\, locus G6F	6p21	4.98E-09

Table 6. 3 Top 20 significant pathways according to combined p-value of GWAS for RA data.

Pathway System	Pathway Title	Gene Count	P-Value
GO Function	MHC class II receptor activity	15	2.16E-19
GO Process	Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	18	6.23E-16
GO Component	MHC class II protein complex	18	6.23E-16
KEGG	Antigen processing and presentation	83	4.79E-15
WFINFLAM	Phagocytosis-Ag presentation	39	8.23E-13
GO Process	Immune response	213	7.71E-12
KEGG	Cell adhesion molecules (CAMs)	133	3.03E-11
GO Process	Antigen processing and presentation	43	1.38E-10
GO Process	Antigen processing and presentation of peptide antigen via MHC class I	10	5.06E-6
GO Function	MHC class I receptor activity	16	2.31E-5
GO Component	MHC class I protein complex	25	9.22E-5
BioCarta	Antigen Processing and Presentation	8	3.45E-4
KEGG	Natural killer cell mediated cytotoxicity	132	0.0011471
WikiPathways	Proteasome Degradation	61	0.0012770
BioCarta	Complement Pathway	16	0.0014370
BioCarta	Cytokines and Inflammatory Response	26	0.0037603
GO Process	Hemopoiesis	27	0.0040472
GO Function	ATPase activity, coupled to transmembrane movement of substances	31	0.0052882
WFINFLAM	Natural Killer Cell Signaling	31	0.0052882
WFINFLAM	Complement Cascade	6	0.0086

As expected, genes that are found significantly associated with RA during our analysis were all located 6p.21 region in chromosome 6. All of the genes listed are MHC based genes which is one of the main underlying pathways in RA etiology. Association of MHC loci with the RA disease has already been established in previous studies as published in the literature [153-161]. Similarly significant pathways are also associated with MHC class genes. Antigen processing and presentation and MHC class pathways are listed on the top ranks of the list, which are also the main focus of current studies on the etiology of RA in the recent publications [64], [162, 163]. Additionally Cell Adhesion Molecules [164, 165] and Immune Response [166, 167] are also labeled as significantly associated genes, which are again well studied pathways for the pathogenesis of RA.

These findings suggest that the newly developed METU-SNP application was able to integrate the second-wave GWAS approach after initial GWAS to reveal significantly associated genes and pathways and can be regarded as an integrated all-in-one tool for GWAS.

6.3 SNP Prioritization

Following GWAS, we have performed the AHP based SNP prioritization approach proposed in this study and compared the prediction performance of the proposed prioritization algorithm with that of SPOT's. METU-SNP based prioritization scheme was explained in greater detail in Section 5.5.4. AHP based prioritization algorithm makes use of genetic information that is held in the METU-SNP database as well as statistical information populated after GWAS run.

Using these, AHP scores are calculated for each SNP. As this step can be very demanding in terms of computational run time, user is given the option to run AHP algorithm (and calculate scores) for a subset of the overall set. This can be done by utilizing the thresholds related with *p*-value and number of SNPs to be selected.

Using the user defined options (number of SNPs for which AHP scores are greater than zero and *p*-value < 0.5) provided through METU-SNP's prioritization tab, we have selected 7,155 SNPs out of the prioritized lists for the 5-fold cross validation by using Naive Bayes as the supervised learning scheme. The results of the 5-fold cross validation has been presented in Table 6.4 and it is evident that AHP based prioritization outperform SPOT in all of the classification measures even though we have used the same number of SNPs in the analysis.

Table 6. 4 5-fold Cross Validation results for AHP and SPOT based list of SNPs over disease trait for RA data.

	Accuracy	Recall	NPV	Precision	Specificity
AHP	**0.786**	**0.733**	**0.800**	**0.767**	**0.828**
SPOT	0.779	0.722	0.793	0.760	0.823

From the chromosomal distribution of the prioritized SNPs it can be seen that the distributions follow an almost parallel structure. However AHP based list signals relatively higher number of SNPs from chromosome 6, which has been shown to be significantly associated loci with RA in the literature. Figure 6.6 and 6.7 depict chromosomal distribution of the prioritized SNPs for AHP and SPOT based list respectively.

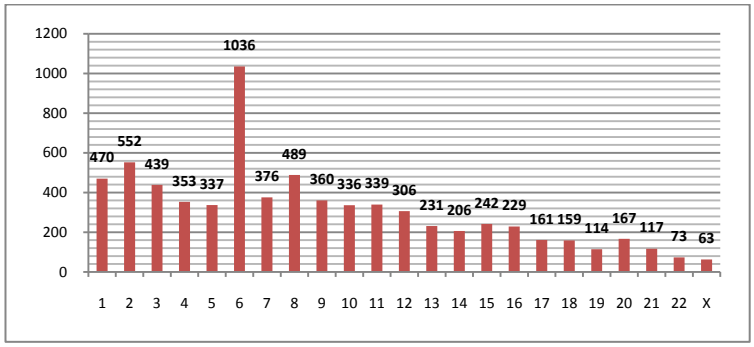


Figure 6. 6 Chromosomal distributions of prioritized SNPs via AHP algorithm for RA data.

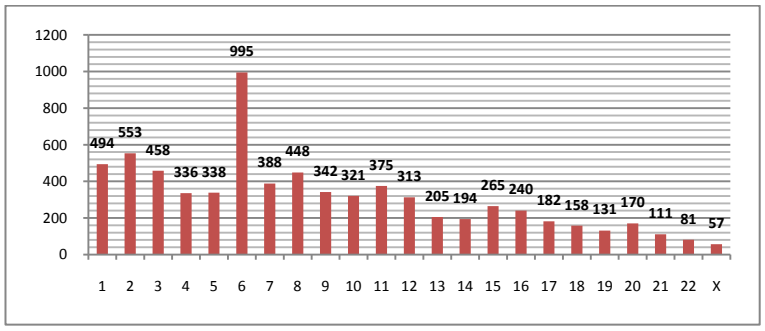


Figure 6. 7 Chromosomal distributions of prioritized SNPs via SPOT for RA data.

Table 6.5 depicts first 20 SNPs chosen by each prioritization scheme. As noted in Chapter 3, SPOT's list heavily depends on individual *p*-values whereas our AHP scheme integrates an evidence based biological functionality for determining relevance to a greater level.

Table 6. 5 Top 20 SNPs according to SPOT and AHP prioritization of GWAS for RA data.

	SPOT Ranking				AHP Ranking			
RANK	**CHR**	**SNP**	**P-value**	**P-Value Rank**	**CHR**	**SNP**	**P-value**	**P-Value Rank**
1	6	rs2395175	1.02E-67	1	6	rs2070600	1.18E-11	89
2	6	rs660895	7.10E-66	2	6	rs2256175	3.37E-4	148
3	6	rs2395163	2.04E-52	3	6	rs3134943	0.001351	248
4	6	rs6910071	1.34E-50	4	6	rs3134940	0.0001395	209
5	6	rs3763309	1.96E-46	5	6	rs3093662	0.0002165	218
6	6	rs3763312	6.22E-46	6	6	rs2256028	0.01838	388
7	6	rs9275224	3.45E-45	7	6	rs2074488	9.82E-8	101
8	6	rs6457617	7.41E-42	8	6	rs1063355	8.27E-05	204
9	6	rs2395185	6.96E-34	9	6	rs9264536	0.3904	5,759
10	6	rs9275555	6.60E-33	10	6	rs2395471	0.2843	2,568
11	6	rs2516049	9.23E-33	11	6	rs1051794	0.3928	5,888
12	6	rs477515	1.56E-32	12	1	rs2476601	3.19E-3	164
13	6	rs9275595	1.53E-31	13	9	rs17611	0.2412	1,872
14	6	rs2076530	2.61E-27	31	6	rs1041981	0.4779	11,098
15	6	rs3817973	1.99E-28	14	9	rs1468673	0.2468	1,941
16	6	rs4424066	3.27E-28	15	9	rs10818500	0.2767	2,413
17	6	rs9275374	3.27E-28	17	6	rs2075800	1.32E-3	161
18	6	rs9275390	3.27E-28	21	6	rs2227956	0.007967	339
19	6	rs9275393	3.27E-28	19	6	rs7383287	0.007007	333
20	6	rs9275406	3.27E-28	16	9	rs7037673	0.0174	384

Finally the biological functionality of the top 20 SNPs listed by AHP versus SPOT based prioritization algorithms are compared through OMIM gene and associated loci analysis. Table 6.6 presents the associations between selected SNPs and OMIM genes which have been experimentally proven to be associated with RA, thus highlighting the algorithm's performance in prioritization of biologically relevant SNPs. As summarized in Table 6.6, it is evident that AHP based prioritization outperforms SPOT's approach in terms of biological relevance. Among top ranking 20 SNPs, only 2% of the SNPs in SPOT list are associated with an OMIM gene, whereas it is 70% for AHP based list. It has also been verified that out of the 20 SNPs in AHP list, 2 of them are experimentally associated with RA, while none of the SNPs in SPOT list has been identified as associated with RA in the literature. Details of our AHP based analysis can be found in Appendix F.

Table 6. 6 Comparison of biological relevance of SPOT and AHP lists for RA.

	SPOT List			AHP List		
RANK	**SNP**	**OMIM Gene**	**RA Association**	**SNP**	**OMIM Gene**	**RA Association**
1	rs2395175			rs2070600	AGER	
2	rs660895			rs2256175		
3	rs2395163			rs3134943	RNF5	
4	rs6910071			rs3134940	AGER	
5	rs3763309			rs3093662	TNF	YES
6	rs3763312			rs2256028		
7	rs9275224			rs2074488		
8	rs6457617			rs1063355	HLA-DQB1	
9	rs2395185			rs9264536		
10	rs9275555			rs2395471		
11	rs2516049			rs1051794	MICA	
12	rs477515			rs2476601	PTPN22	YES
13	rs9275595			rs17611	CC5D	
14	rs2076530	BTNL2		rs1041981	LTA	
15	rs3817973			rs1468673	CC5D	
16	rs4424066			rs10818500	CC5D	
17	rs9275374			rs2075800	HSPA1L	
18	rs9275390			rs2227956	HSPA1L	
19	rs9275393			rs7383287		
20	rs9275406			rs7037673	CC5D	

Interestingly, the top listing SNP in our analysis, rs2070600, has been investigated in a previous study [168] for its relationship with type 2 diabetes mellitus (T2DM) and insulin resistance. The relationship between RA and T2DM has been questioned recently in [169]. Additionally, few of the SNPs (rs1063355, rs2070600, rs3134940) in AHP list have been found to be linked to diabetes mellitus loci previously. So, AHP based SNP prioritization is supporting the recent literature on the novel association suggested between RA and T2DM and insulin resistance.

While AHP based prioritization is offering a more viable prioritized SNP list for subsequent GWAS by integrating significant genes and pathway information from GWAS and backing this up with a scoring mechanism that integrates disease data and disease-disease interactions , here we have also presented its power for suggesting novel associations between SNPs, genes and disease phenotypes. To further prove these suggestions ideally, large-scale, prospective studies are needed. One of the major goals of SNP association studies is to reveal such SNP-phenotype associations to help researcher build new hypothesis and conduct studies.

6.4 SNP Selection

Following prioritization process, as a last step of our study we ran our SA based representative SNP selection algorithm on the training set. Similar to the analysis performed in Chapter 4 we ran the algorithm for each chromosome and merged the selected SNPs as the overall representative SNP set. For computational reasons, the algorithm is stopped if the evaluation function value does not change in 40 SA step in a row. The prediction performance (average accuracy) of the selected SNPs on unselected SNPs and phenotype for each chromosome is presented in Tables 6.7 below for different *w* values (0.3, 0.5 and 0.7, respectively).

Table 6. 7 Prediction performance of representative SNP selection algorithm: t = 10, d = 0.1, cmax = 1,000 for RA data.

	w = 0.3			**w = 0.5**			**w = 0.7**		
CHR	**INI**	**SEL**	**ACA**	**INI**	**SEL**	**ACA**	**INI**	**SEL**	**ACA**
1	470	20	0.613	470	30	0.641	470	63	0.691
2	552	43	0.635	552	34	0.673	552	52	0.686
3	439	26	0.633	439	30	0.651	439	59	0.696
4	353	19	0.62	353	23	0.637	353	49	0.688
5	337	14	0.609	337	19	0.63	337	48	0.694
6	1,036	70	0.699	1,036	85	0.729	1,036	142	0.759
7	376	32	0.676	376	32	0.676	376	52	0.696
8	489	43	0.697	489	43	0.697	489	84	0.746
9	360	29	0.671	360	29	0.671	360	48	0.704
10	336	27	0.655	336	27	0.655	336	48	0.701
11	339	20	0.645	339	20	0.645	339	43	0.701
12	306	22	0.667	306	22	0.667	306	35	0.704
13	231	20	0.679	231	20	0.679	231	40	0.738
14	206	20	0.667	206	20	0.667	206	33	0.722
15	242	18	0.727	242	18	0.727	242	27	0.74
16	229	14	0.649	229	14	0.649	229	31	0.701
17	161	6	0.636	161	6	0.636	161	19	0.686
18	159	10	0.649	159	10	0.649	159	19	0.704
19	114	7	0.62	114	7	0.62	114	11	0.645
20	167	11	0.659	167	11	0.659	167	19	0.686
21	117	6	0.671	117	6	0.671	117	21	0.741
22	73	2	0.632	73	2	0.632	73	9	0.66
X	63	2	0.703	63	2	0.703	63	6	0.735
TOTAL/AVERAGE		**481**	**0.657**		**510**	**0.664**		**958**	**0.705**

CHR: Chromosome number, **SEL:** Selected number, **ACA:** Average classification accuracy.

Using representative SNP selection algorithm we managed to decrease the dimensions considerably. For example for *w* = 0.5 the number of SNPs is decreased from 7,155 to 510 for

RA data. Average classification accuracy for the representative SNP set over unselected for each chromosome is 0.664. Similar to Chapter 4, we did not observe a considerable information loss (in terms of classification accuracy over unselected SNPs). To observe the classification performance of the selected set over the disease phenotype, we compared the performance against Relief-F and Chi-Square. In order to achieve that, we have selected the same set of SNPs for the test sets to that of training sets' and applied a 10-fold Cross Validation run using Naive Bayes classifier as the supervised learning scheme. Results are presented in Table 6.8.

Table 6. 8 Prediction performance comparisons for SA algorithm and WEKA based algorithms.

	w = 0.3, 481 SNPs			w = 0.5, 510 SNPs			w = 0.7, 958 SNPs		
Measure	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F
Accuracy	**0.7160**	0.7306	0.7039	**0.7209**	0.7354	0.7039	0.7136	0.7330	0.7209
Recall	0.6265	0.7108	0.6988	0.6265	0.7169	0.6988	0.6627	0.7108	0.7108
NPV	0.7549	0.7922	0.7768	0.7569	0.7965	0.7768	0.7667	0.7931	0.7885
Precision	**0.6541**	0.6519	0.6170	**0.6624**	0.6575	0.6170	**0.6395**	0.6556	0.6378
Specificity	**0.7764**	0.7439	0.7073	**0.7846**	0.7480	0.7073	**0.7480**	0.7480	0.7276

Table 6.8 reveals that our SA based algorithm outperforms Relief-F in most of the cases. In terms of precision and specificity SA performs way much better than Relief-F and Chi-Square based attribute selection. It is also noteworthy that even if we decrease the dimension considerably, prediction performance of the selected set is still considerably good.

6.5 Conclusion

In this chapter we have presented the performance of integrated METU-SNP application as an all-in-one GWAS over the real life data of Rheumatoid Arthritis consisting of 501,463 SNPs taken from 2,062 patients (868 cases and 1,194 controls). The whole analysis is done by using the newly developed software platform METU-SNP. Analysis also involved use of two novel algorithms developed in the course of this research study: (1) AHP based prioritization and (2) Simulated Annealing based representative SNP selection. Additionally we have integrated state-of-the-art enrichment methods for genes and biological pathways. Results reveal that by using METU-SNP we are able to identify significant signals associated disease phenotypes. We were able to identify significant SNPs as well as genes and pathways, which showed statistically and biologically significant contribution on RA disease phenotype. Our findings are in parallel to current literature on the etiology of Rheumatoid Arthritis. Therefore we conclude that presented software platform METU-SNP is an effective and a reliable tool for GWAS, integrating standardized analysis methods and novel algorithms. We hope that through its easy to use GUI the novel algorithms proposed here will be utilized by SNP biomarker researchers and will lead to discovery of SNP-gene-pathway associations for many more diseases.

CHAPTER 7

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this thesis work, we aimed to find a solution to the problem of identifying genetic variations that are the major reasons of complex diseases. In Chapter 1, we have presented the problem, our proposed methods and major contributions of this dissertation. We have also introduced the experimental data sets that have been utilized for measuring the performance of the algorithms and systems proposed. In Chapter 2, the biological concepts that are utilized for development of the ideas that shaped our research work are introduced. We have explored the human genome and defined the SNPs, which are mutations caused by substitution of one single nucleotide for another at homologous sites in a population, as the most frequently observed genetic variations. Additionally, features of complex diseases are summarized and two complex diseases: (1) Alzheimer's Disease (AD) and (2) Rheumatoid Arthritis (RA), for which we performed extensive analysis are presented. Next, we have showed how GWAS make use of SNPs in an attempt to identify disease causing variants and introduced biological pathways and stating why they are important for the etiology of complex diseases. In Chapter 3, the novel prioritization algorithm for finding biologically and statistically relevant set of SNPs for use in subsequent GWAS has been proposed. The algorithm was based on the well-known multi-criteria decision making method Analytic Hierarchy Process (AHP). We presented a scoring scheme based on the weights assigned to the developed hierarchy tree structure. Based on the calculated final scores, individual SNPs were ranked and prioritization was made accordingly. We showed that our prioritization algorithm outperformed a similar algorithm, SPOT, which was proposed recently, in terms of prediction performance over disease phenotype and biological relevance of the top ranking SNPs. In Chapter 4, we proposed another novel algorithm to tackle the representative SNP selection problem. Our algorithm was based on Simulated Annealing (SA) and it performed reasonably well against two filtering based algorithm presented by WEKA machine learning tool. The newly developed METU-SNP application has been introduced in Chapter 5. METU-SNP is a java based all-in-one software system with MySQL back-end, which is specifically designed for GWAS. METU-SNP database integrated biological information from

major databases and makes use of this valuable information in SNP prioritization process. Well-known 3rd party tools such as PLINK and BEAGLE are also integrated to the application. Integration with these tools enabled METU-SNP to perform comprehensive analysis for GWAS. Quality control based preprocessing, association analysis for individual SNPs and second-wave GWAS through gene set enrichment framework, AHP based prioritization and SA based representative SNP selection are among the analysis options offered by METU-SNP. In Chapter 6, METU-SNP's functionality on real life case control data for RA has been presented. Results showed that METU-SNP was able to identify significant signals towards disease phenotype. It was able to identify significant SNPs, genes and pathways and findings were in parallel to current etiology of RA. In the following sections we will summarize our contributions, state difficulties faced during the course of this research and present young researchers with possible future research directions in this field.

7.1 Contributions

This section aims to provide a summary of contributions of this thesis:

- First and foremost contribution of our research study is our AHP based prioritization scheme for SNP selection after a GWAS. We stated that focusing on the statistical evidence alone is not a valid approach for biomarker discovery in a GWAS setting. It is highly unlikely to spot the true causal polymorphisms by looking at only the *p*-values of association. Therefore, there is a need to prioritize and select a subset of SNPs for subsequent GWAS studies and/or functional experiments. In the literature, alongside with statistical evidence, use of biological information is favored. With this in mind we introduced a novel method for SNP biomarker scoring and prioritizing based on the well-known multi-criteria decision making method Analytic Hierarchy Process. We created a hierarchy tree to structurize the SNP prioritization process. We combined statistical information (*p*-values for SNPs and additionally combined *p*-values for genes and pathways) with SNP's biological importance depending on three different criteria: (1) Evolutionary Conservation, (2) Gene Association and (3) Genomic Location. In order to calculate scores for SNPs we integrate information from major public databases. Our algorithm steps forward as one of few algorithms presented in the literature that aims to prioritize SNPs. We performed a comparative study with one of mostly cited approaches in the literature: SPOT. Our algorithm outperformed SPOT in prediction performance.
- Next, we presented a Simulated Annealing based representative SNP selection algorithm. We stated that the enormous number of SNPs makes it difficult to perform an extensive analysis on all the SNPs in the human genome. Thus, while performing a disease association study, it is desired to work with a subset of the entire SNP set and not all SNPs, thereby considerably saving limited resources. Therefore, selecting a subset of

SNPs that is informative enough to perform association studies but still small enough to reduce the analysis workload, to which we refer as representative SNP selection, has become an important step for disease-gene association studies. The major aim for representative SNP selection is reducing the biological and statistical redundancy from hundreds of thousands of SNPs. To solve this problem, we present an OR class novel feature selection method based on Simulated Annealing (SA). In this method, we try to maximize tagged SNP prediction while minimizing cardinality of the selected SNP subset. We performed a comparative study with two filtering based attribute selection methods offered by WEKA tool: Relief-F and Chi-square. Our algorithm performed well against these algorithms especially if classification accuracy is favored over SNP number.

- Lastly, we have developed a java based integrated desktop software, which is specifically designed as an all-in-one application for use in GWAS: METU-SNP. Our software aims to present researchers with an effective and handy tool to help during their endeavors to find SNPs that are significantly associated with complex diseases. It makes use of data from major public databases such as dbSNP, Entrez Gene, KEGG, Gene Ontology etc. It is equipped with a state-of-the-art Gene Set Enrichment Analysis and pathway discovery methods. Additionally we have implemented the algorithms developed based on AHP based SNP prioritization and Simulated Annealing based representative SNP selection within METU-SNP framework. To us, METU-SNP is the very first version of a much more detailed analysis tool that will be shaped within METU Bioinformatics labs. We hope that it will be a common tool of choice for GWAS and it will ultimately support timely identification of genomic disease biomarkers, and development of personalized medicine approaches and targeted drug discoveries.

7.2 Major Drawbacks

In this section, we list the difficulties faced during the course of this research study:

- Accessing to real experimental genotyping data has been one of the major difficulties. This even became a bottleneck as without knowing the structure of the data set commonly used for GWAS we could not continue developing our algorithms. The data is usually regarded as classified, therefore to get a hand on a real data set we had to wait for a long time until we have convinced the Genetic Analysis Workshop organizers.
- The genotyping data we had gain access to analyze was too large. In many occasions, analyzing the data, which is in gigabytes range, has put a computational burden to our work many times. At some point we had to change the structure of our database and SNP prioritization algorithm as we had to prematurely end a single SQL query that took 5 days to run at an Intel 2 Duo, 2.26 GHz and 2.86 GB RAM machine. Editing the data

was required scripting as common text editors are unable to "even" open the data at hand.

- The databases that needed to be integrated required tens of gigabytes of disk storage. Storing the whole data in a local computer hasn't been feasible and we, unfortunately, could not get access to a dedicated server in METU premises during the course of this study. Therefore we decided to use a compact, integrated SQL dump we acquired from the developers of SNPLogic tool and SPOT. We have optimized the MySQL database to work with our algorithms: however this resulted in a tradeoff for maintainability. One needs to further process and update the data to keep current with literature for new versions of the METU-SNP.

7.3 Future Research

This section aims to guide interested graduate students for possible future research:

- During our analysis we found significant genes and pathways by using enrichment analysis. One could try to integrate haplotype based association and identify significant haplotypes. Significant haplotype information can also be used to restructure the AHP tree. BEAGLE presents users with the functionality of calculating haplotype based associations.
- It is a fact that different individuals diagnosed with the same complex disease may have different reasons for disease development. Therefore one may try to divide case data into different clusters and perform simultaneous association runs on different clusters. That is because some SNPs would cause disease in some of the patients but not in others and catching these not-so-strong signals in the overall data is not an easy task. We think that biclustering[23] will be a good fit for this purpose. One may try to integrate different biclustering algorithms into METU-SNP and apply a biclustering run before GWAS step. It would be interesting to compare the performance of SNP prioritization with and without biclustering application.
- We stated that imputation is important for increasing the quality of GWAS. We have integrated BEAGLE's standard imputation methodology, which is based on hidden markov model. One may try to apply different imputation methods and compare the performance with BEAGLE.
- METU-SNP version 1.0 is a java based desktop application with a MySQL back-end. The database includes biological data incorporated from different major databases such as dbSNP, Entrez Gene and others. The database currently lacks an automated update functioning and it is designed so as to optimize the performance of AHP algorithm. Therefore if one wants to update the database lots of manual tasks are needed. Hence, it

[23] Biclustering, co-clustering, or two-mode clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix.

would be quite useful to implement a similar mechanism to that of dbAutoMaker[24] and run periodical updates for the database. Additionally, disease database that is based on Disease Ontology and GeneRIF dates back to 2008. It is therefore required to integrate an updated version. Lastly, a web based version of the METU-SNP that would be hosted on a dedicated "secure" server within METU premises would help researchers worldwide to use the functionality offered by METU-SNP and would permit collaborations that would initiate productive research.

- For intelligent filtering of the massive SNP data we have used an AHP based prioritization algorithm and we developed a Simulated Annealing based selection algorithm for representative SNP selection. One could work on a different, improved version of the AHP tree and perform Delphi Study with a different and extended group of biology experts. With current version of METU-SNP, it is not possible to change the AHP scores of the leaf nodes. It would be a good add on to allow the user to "fine-tune" the scores according to user's choice. For supervised learning step of selection algorithm we used Naive Bayes classifier because it is fast. It is possible to use different classifiers offered by WEKA, as METU-SNP allows use of other WEKA based classifiers and perform a comparative study. For Simulated Annealing based selection, the performance of our algorithm depends hardly on initial random selection of SNPs. One enhancement would be to integrate AHP scores or any other intelligent method to modify the initialization part of the algorithm so that more relevant SNPs are selected at first step.
- Current version of METU-SNP lacks graphical representation of the analysis results. Haploview[25] is a great tool that offers visualization and plotting of PLINK based analysis. One may try to implement functionality offered by Haploview to complete an important shortcoming of METU-SNP. Another important add-on to the current version of METU-SNP would be giving the user an option to populate gene-disease table in the database with the results of the analysis, which is currently being run. Genes that are found out to be statistically relevant for a particular disease will be favored in subsequent GWAS studies for the very same disease.
- Lastly, a noteworthy enhancement would be to allow use of different input file formats by integrating a file format conversion tool. This way data gathered from different sources can be converted into PLINK ready ped and map files. Additionally, presenting user with the ability to perform other relevant analysis, presented by PLINK, such as family based association or population stratification could be added as functionality.

[24] http://www.csiro.au/science/dbAutoMaker.html
[25] http://www.broadinstitute.org/haploview

REFERENCES

[1] "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299-1320, Oct. 2005.

[2] "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851-861, Oct. 2007.

[3] L. Kruglyak and D. A. Nickerson, "Variation is the spice of life," *Nat Genet*, vol. 27, no. 3, pp. 234-236, Mar. 2001.

[4] M. Corbex et al., "Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction," *Genetic Epidemiology*, vol. 19, no. 1, pp. 64-80, Jul. 2000.

[5] D. Tregouet et al., "Specific haplotypes of the P-selectin gene are associated with myocardial infarction," *Human Molecular Genetics*, vol. 11, no. 17, pp. 2015 -2023, 2002.

[6] N. Tahri-Daizadeh, D. Tregouet, V. Nicaud, N. Manuel, F. Cambien, and L. Tiret, "Automated Detection of Informative Combined Effects in Genetic Association Studies of Complex Traits," *Genome Research*, vol. 13, no. 8, pp. 1952 -1960, 2003.

[7] B. S. Shastry, "SNPs in disease gene mapping, medicinal drug development and evolution," *Journal of Human Genetics*, vol. 52, no. 11, pp. 871-880, 2007.

[8] T. L. Saaty, "How to Make a Decision: The Analytic Hierarchy Process," *Interfaces*, vol. 24, no. 6, pp. 19-43, Dec. 1994.

[9] N. Dalkey and O. Helmer, "An Experimental Application of the DELPHI Method to the Use of Experts," *Management Science*, vol. 9, no. 3, pp. 458-467, Apr. 1963.

[10] F. Azuaje, *Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine*. John Wiley and Sons, 2010.

[11] S. Purcell et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559-575, Sep. 2007.

[12] B. L. Browning and S. R. Browning, "Efficient Multilocus Association Testing for Whole Genome Association Studies Using Localized Haplotype Clustering," *Genetic Epidemiology*, no. 31, pp. 365 - 375, 2007.

[13] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nat Genet*, vol. 39, no. 7, pp. 906-913, Jul. 2007.

[14] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn, "GenABEL: an R library for

genome-wide association analysis," *Bioinformatics*, vol. 23, no. 10, pp. 1294 -1296, May. 2007.

[15] S. T. Sherry et al., "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308-311, Jan. 2001.

[16] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 33, pp. D54-58, Jan. 2005.

[17] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27-30, Jan. 2000.

[18] M. A. Harris et al., "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258-261, Jan. 2004.

[19] R. S. S. Hawley and C. A. Mori, *The Human Genome: A User's Guide*, 1st ed. Academic Press, 1998.

[20] A. Ghosh, M. Bansal, "*A glossary of DNA structures from A to Z*". Acta Crystallogr D Biol Crystallogr 59 (Pt 4): 620–6, 2003.

[21] Wikipedia contributors, "Gene" in *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, 2010.

[22] E. C. Friedberg, G. C. Walker, and W. Siede, DNA repair and mutagenesis. ASM Press, 1995.

[23] Wikipedia contributors, "Single nucleotide polymorphism," in *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, 2010.

[24] N. Malats and F. Calafell, "Basic glossary on genetic epidemiology," *Journal of Epidemiology and Community Health*, vol. 57, no. 7, pp. 480-482, Jul. 2003.

[25] E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry, "dbSNP: a database of single nucleotide polymorphisms," *Nucleic Acids Research*, vol. 28, no. 1, pp. 352 -355, Jan. 2000.

[26] D. M. Altshuler et al., "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52-58, Sep. 2010.

[27] D. F. Conrad et al., "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome," *Nat Genet*, vol. 38, no. 11, pp. 1251-1260, Nov. 2006.

[28] "Single Nucleotide Polymorphisms. Methods and Protocols. Second Edition.," *Anticancer Research*, vol. 30, no. 3, p. 1035, Mar. 2010.

[29] L. X. Shen, J. P. Basilion, and V. P. Stanton, "Single-nucleotide polymorphisms can cause different structural folds of mRNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 14, pp. 7871-7876, Jul. 1999.

[30] L. Diatchenko et al., "Genetic basis for individual variations in pain perception and the development of a chronic pain condition," *Human Molecular Genetics*, vol. 14, no. 1, pp. 135-143, Jan. 2005.

[31] E. T. Wang et al., "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470-476, Nov. 2008.

[32] F. Capon, “A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups,” *Human Molecular Genetics*, vol. 13, no. 20, pp. 2361-2368, 2004.

[33] T. R. Rebbeck, M. Spitz, and X. Wu, “Assessing the function of genetic variants in candidate gene association studies,” *Nature Reviews. Genetics*, vol. 5, no. 8, pp. 589-597, Aug. 2004.

[34] P. H. LEE, “Prioritizing SNPs for Disease-Gene Association Studies: Algorithms and Systems,” 2009.

[35] P. Kwok and M. Xiao, “Single-Molecule Analysis for Molecular Haplotyping,” *Human mutation*, vol. 23, no. 5, pp. 442-446, May. 2004.

[36] Y. Zhang and J. C. Rajapakse, *Machine learning in bioinformatics*. John Wiley and Sons, 2008.

[37] D. E. Reich et al., “Linkage disequilibrium in the human genome,” *Nature*, vol. 411, no. 6834, pp. 199-204, May. 2001.

[38] B. D and N. R, “A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping,” *Genomics*, vol. 29, no. 2, pp. 311-322, Sep. 1995.

[39] D. C. Crawford and D. A. Nickerson, “Definition and clinical importance of haplotypes,” *Annual Review of Medicine*, vol. 56, pp. 303-320, 2005.

[40] A. Kuliev and Y. Verlinsky, “Preimplantation diagnosis: a realistic option for assisted reproduction and genetic practice,” *Current Opinion in Obstetrics & Gynecology*, vol. 17, no. 2, pp. 179-183, Apr. 2005.

[41] A. G. Motulsky, “Genetics of complex diseases,” *Journal of Zhejiang University. Science. B*, vol. 7, no. 2, pp. 167-168, Feb. 2006.

[42] J. Pevsner, *Bioinformatics and functional genomics*. John Wiley and Sons, 2009.

[43] Wikipedia contributors, “Alzheimer's disease,” in *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, 2010.

[44] R. Guttman, R. D. Altman, and N. H. Nielsen, “Alzheimer disease. Report of the Council on Scientific Affairs,” *Archives of Family Medicine*, vol. 8, no. 4, pp. 347-353, Aug. 1999.

[45] “Alzheimer's Association - Risk Factors.” [Online]. Available: http://www.alz.org/alzheimers_disease_causes_risk_factors.asp. [Accessed: 01-Oct-2010].

[46] Wikipedia contributors, “Rheumatoid arthritis,” in *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, 2010.

[47] A. J. Silman and J. E. Pearson, “Epidemiology and genetics of rheumatoid arthritis,” *Arthritis Research*, vol. 4, pp. S265-272, 2002.

[48] K. L. Sewell and D. E. Trentham, “Pathogenesis of rheumatoid arthritis,” *Lancet*, vol. 341, no. 8840, pp. 283-286, Jan. 1993.

[49] E. D. Harris, “Rheumatoid arthritis. Pathophysiology and implications for therapy,” *The*

New England Journal of Medicine, vol. 322, no. 18, pp. 1277-1289, May. 1990.

[50] P. K. Gregersen, "Genetics of rheumatoid arthritis: confronting complexity," *Arthritis Research*, vol. 1, no. 1, pp. 37-44, 1999.

[51] J. S. Lawrence, "Genetics of rheumatoid factor and rheumatoid arthritis," *Clinical and Experimental Immunology*, vol. 2, pp. 769-783, Dec. 1967.

[52] E. Zanelli, F. C. Breedveld, and R. R. de Vries, "HLA class II association with rheumatoid arthritis: facts and interpretations," *Human Immunology*, vol. 61, no. 12, pp. 1254-1261, Dec. 2000.

[53] M. S. Fife et al., "Multipoint linkage analysis of a candidate gene locus in rheumatoid arthritis demonstrates significant evidence of linkage and association with the corticotropin-releasing hormone genomic region," *Arthritis and Rheumatism*, vol. 43, no. 8, pp. 1673-1678, Aug. 2000.

[54] S. John et al., "Linkage of a marker in intron D of the estrogen synthase locus to rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 42, no. 8, pp. 1617-1620, Aug. 1999.

[55] A. Khani-Hanjani et al., "Association between dinucleotide repeat in non-coding region of interferon-gamma gene and susceptibility to, and severity of, rheumatoid arthritis," *Lancet*, vol. 356, no. 9232, pp. 820-825, Sep. 2000.

[56] S. John et al., "Linkage of cytokine genes to rheumatoid arthritis. Evidence of genetic heterogeneity," *Annals of the Rheumatic Diseases*, vol. 57, no. 6, pp. 361-365, Jun. 1998.

[57] A. Myerscough, S. John, J. H. Barrett, W. E. Ollier, and J. Worthington, "Linkage of rheumatoid arthritis to insulin-dependent diabetes mellitus loci: evidence supporting a hypothesis for the existence of common autoimmune susceptibility loci," *Arthritis and Rheumatism*, vol. 43, no. 12, pp. 2771-2775, Dec. 2000.

[58] Y. Ohnishi, T. Tanaka, K. Ozaki, R. Yamada, H. Suzuki, and Y. Nakamura, "A high-throughput SNP typing system for genome-wide association studies," *Journal of Human Genetics*, vol. 46, no. 8, pp. 471-477, 2001.

[59] W. Chung et al., "Medical genetics: A marker for Stevens-Johnson syndrome," *Nature*, vol. 428, no. 6982, p. 486, Apr. 2004.

[60] T. A. Pearson and T. A. Manolio, "How to Interpret a Genome-wide Association Study," *JAMA*, vol. 299, no. 11, pp. 1335-1344, Mar. 2008.

[61] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *The New England Journal of Medicine*, vol. 360, no. 17, pp. 1759-1768, Apr. 2009.

[62] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *The New England Journal of Medicine*, vol. 363, no. 2, pp. 166-176, Jul. 2010.

[63] "Genome.gov | Biological Pathways Fact Sheet." [Online]. Available: http://www.genome.gov/27530687. [Accessed: 14-Nov-2010].

[64] A. Torkamani, E. J. Topol, and N. J. Schork, "Pathway analysis of seven common diseases assessed by genome-wide association," *Genomics*, vol. 92, no. 5, pp. 265-272, Nov. 2008.

[65] R. Huber et al., "Identification of intra-group, inter-individual, and gene-specific

variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane," *Arthritis Research & Therapy*, vol. 10, no. 4, p. R98, 2008.

[66] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440 -9445, 2003.

[67] K. K. Nicodemus, W. Liu, G. A. Chase, Y. Tsai, and M. D. Fallin, "Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms," *BMC Genetics*, vol. 6, no. 1, pp. S78-S78.

[68] S. F. Saccone et al., "Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence," *Bioinformatics*, vol. 24, no. 16, pp. 1805 - 1811, 2008.

[69] S. F. Saccone et al., "SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study," *Nucleic Acids Research*, vol. 38, pp. W201-W209, 2010.

[70] S. Goodswen, C. Gondro, N. Watson-Haigh, and H. Kadarmideen, "FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases," *BMC Bioinformatics*, vol. 11, no. 1, p. 311, 2010.

[71] H. Yuan et al., "FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization," *Nucleic Acids Research*, vol. 34, pp. W635-W641, Jul. 2006.

[72] A. R. Pico et al., "SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system," *Nucleic Acids Research*, vol. 37, pp. D803-D809, 2009.

[73] Z. Xu and J. A. Taylor, "SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies," *Nucleic Acids Research*, vol. 37, pp. W600-W605, 2009.

[74] T. H. Shen, C. S. Carlson, and P. Tarczy-Hornoch, "SNPit: a federated data integration system for the purpose of functional SNP annotation," *Computer methods and programs in biomedicine*, vol. 95, no. 2, pp. 181-189, Aug. 2009.

[75] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *The New England Journal of Medicine*, vol. 360, no. 17, pp. 1759-1768, Apr. 2009.

[76] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," *The American Journal of Human Genetics*, vol. 86, no. 1, pp. 6-22, 2010.

[77] G. S. Sagoo, J. Little, and J. P. T. Higgins, "Systematic reviews of genetic association studies. Human Genome Epidemiology Network," *PLoS Medicine*, vol. 6, no. 3, p. e28, Mar. 2009.

[78] A. Yesupriya, W. Yu, M. Clyne, M. Gwinn, and M. J. Khoury, "The continued need to synthesize the results of genetic associations across multiple studies," *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, vol. 10, no. 8, pp. 633-635, Aug. 2008.

[79] P. I. de Bakker, M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, "Practical aspects of imputation-driven meta-analysis of genome-wide association

studies," *Human Molecular Genetics*, vol. 17, no. 2, pp. R122-R128, 2008.

[80] P. L. De Jager et al., "Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci," *Nature Genetics*, vol. 41, no. 7, pp. 776-782, Jul. 2009.

[81] J. D. Cooper et al., "Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci," *Nat Genet*, vol. 40, no. 12, pp. 1399-1401, Dec. 2008.

[82] A. D. Johnson et al., "Genome-wide association meta-analysis for total serum bilirubin levels," *Human Molecular Genetics*, vol. 18, no. 14, pp. 2700 -2710, Jul. 2009.

[83] R. Fisher, "Combining independent tests of significance," *American Statistician*, no. 2, p. 30, 1948.

[84] N. Soranzo et al., "Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size," *PLoS Genetics*, vol. 5, no. 4, p. e1000445, Apr. 2009.

[85] G. Peng et al., "Gene and pathway-based second-wave analysis of genome-wide association studies," *European Journal of Human Genetics: EJHG*, vol. 18, no. 1, pp. 111-117, Jan. 2010.

[86] N. J. Timpson et al., "Common variants in the region around Osterix are associated with bone mineral density and growth in childhood," *Human Molecular Genetics*, vol. 18, no. 8, pp. 1510-1517, Apr. 2009.

[87] C. M. Lindgren et al., "Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution," *PLoS Genetics*, vol. 5, no. 6, p. e1000508, Jun. 2009.

[88] M. Kolz et al., "Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations," *PLoS Genetics*, vol. 5, no. 6, p. e1000504, Jun. 2009.

[89] T. Tanaka et al., "Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations," *American Journal of Human Genetics*, vol. 84, no. 4, pp. 477-482, Apr. 2009.

[90] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545 -15550, Oct. 2005.

[91] K. Yu et al., "Pathway analysis by adaptive combination of P-values," *Genetic Epidemiology*, vol. 33, no. 8, pp. 700-709, Dec. 2009.

[92] S. E. Baranzini et al., "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis," *Human Molecular Genetics*, vol. 18, no. 11, pp. 2078-2090, Jun. 2009.

[93] M. Hirakawa, T. Tanaka, Y. Hashimoto, M. Kuroda, T. Takagi, and Y. Nakamura, "JSNP: a database of common gene variations in the Japanese population," *Nucleic Acids Research*, vol. 30, no. 1, pp. 158 -162, Jan. 2002.

[94] T. Hubbard et al., "Ensembl 2005," *Nucleic Acids Research*, vol. 33, pp. D447-453, Jan. 2005.

[95] W. J. Kent et al., “The Human Genome Browser at UCSC,” *Genome Research*, vol. 12, no. 6, pp. 996 -1006, Jun. 2002.

[96] B. R. Packer, “SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes,” *Nucleic Acids Research*, vol. 34, no. 90001, pp. D617-D621, 2006.

[97] S. H. Wilson and K. Olden, “The Environmental Genome Project: phase I and beyond,” *Molecular Interventions*, vol. 4, no. 3, pp. 147-156, Jun. 2004.

[98] “SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA (URL:http://pga.gs.washington.edu) [11.10.2010].” [Online]. Available: http://pga.gs.washington.edu/UseStatement.html. [Accessed: 11-Oct-2010].

[99] D. N. Cooper, P. D. Stenson, and N. A. Chuzhanova, “The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms,” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, vol. 1, p. Unit 1.13, Jan. 2006.

[100] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 33, pp. D514-517, Jan. 2005.

[101] J. Osborne et al., “Annotating the human genome with Disease Ontology,” *BMC Genomics*, vol. 10, no. 1, p. S6, 2009.

[102] P. D. Karp et al., “Expansion of the BioCyc collection of pathway/genome databases to 160 genomes,” *NUCLEIC ACIDS RES*, vol. 33, pp. 6083--6089, 2005.

[103] T. Kelder, A. R. Pico, K. Hanspers, M. P. van Iersel, C. Evelo, and B. R. Conklin, “Mining Biological Pathways Using WikiPathways Web Services,” *PLoS ONE*, vol. 4, no. 7, p. e6447, Jul. 2009.

[104] L. Conde et al., “PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes,” *Nucleic Acids Research*, vol. 34, pp. W621-W625, 2006.

[105] M. Doumpos and C. Zopounidis, *Multicriteria decision aid classification methods*. Springer, 2002.

[106] M. Tavana, D. T. Kennedy, J. Rappaport, and Y. J. Ugras, “An AHP-Delphi Group Decision Support System Applied to Conflict Resolution in Hiring Decisions,” *SSRN eLibrary*.

[107] C. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilità,” *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, no. 8, pp. 3-62, 1936.

[108] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289-300, 1995.

[109] V. Bafna, B. V. Halldórsson, R. Schwartz, and A. G. Clark, “Haplotypes and informative SNP selection algorithms: don’t block out information,” *PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON RESEARCH IN COMPUTATIONAL MOLECULAR BIOLOGY*, p. pages, 2003.

[110] E. Halperin, “Tag SNP selection in genotype data for maximizing SNP prediction

accuracy," *Bioinformatics*, vol. 21, no. 1, pp. i195-i203, 2005.

[111] K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun, "A dynamic programming algorithm for haplotype block partitioning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 11, pp. 7335-7339, May. 2002.

[112] K. Zhang, Z. Qin, T. Chen, J. S. Liu, M. S. Waterman, and F. Sun, "HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms," *Bioinformatics*, vol. 21, no. 1, pp. 131 -134, Jan. 2005.

[113] K. Zhang and L. Jin, "HaploBlockFinder: haplotype block analyses," *Bioinformatics (Oxford, England)*, vol. 19, no. 10, pp. 1300-1301, Jul. 2003.

[114] K. Zhang, P. Calabrese, M. Nordborg, and F. Sun, "Haplotype block structure and its applications to association studies: power and study designs," *American Journal of Human Genetics*, vol. 71, no. 6, pp. 1386-1394, Dec. 2002.

[115] E. Lindholm et al., "Polymorphism in the MHC2TA Gene Is Associated with Features of the Metabolic Syndrome and Cardiovascular Mortality," *PLoS ONE*, vol. 1, no. 1, p. e64, Dec. 2006.

[116] B. M. Neale and P. C. Sham, "The Future of Association Studies: Gene-Based Analysis and Replication," *American Journal of Human Genetics*, vol. 75, no. 3, pp. 353-362, Sep. 2004.

[117] F. Dudbridge and B. Koeleman, "Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies," *The American Journal of Human Genetics*, vol. 75, no. 3, pp. 424-435, 2004.

[118] S. R. Seaman and B. Müller-Myhsok, "Rapid Simulation of P Values for Product Methods and Multiple-Testing Adjustment in Association Studies," *American Journal of Human Genetics*, vol. 76, no. 3, pp. 399-408, Mar. 2005.

[119] Y. V. Sun, A. M. Levin, E. Boerwinkle, H. Robertson, and S. L. Kardia, "A scan statistic for identifying chromosomal patterns of SNP association," *Genetic Epidemiology*, vol. 30, no. 7, pp. 627-635, 2006.

[120] A. Dembo and S. Karlin, "Poisson Approximations for r-Scan Processes," *The Annals of Applied Probability*, vol. 2, no. 2, pp. 329-357, May. 1992.

[121] J. Hoh and J. Ott, "Scan statistics to scan markers for susceptibility genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 17, pp. 9615-9617, Aug. 2000.

[122] A. M. Levin, D. Ghosh, K. R. Cho, and S. L. R. Kardia, "A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors," *Bioinformatics*, vol. 21, no. 12, pp. 2867 -2874.

[123] N. Patil et al., "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," *Science (New York, N.Y.)*, vol. 294, no. 5547, pp. 1719-1723, Nov. 2001.

[124] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander, "High-resolution haplotype structure in the human genome," *Nature Genetics*, vol. 29, no. 2, pp. 229-232, Oct. 2001.

[125] J. Hampe, S. Schreiber, and M. Krawczak, "Entropy-based SNP selection for genetic

association studies," *Human Genetics*, vol. 114, no. 1, pp. 36-43, 2003.

[126] G. C. Johnson et al., "Haplotype tagging for the identification of common disease genes," *Nat Genet*, vol. 29, no. 2, pp. 233-237, Oct. 2001.

[127] R. Judson, B. Salisbury, J. Schneider, A. Windemuth, and J. C. Stephens, "How many SNPs does a genome-wide haplotype map require?," *Pharmacogenomics*, vol. 3, no. 3, pp. 379-391, May. 2002.

[128] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag SNPs," *Bioinformatics*, vol. 19, no. 2, pp. 287 -288, Jan. 2003.

[129] B. D. Horne and N. J. Camp, "Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation," *Genetic Epidemiology*, vol. 26, no. 1, pp. 11-21, Jan. 2004.

[130] Z. Lin and R. B. Altman, "Finding haplotype tagging SNPs by use of principal components analysis," *American Journal of Human Genetics*, vol. 75, no. 5, pp. 850-861, Nov. 2004.

[131] Z. Meng, D. V. Zaykin, C. Xu, M. Wagner, and M. G. Ehm, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes," *American Journal of Human Genetics*, vol. 73, no. 1, pp. 115-130, Jul. 2003.

[132] P. Zhang, H. Sheng, and R. Uehara, "A double classification tree search algorithm for index SNP selection," *BMC Bioinformatics*, vol. 5, pp. 89-89.

[133] S. I. Ao et al., "CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs," *Bioinformatics (Oxford, England)*, vol. 21, no. 8, pp. 1735-1736, Apr. 2005.

[134] M. C. Byng, J. C. Whittaker, A. P. Cuthbert, C. G. Mathew, and C. M. Lewis, "SNP subset selection for genetic association studies," *Annals of Human Genetics*, vol. 67, no. 6, pp. 543-556, Nov. 2003.

[135] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *American Journal of Human Genetics*, vol. 74, no. 1, pp. 106-120, Jan. 2004.

[136] X. Ke et al., "A comparison of tagging methods and their tagging space," *Human Molecular Genetics*, vol. 14, no. 18, pp. 2757 -2767, Sep. 2005.

[137] P. H. Lee and H. Shatkay, "BNTagger: improved tagging SNP selection using Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e211 -e219, Jul. 2006.

[138] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science (New York, N.Y.)*, vol. 220, no. 4598, pp. 671-680, May. 1983.

[139] S. Browning and B. Browning, "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 1084-1097, 2007.

[140] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10-18, 2009.

[141] M. Nothnagel, D. Ellinghaus, S. Schreiber, M. Krawczak, and A. Franke, “A comprehensive evaluation of SNP genotype imputation,” *Human Genetics*, vol. 125, no. 2, pp. 163-171, Mar. 2009.

[142] K. Hao, E. Chudin, J. McElwee, and E. Schadt, “Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies.,” *BMC genetics*, vol. 10, no. 1, p. 27, Jun. 2009.

[143] P. I. de Bakker, M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, “Practical aspects of imputation-driven meta-analysis of genome-wide association studies,” *Human Molecular Genetics*, vol. 17, no. 2, pp. R122-R128, 2008.

[144] Y. Li, “Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference.”

[145] P. Scheet and M. Stephens, “A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase,” *American Journal of Human Genetics*, vol. 78, no. 4, pp. 629-644, Apr. 2006.

[146] J. Biernacka et al., “Assessment of genotype imputation methods,” *BMC Proceedings*, vol. 3, no. 7, p. S5, 2009.

[147] D. Karolchik et al., “The UCSC Genome Browser Database: 2008 update,” *Nucleic Acids Research*, vol. 36, pp. D773-D779, 2007.

[148] P. Du et al., “From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations,” *Bioinformatics*, vol. 25, no. 12, pp. i63 -i68, Jun. 2009.

[149] L. Chen, M. Zhong, W. Chen, C. Amos, and R. Fan, “A genome-wide association scan for rheumatoid arthritis data by Hotelling's T2 tests,” *BMC Proceedings*, vol. 3, no. 7, p. S6, 2009.

[150] S. Raychaudhuri, “Recent advances in the genetics of rheumatoid arthritis,” *Current Opinion in Rheumatology*, vol. 22, no. 2, pp. 109-118, 2010.

[151] J. Sundqvist et al., “Endometriosis and autoimmune disease: association of susceptibility to moderate/severe endometriosis with CCL21 and HLA-DRB1,” *Fertility and Sterility*.

[152] O. González-Recio, E. L. de Maturana, A. T. Vega, C. D. Engelman, and K. W. Broman, “Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model,” vol. 3, no. 7, pp. S63-S63.

[153] G. Hein and S. Franke, “Are advanced glycation end-product-modified proteins of pathogenetic importance in fibromyalgia?,” *Rheumatology*, vol. 41, no. 10, pp. 1163 - 1167, Oct. 2002.

[154] S. Drinda et al., “Identification of the advanced glycation end products Nε-carboxymethyllysine in the synovial tissue of patients with rheumatoid arthritis,” *Annals of the Rheumatic Diseases*, vol. 61, no. 6, pp. 488 -492, Jun. 2002.

[155] L. Fugger et al., “Restriction fragment length polymorphism of two HLA-B-associated transcripts genes in five autoimmune diseases,” *Human Immunology*, vol. 30, no. 1, pp. 27-31, Jan. 1991.

[156] D. P. Singal and W. W. Buchanan, “Human leucocyte antigens (HLA) and rheumatic diseases: HLA class ii antigen-associated diseases,” *Inflammopharmacology*, vol. 2, no. 1,

pp. 47-61, 1993.

[157] S. H. Pincus, D. O. Clegg, and J. R. Ward, "Characterization of T cells bearing HLA-DR antigens in rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 28, no. 1, pp. 8-15, Jan. 1985.

[158] D. P. Singal, M. Ye, X. Qiu, and M. D'Souza, "Polymorphisms in the TAP2 gene and their association with rheumatoid arthritis," *Clinical and Experimental Rheumatology*, vol. 12, no. 1, pp. 29-33, Feb. 1994.

[159] H. Ishii, M. Nakazawa, S. Yoshino, H. Nakamura, K. Nishioka, and T. Nakajima, "Expression of Notch homologues in the synovium of rheumatoid arthritis and osteoarthritis patients," *Rheumatology International*, vol. 21, no. 1, pp. 10-14, 2001.

[160] R. A. Kaplan et al., "Metabolism of C4 and factor B in rheumatoid arthritis. Relation to rheumatoid factor," *Arthritis and Rheumatism*, vol. 23, no. 8, pp. 911-920, Aug. 1980.

[161] S. Kumar, S. M. Blake, and J. G. Emery, "Intracellular signaling pathways as a target for the treatment of rheumatoid arthritis," *Current Opinion in Pharmacology*, vol. 1, no. 3, pp. 307-313, Jun. 2001.

[162] O. Werdelin, M. Meldal, and T. Jensen, "Processing of glycans on glycoprotein and glycopeptide antigens in antigen-presenting cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9611 -9613, Jul. 2002.

[163] R. Thomas and P. E. Lipsky, "Presentation of self peptides by dendritic cells: possible implications for the pathogenesis of rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 39, no. 2, pp. 183-190, Feb. 1996.

[164] D. O. Haskard, "Cell adhesion molecules in rheumatoid arthritis," *Current Opinion in Rheumatology*, vol. 7, no. 3, pp. 229-234, May. 1995.

[165] M. D. Smith et al., "Successful treatment of rheumatoid arthritis is associated with a reduction in synovial membrane cytokines and cell adhesion molecule expression," *Rheumatology*, vol. 40, no. 9, pp. 965 -977, 2001.

[166] G. S. Firestein, "Immunologic mechanisms in the pathogenesis of rheumatoid arthritis," *Journal of Clinical Rheumatology: Practical Reports on Rheumatic & Musculoskeletal Diseases*, vol. 11, no. 3, pp. S39-44, Jun. 2005.

[167] R. N. Maini, "Heberden Oration, 1988. Exploring immune pathways in rheumatoid arthritis," *British Journal of Rheumatology*, vol. 28, no. 6, pp. 466-479, Dec. 1989.

[168] A. C. Goulart, S. Germer, K. M. Rexrode, M. Martin, and R. Y. L. Zee, "Polymorphisms in advanced glycosylation end product-specific receptor (AGER) gene, insulin resistance, and type 2 diabetes mellitus," *Clinica Chimica Acta; International Journal of Clinical Chemistry*, vol. 398, no. 1, pp. 95-98, Dec. 2008.

[169] M. Doran, "Rheumatoid arthritis and diabetes mellitus: evidence for an association?," *The Journal of Rheumatology*, vol. 34, no. 3, pp. 460-462, Mar. 2007.

APPENDICES

APPENDIX A: GLOSSARY FOR GWAS

Alleles: Alternate forms of a gene or chromosomal locus that differ in DNA sequence

Annotation catalog: A map denoting the function of specific genomic regions, such as sites to which noncoding RNA or transcription factors bind.

Candidate gene: A gene believed to influence expression of complex phenotypes due to known biological and/or physiological properties of its products, or to its location near a region of association or linkage

Common disease–common variant hypothesis: A theory that many common diseases are caused by common alleles that individually have little effect but in concert confer a high risk.

Complex disease: A disorder in which the cause is considered to be a combination of genetic effects and environmental influences.

Copy number variants: Stretches of genomic sequence of roughly 1 kb to 3 Mb in size that are deleted or are duplicated in varying numbers

Deep resequencing: A technique for sequencing a gene in several thousand subjects, typically with the use of high-throughput sequencing.

Epigenetics: The study of heritable changes to DNA structure that does not alter the underlying sequence; well-known examples are DNA methylation and histone modification.

Exome: All the expressed messenger RNA sequences in any tissue.

False discovery rate: Proportion of significant associations that are actually false positives

False-positive report probability: Probability that the null hypothesis is true, given a statistically significant finding

Fine mapping: The precise mapping of a locus after it has been identified by genetic linkage or association. The initial localization is determined within megabases of DNA in genetic linkage studies and within tens of kilobases in genetic association studies. In genetic association studies, fine mapping implies finding all the variants at the locus and trying to determine which changes may be related to pathogenesis

with the use of statistical, functional, or bioinformatic methods.

Functional studies: Investigations of the role or mechanism of a genetic variant in causation of a disease or trait

Gene-environment interactions: Modification of gene-disease associations in the presence of environmental factors

Genetic association: A relationship that is defined by the nonrandom occurrence of a genetic marker with a trait, which suggests an association between the genetic marker (and/or a marker close to it) and disease pathogenesis.

Genetic linkage: A relationship that is defined by the coinheritance of a genetic marker with disease in a family with multiple disease-affected members.

Genome-wide association study: Any study of genetic variation across the entire human genome

designed to identify genetic association with observable traits or the presence or absence of a disease, usually referring to studies with genetic marker density of 100,000 or more to represent a large proportion of variation in the human genome

Genotyping call rate: Proportion of samples or SNPs for which a specific allele SNP can be reliably identified by a genotyping method

Haplotype: A group of specific alleles at neighboring genes or markers that tends to be inherited together

Haplotypic structure: The general underlying segmentation of the genome. As a result of recombination events occurring throughout the history of a population, contiguous segments of DNA are shared by persons within a population. Chromosomes can thus be broken down into contiguous segments, containing haplotypes common to members of particular populations.

HapMap: Genome-wide database of patterns of common human genetic sequence variation among multiple ancestral population samples

Hardy Weinberg equilibrium: Population distribution of 2 alleles (with frequencies p and q) such that the distribution is stable from generation to generation and genotypes occur at frequencies of p^2, $2pq$, and q^2 for the major allele homozygote, heterozygote, and minor allele homozygote, respectively

Heritability: The proportion of inter-individual differences (variance) in a trait that is the result of genetic factors; often estimated on the basis of parent–offspring correlations for continuous traits or the ratio of the incidence in first-degree relatives of affected persons to the incidence in first-degree relatives of unaffected persons.

High-throughput sequencing: Several new techniques that since 2005 have increased the speed and decreased the cost of DNA sequencing by two orders of magnitude.

Human Genome Project: A coordinated international effort that led to the consensus sequence of the human genome

Intergenic regions: Segments of DNA that do not contain or overlap genes.

Introns: The portions of a gene that are removed (spliced out) before translation to a protein. Introns may contain regulatory information that is critical to appropriate gene expression.

Inversion: A chromosomal segment that has been broken off and reinserted in the same place, but with the genetic sequence in reverse order.

Linkage disequilibrium: Association between 2 alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance

Mendelian disease: Condition caused almost entirely by a single major gene, such as cystic fibrosis orHuntington'sdisease, inwhichdisease ismanifested in only 1 (recessive) or 2 (dominant) of the 3 possible genotype groups

Minor allele: The allele of a biallelic polymorphism that is less frequent in the study population

Minor allele frequency: Proportion of the less common of 2 alleles in a population (with 2 alleles carried by each person at each autosomal locus) ranging from less than 1% to less than 50%

Modest effect: Association between a gene variant and disease or trait that is statistically significant but carries a small odds ratio (usually < 1.5)

Monogenic disease: A disorder caused by a mutation in a single gene (also called a Mendelian disease).

Noncoding RNAs: Segments of RNA that are not translated into amino acid sequences but may be involved in the regulation of gene expression.

Non-Mendelian disease (also "common" or "complex" disease): Condition influenced by multiple genes and environmental factors and not showing Mendelian inheritance patterns

Nonsynonymous SNP: Apolymorphismthat results in achangein theaminoacid sequence of a protein (and therefore may affect the function of the protein)

Platform: Arrays or chips on which high-throughput genotyping is performed

Polymorphic: A gene or site with multiple allelic forms. The term *polymorphism* usually implies a minor allele frequency of at least 1%.

Population attributable risk: Proportion of a disease or trait in the population that is due to a specific cause, such as a genetic variant

Population stratification (also "population structure"): A form of confounding in genetic association studies caused by genetic differencesbetweencasesandcontrols unrelated to disease but due to sampling them from populations of different ancestries

Positional cloning: An approach for determining the position of a gene that, when mutated, causes monogenic disease. In families with disease, genetic markers from every chromosome are typed in both affected and unaffected members. Markers that are co-inherited with disease indicate the chromosomal position of the genetic defect, and then genes at that position are sequenced to find the pathogenic mutation, which in turn indicates the causative gene.

Power: A statistical term for the probability of identifying a difference between 2 groups in a study when a difference truly exists.

Rare variant: A genetic variant with a minor-allele frequency of less than 1%. Rare variants are typically single-nucleotide substitutions but can also be structural variants.

RNA interference: The inhibition of gene expression by noncoding RNA molecules.

Sequence motif: DNA sequences whose functions can be inferred because they are similar to sequences whose function has been biologically determined.

Single-nucleotide polymorphism: Most common form of genetic variation in the genome, in which

a single-base substitution has created 2 forms of a DNA sequence that differ by a single nucleotide.

Structural genomic variation: Variation within the genome that results from deletion or duplication (both referred to as copy-number variation) or from inversion of genomic segments. Although common large variants (of more than one kilobase) exist, the majority of such variants are rare.

Tag SNP: A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms.

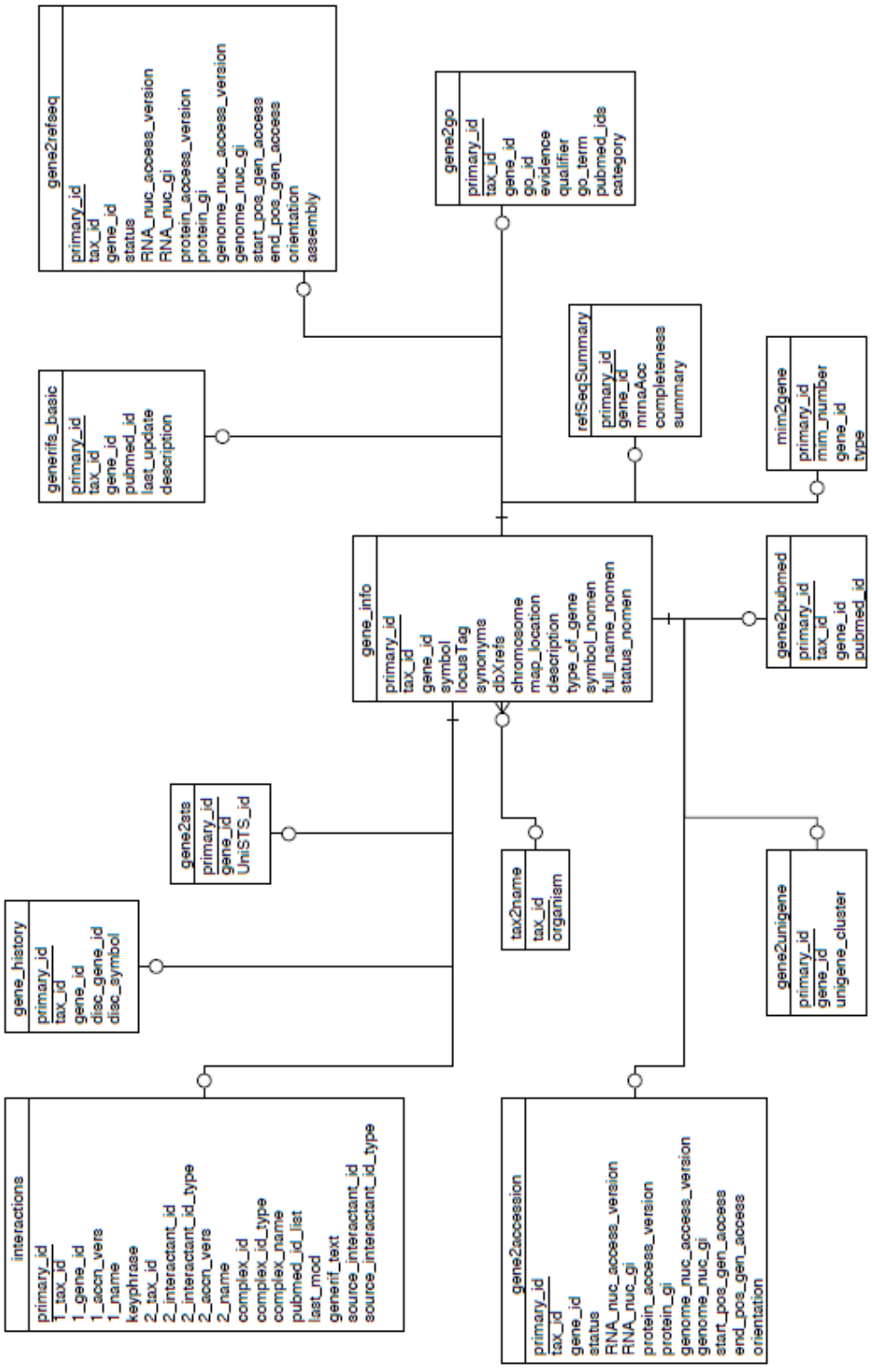
Transcription factor: A protein that binds to gene regulatory regions in DNA and helps to control gene expression.

Translocation: A chromosomal segment that has been broken off and reinserted in a different place in the genome.

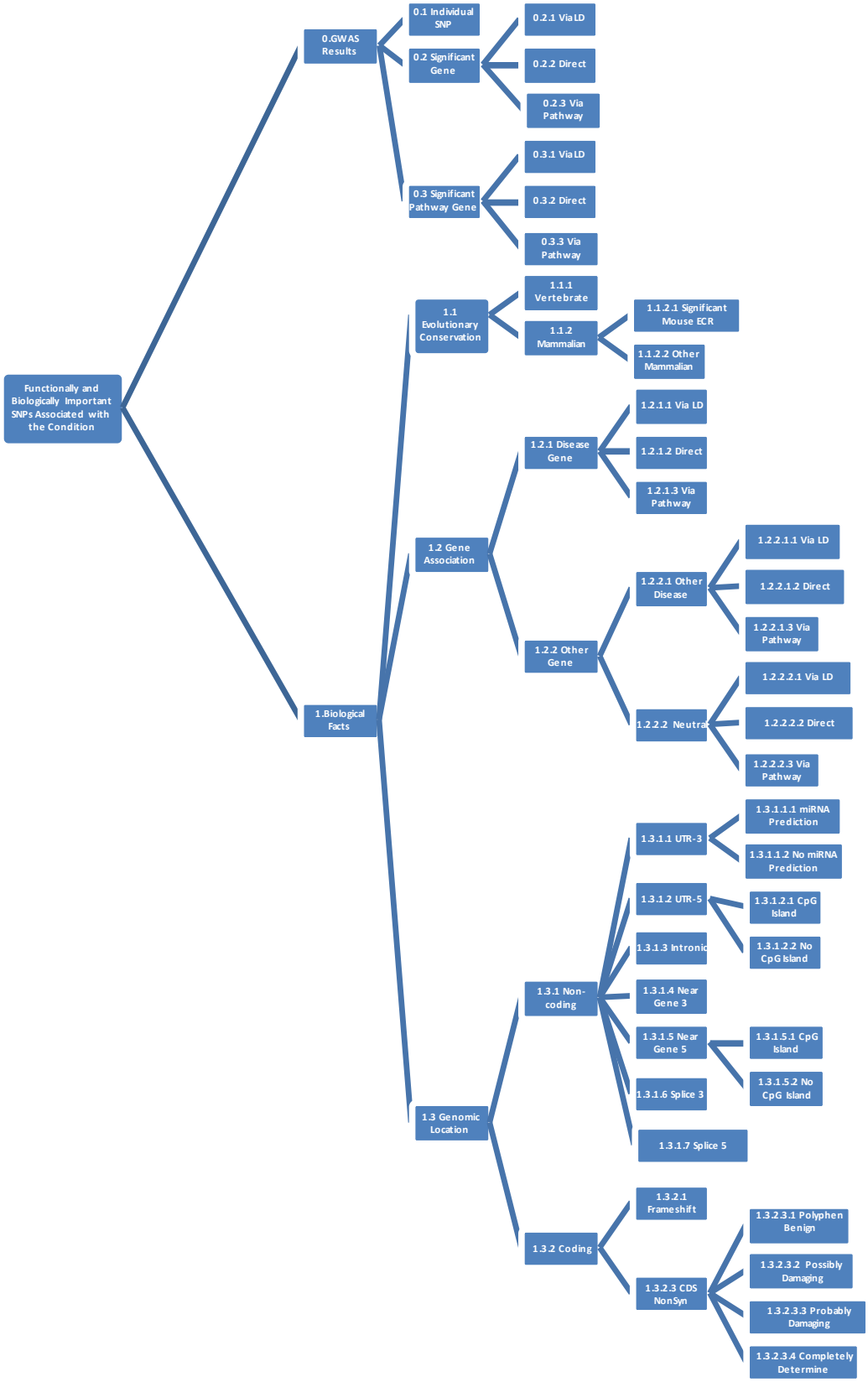
Trio: Genetic study design including an affected offspring and both parents**.**

1000 Genomes Project: A whole-genome resequencing of 1000 subjects from the original and extended HapMap populations, which was started in 2008, with funding from an international research consortium.

APPENDIX B: ENTITY-RELATIONSHIP DIAGRAM OF ENTREZ-GENE



APPENDIX C: AHP-TREE STRUCTURE



APPENDIX D: AHP SCORING DETAILS

Consistency Check

[8], it has been proven that for a consistent comparison matrix, the largest Eigen Value (λ_{max}) is equal to the size of the matrix *n*. Then a measure of consistency, called Consistency Index (CI) has been proposed such that:

$$\boldsymbol{CI = \frac{\lambda_{max} - n}{n - 1}}$$

[8] also suggested that CI should be compared with an appropriate one to check consistency so a Random Consistency Index (RI) is calculated by taking the average of CI for 500 randomly created matrices. RI is depicted in the table below:

n	2	3	4	5	6	7	8	9	10
RI	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

Following that, a measure called Consistency Ratio is proposed such that:

$$CR = \frac{CI}{RI}$$

If the value of Consistency Ratio is smaller or equal to 10%, the inconsistency is acceptable; otherwise one has to revise the judgement.

Expert 1

Pairwise comparisons made by Expert 1 are depicted below:

CRITERIA		MORE IMPORTANT (A OR B)	INTENSITY (SCALE FROM 1 TO 9)
A	**B**		
0	1	B	2,0
0.1	0.2	B	7,0
0.1	0.3	B	5,0
0.2	0.3	A	3,0
0.2.1	0.2.2	B	9,0
0.2.1	0.2.3	B	6,0
0.2.2	0.2.3	A	3,0
0.3.1	0.3.2	B	9,0
0.3.1	0.3.3	B	6,0
0.3.2	0.3.3	A	3,0
1.1	1.2	B	8,0
1.1	1.3	B	6,0
1.2	1.3	A	3,5
1.1.1	1.1.2	B	2,0

1.1.2.1	1.1.2.2	A	2,0
1.2.1	1.2.2	A	9,0
1.2.1.1	1.2.1.2	B	9,0
1.2.1.1	1.2.1.3	B	6,0
1.2.1.2	1.2.1.3	A	3,0
1.2.2.1	1.2.2.2	A	3,0
1.2.2.1.1	1.2.2.1.2	B	9,0
1.2.2.1.1	1.2.2.1.3	B	6,0
1.2.2.1.2	1.2.2.1.3	A	3,0
1.2.2.2.1	1.2.2.2.2	B	9,0
1.2.2.2.1.	1.2.2.2.3	B	6,0
1.2.2.2.2	1.2.2.2.3	A	3,0
1.3.1	1.3.2	B	7,0
1.3.1.1	1.3.1.2	B	4,0
1.3.1.1	1.3.1.3	A	6,0
1.3.1.1	1.3.1.4	A	5,5
1.3.1.1	1.3.1.5	B	3,0
1.3.1.1	1.3.1.6	A	4,0
1.3.1.1	1.3.1.7	A	3,0
1.3.1.1.1	1.3.1.1.2	A	7,0
1.3.1.2	1.3.1.3	A	9,0
1.3.1.2	1.3.1.4	A	8,0
1.3.1.2	1.3.1.5	A	3,0
1.3.1.2	1.3.1.6	A	7,0
1.3.1.2	1.3.1.7	A	5,0
1.3.1.2.1	1.3.1.2.2	A	5,0
1.3.1.3	1.3.1.4	B	2,0
1.3.1.3	1.3.1.5	B	8,0
1.3.1.3	1.3.1.6	B	3,0
1.3.1.3	1.3.1.7	B	5,0
1.3.1.4	1.3.1.5	B	7,0
1.3.1.4	1.3.1.6	B	3,0
1.3.1.4	1.3.1.7	B	5,0
1.3.1.5	1.3.1.6	A	6,0
1.3.1.5	1.3.1.7	A	4,0
1.3.1.5.1	1.3.1.5.2	A	3,0
1.3.1.6	1.3.1.7	B	3,0
1.3.2.1	1.3.2.2	A	9,0
1.3.2.1	1.3.2.3	A	6,0
1.3.2.2	1.3.2.3	B	4,0
1.3.2.3.1	1.3.2.3.2	B	3,0
1.3.2.3.1	1.3.2.3.3	B	7,0
1.3.2.3.1	1.3.2.3.4	B	9,0
1.3.2.3.2	1.3.2.3.3	B	2,0
1.3.2.3.2	1.3.2.3.4	B	3,0
1.3.2.3.3	1.3.2.3.4	B	2,0

Resulting comparison matrices are presented below:

Criteria/Criteria	0	1
0	1	0,5
1	2	1
Total	3	1,5

	0	1	Priority Vector
0	0,333333	0,333333	0,33
1	0,666667	0,666667	0,67

	0.1	0.2	0.3
0.1	1	0,142857	0,2
0.2	7	1	3
0.3	5	0,333333	1
Total	13	1,47619	4,2

	0.1	0.2	0.3	Priority Vector
0.1	0,076923	0,096774	0,047619048	0,07
0.2	0,538462	0,677419	0,714285714	0,64
0.3	0,384615	0,225806	0,238095238	0,28

lambda max	3,031
consistency index (CI)	1,54%
consistency ratio (CR)	2,65%

	0.2.1	0.2.2	0.2.3
0.2.1	1	0,111111	0,16667
0.2.2	9	1	3
0.2.3	6	0,333333	1
Total	16	1,444444	4,16667

	0.2.1	0.2.2	0.2.3	Priority Vector
0.2.1	0,0625	0,076923	0,04	0,06
0.2.2	0,5625	0,692308	0,72	0,66
0.2.3	0,375	0,230769	0,24	0,28

lambda max	3,08
consistency index (CI)	4,00%
consistency ratio (CR)	6,90%

	0.3.1	0.3.2	0.3.3
0.3.1	1	0,111111	0,16667
0.3.2	9	1	3
0.3.3	6	0,333333	1
Total	16	1,444444	4,16667

	0.3.1	0.3.2	0.3.3	Priority Vector
0.3.1	0,0625	0,076923	0,04	0,06
0.3.2	0,5625	0,692308	0,72	0,66
0.3.3	0,375	0,230769	0,24	0,28

lambda max	3,08
consistency index (CI)	4,00%
consistency ratio (CR)	6,90%

	1.1	1.2	1.3
1.1	1	0,125	0,16667
1.2	8	1	3,5
1.3	6	0,285714	1
Total	15	1,410714	4,66667

	1.1	1.2	1.3	Priority Vector
1.1	0,066667	0,088608	0,035714286	0,06
1.2	0,533333	0,708861	0,75	0,66
1.3	0,4	0,202532	0,214285714	0,27

lambda max	3,091
consistency index (CI)	4,55%
consistency ratio (CR)	7,85%

	1.1.1	1.1.2
1.1.1	1	0,5
1.1.2	2	1
Total	3	1,5

	1.1.1	1.1.2	Priority Vector
1.1.1	0,333333	0,333333	0,33
1.1.2	0,666667	0,666667	0,67

	1.1.2.1	1.1.2.2
1.1.2.1	1	2
1.1.2.2	0,5	1
Total	1,5	3

	1.1.2.1	1.1.2.2	Priority Vector
1.1.2.1	0,666667	0,666667	0,67
1.1.2.2	0,333333	0,333333	0,33

	1.2.1	1.2.2
1.2.1	1	9
1.2.2	0,111111	1
Total	1,111111	10

	1.2.1	1.2.2	Priority Vector
1.2.1	0,9	0,9	0,9
1.2.2	0,1	0,1	0,1

	1.2.1.1	1.2.1.2	1.2.1.3
1.2.1.1	1	0,111111	0,16667
1.2.1.2	9	1	3
1.2.1.3	6	0,333333	1
Total	16	1,444444	4,16667

	1.2.1.1	1.2.1.2	1.2.1.3	Priority Vector
1.2.1.1	0,0625	0,076923	0,04	0,06
1.2.1.2	0,5625	0,692308	0,72	0,66
1.2.1.3	0,375	0,230769	0,24	0,28

lambda max	3,08
consistency index (CI)	4,00%
consistency ratio (CR)	6,90%

	1.2.2.1	1.2.2.2
1.2.2.1	1	3
1.2.2.2	0,333333	1
Total	1,333333	4

	1.2.2.1	1.2.2.2	Priority Vector
1.2.2.1	0,75	0,75	0,75
1.2.2.2	0,25	0,25	0,25

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3
1.2.2.1.1	1	0,111111	0,16667
1.2.2.1.2	9	1	3
1.2.2.1.3	6	0,333333	1
Total	16	1,444444	4,16667

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3	Priority Vector
1.2.2.1.1	0,0625	0,076923	0,04	0,06
1.2.2.1.2	0,5625	0,692308	0,72	0,66
1.2.2.1.3	0,375	0,230769	0,24	0,28

lambda max	3,08
consistency index (CI)	4,00%
consistency ratio (CR)	6,90%

	1.3.1	1.3.2
1.3.1	1	0,142857
1.3.2	7	1
Total	8	1,142857

	1.3.1	1.3.2	Priority Vector
1.3.1	0,125	0,125	0,13
1.3.2	0,875	0,875	0,88

	1.3.1.1	1.3.1.2	1.3.1.3	.3.1.	1.3.1.5	1.3.1.6	1.3.1.7
1.3.1.1	1	0,25	6	5,5	0,333333	4	3
1.3.1.2	4	1	9	8	3	7	5
1.3.1.3	0,166667	0,111111	1	0,5	0,125	0,333333	0,2
1.3.1.4	0,181818	0,125	2	1	0,142857	0,333333	0,2
1.3.1.5	3	0,333333	8	7	1	6	4
1.3.1.6	0,25	0,142857	3	3	0,166667	1	0,333333
1.3.1.7	0,333333	0,2	5	5	0,25	3	1
Total	8,931818	2,162302	34	30	5,017857	21,66667	13,73333

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7	Priority Vector
1.3.1.1	0,111959	0,115617545	0,176470588	0,183333333	0,066429	0,184615	0,218446602	0,15
1.3.1.2	0,447837	0,462470178	0,264705882	0,266666667	0,597865	0,323077	0,36407767	0,39
1.3.1.3	0,01866	0,051385575	0,029411765	0,016666667	0,024911	0,015385	0,014563107	0,02
1.3.1.4	0,020356	0,057808772	0,058823529	0,033333333	0,02847	0,015385	0,014563107	0,03
1.3.1.5	0,335878	0,154156726	0,235294118	0,233333333	0,199288	0,276923	0,291262136	0,25
1.3.1.6	0,02799	0,066067168	0,088235294	0,1	0,033215	0,046154	0,024271845	0,06
1.3.1.7	0,03732	0,092494036	0,147058824	0,166666667	0,049822	0,138462	0,072815534	0,1

lambda max	7,690868
consistency index (CI)	11,51%
consistency ratio (CR)	8,72%

	1.3.1.1.1	1.3.1.1.2
1.3.1.1.1	1	7
1.3.1.1.2	0,142857	1
Total	1,142857	8

	1.3.1.1.1	1.3.1.1.2	Priority Vector
1.3.1.1.1	0,875	0,875	0,88
1.3.1.1.2	0,125	0,125	0,13

	1.3.1.2.1	1.3.1.2.2
1.3.1.2.1	1	5
1.3.1.2.2	0,2	1
Total	1,2	6

	1.3.1.2.1	1.3.1.2.2	Priority Vector
1.3.1.2.1	0,833333	0,833333	0,83
1.3.1.2.2	0,166667	0,166667	0,17

	1.3.1.5.1	1.3.1.5.2
1.3.1.5.1	1	3
1.3.1.5.2	0,333333	1
Total	1,333333	4

	1.3.1.5.1	1.3.1.5.2	Priority Vector
1.3.1.5.1	0,75	0,75	0,75
1.3.1.5.2	0,25	0,25	0,25

	1.3.2.1	1.3.2.2	1.3.2.3
1.3.2.1	1	9	6
1.3.2.2	0,111111	1	0,25
1.3.2.3	0,166667	4	1
Total	1,277778	14	7,25

	1.3.2.1	1.3.2.2	1.3.2.3	Priority Vector
1.3.2.1	0,782609	0,642857	0,827586207	0,75
1.3.2.2	0,086957	0,071429	0,034482759	0,06
1.3.2.3	0,130435	0,285714	0,137931034	0,18

lambda max	3,1033
consistency index (CI)	5,17%
consistency ratio (CR)	8,91%

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4
1.3.2.3.1	1	0,333333	0,1428571	0,111111111
1.3.2.3.2	3	1	0,5	0,333333333
1.3.2.3.3	7	2	1	0,5
1.3.2.3.4	9	3	2	1
Total	20	6,333333	3,6428571	1,944444444

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	Priority Vector
1.3.2.3.1	0,05	0,052631579	0,039215686	0,057143	0,05
1.3.2.3.2	0,15	0,157894737	0,137254902	0,171429	0,15
1.3.2.3.3	0,35	0,315789474	0,274509804	0,257143	0,3
1.3.2.3.4	0,45	0,473684211	0,549019608	0,514286	0,5

lambda max	4,015079
consistency index (CI)	0,50%
consistency ratio (CR)	0,56%

Expert 2

Pairwise comparisons made by Expert 2 are depicted below:

CRITERIA		MORE IMPORTANT (A OR B)	INTENSITY (SCALE FROM 1 TO 9)
A	**B**		
0	1	B	3

0.1	0.2	B	7
0.1	0.3	B	5
0.2	0.3	A	3
0.2.1	0.2.2	B	5
0.2.1	0.2.3	B	3
0.2.2	0.2.3	A	3
0.3.1	0.3.2	B	5
0.3.1	0.3.3	B	3
0.3.2	0.3.3	A	3
1.1	1.2	B	3
1.1	1.3	B	4
1.2	1.3	B	2
1.1.1	1.1.2	B	5
1.1.2.1	1.1.2.2	A	5
1.2.1	1.2.2	A	5
1.2.1.1	1.2.1.2	B	5
1.2.1.1	1.2.1.3	B	3
1.2.1.2	1.2.1.3	A	3
1.2.2.1	1.2.2.2	A	7
1.2.2.1.1	1.2.2.1.2	B	5
1.2.2.1.1	1.2.2.1.3	B	3
1.2.2.1.2	1.2.2.1.3	A	3
1.2.2.2.1	1.2.2.2.2	B	5
1.2.2.2.1.	1.2.2.2.3	B	3
1.2.2.2.2	1.2.2.2.3	A	3
1.3.1	1.3.2	B	7
1.3.1.1	1.3.1.2	B	3
1.3.1.1	1.3.1.3	A	3
1.3.1.1	1.3.1.4	B	3
1.3.1.1	1.3.1.5	B	3
1.3.1.1	1.3.1.6	B	5
1.3.1.1	1.3.1.7	B	5
1.3.1.1.1	1.3.1.1.2	A	7
1.3.1.2	1.3.1.3	A	5
1.3.1.2	1.3.1.4	A	3
1.3.1.2	1.3.1.5	B	4
1.3.1.2	1.3.1.6	B	4
1.3.1.2	1.3.1.7	B	4
1.3.1.2.1	1.3.1.2.2	A	5
1.3.1.3	1.3.1.4	B	4
1.3.1.3	1.3.1.5	B	4
1.3.1.3	1.3.1.6	B	6
1.3.1.3	1.3.1.7	B	6
1.3.1.4	1.3.1.5	B	1
1.3.1.4	1.3.1.6	B	5
1.3.1.4	1.3.1.7	B	5

1.3.1.5	1.3.1.6	B	5
1.3.1.5	1.3.1.7	B	5
1.3.1.5.1	1.3.1.5.2	A	5
1.3.1.6	1.3.1.7	A	1
1.3.2.1	1.3.2.2	A	7
1.3.2.1	1.3.2.3	A	2
1.3.2.2	1.3.2.3	B	7
1.3.2.3.1	1.3.2.3.2	B	5
1.3.2.3.1	1.3.2.3.3	B	7
1.3.2.3.1	1.3.2.3.4	B	8
1.3.2.3.2	1.3.2.3.3	B	3
1.3.2.3.2	1.3.2.3.4	B	5
1.3.2.3.3	1.3.2.3.4	B	3

Resulting comparison matrices are presented below:

Criteria/Criteria	0	1
0	1	0,333333
1	3	1
Total	4	1,333333

	0	1	Priority Vector
0	0,25	0,25	0,25
1	0,75	0,75	0,75

	0.1	0.2	0.3
0.1	1	0,142857	0,2
0.2	7	1	3
0.3	5	0,333333	1
Total	13	1,47619	4,2

	0.1	0.2	0.3	Priority Vector
0.1	0,076923	0,096774	0,047619048	0,07
0.2	0,538462	0,677419	0,714285714	0,64
0.3	0,384615	0,225806	0,238095238	0,28

lambda max	3,031
consistency index (CI)	1,54%
consistency ratio (CR)	2,65%

	0.2.1	0.2.2	0.2.3
0.2.1	1	0,2	0,33
0.2.2	5	1	3
0.2.3	3	0,333333	1
Total	9	1,533333	4,33

	0.2.1	0.2.2	0.2.3	Priority Vector
0.2.1	0,111111	0,130435	0,076923077	0,11
0.2.2	0,555556	0,652174	0,692307692	0,63
0.2.3	0,333333	0,217391	0,230769231	0,26

lambda max	3,083
consistency index (CI)	4,13%
consistency ratio (CR)	7,13%

	0.3.1	0.3.2	0.3.3
0.3.1	1	0,2	0,33
0.3.2	5	1	3
0.3.3	3	0,333333	1
Total	9	1,533333	4,33

	0.3.1	0.3.2	0.3.3	Priority Vector
0.3.1	0,111111	0,130435	0,076923077	0,11
0.3.2	0,555556	0,652174	0,692307692	0,63
0.3.3	0,333333	0,217391	0,230769231	0,26

lambda max	3,083
consistency index (CI)	4,13%
consistency ratio (CR)	7,13%

	1.1	1.2	1.3
1.1	1	0,333333	0,25
1.2	3	1	0,5
1.3	4	2	1
Total	8	3,333333	1,75

	1.1	1.2	1.3	Priority Vector
1.1	0,125	0,1	0,142857143	0,12
1.2	0,375	0,3	0,285714286	0,32
1.3	0,5	0,6	0,571428571	0,56

lambda max	3,007
consistency index (CI)	0,33%
consistency ratio (CR)	0,57%

	1.1.1	1.1.2
1.1.1	1	0,2
1.1.2	5	1
Total	6	1,2

	1.1.1	1.1.2	Priority Vector
1.1.1	0,166667	0,166667	0,17
1.1.2	0,833333	0,833333	0,83

	1.1.2.1	1.1.2.2
1.1.2.1	1	5
1.1.2.2	0,2	1
Total	1,2	6

	1.1.2.1	1.1.2.2	Priority Vector
1.1.2.1	0,833333	0,833333	0,83
1.1.2.2	0,166667	0,166667	0,17

	1.2.1	1.2.2
1.2.1	1	5
1.2.2	0,2	1
Total	1,2	6

	1.2.1	1.2.2	Priority Vector
1.2.1	0,833333	0,833333	0,83
1.2.2	0,166667	0,166667	0,17

	1.2.1.1	1.2.1.2	1.2.1.3
1.2.1.1	1	0,2	0,3333
1.2.1.2	5	1	3
1.2.1.3	3	0,333333	1
Total	9	1,533333	4,3333

	1.2.1.1	1.2.1.2	1.2.1.3	Priority Vector
1.2.1.1	0,111111	0,130435	0,076923077	0,11
1.2.1.2	0,555556	0,652174	0,692307692	0,63
1.2.1.3	0,333333	0,217391	0,230769231	0,26

lambda max	3,083
consistency index (CI)	4,13%
consistency ratio (CR)	7,13%

	1.2.2.1	1.2.2.2
1.2.2.1	1	7
1.2.2.2	0,142857	1
Total	1,142857	8

	1.2.2.1	1.2.2.2	Priority Vector
1.2.2.1	0,875	0,875	0,88
1.2.2.2	0,125	0,125	0,13

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3
1.2.2.1.1	1	0,2	0,3333
1.2.2.1.2	5	1	3
1.2.2.1.3	3	0,333333	1
Total	9	1,533333	4,3333

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3	Priority Vector
1.2.2.1.1	0,111111	0,130435	0,076923077	0,11
1.2.2.1.2	0,555556	0,652174	0,692307692	0,63
1.2.2.1.3	0,333333	0,217391	0,230769231	0,26

lambda max	3,083
consistency index (CI)	4,13%
consistency ratio (CR)	7,13%

	1.3.1	1.3.2
1.3.1	1	0,142857
1.3.2	7	1
Total	8	1,142857

	1.3.1	1.3.2	Priority Vector
1.3.1	0,125	0,125	0,13
1.3.2	0,875	0,875	0,88

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7
1.3.1.1	1	0,333333	3	0,333333333	0,333333	0,2	0,2
1.3.1.2	3	1	5	3	0,25	0,25	0,25
1.3.1.3	0,333333	0,2	1	0,25	0,25	0,166667	0,166667
1.3.1.4	3	0,333333	4	1	1	0,2	0,2
1.3.1.5	3	4	4	1	1	0,2	0,2
1.3.1.6	5	4	6	5	5	1	1
1.3.1.7	5	4	6	5	5	1	1
Total	20,33333	13,86667	29	15,58333333	12,83333	3,016667	3,016667

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7	Priority Vector
1.3.1.1	0,04918	0,024038	0,103448	0,02139	0,025974	0,066298	0,066298	0,05
1.3.1.2	0,147541	0,072115	0,172414	0,192513	0,019481	0,082873	0,082873	0,11
1.3.1.3	0,016393	0,014423	0,034483	0,016043	0,019481	0,055249	0,055249	0,03
1.3.1.4	0,147541	0,024038	0,137931	0,064171	0,077922	0,066298	0,066298	0,08
1.3.1.5	0,147541	0,288462	0,137931	0,064171	0,077922	0,066298	0,066298	0,12
1.3.1.6	0,245902	0,288462	0,206897	0,320856	0,38961	0,331492	0,331492	0,3
1.3.1.7	0,245902	0,288462	0,206897	0,320856	0,38961	0,331492	0,331492	0,3

lambda max	8,008667
consistency index (CI)	16,81%
consistency ratio (CR)	12,74%

	1.3.1.1.1	1.3.1.1.2
1.3.1.1.1	1	7
1.3.1.1.2	0,142857	1
Total	1,142857	8

	1.3.1.1.1	1.3.1.1.2	Priority Vector
1.3.1.1.1	0,875	0,875	0,88
1.3.1.1.2	0,125	0,125	0,13

	1.3.1.2.1	1.3.1.2.2
1.3.1.2.1	1	5
1.3.1.2.2	0,2	1
Total	1,2	6

	1.3.1.2.1	1.3.1.2.2	Priority Vector
1.3.1.2.1	0,833333	0,833333	0,83
1.3.1.2.2	0,166667	0,166667	0,17

	1.3.1.5.1	1.3.1.5.2
1.3.1.5.1	1	5
1.3.1.5.2	0,2	1
Total	1,2	6

	1.3.1.5.1	1.3.1.5.2	Priority Vector
1.3.1.5.1	0,833333	0,833333	0,83
1.3.1.5.2	0,166667	0,166667	0,17

	1.3.2.1	1.3.2.2	1.3.2.3
1.3.2.1	1	7	2
1.3.2.2	0,142857	1	0,14286
1.3.2.3	0,5	7	1
Total	1,642857	15	3,14286

	1.3.2.1	1.3.2.2	1.3.2.3	Priority Vector
1.3.2.1	0,608696	0,466667	0,636363636	0,57
1.3.2.2	0,086957	0,066667	0,045454545	0,07
1.3.2.3	0,304348	0,466667	0,318181818	0,36

lambda max	3,117857
consistency index (CI)	5,89%
consistency ratio (CR)	10,16%

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4
1.3.2.3.1	1	0,2	0,142857	0,125
1.3.2.3.2	5	1	0,333333	0,2
1.3.2.3.3	7	3	1	0,333333
1.3.2.3.4	8	5	3	1
Total	21	9,2	4,47619	1,658333

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	Priority Vector
1.3.2.3.1	0,047619	0,02173913	0,031914894	0,075377	0,04
1.3.2.3.2	0,238095	0,108695652	0,074468085	0,120603	0,14
1.3.2.3.3	0,333333	0,326086957	0,223404255	0,201005	0,27
1.3.2.3.4	0,380952	0,543478261	0,670212766	0,603015	0,55

lambda max	4,248655
consistency index (CI)	8,29%
consistency ratio (CR)	9,21%

Expert 3

Pairwise comparisons made by Expert 3 are depicted below:

CRITERIA		MORE IMPORTANT (A OR B)	INTENSITY (SCALE FROM 1 TO 9)
A	B		
0	1	A	5,00
0.1	0.2	B	3,00
0.1	0.3	B	7,00
0.2	0.3	B	5,00
0.2.1	0.2.2	B	9,00
0.2.1	0.2.3	B	3,00
0.2.2	0.2.3	A	5,00
0.3.1	0.3.2	B	9,00
0.3.1	0.3.3	B	3,00
0.3.2	0.3.3	A	3,00
1.1	1.2	B	3,00
1.1	1.3	A	3,00
1.2	1.3	A	5,00
1.1.1	1.1.2	A	9,00
1.1.2.1	1.1.2.2	B	3,00
1.2.1	1.2.2	A	7,00
1.2.1.1	1.2.1.2	B	8,00
1.2.1.1	1.2.1.3	B	2,00
1.2.1.2	1.2.1.3	A	5,00
1.2.2.1	1.2.2.2	A	3,00
1.2.2.1.1	1.2.2.1.2	B	9,00
1.2.2.1.1	1.2.2.1.3	B	3,00
1.2.2.1.2	1.2.2.1.3	A	3,00
1.2.2.2.1	1.2.2.2.2	A	1,00
1.2.2.2.1.	1.2.2.2.3	A	1,00
1.2.2.2.2	1.2.2.2.3	A	1,00
1.3.1	1.3.2	B	5,00
1.3.1.1	1.3.1.2	B	2,00

1.3.1.1	1.3.1.3	A	3,00
1.3.1.1	1.3.1.4	A	2,00
1.3.1.1	1.3.1.5	B	2,00
1.3.1.1	1.3.1.6	B	3,00
1.3.1.1	1.3.1.7	B	3,00
1.3.1.1.1	1.3.1.1.2	B	7,00
1.3.1.2	1.3.1.3	A	5,00
1.3.1.2	1.3.1.4	A	3,00
1.3.1.2	1.3.1.5	A	2,00
1.3.1.2	1.3.1.6	B	5,00
1.3.1.2	1.3.1.7	B	5,00
1.3.1.2.1	1.3.1.2.2	A	5,00
1.3.1.3	1.3.1.4	B	2,00
1.3.1.3	1.3.1.5	B	4,00
1.3.1.3	1.3.1.6	B	9,00
1.3.1.3	1.3.1.7	B	9,00
1.3.1.4	1.3.1.5	B	2,00
1.3.1.4	1.3.1.6	B	2,00
1.3.1.4	1.3.1.7	B	5,00
1.3.1.5	1.3.1.6	B	2,00
1.3.1.5	1.3.1.7	B	2,00
1.3.1.5.1	1.3.1.5.2	A	3,00
1.3.1.6	1.3.1.7	B	1,00
1.3.2.1	1.3.2.2	A	9,00
1.3.2.1	1.3.2.3	A	4,00
1.3.2.2	1.3.2.3	B	2,00
1.3.2.3.1	1.3.2.3.2	B	2,00
1.3.2.3.1	1.3.2.3.3	B	3,00
1.3.2.3.1	1.3.2.3.4	B	9,00
1.3.2.3.2	1.3.2.3.3	B	3,00
1.3.2.3.2	1.3.2.3.4	B	7,00
1.3.2.3.3	1.3.2.3.4	B	5,00

Resulting comparison matrices are presented below:

Criteria/Criteria	0	1
0	1	5
1	0,2	1
Total	1,2	6

	0	1	Priority Vector
0	0,833333	0,833333	0,83
1	0,166667	0,166667	0,17

	0.1	0.2	0.3
0.1	1	0,333333	0,1428571
0.2	3	1	0,2
0.3	7	5	1
Total	11	6,333333	1,3428571

	0.1	0.2	0.3	Priority Vector
0.1	0,090909	0,052632	0,106382979	0,08
0.2	0,272727	0,157895	0,14893617	0,19
0.3	0,636364	0,789474	0,744680851	0,72

lambda max	3,05019
consistency index (CI)	2,51%
consistency ratio (CR)	4,33%

	0.2.1	0.2.2	0.2.3
0.2.1	1	0,111111	0,3333333
0.2.2	9	1	5
0.2.3	3	0,2	1
Total	13	1,311111	6,3333333

	0.2.1	0.2.2	0.2.3	Priority Vector
0.2.1	0,076923	0,084746	0,052631579	0,07
0.2.2	0,692308	0,762712	0,789473684	0,75
0.2.3	0,230769	0,152542	0,157894737	0,18

lambda max	3,033333
consistency index (CI)	1,67%
consistency ratio (CR)	2,87%

	0.3.1	0.3.2	0.3.3
0.3.1	1	0,111111	0,3333333
0.3.2	9	1	3
0.3.3	3	0,333333	1
Total	13	1,444444	4,3333333

	0.3.1	0.3.2	0.3.3	Priority Vector
0.3.1	0,076923	0,076923	0,076923077	0,08
0.3.2	0,692308	0,692308	0,692307692	0,69
0.3.3	0,230769	0,230769	0,230769231	0,23

lambda max	3,033333
consistency index (CI)	1,67%
consistency ratio (CR)	2,87%

	1.1	1.2	1.3
1.1	1	0,333333	3
1.2	3	1	5
1.3	0,333333	0,2	1
Total	4,333333	1,533333	9

	1.1	1.2	1.3	Priority Vector
1.1	0,230769	0,217391	0,333333333	0,26
1.2	0,692308	0,652174	0,555555556	0,63
1.3	0,076923	0,130435	0,111111111	0,11

lambda max	3,082667
consistency index (CI)	4,13%
consistency ratio (CR)	7,13%

	1.1.1	1.1.2
1.1.1	1	9
1.1.2	0,111111	1
Total	1,111111	10

	1.1.1	1.1.2	Priority Vector
1.1.1	0,9	0,9	0,9
1.1.2	0,1	0,1	0,1

	1.1.2.1	1.1.2.2
1.1.2.1	1	0,333333
1.1.2.2	3	1
Total	4	1,333333

	1.1.2.1	1.1.2.2	Priority Vector
1.1.2.1	0,25	0,25	0,25
1.1.2.2	0,75	0,75	0,75

	1.2.1	1.2.2
1.2.1	1	7
1.2.2	0,142857	1
Total	1,142857	8

	1.2.1	1.2.2	Priority Vector
1.2.1	0,875	0,875	0,88
1.2.2	0,125	0,125	0,13

	1.2.1.1	1.2.1.2	1.2.1.3
1.2.1.1	1	0,125	0,5
1.2.1.2	8	1	5
1.2.1.3	2	0,2	1
Total	11	1,325	6,5

	1.2.1.1	1.2.1.2	1.2.1.3	Priority Vector
1.2.1.1	0,090909	0,09434	0,076923077	0,09
1.2.1.2	0,727273	0,754717	0,769230769	0,75
1.2.1.3	0,181818	0,150943	0,153846154	0,16

lambda max	3,02375
consistency index (CI)	1,19%
consistency ratio (CR)	2,05%

	1.2.2.1	1.2.2.2
1.2.2.1	1	3
1.2.2.2	0,333333	1
Total	1,333333	4

	1.2.2.1	1.2.2.2	Priority Vector
1.2.2.1	0,75	0,75	0,75
1.2.2.2	0,25	0,25	0,25

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3
1.2.2.1.1	1	0,111111	0,3333333
1.2.2.1.2	9	1	3
1.2.2.1.3	3	0,333333	1
Total	13	1,444444	4,3333333

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3	Priority Vector
1.2.2.1.1	0,076923	0,076923	0,076923077	0,08
1.2.2.1.2	0,692308	0,692308	0,692307692	0,69
1.2.2.1.3	0,230769	0,230769	0,230769231	0,23

lambda max	3,033333
consistency index (CI)	1,67%
consistency ratio (CR)	2,87%

	1.3.1	1.3.2
1.3.1	1	0,2
1.3.2	5	1
Total	6	1,2

	1.3.1	1.3.2	Priority Vector
1.3.1	0,166667	0,166667	0,17
1.3.2	0,833333	0,833333	0,83

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7
1.3.1.1	1	0,5	3	2	0,5	0,333333	0,333333
1.3.1.2	2	1	5	3	2	0,2	0,2
1.3.1.3	0,333333	0,2	1	0,5	0,25	0,111111	0,111111
1.3.1.4	0,5	0,333333	2	1	0,5	0,5	0,2
1.3.1.5	2	0,5	4	2	1	0,5	0,5
1.3.1.6	3	5	9	2	2	1	1
1.3.1.7	3	5	9	5	2	1	1
Total	11,83333	12,53333	33	15,5	8,25	3,644444	3,344444

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7	Priority Vector
1.3.1.1	0,084507	0,039894	0,0909091	0,129032	0,060606	0,091463	0,099668	0,09
1.3.1.2	0,169014	0,079787	0,1515152	0,193548	0,242424	0,054878	0,059801	0,14
1.3.1.3	0,028169	0,015957	0,030303	0,032258	0,030303	0,030488	0,033223	0,03
1.3.1.4	0,042254	0,026596	0,0606061	0,064516	0,060606	0,137195	0,059801	0,06
1.3.1.5	0,169014	0,039894	0,1212121	0,129032	0,121212	0,137195	0,149502	0,12
1.3.1.6	0,253521	0,398936	0,2727273	0,129032	0,242424	0,27439	0,299003	0,27
1.3.1.7	0,253521	0,398936	0,2727273	0,322581	0,242424	0,27439	0,299003	0,29

lambda max	7,683556
consistency index (CI	11,39%
consistency ratio (CR)	8,63%

	1.3.1.1.1	1.3.1.1.2
1.3.1.1.1	1	0,142857
1.3.1.1.2	7	1
Total	8	1,142857

	1.3.1.1.1	1.3.1.1.2	Priority Vector
1.3.1.1.1	0,125	0,125	0,13
1.3.1.1.2	0,875	0,875	0,88

	1.3.1.2.1	1.3.1.2.2
1.3.1.2.1	1	5
1.3.1.2.2	0,2	1
Total	1,2	6

	1.3.1.2.1	1.3.1.2.2	Priority Vector
1.3.1.2.1	0,833333	0,833333	0,83
1.3.1.2.2	0,166667	0,166667	0,17

	1.3.1.5.1	1.3.1.5.2
1.3.1.5.1	1	3
1.3.1.5.2	0,333333	1
Total	1,333333	4

	1.3.1.5.1	1.3.1.5.2	Priority Vector
1.3.1.5.1	0,75	0,75	0,75
1.3.1.5.2	0,25	0,25	0,25

	1.3.2.1	1.3.2.2	1.3.2.3
1.3.2.1	1	9	4
1.3.2.2	0,111111	1	0,5
1.3.2.3	0,25	2	1
Total	1,361111	12	5,5

	1.3.2.1	1.3.2.2	1.3.2.3	Priority Vector
1.3.2.1	0,734694	0,75	0,727272727	0,74
1.3.2.2	0,081633	0,083333	0,090909091	0,09
1.3.2.3	0,183673	0,166667	0,181818182	0,18

lambda max	3,077222
consistency index (CI)	3,86%
consistency ratio (CR)	6,66%

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4
1.3.2.3.1	1	0,5	0,333333333	0,111111111
1.3.2.3.2	2	1	0,333333333	0,142857143
1.3.2.3.3	3	3	1	0,2
1.3.2.3.4	9	7	5	1
Total	15	11,5	6,666666667	1,453968254

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	Priority Vector
1.3.2.3.1	0,066667	0,043478261	0,05	0,076419	0,06
1.3.2.3.2	0,133333	0,086956522	0,05	0,098253	0,09
1.3.2.3.3	0,2	0,260869565	0,15	0,137555	0,19
1.3.2.3.4	0,6	0,608695652	0,75	0,687773	0,66

lambda max	4,161286
consistency index (CI)	5,38%
consistency ratio (CR)	5,97%

Expert 4

Pairwise comparisons made by Expert 4 are depicted below:

CRITERIA		MORE IMPORTANT (A OR B)	INTENSITY (SCALE FROM 1 TO 9)
A	**B**		
0	1	B	6
0.1	0.2	B	5
0.1	0.3	A	2
0.2	0.3	A	8
0.2.1	0.2.2	B	9
0.2.1	0.2.3	B	2
0.2.2	0.2.3	A	9
0.3.1	0.3.2	B	5
0.3.1	0.3.3	B	2
0.3.2	0.3.3	A	5
1.1	1.2	B	2
1.1	1.3	A	4
1.2	1.3	A	8
1.1.1	1.1.2	B	7
1.1.2.1	1.1.2.2	A	3
1.2.1	1.2.2	A	9
1.2.1.1	1.2.1.2	B	4

1.2.1.1	1.2.1.3	A	4
1.2.1.2	1.2.1.3	A	6
1.2.2.1	1.2.2.2	A	3
1.2.2.1.1	1.2.2.1.2	A	2
1.2.2.1.1	1.2.2.1.3	A	2
1.2.2.1.2	1.2.2.1.3	A	1
1.2.2.2.1	1.2.2.2.2	A	1
1.2.2.2.1.	1.2.2.2.3	A	1
1.2.2.2.2	1.2.2.2.3	A	1
1.3.1	1.3.2	B	9
1.3.1.1	1.3.1.2	A	2
1.3.1.1	1.3.1.3	A	1
1.3.1.1	1.3.1.4	B	3
1.3.1.1	1.3.1.5	A	2
1.3.1.1	1.3.1.6	B	3
1.3.1.1	1.3.1.7	B	3
1.3.1.1.1	1.3.1.1.2	A	1
1.3.1.2	1.3.1.3	A	1
1.3.1.2	1.3.1.4	B	3
1.3.1.2	1.3.1.5	B	5
1.3.1.2	1.3.1.6	B	3
1.3.1.2	1.3.1.7	B	3
1.3.1.2.1	1.3.1.2.2	B	1
1.3.1.3	1.3.1.4	B	3
1.3.1.3	1.3.1.5	B	5
1.3.1.3	1.3.1.6	B	2
1.3.1.3	1.3.1.7	B	3
1.3.1.4	1.3.1.5	B	2
1.3.1.4	1.3.1.6	B	3
1.3.1.4	1.3.1.7	B	3
1.3.1.5	1.3.1.6	B	3
1.3.1.5	1.3.1.7	B	3
1.3.1.5.1	1.3.1.5.2	A	5
1.3.1.6	1.3.1.7	A	1
1.3.2.1	1.3.2.2	A	9
1.3.2.1	1.3.2.3	A	4
1.3.2.2	1.3.2.3	B	5
1.3.2.3.1	1.3.2.3.2	B	2
1.3.2.3.1	1.3.2.3.3	B	5
1.3.2.3.1	1.3.2.3.4	B	9
1.3.2.3.2	1.3.2.3.3	B	2
1.3.2.3.2	1.3.2.3.4	B	8
1.3.2.3.3	1.3.2.3.4	B	7

Resulting comparison matrices are presented below:

Criteria/Criteria	0	1
0	1	0,166667
1	6	1
Total	7	1,166667

	0	1	Priority Vector
0	0,142857	0,142857	0,14
1	0,857143	0,857143	0,86

	0.1	0.2	0.3
0.1	1	0,2	2
0.2	5	1	7,73
0.3	0,5	0,129366	1
Total	6,5	1,329366	10,73

	0.1	0.2	0.3	Priority Vector
0.1	0,153846	0,150448	0,18639329	0,16
0.2	0,769231	0,752238	0,720410065	0,75
0.3	0,076923	0,097314	0,093196645	0,09

lambda max	3,0027
consistency index (CI)	0,14%
consistency ratio (CR)	0,23%

	0.2.1	0.2.2	0.2.3
0.2.1	1	0,111111	0,5
0.2.2	9	1	9
0.2.3	2	0,111111	1
Total	12	1,222222	10,5

	0.2.1	0.2.2	0.2.3	Priority Vector
0.2.1	0,083333	0,090909	0,047619048	0,07
0.2.2	0,75	0,818182	0,857142857	0,81
0.2.3	0,166667	0,090909	0,095238095	0,12

lambda max	3,09
consistency index (CI)	4,50%
consistency ratio (CR)	7,76%

	0.3.1	0.3.2	0.3.3
0.3.1	1	0,2	0,4545455
0.3.2	5	1	5
0.3.3	2,2	0,2	1
Total	8,2	1,4	6,4545455

	0.3.1	0.3.2	0.3.3	Priority Vector
0.3.1	0,121951	0,142857	0,070422535	0,11
0.3.2	0,609756	0,714286	0,774647887	0,7
0.3.3	0,268293	0,142857	0,154929577	0,19

lambda max	3,1084
consistency index (CI)	5,42%
consistency ratio (CR)	9,34%

	1.1	1.2	1.3
1.1	1	0,5	4
1.2	2	1	8
1.3	0,25	0,125	1
Total	3,25	1,625	13

	1.1	1.2	1.3	Priority Vector
1.1	0,307692	0,307692	0,307692308	0,31
1.2	0,615385	0,615385	0,615384615	0,62
1.3	0,076923	0,076923	0,076923077	0,08

lambda max	3,055
consistency index (CI)	2,75%
consistency ratio (CR)	4,74%

	1.1.1	1.1.2
1.1.1	1	0,142857
1.1.2	7	1
Total	8	1,142857

	1.1.1	1.1.2	Priority Vector
1.1.1	0,125	0,125	0,13
1.1.2	0,875	0,875	0,88

	1.1.2.1	1.1.2.2
1.1.2.1	1	3
1.1.2.2	0,333333	1
Total	1,333333	4

	1.1.2.1	1.1.2.2	Priority Vector
1.1.2.1	0,75	0,75	0,75
1.1.2.2	0,25	0,25	0,25

	1.2.1	1.2.2
1.2.1	1	9
1.2.2	0,111111	1
Total	1,111111	10

	1.2.1	1.2.2	Priority Vector
1.2.1	0,9	0,9	0,9
1.2.2	0,1	0,1	0,1

	1.2.1.1	1.2.1.2	1.2.1.3
1.2.1.1	1	0,25	3,7
1.2.1.2	4	1	6,3
1.2.1.3	0,27027	0,15873	1
Total	5,27027	1,40873	11

	1.2.1.1	1.2.1.2	1.2.1.3	Priority Vector
1.2.1.1	0,189744	0,177465	0,336363636	0,23
1.2.1.2	0,758974	0,709859	0,572727273	0,68
1.2.1.3	0,051282	0,112676	0,090909091	0,08

lambda max	3,0501
consistency index (CI)	2,50%
consistency ratio (CR)	4,32%

	1.2.2.1	1.2.2.2
1.2.2.1	1	3
1.2.2.2	0,333333	1
Total	1,333333	4

	1.2.2.1	1.2.2.2	Priority Vector
1.2.2.1	0,75	0,75	0,75
1.2.2.2	0,25	0,25	0,25

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3
1.2.2.1.1	1	1,5	2
1.2.2.1.2	0,666667	1	1
1.2.2.1.3	0,5	1	1
Total	2,166667	3,5	4

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3	Priority Vector
1.2.2.1.1	0,461538	0,428571	0,5	0,46
1.2.2.1.2	0,307692	0,285714	0,25	0,28
1.2.2.1.3	0,230769	0,285714	0,25	0,26

lambda max	3,0167
consistency index (CI)	0,83%
consistency ratio (CR)	1,44%

	1.3.1	1.3.2
1.3.1	1	0,111111
1.3.2	9	1
Total	10	1,111111

	1.3.1	1.3.2	Priority Vector
1.3.1	0,1	0,1	0,1
1.3.2	0,9	0,9	0,9

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7
1.3.1.1	1	1,6	1	0,333333333	2	0,333333	0,333333
1.3.1.2	0,625	1	1	0,333333333	0,2	0,333333	0,333333
1.3.1.3	1	1	1	0,333333333	0,2	0,5	0,4
1.3.1.4	3	3	3	1	0,5	0,333333	0,333333
1.3.1.5	0,5	5	5	2	1	0,333333	0,333333
1.3.1.6	3	3	2	3	3	1	1
1.3.1.7	3	3	2,5	3	3	1	1
Total	12,125	17,6	15,5	10	9,9	3,833333	3,733333

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7	Priority Vector
1.3.1.1	0,082474	0,090909	0,064516	0,0333	0,20202	0,086957	0,089286	0,09
1.3.1.2	0,051546	0,056818	0,064516	0,0333	0,020202	0,086957	0,089286	0,06
1.3.1.3	0,082474	0,056818	0,064516	0,0333	0,020202	0,130435	0,107143	0,07
1.3.1.4	0,247423	0,170455	0,193548	0,1	0,050505	0,086957	0,089286	0,13
1.3.1.5	0,041237	0,284091	0,322581	0,2	0,10101	0,086957	0,089286	0,16
1.3.1.6	0,247423	0,170455	0,129032	0,3	0,30303	0,26087	0,267857	0,24
1.3.1.7	0,247423	0,170455	0,16129	0,3	0,30303	0,26087	0,267857	0,24

lambda max	7,93225
consistency index (CI)	15,54%
consistency ratio (CR)	11,77%

	1.3.1.1.1	1.3.1.1.2
1.3.1.1.1	1	1
1.3.1.1.2	1	1
Total	2	2

	1.3.1.1.1	1.3.1.1.2	Priority Vector
1.3.1.1.1	0,5	0,5	0,5
1.3.1.1.2	0,5	0,5	0,5

	1.3.1.2.1	1.3.1.2.2
1.3.1.2.1	1	0,714286
1.3.1.2.2	1,4	1
Total	2,4	1,714286

	1.3.1.2.1	1.3.1.2.2	Priority Vector
1.3.1.2.1	0,416667	0,416667	0,42
1.3.1.2.2	0,583333	0,583333	0,58

	1.3.1.5.1	1.3.1.5.2
1.3.1.5.1	1	5
1.3.1.5.2	0,2	1
Total	1,2	6

	1.3.1.5.1	1.3.1.5.2	Priority Vector
1.3.1.5.1	0,833333	0,833333	0,83
1.3.1.5.2	0,166667	0,166667	0,17

	1.3.2.1	1.3.2.2	1.3.2.3
1.3.2.1	1	9	4
1.3.2.2	0,111111	1	0,2
1.3.2.3	0,25	5	1
Total	1,361111	15	5,2

	1.3.2.1	1.3.2.2	1.3.2.3	Priority Vector
1.3.2.1	0,734694	0,6	0,769230769	0,7
1.3.2.2	0,081633	0,066667	0,038461538	0,06
1.3.2.3	0,183673	0,333333	0,192307692	0,24

lambda max	3,1008
consistency index (CI)	5,04%
consistency ratio (CR)	8,69%

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4
1.3.2.3.1	1	0,5	0,2	0,111111111
1.3.2.3.2	2	1	0,588235294	0,128205128
1.3.2.3.3	5	1,7	1	0,145560408
1.3.2.3.4	9	7,8	6,87	1
Total	17	11	8,658235294	1,384876647

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	Priority Vector
1.3.2.3.1	0,058824	0,045454545	0,023099395	0,080232	0,05
1.3.2.3.2	0,117647	0,090909091	0,067939398	0,092575	0,09
1.3.2.3.3	0,294118	0,154545455	0,115496977	0,105107	0,17
1.3.2.3.4	0,529412	0,709090909	0,79346423	0,722086	0,69

lambda max	4,267465
consistency index (CI)	8,92%
consistency ratio (CR)	9,91%

Expert 5

Pairwise comparisons made by Expert 5 are depicted below:

CRITERIA		MORE IMPORTANT (A OR B)	INTENSITY (SCALE FROM 1 TO 9)
A	**B**		

0	1	B	1,75
0.1	0.2	B	5,00
0.1	0.3	B	9,00
0.2	0.3	B	3,00
0.2.1	0.2.2	B	3,00
0.2.1	0.2.3	B	5,00
0.2.2	0.2.3	B	1,70
0.3.1	0.3.2	B	3,00
0.3.1	0.3.3	B	5,00
0.3.2	0.3.3	B	1,70
1.1	1.2	B	5,00
1.1	1.3	B	3,00
1.2	1.3	A	2,00
1.1.1	1.1.2	B	3,00
1.1.2.1	1.1.2.2	A	4,00
1.2.1	1.2.2	A	7,00
1.2.1.1	1.2.1.2	B	3,00
1.2.1.1	1.2.1.3	B	5,00
1.2.1.2	1.2.1.3	B	1,70
1.2.2.1	1.2.2.2	A	7,00
1.2.2.1.1	1.2.2.1.2	B	3,00
1.2.2.1.1	1.2.2.1.3	B	5,00
1.2.2.1.2	1.2.2.1.3	B	1,70
1.2.2.2.1	1.2.2.2.2	B	3,00
1.2.2.2.1.	1.2.2.2.3	B	5,00
1.2.2.2.2	1.2.2.2.3	B	1,70
1.3.1	1.3.2	B	7,00
1.3.1.1	1.3.1.2	B	2,00
1.3.1.1	1.3.1.3	A	4,00
1.3.1.1	1.3.1.4	A	4,00
1.3.1.1	1.3.1.5	B	3,00
1.3.1.1	1.3.1.6	B	2,00
1.3.1.1	1.3.1.7	B	2,50
1.3.1.1.1	1.3.1.1.2	A	9,00
1.3.1.2	1.3.1.3	A	7,00
1.3.1.2	1.3.1.4	A	5,00
1.3.1.2	1.3.1.5	A	3,00
1.3.1.2	1.3.1.6	B	1,40
1.3.1.2	1.3.1.7	B	1,80
1.3.1.2.1	1.3.1.2.2	A	9,00
1.3.1.3	1.3.1.4	B	1,30
1.3.1.3	1.3.1.5	B	6,00
1.3.1.3	1.3.1.6	B	7,00
1.3.1.3	1.3.1.7	B	8,00
1.3.1.4	1.3.1.5	B	5,00
1.3.1.4	1.3.1.6	B	5,50

1.3.1.4	1.3.1.7	B	7,50
1.3.1.5	1.3.1.6	B	2,00
1.3.1.5	1.3.1.7	B	2,00
1.3.1.5.1	1.3.1.5.2	A	9,00
1.3.1.6	1.3.1.7	B	3,00
1.3.2.1	1.3.2.2	A	5,00
1.3.2.1	1.3.2.3	A	3,00
1.3.2.2	1.3.2.3	B	2,00
1.3.2.3.1	1.3.2.3.2	B	3,00
1.3.2.3.1	1.3.2.3.3	B	4,00
1.3.2.3.1	1.3.2.3.4	B	9,00
1.3.2.3.2	1.3.2.3.3	B	3,00
1.3.2.3.2	1.3.2.3.4	B	5,00
1.3.2.3.3	1.3.2.3.4	B	4,00

Resulting comparison matrices are presented below:

Criteria/Criteria	0	1
0	1	0,571428571
1	1,75	1
Total	2,75	1,571428571

	0	1	Priority Vector
0	0,363636364	0,363636364	0,36
1	0,636363636	0,636363636	0,64

	0.1	0.2	0.3
0.1	1	0,2	0,11
0.2	5	1	0,33
0.3	9	3	1
Total	15	4,2	1,44

	0.1	0.2	0.3	Priority Vector
0.1	0,066666667	0,047619048	0,076923077	0,06
0.2	0,333333333	0,238095238	0,230769231	0,27
0.3	0,6	0,714285714	0,692307692	0,67

lambda max	3,00
consistency index (CI)	0,09%
consistency ratio (CR)	0,15%

	0.2.1	0.2.2	0.2.3
0.2.1	1	0,333333333	0,2
0.2.2	3	1	0,59
0.2.3	5	1,7	1
Total	9	3,033333333	1,79

	0.2.1	0.2.2	0.2.3	Priority Vector
0.2.1	0,111111111	0,10989011	0,111842105	0,11
0.2.2	0,333333333	0,32967033	0,328947368	0,33
0.2.3	0,555555556	0,56043956	0,559210526	0,56

lambda max	3
consistency index (CI)	0,00%
consistency ratio (CR)	0,00%

	0.3.1	0.3.2	0.3.3
0.3.1	1	0,333333333	0,2
0.3.2	3	1	0,59
0.3.3	5	1,7	1
Total	9	3,033333333	1,79

	0.3.1	0.3.2	0.3.3	Priority Vector
0.3.1	0,111111111	0,10989011	0,111842105	0,11
0.3.2	0,333333333	0,32967033	0,328947368	0,33
0.3.3	0,555555556	0,56043956	0,559210526	0,56

lambda max	3
consistency index (CI)	0,00%
consistency ratio (CR)	0,00%

	1.1	1.2	1.3
1.1	1	0,2	0,33
1.2	5	1	2
1.3	3	0,5	1
Total	9	1,7	3,33

	1.1	1.2	1.3	Priority Vector
1.1	0,111111111	0,117647059	0,1	0,11
1.2	0,555555556	0,588235294	0,6	0,58
1.3	0,333333333	0,294117647	0,3	0,31

lambda max	3,01
consistency index (CI)	0,47%
consistency ratio (CR)	0,80%

	1.1.1	1.1.2
1.1.1	1	0,333333333
1.1.2	3	1
Total	4	1,333333333

	1.1.1	1.1.2	Priority Vector
1.1.1	0,25	0,25	0,25
1.1.2	0,75	0,75	0,75

	1.1.2.1	1.1.2.2
1.1.2.1	1	4
1.1.2.2	0,25	1
Total	1,25	5

	1.1.2.1	1.1.2.2	Priority Vector
1.1.2.1	0,8	0,8	0,8
1.1.2.2	0,2	0,2	0,2

	1.2.1	1.2.2
1.2.1	1	7
1.2.2	0,142857143	1
Total	1,142857143	8

	1.2.1	1.2.2	Priority Vector
1.2.1	0,875	0,875	0,88
1.2.2	0,125	0,125	0,13

	1.2.1.1	1.2.1.2	1.2.1.3
1.2.1.1	1	0,333333333	0,2
1.2.1.2	3	1	0,588235294
1.2.1.3	5	1,7	1
Total	9	3,033333333	1,788235294

	1.2.1.1	1.2.1.2	1.2.1.3	Priority Vector
1.2.1.1	0,111111111	0,10989011	0,111842105	0,11
1.2.1.2	0,333333333	0,32967033	0,328947368	0,33
1.2.1.3	0,555555556	0,56043956	0,559210526	0,56

lambda max	3
consistency index (CI)	0,00%
consistency ratio (CR)	0,00%

	1.2.2.1	1.2.2.2
1.2.2.1	1	7
1.2.2.2	0,142857143	1
Total	1,142857143	8

	1.2.2.1	1.2.2.2	Priority Vector
1.2.2.1	0,875	0,875	0,88
1.2.2.2	0,125	0,125	0,13

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3
1.2.2.1.1	1	0,333333333	0,2
1.2.2.1.2	3	1	0,588235294
1.2.2.1.3	5	1,7	1
Total	9	3,033333333	1,788235294

	1.2.2.1.1	1.2.2.1.2	1.2.2.1.3	Priority Vector
1.2.2.1.1	0,111111111	0,10989011	0,111842105	0,11
1.2.2.1.2	0,333333333	0,32967033	0,328947368	0,33
1.2.2.1.3	0,555555556	0,56043956	0,559210526	0,56

lambda max	3
consistency index (CI)	0,00%
consistency ratio (CR)	0,00%

	1.3.1	1.3.2
1.3.1	1	0,142857143
1.3.2	7	1
Total	8	1,142857143

	1.3.1	1.3.2	Priority Vector
1.3.1	0,125	0,125	0,13
1.3.2	0,875	0,875	0,88

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7
1.3.1.1	1	0,5	4	4	0,333333333	0,5	0,4
1.3.1.2	2	1	7	5	3	0,714285714	0,555555556
1.3.1.3	0,25	0,14285714	1	0,7692308	0,166666667	0,142857143	0,125
1.3.1.4	0,25	0,2	1,3	1	0,2	0,181818182	0,133333333
1.3.1.5	3	0,33333333	6	5	1	0,5	0,5
1.3.1.6	2	1,4	7	5,5	2	1	0,333333333
1.3.1.7	2,5	1,8	8	7,5	2	3	1
Total	11	5,37619048	34,3	28,769231	8,7	6,038961039	3,047222222

	1.3.1.1	1.3.1.2	1.3.1.3	1.3.1.4	1.3.1.5	1.3.1.6	1.3.1.7	Priority Vector
1.3.1.1	0,090909091	0,093002657	0,116618076	0,139037433	0,038314176	0,082795699	0,131267092	0,1
1.3.1.2	0,181818182	0,186005314	0,204081633	0,173796791	0,344827586	0,11827957	0,182315406	0,2
1.3.1.3	0,022727273	0,026572188	0,029154519	0,026737968	0,019157088	0,023655914	0,041020966	0,03
1.3.1.4	0,022727273	0,037201063	0,037900875	0,034759358	0,022988506	0,030107527	0,043755697	0,03
1.3.1.5	0,272727273	0,062001771	0,174927114	0,173796791	0,114942529	0,082795699	0,164083865	0,15
1.3.1.6	0,181818182	0,26040744	0,204081633	0,191176471	0,229885057	0,165591398	0,109389243	0,19
1.3.1.7	0,227272727	0,334809566	0,233236152	0,260695187	0,229885057	0,496774194	0,32816773	0,3

lambda max	7,433884
consistency index (CI)	7,23%
consistency ratio (CR)	5,48%

	1.3.1.1.1	1.3.1.1.2
1.3.1.1.1	1	9
1.3.1.1.2	0,111111111	1
Total	1,111111111	10

	1.3.1.1.1	1.3.1.1.2	Priority Vector
1.3.1.1.1	0,9	0,9	0,9
1.3.1.1.2	0,1	0,1	0,1

	1.3.1.2.1	1.3.1.2.2
1.3.1.2.1	1	9
1.3.1.2.2	0,111111111	1
Total	1,111111111	10

	1.3.1.2.1	1.3.1.2.2	Priority Vector
1.3.1.2.1	0,9	0,9	0,9
1.3.1.2.2	0,1	0,1	0,1

	1.3.1.5.1	1.3.1.5.2
1.3.1.5.1	1	9
1.3.1.5.2	0,111111111	1
Total	1,111111111	10

	1.3.1.5.1	1.3.1.5.2	Priority Vector
1.3.1.5.1	0,9	0,9	0,9
1.3.1.5.2	0,1	0,1	0,1

lambda max	3
consistency index (CI)	0,00%
consistency ratio (CR)	0,00%

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4
1.3.2.3.1	1	0,33333333	0,25	0,1111111
1.3.2.3.2	3	1	0,33333	0,2
1.3.2.3.3	4	3	1	0,25
1.3.2.3.4	9	5	4	1
Total	17	9,33333333	5,58333	1,5611111

	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	Priority Vector
1.3.2.3.1	0,058823529	0,035714286	0,044776119	0,071174377	0,05
1.3.2.3.2	0,176470588	0,107142857	0,059701493	0,128113879	0,12
1.3.2.3.3	0,235294118	0,321428571	0,179104478	0,160142349	0,22
1.3.2.3.4	0,529411765	0,535714286	0,71641791	0,640569395	0,61

lambda max	4,150611111
consistency index (CI)	5,02%
consistency ratio (CR)	5,58%

Summary Statistics

The summary statistics related to the performed AHP study is as follows:

Node	Description	Arithmetic Means	Geometric Means	Min	Median	Max
0	Gwas Results	0,382	0,322	0,14	0,33	0,83
1	Biological Facts	0,618	0,543	0,17	0,67	0,86
0.1	Individual SNP	0,088	0,082	0,06	0,07	0,16
0.2	Significant Gene	0,498	0,436	0,19	0,64	0,75
0.2.1	Significant Gene - Via LD	0,084	0,081	0,06	0,07	0,11
0.2.2	Significant Gene - Via Direct	0,636	0,608	0,33	0,66	0,81
0.2.3	Significant Gene - Via Pathway	0,28	0,245	0,12	0,26	0,56
0.3	Significant Pathway Gene	0,408	0,321	0,09	0,28	0,72
0.3.1	Significant Pathway Gene - Via LD	0,094	0,091	0,06	0,11	0,11
0.3.2	Significant Pathway Gene - Via Direct	0,602	0,581	0,33	0,66	0,7
0.3.3	Significant Pathway Gene - Via Pathway	0,304	0,282	0,19	0,26	0,56
1.1	Evolutionary Conservation	0,172	0,145	0,06	0,12	0,31
1.1.1	Vertebrate	0,356	0,277	0,13	0,25	0,9
1.1.2	Mammalian	0,646	0,516	0,1	0,75	0,88
1.1.2.1	Mammalian - Significant Mouse ECR	0,66	0,608	0,25	0,75	0,83
1.1.2.2	Mammalian - Other Mammalian	0,34	0,291	0,17	0,25	0,75
1.2	Gene Association	0,562	0,544	0,32	0,62	0,66
1.2.1	Disease Gene	0,878	0,878	0,83	0,88	0,9
1.2.1.1	Disease Gene - Via LD	0,12	0,108	0,06	0,11	0,23
1.2.1.2	Disease Gene - Via Direct	0,61	0,587	0,33	0,66	0,75
1.2.1.3	Disease Gene - Via Pathway	0,268	0,221	0,08	0,26	0,56
1.2.2	Other Gene	0,126	0,124	0,1	0,13	0,17
1.2.2.1	Other Gene - Other Disease	0,802	0,800	0,75	0,75	0,88
1.2.2.1.1	Other Gene - Other Disease - Via LD	0,164	0,122	0,06	0,11	0,46
1.2.2.1.2	Other Gene - Other Disease - Via Direct	0,518	0,484	0,28	0,63	0,69
1.2.2.1.3	Other Gene - Other Disease - Via Pathway	0,318	0,300	0,23	0,26	0,56
1.2.2.2	Other Gene - Neutral	0,202	0,192	0,13	0,25	0,25
1.2.2.2.1	Other Gene - Neutral - Via LD	0,164	0,122	0,06	0,11	0,46
1.2.2.2.2	Other Gene - Neutral - Via Direct	0,518	0,484	0,28	0,63	0,69
1.2.2.2.3	Other Gene - Neutral - Via Pathway	0,318	0,300	0,23	0,26	0,56
1.3	Genomic Location	0,266	0,210	0,08	0,27	0,56
1.3.1	Non-Coding	0,132	0,130	0,1	0,13	0,17
1.3.1.1	Non-Coding- UTR-3	0,08	0,078	0,05	0,09	0,1
1.3.1.1.1	Non-Coding- UTR-3 - MiRNA Prediction	0,658	0,539	0,13	0,88	0,9
1.3.1.1.2	Non-Coding- UTR-3 - No MiRNA Prediction	0,348	0,237	0,1	0,13	0,88
1.3.1.2	Non-Coding- UTR-5	0,122	0,113	0,06	0,11	0,2
1.3.1.2.1	Non-Coding- UTR-5 - CpG Island	0,762	0,736	0,42	0,83	0,9
1.3.1.2.2	Non-Coding- UTR-5 - No CpG Island	0,238	0,195	0,1	0,17	0,58
1.3.1.3	Non-Coding - Intronic	0,038	0,036	0,03	0,03	0,07
1.3.1.4	Non-Coding - Near Gene 3	0,068	0,060	0,03	0,06	0,13
1.3.1.5	Non-Coding - Near Gene 5	0,14	0,139	0,12	0,15	0,16
1.3.1.5.1	Non-Coding - Near Gene 5 - CpG Island	0,812	0,810	0,75	0,83	0,9
1.3.1.5.2	Non-Coding - Near Gene 5 - No CpG Island	0,188	0,178	0,1	0,17	0,25
1.3.1.6	Non-Coding - Splice3	0,244	0,241	0,19	0,24	0,3
1.3.1.7	Non-Coding - Splice 5	0,304	0,300	0,24	0,3	0,39
1.3.2	Coding	0,874	0,874	0,83	0,88	0,9
1.3.2.1	Coding - Frameshift	0,722	0,719	0,6	0,74	0,8
1.3.2.3	Coding - CDS Non Syn	0,278	0,271	0,2	0,26	0,4
1.3.2.3.1	Coding - CDS Non Syn - Polyphen Benign	0,05	0,050	0,04	0,05	0,06
1.3.2.3.2	Coding - CDS Non Syn - Possibly Damaging	0,118	0,115	0,09	0,12	0,15
1.3.2.3.3	Coding - CDS Non Syn - Probably Damaging	0,23	0,225	0,17	0,22	0,3
1.3.2.3.4	Coding - CDS Non Syn - Completely Determi	0,602	0,598	0,5	0,61	0,69

APPENDIX E: MATHEMATICAL AND STATISTICAL BACKGROUND FOR SNP - COMPLEX DISEASE ASSOCIATION ANALYSIS

Hardy-Weinberg Equilibrium

Suppose we have a marker with alleles 1,…,*k* having frequencies $p_1,\ldots,p_k$. We may write the genotype count for alleles *i* and *j* as n_{ij}. Due to phase ambiguity, if $i \neq j$, we count occurrences of allele *i* on the first chromosome and allele *j* on the second chromosome, along with occurrences of allele *j* on the first chromosome and allele *i* on the second chromosome in both the notations n_{ij} and n_{ji}.

Thus, we may write the count for allele *i* as $n_i = 2n_{ii} + \sum_{j=1,j\neq i}^{k} n_{ij}$. We may also express the genotype frequency for allele *i* occurring homozygously as $p_{ii} = \frac{n_{ii}}{n}$ and the genotype frequency for heterozygous alleles *i* and *j* as $p_{iJ} = \frac{n_{iJ}}{n}$, where *n* is the population count. The frequency of allele *i* may be expressed as $p_i = \frac{n_i}{2n} = p_{ii} + \frac{1}{2}\sum_{j=1,j\neq i}^{k} p_{ij}$.

We wish to check the agreement of p_{ii} with p_i^2 and the agreement of p_{ij}, where $i \neq j$, with $2p_ip_j$. We multiply by two because of how we deal with the phase ambiguity (see above). Thus, we will define the Hardy-Weinberg equilibrium coefficient D_{ii} or D_{ij} for alleles *i* and *j* such that

$p_{ii} = p_i^2 + D_{ii}$

$p_{ij} = 2p_ip_j - 2D_{ij}$ (for $i \neq j$).

(It may be shown that for a bi-allelic marker, $D_{11} = D_{12} = D_{22}$.)

Call Rate

The call rate is the fraction of genotypes present and not missing for the given marker:

$$call\ rate = \frac{number\ of\ complete\ and\ non-missing\ genotypes}{total\ number\ of\ genotypes}$$

Minor Allele Frequency (MAF)

The minor allele frequency is the fraction of the total alleles of the given marker that are minor alleles:

$$MAF = \frac{minor\ allele\ count}{total\ allele\ count}$$

Chi-Squared Test (Pearson)

This is the most-often used way to obtain a p-value for (the extremeness of) an (unordered) $m \times n$ contingency table, to know whether to reject the null hypothesis that the proportions in the rows and columns of the table differ from the proportions of the margin column totals and the margin row totals, respectively as much as they do by chance alone. If the contingency table with elements x_{ij} has N observations, we make an "expected" contingency table based on the marginal totals as $e_{ij} = \frac{r_i c_j}{N}$. We then obtain a p-value from the fact that

$$\chi^2 = \sum \frac{\left(x_{ij} - e_{ij}\right)^2}{e_{ij}}$$

approximates a chi-squared distribution with $(m - 1)(n - 1)$ degrees of freedom.

Fisher's Exact Test

The output of this test is the sum of the probabilities of all contingency tables whose marginal sums are the same as those of the observed contingency table and which are as extreme or more extreme (equally probable or less probable) than the observed contingency table. The probability of a $2 \times r$ contingency table with elements x_{rc} and row totals r_c and column totals c_r and N elements is given by

$$p = \frac{(r_1!\ r_2!\ )(c_1!\, c_2! \ldots c_r!)}{x_{11}!\, x_{12}! \ldots x_{2r}!\, N!}$$

Odds Ratio

For the purposes of this method's description, we define a 2×2 contingency table as being organized as "(Case/Control) vs. (Yes/No)" demonstrated in the table below.

	Yes	No	Total
Case	y_{case}	n_{case}	$y_{case} + n_{case}$
Control	$y_{control}$	$n_{control}$	$y_{control} + n_{control}$
Total	$y_{case} + y_{control}$	$n_{case} + n_{control}$	N

The odds ratio is defined as the ratio of the odds for "Case" among the Yes's to the odds for "Case" among the No's, or equivalently the ratio of the odds for "Yes" among the cases to the odds for "Yes" among the controls, or equivalently

$$OR = \frac{y_{case} n_{control}}{n_{case} y_{control}}$$

False Discovery Rate

When testing multiple hypotheses, there is always the possibility one or more tests have appeared significant just by chance. Various techniques have been proposed to adjust the p-values or to otherwise correct for multiple testing issues. Among these are the Bonferroni adjustment and the False Discovery Rate.

Suppose that *m* hypotheses are tested, and *R* of them is rejected (positive results). Of the rejected hypotheses, suppose that *V* of them are really false positive results, that is *V* is the number of type I errors. The False Discovery Rate is defined as $FDR = E\left(\frac{V}{R}\middle|R > 0\right)\Pr(R > 0)$, that is, the expected proportion of false positive findings among all rejected hypotheses times the probability of making at least one rejection.

Suppose we are rejecting (the null hypothesis) on the basis of the *p*-values $p_1,\ldots,p_m$ from these *m* tests, specifically, when a *p*-value is less than a parameter γ. If we can treat the *p*-values as being independent, then we can estimate $\Pr(p \leq \gamma)$ as $\widehat{\Pr}\,(P \leq \gamma) = \frac{\max(R(\gamma),1)}{\mathrm{m}}$ where $R(\gamma)$ is the number of p_i less than or equal to γ, and use this to estimate the False Discovery Rate FDR as $\widehat{FDR}(\gamma) = \frac{\gamma}{\widehat{\Pr}(\mathrm{P} \leq \gamma)}$

When this is computed for γ equal to any particular p-value, these expressions simplify to $\widehat{\Pr}\,(P \leq \gamma) = \frac{\mathrm{R}(\gamma)}{\mathrm{m}}$ and $\widehat{FDR}(\gamma) = \frac{m\gamma}{\mathrm{R}(\gamma)} = \frac{m\gamma}{\mathrm{j}}$ where *j* is the number of *p*-values less than or equal to γ.

APPENDIX F: METU-SNP BASED ANALYSIS

Alzheimer's Disease

AHP Scoring Details

								GWAS RELATED										
RSID\Score	0,033616	0,01598	0,12099	0,053266	0,01465	0,093825	0,04738	0,036593	0,186016	0,081725	0,005756	0,01818	0,011161	0,00145	0,004579	0,002811	TOTAL
10808738	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
4395923	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
4936637	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
2070045	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
4652769	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
4651138	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
6424883	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
10752893	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,727978
2966952	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
17561	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1606659	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2280294	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1986181	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
9881879	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1010158	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2830052	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1800464	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1532268	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
4486246	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
3779870	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549

	GENERIC																		
	1.1.1	1.1.2.1	1.1.2.2	1.3.1.1.1	1.3.1.1.2	1.3.1.2.1	1.3.1.2.2	1.3.1.3	1.3.1.4	1.3.1.5.1	1.3.1.5.2	1.3.1.6	1.3.1.7	1.3.2.1	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	
RSID\Score	0,037841	0,04532	0,02335	0,001142	0,0006	0,002017	0,00063	0,000825	0,001476	0,002467	0,000571	0,0053	0,006597	0,10373	0,001997	0,004713	0,009187	0,024045	TOTAL
3779870	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,083986
1800464	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,063185
2070045	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,063185
4651138	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,062013
1606659	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,038666
17561	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0,02806
1532268	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0,02806
2966952	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0,02806
1010158	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
1986181	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
2280294	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
2830052	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
4486246	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
4652769	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
9881879	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,024172
4395923	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
4936637	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
6424883	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
10752893	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
10808738	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825

5-fold Cross Validation Performance of AHP Based List of SNPs

Cross validate train data----------

Correctly Classified Instances	228	97.8541 %
Incorrectly Classified Instances	5	2.1459 %
Kappa statistic	0.9569	
Mean absolute error	0.0215	
Root mean squared error	0.1465	
Relative absolute error	4.2979 %	
Root relative squared error	29.3177 %	
Total Number of Instances	233	

=== Confusion Matrix ===

a	b	<-- classified as
121	0	a = control
5	107	b = case

Train and test ----------

Correctly Classified Instances	28	57.1429 %
Incorrectly Classified Instances	21	42.8571 %
Kappa statistic	0.1517	
Mean absolute error	0.4302	
Root mean squared error	0.6499	
Relative absolute error	86.9175 %	
Root relative squared error	130.6521 %	
Total Number of Instances	49	

=== Confusion Matrix ===

a	b	<-- classified as
14	13	a = control
8	14	b = case

5-fold Cross Validation Performance of SPOT Based List of SNPs

Cross validate train data----------

Correctly Classified Instances	229	98.2833 %
Incorrectly Classified Instances	4	1.7167 %
Kappa statistic	0.9656	
Mean absolute error	0.0172	
Root mean squared error	0.131	
Relative absolute error	3.4383 %	
Root relative squared error	26.2225 %	
Total Number of Instances	233	

=== Confusion Matrix ===

a	b	<-- classified as
121	0	a = control
4	108	b = case

Train and test ----------

Correctly Classified Instances	24	48.9796 %
Incorrectly Classified Instances	25	51.0204 %
Kappa statistic	-0.0099	
Mean absolute error	0.5007	
Root mean squared error	0.7025	
Relative absolute error	101.1489 %	
Root relative squared error	141.236 %	
Total Number of Instances	49	

=== Confusion Matrix ===

a	b	<-- classified as
12	15	a = control
10	12	b = case

Rheumatoid Arthritis

AHP Scoring Details

							GWAS RELATED								
RSID\Score	0.1	0.2.1	0.2.2	0.2.3	0.3.1	0.3.2	0.3.3	1.2.1.1	1.2.1.2	1.2.1.3	1.2.1.1.1	1.2.1.1.2	1.2.1.1.3	1.2.2.1.1	1.2.2.2.2	1.2.2.2.3	TOTAL
	0,033616	0,01398	0,12099	0,053266	0,01465	0,093825	0,04738	0,036593	0,186016	0,081725	0,005756	0,01818	0,011161	0,00145	0,004579	0,002811	
2070600	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2074488	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2256175	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
1063355	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
3134940	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
3093662	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
3134943	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2256028	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,653549
2395471	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,619933
9264536	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,619933
1051794	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0,619933
1468673	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0,557392
10818500	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0,557392
7037673	1	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0,532559
2476601	1	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1	0,506458
17611	0	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0,498943
1041981	0	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0,498943
2075800	1	0	1	1	0	1	1	0	0	1	0	1	1	0	1	1	0,467533
7383287	1	0	1	1	0	1	1	0	0	1	0	1	1	0	1	1	0,467533
2227956	1	0	1	1	0	1	1	0	0	1	0	1	1	0	1	1	0,467533

	GENERIC																	
	1.1.1	1.1.2.1	1.1.2.2	1.3.1.1.1	1.3.1.1.2	1.3.1.2.1	1.3.1.2.2	1.3.1.3	1.3.1.4	1.3.1.5.1	1.3.1.5.2	1.3.1.6	1.3.1.7	1.3.2.1	1.3.2.3.1	1.3.2.3.2	1.3.2.3.3	1.3.2.3.4	
RSID\Score	0,037841	0,04532	0,023347	0,001142	0,000604	0,002017	0,00063	0,000825	0,001476	0,002467	0,000571	0,005295	0,006597	0,103733	0,001997	0,004713	0,009187	0,024045	TOTAL
2075800	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0,087874
2227956	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,085158
7383287	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,083161
2476601	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0,067898
2070600	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0,067377
17611	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,063185
1041981	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,063185
2256175	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0,002301
3134943	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0,002301
9264536	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0,001476
1468673	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
2256028	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
3093662	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
3134940	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
7037673	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
10818500	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0,000825
1051794	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1063355	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2074488	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2395471	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5-fold Cross Validation Performance of AHP Based List of SNPs

Cross validate train data----------

Correctly Classified Instances	1162	93.8611 %
Incorrectly Classified Instances	76	6.1389 %
Kappa statistic	0.875	
Mean absolute error	0.0617	
Root mean squared error	0.2469	
Relative absolute error	12.6496 %	
Root relative squared error	50.0004 %	
Total Number of Instances	1238	

=== Confusion Matrix ===

a	b	<-- classified as
664	52	a = control
24	498	b = case

Train and test ----------

Correctly Classified Instances	324	78.6408 %
Incorrectly Classified Instances	88	21.3592 %
Kappa statistic	0.5637	
Mean absolute error	0.2146	
Root mean squared error	0.4592	
Relative absolute error	43.6164 %	
Root relative squared error	92.5865 %	
Total Number of Instances	412	

=== Confusion Matrix ===

a	b	<-- classified as
192	40	a = control
48	132	b = case

5-fold Cross Validation Performance of SPOT Based List of SNPs

Cross validate train data----------

Correctly Classified Instances	1155	93.2956 %
Incorrectly Classified Instances	83	6.7044 %
Kappa statistic	0.8636	
Mean absolute error	0.0661	
Root mean squared error	0.2533	
Relative absolute error	13.5493 %	
Root relative squared error	51.2996 %	
Total Number of Instances	1238	

=== Confusion Matrix ===

a	b	<-- classified as
660	56	a = control
27	495	b = case

Train and test ----------

Correctly Classified Instances	321	77.9126 %
Incorrectly Classified Instances	91	22.0874 %
Kappa statistic	0.5486	
Mean absolute error	0.2188	
Root mean squared error	0.4638	
Relative absolute error	44.4693 %	
Root relative squared error	93.5058 %	
Total Number of Instances	412	

=== Confusion Matrix ===

a	b	<-- classified as
191	41	a = control
50	130	b = case

VITA

PERSONAL INFORMATION

Surname, Name: Üstünkar, Gürkan
Nationality: Turkish (TC)
Date and Place of Birth: 18 June 1980 , İzmir
Marital Status: Single
Phone: +90 232 360 11 52, +90 536 787 55 36
Fax: +90 236 233 15 45
E-mail: ustunkar@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
M.Eng.	Texas A&M University Industrial Engineering	2004
BS	Bilkent Industrial Engineering	2002
High School	İzmir Atatürk High School	1998

WORK EXPERIENCE

Year	Company	Position
2009 September	VESTEL WHITE	Purchasing Responsible
2007 July	TÜBİTAK UZAY	Project Management Specialist
2005 December	AYESAŞ	Project Management Specialist
2004 May	ASELSAN	Strategic Planning Engineer

FOREIGN LANGUAGES

English (fluent)

PUBLICATIONS & TALKS

- **G. Üstünkar**, S. Özöğür-Akyüz, U. Sezerman, G.-W. Weber and N. Baykal, *Application of Advanced Machine Learning Methods for SNP Discovery In Complex Disease Association Studies*, 14th International Congress on Computational and Applied Mathematics (ICCAM), Antalya, Turkey, 2009.

- **G. Üstünkar**, S. Özöğür-Akyüz, G.-W. Weber and N. Baykal, *Application of Advanced Machine Learning Methods for Tag SNP Selection in Complex Disease Association* Studies, 24th European Conference on Operational Research (EURO), Lisbon, Portugal, 2010.

- **G. Üstünkar**, S. Özöğür-Akyüz, G.-W. Weber and Y. Aydın Son, *Analysis of SNP-Complex Disease Association by a Novel Feature Selection Method*, International Conference Operations Research (OR), Munich, 2010 (**accepted for inclusion in the post proceedings of the conference - to be published by Springer**).

- **G. Üstünkar**, S. Özöğür-Akyüz, G.-W. Weber, *SNP-Complex Disease Association by Simmulated Annealing Approach*, 8th International Conference on Optimization Techniques and Applications (ICOTA), Shanghai 2010.

- **G. Üstünkar**, S. Özöğür-Akyüz, G.-W. Weber and Y. Aydın Son,C. M. Friedrich, *Selection of Representative SNP Sets for Genome-Wide Association Studies: A Metaheuristic Approach*, Optimization Letters, 2010 (**submitted**)