PRIORITIZING SNPS FOR DISEASE-GENE ASSOCIATION STUDIES: ALGORITHMS AND SYSTEMS

by

PHIL HYOUN LEE

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
June 2009

Abstract

Identifying single nucleotide polymorphisms (SNPs) that are involved in common and complex diseases, such as cancer, is a major challenge in current molecular epidemiology. Knowledge of such SNPs is expected to enable timely diagnosis, effective treatment, and, ultimately, prevention of human disease. However, the tremendous number of SNPs on the human genome, which is estimated at more than eleven million, poses challenges to obtain and analyze the information of all the SNPs.

In this thesis we address the problem of selecting representative SNP markers for supporting effective disease-gene association studies. Our goal is to facilitate the genotyping and analysis procedure, associated with such studies, by providing effective prioritization methods for SNP markers based on both their allele information and functional significance. However, the problem of SNP selection has been proven to be NP-hard in general, and current selection methods impose certain restrictions and use heuristics for reducing the complexity of the problem. We thus aim to develop new heuristic algorithms and systems to advance the state-of-the-art, while relaxing the restrictions. To address this challenge, we formulate several SNP selection problems and present novel algorithms and a database system based on the two major SNP selection approaches: tag SNP selection and functional SNP selection. Furthermore, we describe an innovative approach to combine both tag SNP selection and functional SNP selection into one unified selection process. We demonstrate

the improved performance of all the proposed methods using comparative studies.

Co-Authorship

Section 3.1 is based on the book chapter "P. H. Lee and H. Shatkay. *Machine Learning for Computational Haplotype Analysis*. in Y. Zhang and J. C. Rajapakse, editors, Machine Learning in Bioinformatics, pages 367-388. Wiley, November 2008".

Chapter 4 is based on the paper, "P. H. Lee and H. Shatkay. *BNTagger: Improved Tagging SNP Selection using Bayesian Networks*. ISMB 2006 (Supplement of Bioinformatics) 22(14):e211-219".

Chapter 5 is based on the paper, "P. H. Lee and H. Shatkay. *F-SNP: Computationally Predicted Functional SNPs for Disease Association Studies*. Nucleic Acids Research 36(Database Issue):D820-824 (2008)".

Chapter 6 is based on the two papers, "P. H. Lee and H. Shatkay. *Ranking Single Nucleotide Polymorphisms by their Putative Deleterious Effects*. AMIA 2008, pages 667-671" and "P. H. Lee and H. Shatkay. *An Integrative Scoring System for Ranking SNPs by their Potential Deleterious Effects*. Bioinformatics 25(8):1048-1055 (2009)".

Chapter 7 is based on the paper, "P. H. Lee and H. Shatkay. *Two Birds, One Stone: Selecting Functionally Informative Tag SNPs for Disease Association Studies*. WABI 2007, pages 61-72".

Acknowledgments

I would like to express my deepest gratitude to my advisor, Hagit Shatkay. She has been a great source of inspiration and a genuine role model to me. Throughout the years of my doctoral study, she has heartfully supported, encouraged, advised, and guided this thesis work. None of my work would have been possible without her years of painstaking efforts to teach me how to set up, examine, and present research ideas.

I am also grateful to professors Dorothea Blostein and Janice Glasgow, for their wonderful guidance and encouragement as my Ph.D. committee members. They have served as my role model and mentor throughout my doctoral study, and their suggestions and insights have shaped this thesis work. I am also thankful to professors Jeanette J.A. Holden, Jim Cordy, Vineet Bafna for suggesting thoughtful comments and future directions on my thesis work.

Studying at Queen's University has been a great experience to me. In particular, I would like to thank my lab members, Jess Shen, Fengxia Pan, Scott Brady, Yin Lam, Rob Denroche, Sara Salehi, Na Harrington, Andrew Wong for giving their valuable comments and interesting ideas on my work. I am also thankful to my colleagues and friends in the department, Hung Tam, Amber Simpson, Debby Robertson, and professor Parvin Mousavi for their kind support and encouragement.

My family are an integral part of this dissertation. Especially, I would like to thank

my parents Man-Jong Lee and Sun-Hee Yu and my in-laws Hye-kwan Jung and Soon-ki Shin for their everlasting love and support. I am also very grateful to my lovely sisters Hyun-Mi, Hyun-Ah, and Hyun-Yu, and my hearty brother Youn-Choul for being always with me during good times and bad times. Finally, I deeply thank my husband Jan-Yoon and my two sons Hyun-Woo and Hyun-Joon for giving me enduring support and love to make this thesis work possible. Jae-Yoon has always been my greatest friend, advisor, and supporter, and Hyun-Woo and Hyun-Joon have brought wonderful joy and delight to my life. I dedicate this thesis to all of my family, who have constantly inspired and supported me to pursue my academic career.

A Statement of Originality

I, Phil Hyoun Lee, state that the research work presented in this thesis is my own and was conducted under the supervision of Dr. Hagit Shatkay. All references to the work of other researchers are properly cited. I gratefully acknowledge the contribution of Dr. Jae-Yoon Jung, who has introduced me the theory of Pareto optimality.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Identifying genetic variations that underlie the etiology of common and complex diseases is of primary interest in current molecular epidemiology, medicine, and pharmarcogenomics. Nevertheless, our understanding of the genetic etiology of human disease is still limited due to the enormous number of genetic variations on the human genome, as well as the complex interplay of multiple genes and environmental factors underlying disease.

In this dissertation we address the challenge of selecting representative genetic variation markers, called SNPs, for supporting disease-gene association studies. Our goal is to facilitate the genotyping and analysis procedure associated with such studies, through prioritizing SNP markers based on their allele information and functional significance. However, the problem of SNP selection has been proven NP-hard in the general case, and current selection methods impose certain restrictions and use heuristics for reducing the complexity of the problem. We thus aim to develop new systems and heuristic algorithms that advance the-state-of-the-art. To address this challenge, we formulate several SNP selection problems, present novel algorithms to address the problems, and demonstrate the improved performance of the algorithms using comparative studies. In this chapter, we begin with a

brief overview of the dissertation work.

1.1 Representative SNP Selection

Understanding the genomic differences in the human population is one of the primary challenges of current genomics research [100, 191, 60, 49]. The human genome can be viewed as a sequence of 3.3 billion letters over the nucleotide-alphabet $\{A, C, G, T\}$; this sheer amount of data requires massive computational analysis for deciphering the genetic blueprint for human life. In more than 99 percent of the positions on the genome, the same nucleotide is shared across the population. However, people possess a unique genetic composition in about one percent of their genome. Those genetic variations include different nucleotide occurrences, called *single nucleotide polymorphisms* (SNPs - pronounced 'snips'), deletion/insertion of one or more nucleotides, or variations in the number of multiple nucleotide repetitions. Thus, differences in human traits, as obvious as physical appearance or as subtle as susceptibility to disease, may originate from these variations in the human DNA.

In particular, much current interest is focused on the search for genetic variations that can affect an individual's susceptibility to common and complex diseases and response to medical treatment [122, 148, 93, 25, 180, 45, 51, 165]. Simple Mendelian diseases (such as Huntington disease or Sickle Cell Anemia) are caused by an abnormal alteration of a single gene. However, most current common diseases (such as cancer, heart disease and many others) are known to be affected by a combination of multiple mutated genes, along with certain environmental factors. Thus, these conditions are often referred to as *complex* diseases. To identify the relations among mutations in multiple genes, at a statistically significant level, it is necessary to obtain genetic information from a large population [20].

Thus, instead of family-based studies, which have been successfully used for studying simple Mendelian diseases, large-scale population-based association studies are typically used for identifying the genetic variations underlying common and complex human diseases.

Such association studies typically involve single nucleotide polymorphisms (SNPs), as they are the most common form of genetic variations. The number of SNPs on the human genome is estimated at more than eleven million[1] [167], and, as such, SNPs can represent an individual's genetic variability at the finest level of detail [166]. However, the tremendous number of SNPs makes it neither practical nor feasible to obtain and analyze the information of all the SNPs on the human genome. Thus, selecting a subset of SNPs that is sufficiently informative to conduct association studies but still small enough to reduce the experimental and analysis overhead, to which we refer as *representative SNP selection*, has become an important step toward effective disease-gene association studies.

1.2 Thesis Statement

The primary goal of this research is to develop new SNP selection methods that improve upon currently available ones, and as such, to advance the state-of-the-art in the area. In particular, this dissertation introduces the following methods, all of which have shown improved performance over existing state-of-the-art methods through comparative studies:

1. A new tag SNP selection method based on Bayesian networks [107],

2. A new scoring scheme for prioritizing SNPs based on their potential deleterious functional effects [109, 110, 111],

[1] As of May 2009, dbSNP build 130 [167] provides information about 17,804,034 SNPs including deletions/insertions. Among the SNPs, 6,573,584 SNPs have been validated, while 7,344,853 SNPs occur within gene regions.

3. The first multi-objective optimization framework that combines tag SNP selection and functional SNP selection into one unified selection process [108], and

4. An additional multi-objective optimization framework based on the Pareto optimality for selecting functionally informative tag SNPs [106].

Methods 1 and 2 are based on two major SNP selection approaches called tag SNP selection and functional SNP selection, respectively. Methods 3 and 4 are proposed to support both tag SNP selection and functional SNP selection within one unified selection framework. In the following sections, we briefly introduce the key ideas of each selection method.

1.2.1 Tag SNP Selection using Bayesian Networks

Our first SNP selection method [107] is based on the tag SNP selection approach, which is motivated by the non-random association among SNPs, called *linkage disequilibrium* (LD) [21, 137, 38, 58, 163, 85]. When high LD exists between SNPs, the nucleotide information of one can usually be inferred from that of the others. Thus, we can select a relatively small subset of SNPs that still retains most of the nucleotide information of the original set. The selected SNPs are called *tag* SNPs, while the remaining, unselected SNPs are called *tagged* SNPs. Under the tag SNP selection approach, possible association between a disease phenotype and the unselected tagged SNPs is assumed to be *indirectly* captured through the selected tag SNPs.

In recent years, numerous methods have been proposed for tag SNP selection, and we introduce the state-of-the-art in Section 3.1. The utility of current tag SNP selection methods has been empirically demonstrated by simulation studies [62, 90, 91, 193, 17] or by association studies for many human diseases [152, 120, 13, 51, 80]. However, several pitfalls still exist in current tag SNP selection methods. For instance, the performance

of current tag SNP selection methods is limited by certain restrictions such as the small-bounded location [7] or the fixed number of predictive tag SNPs [21, 67, 112]. Moreover, most methods can only be applied to bi-allelic[2] SNPs or require an additional imputation-procedure as pre-processing.

We aim to address these limitations and to improve the performance of currently available predictive tag SNP selection methods. That is, our method is neither limited to bi-allelic SNPs, nor requires an additional imputation-procedure. Moreover, we allow the number or the location of predictive tag SNPs to vary for each tagged SNP, which improves prediction performance over that of state-of-the-art predictive methods. To reduce the complexity of the SNP selection problem while accommodating these variabilities, both the dependence and the conditional independence relationships between SNPs are exploited using the framework of Bayesian networks. A comparative study based on multiple SNP datasets demonstrates the improved predictive power of the new method over existing state-of-the-art methods [107].

1.2.2 Functional SNP Selection using an Integrative Scoring System

Our second SNP selection method [109, 110, 111] is based on the functional SNP selection approach, which aims to *directly* select a subset of SNPs that are likely to have deleterious functional effects, and as such, more likely to be involved in disease [162, 16, 56, 87, 104, 109]. For example, SNPs occurring in exonic regions may radically change the amino acid composition of a translated protein. As such, they are highly likely to cause functional distortions of that protein, and are therefore more likely to underlie disease [132]. Another example is SNPs occurring in regulatory regions, such as transcription factor binding

[2]"Bi-allelic" means that people possess only one of two different nucleotides among $\{a, g, c, t\}$ at a position in which a SNP occurs. We provide basic genetic concepts in Chapter 2.

sites. The SNPs can alter the binding affinity of transcription factors, and as such, can deleteriously affect gene expression, tissue specificity, and cellular activity of the regulated protein [26]. The key step in functional SNP selection is therefore to effectively assess the putative deleterious effects of SNPs, so that SNPs can be prioritized according to their functional significance.

Indeed, a variety of web services and public databases have been introduced to prioritize SNPs by their putative deleterious effects on major bio-molecular functions. (we provide a literature review in Section 3.2.) Yet, such tools and systems still suffer from several limitations. For example, many of them focus on only a single biological function, such as either protein coding or splicing regulation (but not both). As a result, researchers need to spend much time and effort to separately apply multiple tools, and interpret/integrate their often conflicting predictions. Moreover, most tools only classify SNPs into qualitative subgroups (such as either 'deleterious' or 'neutral'), but do not quantify the functional significance of SNPs. As such, it is not straightforward to select a specific number of the most functionally significant SNPs without additional ranking information.

In this thesis, we propose a new integrative scoring system for assessing the putative deleterious functional effects of SNPs. The system combines the assessment results from multiple independent computational tools, while taking into account the certainty of each prediction as well as the reliability of different tools. The main contributions of this work include:

1. presenting a new integrative scoring approach for quantifying the functional significance of SNPs within a probabilistic framework;

2. demonstrating the utility of this new approach based on known disease-related SNPs [110];

3. showing improved performance with respect to state-of-the-art methods for functional SNP prioritization [111]; and

4. developing a new public web-based database service, F-SNP, that provides the assessed functional information [109].

1.2.3 Combining Tag SNP Selection and Functional SNP Selection

Finally, we propose two multi-objective optimization algorithms for combining both tag SNP selection and functional SNP selection into one unified selection process [108, 106]. As of yet, the identification of predictive tag SNPs and of functionally significant SNPs have been considered as two distinct problems. Consequently, current systems that support both tag SNP selection and functional SNP selection [184, 73, 30] address each selection problem independently; that is, they separately perform tag SNP selection and function-based SNP selection, and combine the two selected sets as a last step.

We hypothesize that simultaneously identifying SNPs that are both informative and carry a deleterious functional effect is possible by taking a multi-objective optimization approach. We also hypothesize that the new approach improves upon the separate optimization approach (currently employed by other systems), in terms of both tagging-informativeness and functional significance of the selected SNP set. The main contributions of this part of the work include:

1. formulating the SNP selection problem as a multi-objective optimization problem;

2. introducing two new heuristic algorithms to address the problem; and

3. demonstrating their improved performance with respect to existing systems through comparative studies.

In conclusion, we expect the new methods, introduced throughout the thesis, to provide advanced SNP selection framework for facilitating disease-gene association studies, in terms of improved tagged SNP prediction accuracy, enhanced way of quantifying the biological significance of SNPs, and finally, integration of two major SNP selection criteria into one unified selection process. The ultimate application of this research is the support of 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.

1.3 Thesis Organization

This chapter has introduced the representative SNP selection problem, and has outlined the goal and major contributions of this dissertation. The rest of the thesis is organized as follows: Chapter 2 provides biological background relevant to genetic variation studies. Chapter 3 provides an overview of major SNP selection approaches, and summarizes related work based on each selection approach. The following five chapters present our SNP selection systems and algorithms, developed throughout the dissertation work. Specifically, Chapter 4 describes a new tag SNP selection method using Bayesian networks. Chapter 5 introduces a web-based public database service for providing functional information about SNPs and its classification system for supporting functional SNP selection. Chapter 6 presents an integrative scoring system for quantitatively assessing the deleterious functional effects of SNPs. Chapter 7 describes our first multi-objective SNP selection algorithm that combines tag SNP selection and functional SNP selection into one unified selection process, using a weighted sum of a single objective function, while Chapter 8 presents our second multi-objective SNP selection system based on the game-theoretic notion of Pareto

optimality. Finally, Chapter 9 concludes the dissertation work and outlines possible directions for future research.

Chapter 2

Biological Background for Genetic Variation Studies

This chapter introduces biological background concerning genetic variation studies. In particular, we focus on defining basic concepts in molecular epidemiology and genetics that are relevant to the problem of SNP selection. Genetics and molecular epidemiology investigate the potential contribution of genetic and environmental risk factors affecting the etiology of disease [139]. Thus, they provide the basis for common and complex disease-gene association studies and for the selection of SNP markers for these studies [201].

2.1 SNPs, Haplotypes, Genotypes, and Phenotypes

As presented in Chapter 1, there are several types of genetic variations on the human genome. In this thesis work, we focus on *single nucleotide polymorphisms* (SNPs), which are the substitutions of single nucleotides at a specific position on the genome, observed in at least 1% of the human population. 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. For example, if 80 percent of a population has the nucleotide A at a certain position on the genome while 20 percent of the population has the nucleotide T at the same position, then A is the major allele for the SNP, and T is the minor allele. Generally, when a SNP occurs in at least a relatively large percentage of a population (typically around 5-10%), it is considered a *common* SNP. To date, millions of common SNPs have been identified and are accessible in public databases, such as dbSNP [167] or Ensembl [78].

Several other terms related to SNPs and to disease-gene association studies are *locus*, *markers*, *haplotypes*, *genotypes*, and *phenotypes*. Suppose that we have chromosome samples from six individuals. Three of them have lung cancer and the other three do not. Using the DNA sequences of the chromosome samples, we aim to identify a set of SNPs that is likely to be associated with lung cancer. Due to experimental cost and time, only a limited region of the chromosome, which was previously suggested to be related to lung cancer by other molecular experiments, is examined. The chromosomal location of the target region is referred to as the *locus*. A locus can be as large as a whole chromosome or as small as a part of a gene. In this example, SNPs are used as *markers*, which are a specific type of DNA sequences that are used in association studies to identify the genetic traits of diseases.

Let us look at the chromosome samples in detail. All species that reproduce sexually have two sets of chromosomes: one inherited from the father and the other inherited from the mother. Thus, for each SNP on the chromosome, every individual in our sample also has two alleles, one on the paternal chromosome and the other on the maternal chromosome. For each SNP, these two alleles can either be identical or be different from each other. When the alleles are both the same, the SNP is referred to as *homozygous* for the individual. When

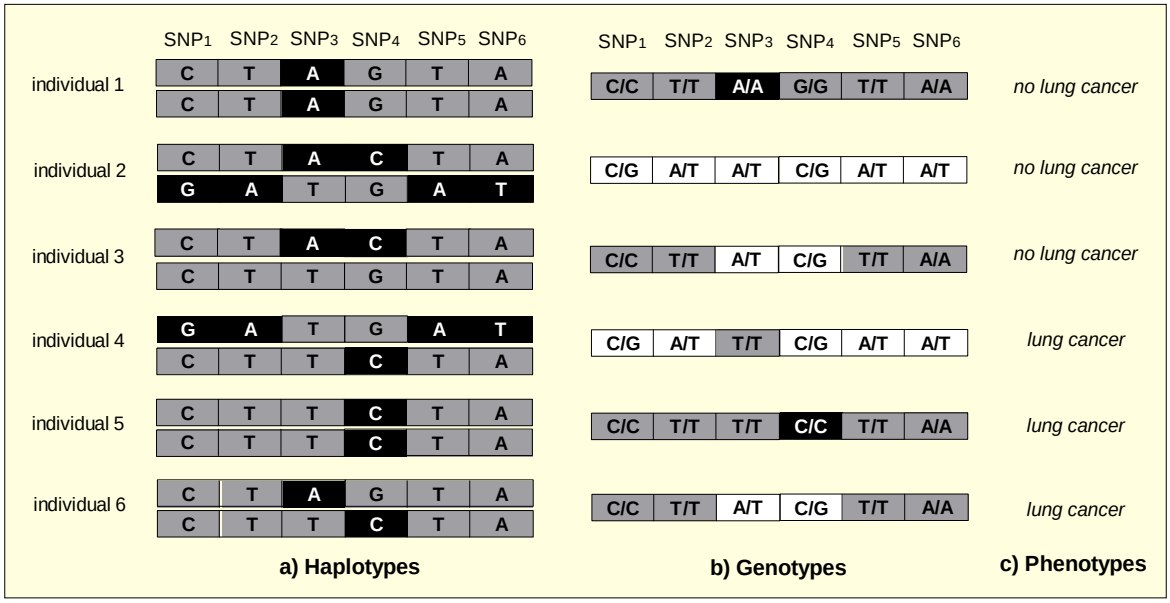


Figure 2.1: **Haplotypes, genotypes, and phenotypes.**

they are different, the SNP is referred to as *heterozygous*.

For instance, suppose that the target locus contains six SNPs, and each SNP has only two different alleles (that is, SNPs are assumed here to be *bi-allelic*). The allele information is as shown in Figure 2.1-a. The major allele of the SNP is colored gray, while the minor is colored black. Each individual has *two* sets of allele information for the six SNPs. A set of consecutive SNPs present on the same chromosome is referred to as a *haplotype* [36]. Notice that, in the above example, there are 12 haplotypes stemming from the six pairs of chromosomal samples, where each pair is associated with one individual.

Several bio-molecular methods can directly identify the haplotype information from chromosomes, but due to high cost and lengthy procedure time, these methods are limited to 10 to 20 kilobase pairs of DNA [102]. For large-scale association studies (typically from hundreds to thousands of individuals), high-throughput bio-molecular methods are typically used to identify the alleles of the target locus for each individual. The main

limitation of current high-throughput methods lies in their lack of ability to distinguish the source chromosome of each allele. Typically, such methods simply associate the two alleles with the SNP position, but do not determine which of the two chromosomes gave rise to which allele. The combined allele information of a target locus is called a *genotype*, and the experimental procedure obtaining the genotype information is called *genotyping*.

Figure 2.1-b displays the genotype information for our example. When the combined allele information of the SNP consists of two major alleles, it is colored gray. SNPs with two minor alleles are colored black, and with one major and one minor allele are colored white. The number of genotypes is six, the same as the number of individuals.

While haplotypes and genotypes represent the allele information of a target locus on chromosomes, a *phenotype* is the physical, observed manifestation of a genetic trait. In this example, the phenotype of an individual is either *lung cancer* or *no lung cancer*. In general, the individuals carrying the disease are referred to as *cases*, while the ones not known to carry the disease are referred to as *controls*. Figure 2.1-c displays the phenotype information for our sample.

2.2 Linkage Disequilibrium and Block Structure of the Human Genome

One interesting feature of a haplotype is the non-random association among the SNPs comprising it, called *linkage disequilibrium* (LD) [61]. As mentioned earlier, humans possess two copies of each chromosome: paternal and maternal. Each of these two chromosomes is generated by *recombination* of the parents' own two copies of chromosomes, and is passed by inheritance to the offspring. Figure 2.2 illustrates this recombination and inheritance

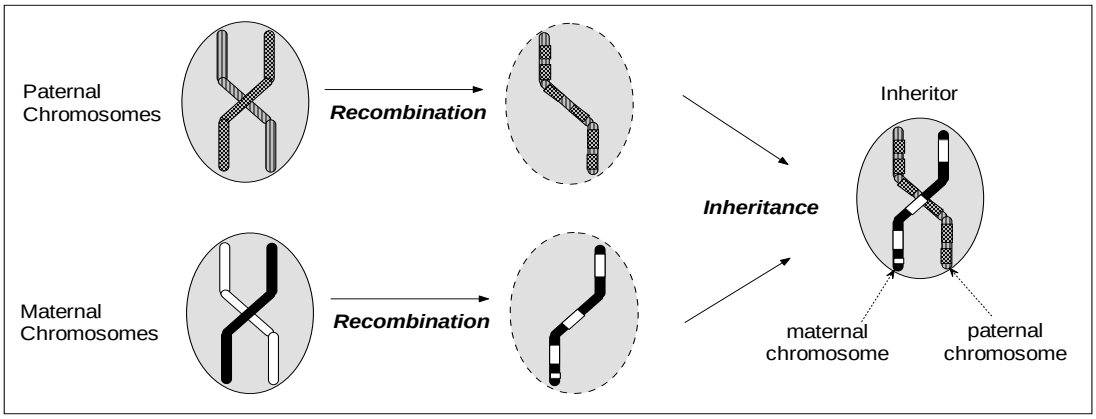


Figure 2.2: **Recombination and inheritance.**

process.

Theoretically, recombination can occur at any position along the two chromosomes any number of times. Thus, a SNP on one chromosome can originate from either copy of the parents' two chromosomes with an equal probability, and the origin of one SNP is not affected by the origin of the others. This characteristic of *independence* among SNPs is called *linkage equilibrium*.

Consider two SNPs, SNP_1 and SNP_2. Let $|\text{SNP}_1|$ and $|\text{SNP}_2|$ denote the number of alleles that the SNPs, SNP_1 and SNP_2 have, respectively. Let s_{1i} denote the i^{th} allele of SNP_1, and s_{2j} denote the j^{th} allele of SNP_2, where $i = 1, ..., |\text{SNP}_1|$ and $j = 1, ..., |\text{SNP}_2|$. Under *linkage equilibrium*, the joint probability of two alleles, s_{1i} and s_{2j}, to occur is expected to be equal to the product of the alleles' individual probabilities, since SNP_1 and SNP_2 are independent. Thus, under the independence assumption:

$$\forall_{i,j} \;\; Pr(s_{1i}, s_{2j}) \;=\; Pr(s_{1i}) \cdot Pr(s_{2j}). \tag{2.1}$$

When Equation 2.1 does not hold, that is, when the two alleles are not independent, we consider them to be in a state of *linkage disequilibrium* (LD). When the dependence between

two SNPs is high[1], the two SNPs are considered to be in a state of *high* LD.

In general, SNPs within close physical proximity are assumed to be in a state of high LD. That is, the probability of recombination increases with the distance between two SNPs [36]. Thus, SNPs within close proximity tend to be passed together from an ancestor to his/her descendants. As a result, their alleles are often highly correlated with each other, and the number of distinct haplotypes involving these SNPs is much smaller than expected under the independence assumption.

Recently, large-scale LD studies [137, 38, 58] have been conducted to understand the comprehensive LD structure of the human genome. The results strongly support the hypothesis that genomic DNA can be partitioned into discrete regions, known as *blocks*, such that recombination has been very rare (i.e., high LD) within the block, and very common (i.e., low LD) between the blocks. As a result, high LD exists between SNPs within a block, and the number of distinct haplotypes consisting of the SNPs is strikingly small across a population. This observation is referred to as the *block structure of the human genome*. At this point, there is no agreed upon way to define blocks on the genome [163, 42]. However, there seems to be no disagreement that the human genome indeed has the block structure regardless of our ability to uniquely identify the blocks.

High LD among SNPs within close physical proximity, along with the limited number of haplotypes due to the block structure of the human genome, has provided the basis for tag SNP selection, which we introduce in detail in Chapter 3. We conclude this chapter by introducing the concept of haplotype analysis and the need for computational haplotype phasing in the next section.

[1]The absolute threshold differs in each LD measure. For details, refer to LD review articles [85, 40]

2.3 Haplotype Analysis and Phasing

The ultimate goal of disease-gene association studies is to identify a set of DNA variations that is highly associated with a specific disease. Haplotype, genotype, or single-SNP information can be used for examining the association of genetic variation with a target disease. For simplicity, when haplotype information is used for examining its association with a target disease phenotype, we refer to the disease-gene association study as *haplotype analysis*. *Single-SNP analysis* and *Genotype analysis* refer to the studies that use single-SNP information and genotype information, respectively.

Haplotype analysis has several advantages compared to single-SNP analysis and genotype analysis. Single-SNP analysis cannot identify the association between variations and a disease in cases where a combination of several SNPs on one chromosome (i.e., a haplotype) is required to affect the phenotype of an individual [193, 38, 2]. Figure 2.3 exemplifies this case. All and only the three individuals with lung cancer share the haplotype *CTTCTA*, marked by a solid box in Figure 2.3-a. Thus, we can conclude that the lung-cancer phenotype is likely to be associated with the haplotype *CTTCTA*. However, if we examine each of the six SNPs individually, no direct association is found between any one of them and the lung-cancer phenotype. For example, both individuals with lung cancer and individuals with no lung cancer have the allele *C* or the allele *G* on the first SNP, the allele *T* or the allele *A* on the second SNP, and so on.

Genotypes do not contain information about the source chromosome, known as *phase*, thus they often hide the obvious association between a haplotype and a target disease. For example, in Figure 2.3-a, each individual with lung cancer (i.e., case) has two haplotypes; one haplotype is *CTTCTA*, which is associated with the lung cancer phenotype, while the other one is unique for each case. Although all cases share the exact same haplotype

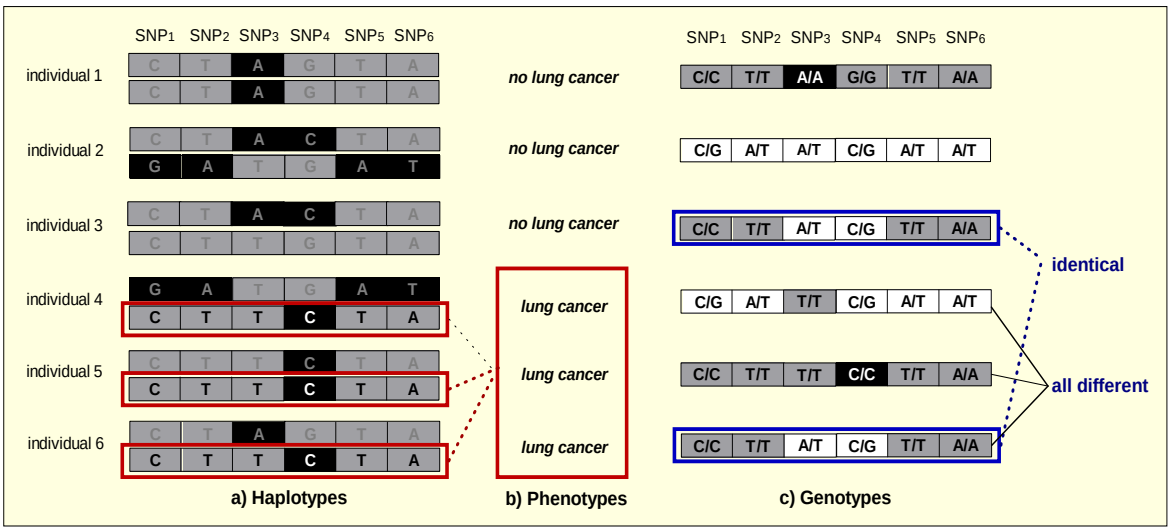


Figure 2.3: **Difference between haplotype analysis and genotype analysis.**

CTTCTA, their genotypes, in Figure 2.3-c, are all distinct due to their unique haplotype. Worse, the genotype of individual 6, who does have lung cancer, is identical to that of individual 3, who does not have lung cancer. Thus, we cannot identify a specific genotype that is highly associated with lung cancer, and as a result, miss the real association between the haplotype *CTTCTA* and lung cancer.

Despite its advantages, the use of haplotype analysis has been limited, due to the high cost and lengthy procedure time of bio-molecular methods for directly obtaining the haplotype information. However, a computational procedure, called *haplotype phasing*, addresses this problem, and greatly promotes the use of haplotype information in disease-gene association studies [14, 52, 10, 12, 27, 43]. Numerous computational and/or statistical methods have been developed for addressing the problem of haplotype phasing, and have been widely used for disease-gene association studies (for review, refer to the work by Lee [105], Niu [134], or Salem *et al*. [157]).

In summary, a typical disease-gene association study consists of SNP selection, haplotype phasing, and statistical association tests along with genotyping experiments. Initially, SNP selection algorithms are used to select a small subset of SNPs on the haplotypes, either based on the tag SNP selection approach or based on the functional SNP selection approach, which we briefly introduced in Chapter 1 and discuss in more detail in Chapter 3. Then, genotyping of selected individuals from a target population is performed, and their haplotypes are inferred from the obtained genotypes using haplotype phasing algorithms. Finally, statistical association tests are performed on the haplotype information, to identify the association of a haplotype or a set of haplotypes with a target disease.

Chapter 3

Literature Review of the Related Work

This chapter reviews the state-of-the-art in current SNP selection approaches. The surveyed methods are grouped into three major categories that support: (1) tag SNP selection; (2) functional SNP selection; and (3) both tag SNP selection and functional SNP selection. In each of the following sections, we first give a brief introduction of each approach, present current state-of-the-art methods based on the selection approach, and conclude with a discussion of open problems and future directions.

3.1 Tag SNP Selection

Tag SNP selection was motivated by *linkage disequilibrium* (LD) among SNPs. As introduced in Section 2.2, LD refers to the non-random association among SNPs within close physical proximity. When high LD exists between SNPs, their allele information is highly correlated, and as such the SNPs can act as representatives for each other with respect to their allele information.

Thus, given a large set of SNPs in a candidate region and the maximum number, k,

of SNPs that can be selected, tag SNP selection aims to find a subset of no more than *k* SNPs whose allele information can best retain the allele information of all the SNPs in the candidate region. This way, the loss of information, incurred by not using all the SNPs in association studies can be reduced. The selected SNPs are called *tag* SNPs, while the remaining, unselected SNPs are called *tagged* SNPs.

Formally, we define the problem of tag SNP Selection as follows: Let $V = \{$ SNP$_1$, ..., SNP$_p$ $\}$ be a set of p SNPs in a candidate region, and $D = \{h_1, ..., h_n\}$ be a data set of n haplotypes, where each haplotype h_i consists of the consecutive allele information of the p SNPs, SNP$_1$, ..., SNP$_p$. For simplicity, we represent $h_i \in D$ as a vector of size m whose vector element is 0 when the allele of a SNP is *major* and 1 when it is *minor*. (Recall that the nucleotide with the dominant occurrence within a population is called the *major* allele for a SNP, while the others are called the *minor* alleles.) Suppose that the maximum number of SNPs that can be selected is k, and a function $f(T'|D)$ evaluates how well the allele information of SNPs in subset $T' \subset V$ retains the allele information of all SNPs in V based on the haplotype data D.

The tag SNP selection problem can then be stated as follows:

Problem : Tag SNP Selection.
Input : A set of SNPs V;
A set of haplotypes D;
The maximum number of tag SNPs k.
Output : A set of tag SNPs $T = \underset{T' \ s.t. \ T' \subset V \ \& |T'| \leq k}{argmax} f(T'|D)$.

In brief, to solve the tag SNP selection problem, one needs to find an optimal subset of SNPs, T, of size $\leq k$ based on the given evaluation function $f(T'|D)$, among all possible subsets of the original SNPs.

Researchers have proposed a variety of objective functions, $f(T'|D)$, to best represent

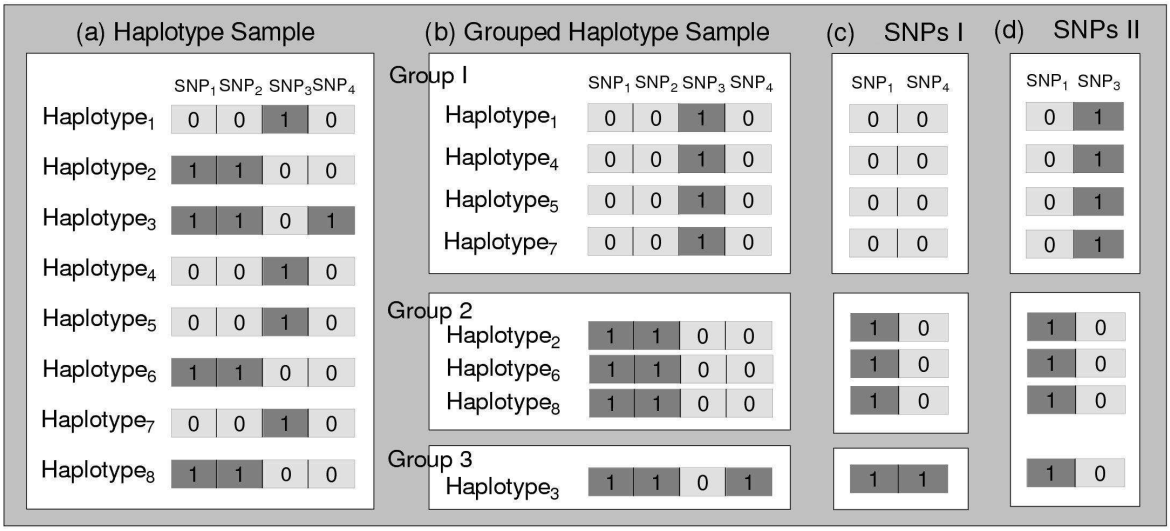


Figure 3.1: **Tag SNP selection based on limited haplotype diversity. A subset of SNPs are selected such that the selected SNPs can distinguish common haplotypes.**

the allele information of haplotypes in D using SNPs in T', and have tried to identify the subset of SNPs that optimizes the function f. We group here the surveyed algorithms for tag SNP selection into three categories based on the approach they take to measure the allele information of haplotypes: (1) haplotype diversity; (2) pairwise association among SNPs; and (3) tagged SNP prediction. In the following sections, we introduce each of them.

3.1.1 Haplotype Diversity

Recent observation of *the block structure of the human genome* [137, 38, 58, 84] demonstrates that the human genome can be partitioned into discrete blocks such that within each block, a small number of common haplotypes (i.e., 3-5 haplotypes) are shared by most of the population (i.e., 80-90%). Based on this assumption, early tag SNP selection research aimed to find *a subset of SNPs that can capture the limited haplotype diversity in the original data*.

Figure 3.1 illustrates how a set of tag SNPs can be selected based on the limited diversity of haplotypes. Suppose that our sample consists of eight haplotypes with four SNPs, as shown in Figure 3.1-a. The major allele of a SNP is coded as 0 in light gray, and the minor allele is shown as 1 in dark gray. Since each allele must be either major or minor, the possible number of *distinct* haplotypes consisting of four SNPs is 2^4. However, the *observed* number of distinct haplotypes in the sample is only 3 as shown in Figure 3.1-b. Therefore, information about 2 SNPs may be sufficient to uniquely identify the limited number of distinct haplotypes. In principle, we can try every possible combination of two SNPs to quantify how well they can distinguish the diverse haplotypes in the original data. Then, the pair that provides the most distinguishing power is selected as tag SNPs.

A variety of haplotype diversity measures were proposed. Some [137, 89] use *the number of haplotypes that are uniquely distinguishable by the candidate subset* T' as a measure of the haplotype diversity captured by T'. For example, in Figures 3.1-c and 3.1-d, SNP_1 and SNP_4 successfully partition all 8 haplotypes into 3 different groups, while SNP_1 and SNP_3 put only 4 of the haplotypes into a truly distinct set (the other 4 haplotypes are placed together despite their differences). Thus, the haplotype diversity captured by the subset $\{\text{SNP}_1, \text{SNP}_4\}$ is 8, while for $\{\text{SNP}_1, \text{SNP}_3\}$, this measure is only 4.

Johnson *et al*. [84] define the haplotype diversity *not* captured by the candidate subset T' (that is, the *residual* haplotype diversity of T') as *the number of allele differences between every haplotype pair in the same group based on T'*. If the candidate subset T' successfully partitions all distinct haplotypes into different groups as shown in Figure 3.1-c, its *residual* haplotype diversity will be 0. Otherwise, originally distinct haplotypes will be placed in the same group, as shown on the bottom of Figure 3.1-d, which makes its *residual* haplotype diversity greater than 0. Thus, T' with the *smallest residual* haplotype diversity is selected

as the set of tag SNPs.

Another popular haplotype diversity measure is *Shannon's Entropy* (H) [86, 6, 1, 68]. Let n' be the number of *distinct* haplotypes in the haplotype data set D, and p_i be the relative frequency of the i^{th} distinct haplotype. The haplotype diversity of D can be computed as its Entropy H:

$$H(D) = -\sum_{i=1}^{n'} p_i \log_2 p_i.$$

Like other methods introduced earlier, for each candidate tag SNP set T', haplotypes are partitioned into groups so that the ones in the same group share the same alleles at the SNPs included in the subset T'. The entropy of the data set D is measured based on this partition. The haplotypes that are placed in the same group are considered identical. The number of distinct haplotypes, n', thus becomes the number of groups, and the relative frequency of the i^{th} distinct haplotype, p_i, is estimated as the ratio between the number of haplotypes in the i^{th} group and the total number of haplotypes. The more groups the candidate subset T' recognizes, the larger the entropy of the data set D based on the grouping. Thus, the candidate set T' with the *largest* entropy is selected as the solution.

The methods introduced above [137, 19, 89, 84, 29, 86, 6, 1, 68] exhaustively examine all subsets of the original SNP set V, limiting their applicability to only a small number of SNPs. To overcome this problem, several heuristics and efficient search methods were proposed using: a greedy algorithm [199], a branch-and-bound rule [41], dynamic programming [194, 193, 192, 195, 197, 197, 196], and principal component analysis (PCA) [75, 112, 123].

Haplotype diversity-based methods are intuitive and straightforward. However, to ensure that haplotype diversity is indeed limited, block-partitioning must first be conducted on the target locus, and tag SNP selection is done block by block. The possible limitation

of this block-dependent approach lies in the possibility that the union of the optimal sets of tag SNPs from each block might not be the optimal set of tag SNPs for a whole region [62]. Furthermore, as discussed in Section 2.2, regions of low linkage disequilibrium exist *between* blocks [36]. Thus, certain regions of the target locus may demonstrate a large number of diverse haplotypes, deeming the above methods impractical. In addition, as of yet there is no agreed upon way to define blocks on the genome. Thus, the selection of tag SNPs depends on the block-partitioning method used [163, 41, 46].

3.1.2 Pairwise Association

Pairwise association-based methods rely on the idea that a set of tag SNPs should be *the smallest subset of available SNPs that are capable of predicting a disease-causal variant* on the genomic region [62, 19, 182, 21, 4, 91, 146]. However, the disease-causal variant is generally the one we are looking for, and is not known ahead of time. Thus, pairwise association between SNPs is used as an estimate for the predictive power with respect to the disease locus.

In principle, a set of tag SNPs is selected such that *all SNPs on the locus are highly associated with at least one of the tag SNPs*. This way, although the SNP that is relevant to a disease phenotype may not be selected as a tag SNP, the association of the target disease with that SNP can be indirectly deduced from the tag SNP that is highly associated with the unselected SNP. In most studies, non-random association of SNPs (that is, linkage disequilibrium (LD)), introduced in Section 2.2, is used to estimate the pairwise association.

Byng *et al*. [19] first proposed to use cluster analysis for pairwise association-based tag SNP selection. The original set of SNPs is partitioned into hierarchical clusters, where SNPs within the same cluster have at least a pre-specified level, σ, (typically $\sigma > 0.6$-0.8)

of pairwise LD with *at least one* of the other SNPs. After clustering is performed, they recommend to select one SNP from each cluster based on practical feasibility such as ease of genotyping, importance of physical location, or significance of the SNP mutation.

Others [182, 21, 4] proposed that a tag SNP should be selected as the one whose pairwise LD is greater than the fixed level, σ, with respect to *all* the other SNPs in the cluster. To identify the tag SNPs, *minimax clustering* [4] and *greedy binning algorithm* [182, 21] were proposed.

In minimax clustering, the *minimax* distance between two clusters C_i and C_j is defined as $D_{minimax}(C_i, C_j) = \underset{s \in (C_i \cup C_j)}{min}(D_{max}(s))$, where $D_{max}(s)$ is the maximum distance between the SNP s and all the other SNPs in the two clusters. Initially, every SNP constitutes its own cluster. The two closest clusters (according to the minimax distance) are then merged iteratively. The merging stops when the smallest distance between two clusters is larger than pre-specified level σ. Finally, the SNP that defines the minimax distance for each merged cluster is selected as the cluster representative.

The greedy binning algorithm works as follows: First, it examines all pairwise LD relationships between SNPs, and for each SNP, counts the number of other SNPs whose pairwise LD with the SNP is greater than a pre-specified level σ. The SNP that has the largest counting number is then clustered together with its associated SNPs, and becomes the tag SNP for the cluster. This procedure is iterated with the remaining SNPs until all the SNPs are clustered. The SNPs whose pairwise LD is not greater than σ with respect to any other SNPs are considered singleton clusters.

All pairwise association-based methods have a complexity of $O(cnp^2)$, where the number of clusters is c, the number of haplotypes is n, and the number of SNPs is p. Thus,

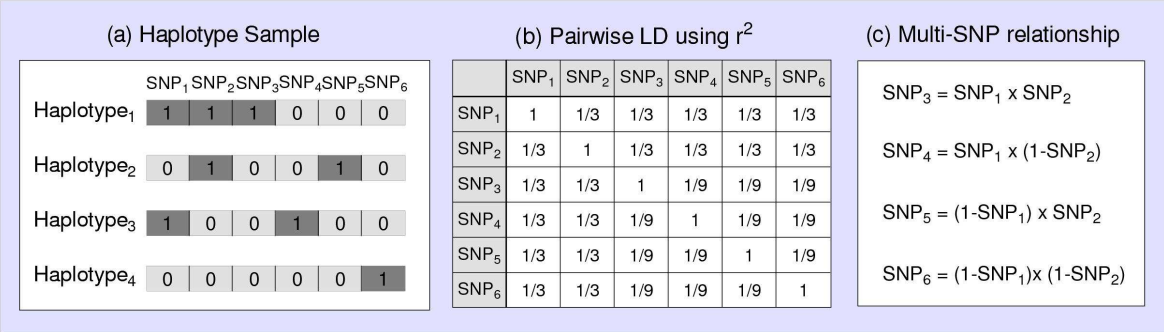


Figure 3.2: **Pairwise linkage disequilibrium (LD) among SNPs and multi-SNP dependencies.**

in general, they run faster than the methods based on haplotype diversity, and do not require a prior block-partitioning procedure. Arguably, pairwise association-based methods are currently the most widely used tag SNP selection methods.

The major shortcoming of pairwise association-based methods lies in their lack of ability to capture multi-SNP dependencies [7] and in a tendency to select more tag SNPs than other methods [91, 163, 62, 17]. Figure 3.2 illustrates this weakness of pairwise association-based methods.

Suppose that our sample consists of four haplotypes with six SNPs, as shown in Figure 3.2-a. If we measure pairwise LD between the SNPs using one of the most commonly used LD measures, correlation coefficient r^2 [62], no two SNPs have pairwise LD greater than 0.5, as shown in Figure 3.2-b. Thus, pairwise association-based methods will select all six SNPs as tag SNPs. However, as shown in Figure 3.2-c, the allele of SNP_3, SNP_4, SNP_5, and SNP_6 can be perfectly represented by the alleles of SNP_1 and SNP_2. Thus, if we consider multi-SNP dependencies, only two SNPs, namely SNP_1 and SNP_2, are sufficient to represent all the six SNPs. The next tag SNP selection approach, referred to as *tagged SNP prediction-based*, uses this multiple SNP dependencies to represent unselected tagged SNPs.

3.1.3 Tagged SNP Prediction

Tagged SNP prediction-based methods consider tag SNP selection as a reconstruction problem of the original haplotype data using only the allele information of the selected tag SNPs. Thus, they aim to select *a set of tag SNPs that can predict the unselected (i.e., tagged) SNPs with the least error*. Unlike the pairwise association-based methods, introduced in the previous section, tagged SNP prediction-based methods use *multiple* tag SNPs to predict the allele information of unselected, tagged SNPs. Therefore, these methods also present a prediction rule for tagged SNPs along with the selected set of tag SNPs.

Bafna *et al*. [7, 65] first proposed to select tag SNPs based on their accuracy in predicting the tagged SNPs. Let $E^t_{i,j}$ be the event that haplotypes h_i and h_j have a different allele at SNP t, and $E^T_{i,j}$ be the event that haplotypes h_i and h_j have a different allele at some SNP in T. To measure how well a set of SNPs, $T = \{\text{SNP}_1, ...,\text{SNP}_k\}$, can predict the SNP, t, Bafna *et al*. define a measure called *informativeness* as:

$$I(T,t) = Pr_{i\neq j}(E^T_{i,j}|E^t_{i,j}).$$

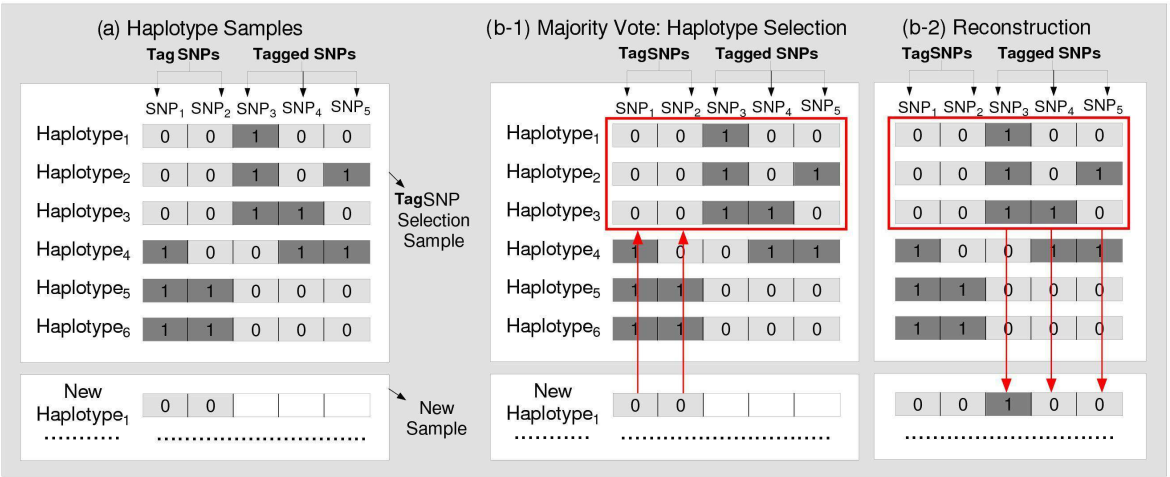


Figure 3.3: **Majority vote in tagged SNP prediction-based methods.**

Based on the proposed measure, an optimal subset of SNPs that can best predict the remaining ones is identified using dynamic programming. Bafna *et al*. restrict the predictive tag SNPs of each tagged SNP to those that are within a relatively close physical proximity w to the predicted SNP. However, the exponential complexity $O(nk2^w)$ of the proposed dynamic programming algorithm needs to be reduced.

Recently, Halperin *et al*. [67] proposed a polynomial time dynamic programming algorithm, but, in principle, their improvement results from fixing the number of tag SNPs for each tagged SNP to be 2. Halperin *et al*. also proposed a prediction rule for tagged SNP alleles based on a *majority vote*. Figure 3.3 illustrates the prediction procedure.

Suppose that our sample consists of six haplotypes with five SNPs, and SNP_1 and SNP_2 are selected as tag SNPs as shown in Figure 3.3-a. We call this sample the tag SNP selection sample. As discussed in Section 2.3, genotyping is conducted to obtain the allele information of the *selected tag SNPs* for studied individuals. To predict the ungenotyped alleles (that is, the tagged SNPs) in this new sample, first, the haplotypes whose tag SNP alleles are the same as those of the new haplotype are identified in the tag SNP selection sample. In Figure 3.3-b-1, these haplotypes are marked by a solid box. Each tagged SNP in the new haplotype is assigned the allele that occurs most often in the haplotypes identified above, as shown in Figure 3.3-b-2. As a result, this majority-vote-rule tends to assign common alleles rather than rare ones to a new haplotype.

Unlike pairwise association-based methods, tagged SNP prediction-based methods use *multi*-SNP dependencies to select the set of tag SNPs. As a result, the number of selected tag SNPs is often smaller than that selected by pairwise association-based methods [8]. In addition, all dynamic programming methods [7, 66, 67] guarantee to find a global optimum with respect to the given measure.

However, the effectiveness of these methods is still limited by some restrictions such as the small-bounded location or the fixed number of tag SNPs. Moreover, tagged SNP prediction-based methods do not reduce the number of SNPs to be examined in subsequent association studies. That is, after the selected tag SNPs are genotyped, the alleles of the tagged SNPs are reconstructed using the alleles of the tag SNPs, and disease-gene association is examined using the *reconstructed full* haplotype data. Therefore, these methods may not be appropriate for large-scale association studies.

3.1.4 Discussion

The feasibility of tag SNP selection has been empirically demonstrated by simulation studies [62, 90, 91, 193, 17] and by association studies [152, 120, 13, 51, 80]. Most importantly, Zhang *et al*. [193] demonstrate that tag SNP selection shows little loss of power[1] in subsequent association studies. Based on 1000 simulated data sets, the average difference in power between a whole set of SNPs and a set of tag SNPs whose size is $1/4$ of the original SNP set is only 4 percent. Other studies also suggest that tag SNP selection can yield about 2-50 fold savings in the genotyping efforts.

However, several pitfalls still exist:

1) Most tag SNP selection algorithms focus on covering common haplotypes or common SNPs rather than rare ones [201]. Common variations are of interest because many common human diseases have been explained by common DNA variations rather than by rare ones [45, 144, 91]. Furthermore, practically, a much larger sample size is needed to identify rare haplotypes [100]. However, it is still an open question whether common variations or rare ones influence the susceptibility to common and complex disease [36].

[1]The power of association tests is the probability that the test rejects the *false* null hypotheses [135].

2) Many algorithms require haplotype data rather than genotype data. When only genotype data are available, *Haplotype Phasing* is performed on the genotype data, and the identified haplotype information is used. However, Haplotype Phasing may lead to incorrect resolution. To address this, some statistical algorithms produce multiple solutions along with their uncertainty [116], or the distribution of haplotype pairs for each genotype rather than a single resolved pair [201]. Until now, no tag SNP selection methods consider this uncertainty of inferred haplotype data.

3) All the algorithms described above assume that the set of tag SNPs selected from a given sample will work well to characterize another sample from the same population. However, the effectiveness of tag SNPs is known to depend on the sample size, allele frequencies, and the SNP density in the haplotype/genotype dataset that is used to select the tag SNPs [126]. Moreover, sufficient number of individuals should be sampled and used in the dataset on which tag SNP selection is performed to ensure that selected tag SNPs work well to characterize other samples. For example, Goldstein *et al*. [62] reported that at least 100 chromosomes, that is, 200 haplotypes, should be used for tag SNP selection when the number of SNPs on the target genomic locus is about 20. Therefore, tag SNP selection should be applied only when a sufficient number of individuals can be sampled. In addition, methods that can avoid over-fitting of the given data set are needed when sample data are insufficient.

3.2 Functional SNP Selection

Functional SNP selection aims to prioritize SNP markers based on their functional significance [174, 162, 16, 56, 87, 104, 26, 11]. As introduced in Chapter 1, a significant fraction of genetic pathology is likely attributable to the deleterious effects of SNPs on protein

function or through alterations in the regulation of genes [148].

For instance, DNA binding properties of transcription regulatory proteins [26], molecular structures of pre-mRNAs [95], signal transduction activities of transmembrane receptors [166], subcellular localization of proteins [83], kinetic parameters of enzymes [149], and conformation of structural proteins [132] are all susceptible to perturbation by SNPs. In other words, some SNPs are highly likely to disrupt major bio-molecular functions of genomic regions where they occur, and as such, are more likely to underlie the genetic basis of human disease [166]. Directly genotyping and analyzing these possibly disease-causal SNPs is expected to increase the chance of finding biologically plausible associations, while reducing false positive or false negative findings in association studies [149, 174].

The key issue in functional SNP selection is therefore the effective assessment of putative deleterious effects of SNPs, so that SNPs can be prioritized according to their functional significance. Indeed, since the initial sequencing of the human genome, numerous public databases have been introduced to *provide functional annotation* about SNPs [167, 78, 178, 155, 23, 69, 170, 55, 128, 168]. A variety of computational tools have been also developed to *predict the potential functional effects of SNPs* with respect to major bio-molecular functions, such as protein coding or transcriptional regulation [132, 147, 153, 88, 189, 9, 53, 175, 172, 129, 44, 156, 74, 94, 140, 203, 145, 186, 15, 50, 187, 22, 200, 63, 92, 119, 158, 121, 202, 159, 99, 3, 76, 114, 143]. More recent systems have focused on *integrating heterogeneous biological databases and prediction tools* for SNPs in order to comprehensively analyze the functional significance of SNPs [32, 31, 30, 188, 24, 181].

Here, we group the surveyed methods into three categories according to the way in which they provide the functional information about SNPs: (1) providing functional annotation; (2) predicting potential functional effects; and (3) integrating existing databases/tools

for comprehensive function-assessment. In the following sections, we introduce the methods within each category.

3.2.1 Providing Functional Annotation

Public SNP databases, such as dbSNP [167] and Ensemble [78], are the simplest form of resources that can be used for selecting functionally significant SNPs. Along with the primary information about SNPs (such as allele information and chromosomal location), the databases provide functional annotation for SNPs. In particular, the annotation called "*the functional type of SNPs*" designates "a *bio-molecular function* of the genomic region where each SNP occurs as well as the *phenotypic consequence* of the SNP mutation to an encoded protein, if available" [78].

Currently used functional types of SNPs and their meaning are as follows:

- *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 [2].

- *Stop lost* (also known as *nonsense* mutation) - SNPs that are located in protein coding regions, and result in the loss of a stop codon.

[2]Frameshift mutation (also called a framing error) is a genetic mutation caused by insertion or deletion of nucleotides, which can disrupt the reading frame, resulting in a completely different translation from the original [167].

- *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.
- *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.

Other SNP databases also provide similar functional annotation as described above. The GeneSNPs database [178] provides a graphical view of SNP data, and uses multiple

colors to designate the different functional types of SNPs. The SNPper database [155] allows users to search SNPs based on their functional type, while the MutDB [128] database provides functional annotations relevant to a protein structure. The PicSNP database [23] provides a search interface to select non-synonymous SNPs based on the function of the gene in which the SNPs are located. Other databases such as OMIM (Online Mendelian Inheritance in Man) [69], HGMD (Human Gene Mutation Database) [170], HGVBase [55], and MutationView [168] provide information about SNPs that have been already identified as disease-associated or disease-causing based on the literature.

These annotation-based SNP databases are typically used for selecting SNPs based on the importance of their genomic region. For example, SNPs in protein coding regions, splice sites, or regulatory regions are considered to be more important than ones in introns or intergenic regions. In particular, much focus has been given to SNPs occurring in protein coding regions with phenotypic consequences, such as *non-synonymous* SNPs and the SNPs leading to *stop gain* or *stop loss*. These SNPs are most likely to damage the function of an encoded protein, and their deleterious effects are relatively easier to verify with bio-molecular experiments [127].

Still, the relative risk[3] of phenotypes generated from the same functional type of SNPs varies greatly. In particular, the relative risk stemming from non-synonymous SNPs is known to vary, from extremely low to very high, depending on their location on protein domains or on the extent of sequence conservation of the genomic location among multiple species [174]. However, none of the annotation-based databases provide users with possibly different phenotypic effects of SNPs of the same functional type.

To address this limitation, computational tools and web-services have been developed

[3]Relative risk refers to the ratio of the risk of having the phenotype among individuals with a particular exposure, genotype or haplotype to the risk among those without that exposure, genotype, or haplotype [70].

to further prioritize the SNPs occurring in functionally important genomic regions. Typically, the tools *predict* the *putative deleterious* effects of SNPs with respect to major bio-molecular functions. We introduce such prediction systems in the following section.

3.2.2 Predicting Functional Effects

In this section, we review publicly available computational tools that *assess the putative deleterious functional effects* of SNPs. Specifically, we focus on the impact of SNPs with respect to the following three major bio-molecular functional categories: 1) *protein coding*; 2) *splicing regulation*; and 3) *transcriptional regulation*.

Protein Coding SNPs in protein coding regions have been most extensively studied due to their direct impact on the function of an encoded protein [127]. In particular, numerous tools have been developed to predict the putative deleterious effects of *non-synonymous* SNPs that cause an amino acid substitution of a translated protein; The substitution may affect protein folding, proper activity of binding or interaction sites, solubility, structure, or stability of the protein [149]. To estimate these effects, current tools mainly rely on structural features of protein domains or evolutionary properties derived from sequence homology.

SIFT (Sorting Intolerant From Tolerant) [132] and PolyPhen (Polymorphism Phenotyping) [147] are the two most widely used tools for assessing the functional impact of non-synonymous SNPs. The SIFT tool takes a comparative genomics and evolutionary approach. It assumes that amino acid positions that play a critical role in protein function are conserved across the protein family and across evolutionary history. It thus uses multiple

alignment information of protein sequences to estimate whether an amino acid substitution would be tolerated or deleterious to protein function. Using SIFT, about 25 percent of non-synonymous SNPs available in dbSNP have been predicted to disrupt protein function.

In addition to employing an evolutionary approach similar to SIFT, PolyPhen uses the structural features of proteins. It maps the position of a substituted amino acid onto the 3D structure of a protein, and examines whether the substitution is likely to destroy the hydrophobic core of the protein, solvent accessibility, beta strands or active sites, electrostatic interactions, interactions with ligands, or other structural features of the protein. Based on the structural parameters as well as the sequence-based profile analysis of homologous sequences, PolyPhen presents empirically derived rules to predict non-synonymous SNPs that possibly damage protein function.

Other web-based tools for examining non-synonymous SNPs include: SNPeffect [153], LS-SNP [88], SNPs3D [189], nsSNPAnalyzer [9], PMUT [53], PARSESNP [175], and TopoSNP [172]. The SNPeffect tool examines the possible deleterious effects of SNPs with respect to protein stability, integrity of functional sites, protein phosphorylation and glycosylation, subcellular localization, protein turnover rates, protein aggregation, amyloidosis and chaperone interaction. LS-SNP uses data on protein sequences, functional pathways and comparative protein structure models to predict positions where non-synonymous SNPs destabilize proteins, interrupt domain-domain interaction, or impact protein-ligand binding. SNPs3D is another resource for inferring the deleterious effects of non-synonymous SNPs, using structural, systems biology and evolutionary information. The remaining tools are also based on the similar structural or sequence-based features, but employ different machine-learning methods, such as Random Forests [9], Neural Networks [53], Decision Trees [161], Support Vector Machines [88, 189], or Hidden Markov Models [172].

Splicing Regulation If SNPs occur within a splice site, noncoding introns may not be spliced out of a transcribed sequence, or exons may be removed from the transcribed sequence. This inadvertent *exon skipping* or *intron retention* will result in an unstable mRNA transcript, and furthermore, can lead to the loss of the protein function. SNPs may also occur within *exonic* splicing enhancers or silencers (ESEs/ESSs). ESEs and ESSs are typically 6-8 consecutive nucleotide sequences in an exonic region, where various components of the splicing machinery localize to splice pre-mRNAs. Like the SNPs occurring in splice sites, those within ESEs or ESSs can result in deleterious intron retention or exon skipping.

To identify SNPs occurring within splice sites, primary databases of SNPs like dbSNP [167] or Ensemble [78] can be used. There are also public databases that are specifically designed to provide information about splicing regulation. For example, ASD (alternative splicing database) provides computationally and experimentally proven data on alternative splicing. Its data include alternatively spliced introns/exons, splicing regulatory elements, splicing signals, expression states, and SNP-mediated splicing. Similar resources include ASAP [94], HOLLYWOOD [74], HMAASE [203], GeneSplicer [140], MaxEntScan [186] and NetGene2 [15].

Several studies have directly examined the phenotypic effects of SNPs on mRNA splicing. Nalla and Rogan [129] studied the effects of sequence changes that alter mRNA splicing in human diseases. Based on information theory, they designed a system to evaluate changes in the specificity of splice sites called Automated Splicing Mutation Analysis. This system can detect cryptic splice sites and associated splicing regulatory sites activated by SNPs. Sahashi *et al*. [156] proposed an algorithm to predict splicing consequences of SNPs affecting the 5' splice site, while Houdayer *et al*. [76] examined the performance of six computational tools for predicting disruption/creation of splice site consensus sequences

by SNPs. Most recently, Divina *et al*. [44] have proposed a computational method based on a multivariate logistic discrimination to predict cryptic splice-site activation and exon skipping induced by SNPs. Their web-based service, CRYP-SKIP (available at http://cryp-skip.img.cas.cz/) provides the probability of altered splicing patterns, the location of predicted cryptic splice sites and their intrinsic specificity.

There are also web-based services for predicting potential splicing enhancers and silencers within exonic regions. These services include RESCUE-ESE [50, 187], PESX [200], ESEfinder [22], and ESRSearch [63]. The RESCUE-ESE tool is based on statistical analysis of exonic splicing regulator sequences; It searches 8-mer sequences that occur more often in exons than in introns. It also searches 8-mers that are enriched in weak splice site flanking regions compared to strong splice site regions. The PESX tool uses similar statistical analysis, but it focuses on comparing the frequency of 8-mer sequences in internal non-coding exons versus un-spliced pseudo exons. The ESEfinder tool focuses on exonic splicing enhancer (ESE) motifs that have been identified through biological experiments, while the ESRSearch tool employs a comparative genomics method based on sequence conservation between human and mouse.

Transcriptional Regulation If SNPs occur within transcriptional regulatory regions (such as transcription factor binding sites, CpG islands, and microRNAs), they may alter the binding affinity of the regions, remove the recognition sites, or create new binding sites for other regulatory proteins. All of these alterations can lead to changes in the level, timing, and localization of gene expression [145]. Recent studies have also reported that SNPs occurring in transcriptional regulatory regions have modified gene expression, and thus underlie the genetic basis of several disease phenotypes, including cancer [160] and autoimmune diseases [177].

To identify SNPs that can alter gene expression, databases that provide information of *experimentally verified* transcriptional regulatory regions can be used. Such databases include TRANSFAC [121], JASPAR [158], and TRRD [99]. The TRANSFAC database provides information on transcription factors and their experimentally proven binding sites and regulated genes. The TRANSFAC database covers a wide variety of species from yeast to human, and arguably, is the most widely used database for studying transcriptional regulation of genes. The JASPAR database is a catalog of a curated, non-redundant set of transcription factor binding profiles. These profiles have been derived from *experimentally defined* transcription factor binding sites for multicellular eukaryotes, obtained from the literature. The TRRD database provides information about structural and functional organization of transcription regulatory regions for eukaryotic genes. Like the other databases, only experimentally verified information is included in TRRD.

In contrast, web-based computational tools, such as Consite [159], Promolign [202], TFSearch [3], rVISTA [114], HGVbase [55], MAPPER [119], rSNP_Guide [143], and Match [92] focus on predicting *potential*, *un-identified* regulatory elements. These tools obtain the sequence information about experimentally verified transcriptional regulatory regions, from databases like TRANSFAC, JASPAR, or TRRD, and apply various techniques to identify genomic regions with similar sequences to the proven regulatory regions. The prediction-based tools also search non-protein-coding regions that are conserved across multiple species as putative regulatory elements. As discussed previously, the basic assumption is that the conserved genomic sequences have been retained due to their functional importance rather than merely by a chance [11, 54].

3.2.3 Integrating Heterogeneous Functional Information

In the previous sections, we introduced a variety of databases and prediction tools for assessing the putative deleterious effects of SNPs with respect to a *single* bio-molecular function. In this section, we review systems that provide more comprehensive functional information about SNPs by *integrating* heterogenous biological databases and prediction tools. The integrative systems enable users to prioritize functionally significant SNPs without applying separate tools. PupaSuite [30], SNP Function Portal [181], SNPnexus [24], and FastSNP [188] are among the publicly available integrative systems.

The PupaSuite [30] service aims to prioritize functional SNPs for large-scale association studies. Given a list of SNPs, it first examines the functional type of the genomic regions where the SNPs occur. For non-coding regions, it uses a variety of tools to select SNPs occurring in: 1) transcription factor binding sites; 2) canonical splice sites; 3) exonic splicing enhancers; and 4) triplex-forming oligonucleotide sequences [4]. In case of SNPs occurring on coding regions, it uses the PMUT [53] and SNPeffect [153] programs to identify nonsynonymous SNPs with potential deleterious effects. PupaSuite also identifies SNPs on genomic regions that are conserved between human and mouse, based on the assumption that these regions are likely to be functional.

The other systems take a similar approach to that of PupaSuite; They use a variety of prediction tools and databases to provide comprehensive functional information about SNPs, mainly focusing on protein coding, splicing regulation and transcriptional regulation. Briefly, the SNP Function Portal service provides functional annotations in six major

[4]Triplex-forming oligonucleotide sequences in the human genome are targets for triplex formation. As they are mostly found in regulatory regions, especially in promoter zones, triplex-forming oligonucleotide sequences have been suggested to involve in gene expression [81].

categories including genomic elements, transcription regulation, protein function, pathway, disease and population genetics. The SNPnexus service integrates five major SNP databases to provide functional information about SNPs. It also provides information on potential regulatory elements or structural variations as well as genetic diseases related to SNPs.

The FastSNP service integrates 11 external web servers to prioritize SNPs according to their phenotypic risks and putative functional effects. Following the SNP prioritization approach proposed by Tabor *et al*. [174], this service estimates the relative risk of functional changes by SNPs using a quantitative ranking scheme. It classifies SNPs into 13 categories of putative phenotypic risks, each of which is assigned a risk rank between 0 (which means no known effect) and 5 (which means very high risk). Therefore, SNPs can be prioritized based on their relative risk of functional changes using the risk rank. Bhatii *et al*. [11] also proposed a similar ranking strategy for prioritizing functionally significant SNPs. However, a publicly available service is not provided.

3.2.4 Discussion

Functional SNP selection has two major merits compared to tag SNP selection. First, most *known* disease-relevant mutations are attributable to changes in the function of a protein, gene expression, or mRNA splicing [34, 170, 69]. As such, it is likely that SNPs affecting the risk of disease phenotypes (but not yet identified) are also ones with deleterious functional effects. It is therefore biologically plausible to study SNPs that have the potential to deleteriously affect molecular processes in search for disease-causal variants.

Second, as previously discussed, effectiveness of tag SNPs is known to depend on the

haplotype/genotype dataset that is used to select the tag SNPs [126]. However, the functional significance of SNPs does not depend on any specific haplotype/genotype dataset. In the same context, functional SNP selection does not require prior genotyping of SNPs as tag SNP selection does.

We also note that functional SNP selection methods are applicable not only to prior selection of SNP markers but also to post evaluation of SNP markers after association with disease is identified. Once association signal is detected in association studies, both *in silico* and bio-molecular experiments are needed to confirm biologically relevant variations on the genomic locus [11]. Tabor *et al*. [174] also emphasized that the biological plausibility of association and its consistency with established knowledge about disease etiology are important evaluation criteria for genetic association studies for complex disease traits.

Nevertheless, functional SNP selection also has several pitfalls. First, functional SNP selection is based on the ability of the function-assessment tools used for predicting functional candidate genes and variants. However, our knowledge and understanding concerning the genetic mechanism of disease and the bio-molecular function of genomic regions are still limited. As such, there could be many SNPs with deleterious functional effects that cannot be captured by current prediction tools. Moreover, when a large range of possible risk factors are available for testing and analysis, it is not straightforward to assess the exact biological impact of SNPs.

Second, Mendelian disease-relevant SNPs tend to have severe impact on the function of the protein [147, 176, 190], while complex disease-associated SNPs may not. Indeed, Thomas and Kejariwal [176] suggested that on average, the molecular effects of SNPs in complex disease might be more subtle than the severe functional changes associated with most Mendelian disease SNPs. This issue is further complicated as some complex diseases

are associated with SNPs with severe impact [173, 125]. Therefore, it is still an open question whether we need to focus on SNPs with modest deleterious effects or on ones with the most deleterious effects for studying complex diseases.

3.3 Supporting Both Tag SNP Selection and Functional SNP Selection

Currently, there are only a limited number of systems that support both tag SNP selection and functional SNP selection. Namely, these systems are TAMAL [73], SNPselector [184], and SNPLogic [142]. They share the same goal: selecting an optimal set of SNP markers for genetic association studies. All of the systems also emphasize the importance of both tag SNP selection and functional SNP selection approaches for prioritizing SNPs. The systems also take a similar integrative approach; instead of developing their own prioritization method, they aim to serve as a comprehensive resource for SNP selection by integrating a variety of existing databases and publicly available tools.

The TAMAL service identifies haplotype tag SNPs and functional SNPs. To select haplotype tag SNPs, it uses Gabriel's method [58] or the TAGGER method [39]. In the case of functional SNPs, users can choose SNPs leading to an amino acid change (that is, non-synonymous SNPs), altering splice sites, or occurring in promoter regions, regions with regulatory potential, or transcription factor binding sites. SNPs occurring in conserved genomic regions across human, chimpanzee, rat, mouse, and chicken can be identified, as well. A major limitation of TAMAL is that it only allows users to input gene symbols to select SNPs. Therefore, SNPs in intergenic regions cannot be assessed. Moreover, only a single tool is used to examine the functional effects of SNPs with respect to each functional

category.

The SNPselector service takes a list of gene names or genomic regions as input, and prioritizes SNPs based on their tagging ability, SNP allele frequencies, and functional significance. To select tag SNPs, it calculates a linkage disequilibrium (LD) score for each SNP, similar to the LD-based SNP selection approach by Carlson *et al*. [21]. It also assigns a function score between 1.0 and 0.6 to each SNP; A higher score is assigned to SNPs affecting gene transcript structure or protein product such as non-synonymous SNPs and SNPs affecting canonical splice sites. However, SNPselector does not examine the possibly different functional effects of non-synonymous SNPs, and thus cannot prioritize them further. This is a major limitation that needs to be addressed, as discussed in Section 3.2.1.

The SNPLogic service has been most recently developed. It integrates functional information about SNPs from diverse resources, mainly focusing on SNPs occurring within transcription factor binding sites, splicing sites, microRNAs and evolutionarily conserved regions. It also provides information on biological pathways, gene ontology terms and OMIM disease terms relevant to SNPs. Another distinguishing feature of SNPLogic is that it enables users to define their own scoring scheme for SNPs. Users can designate scoring conditions for maximum 6 features that the SNPLogic service provides, and select SNPs based on the defined scoring function. The SNPLogic service also provides information on whether selected SNPs are covered by commercial SNP arrays (such as ParAllele, Affymetrix and Illumina).

All of the services provide a user-friendly web service interface, contain a comprehensive collection of functional information about SNPs, and enable tag SNP selection as well. However, they consider tag SNP selection and functional SNP selection separately. That is, they separately conduct tag SNP selection and functional SNP selection, and present the

two selected sets as a last step. A major shortcoming of such approach is that, in addition to the ad-hoc nature of the combination, the number of selected SNPs can be much larger than necessary.

Chapter 4

Improved Tag SNP Selection using Bayesian Networks

This chapter introduces our tag SNP selection method using the framework of Bayesian networks. We first provide the motivation for the proposed method in Section 4.1. In Section 4.2, we formulate the problem of tag SNP selection in terms of optimizing prediction accuracy, and introduce the basic notation used throughout this chapter. Section 4.3 provides the necessary background on Bayesian networks, focusing on the concepts most relevant to our algorithm. Section 4.4 describes the proposed selection and haplotype reconstruction algorithms, and Section 4.5 reports the evaluation results. Section 4.6 summarizes our findings and outlines future directions.

4.1 Motivation and Objectives

We propose a new tag SNP selection method that aims to optimize the *prediction accuracy* for *unselected tagged* SNPs. As stated in Section 3.1.3, this prediction-based selection approach has several advantages over other tag SNP selection approaches that are based on haplotype diversity or pairwise association. First, unlike haplotype diversity-based selection methods (introduced in Section 3.1.1), prediction-based methods do not depend on the block structure of the human genome. Thus, they neither require prior block partitioning nor rely on the limited diversity of haplotypes. Furthermore, prediction-based methods use a combination of several tag SNPs to predict each tagged SNP. Therefore, they typically select a smaller number of tag SNPs than pairwise association-based methods (introduced in Section 3.1.2) [8].

However, despite their advantages, current prediction-based methods [112, 164, 71, 66, 7, 67] still suffer from several limitations. First, they can only be applied to *bi-allelic* SNPs (i.e., SNPs taking only two different nucleotides among $\{a, g, c, t\}$ at the SNP position). While most SNPs are indeed bi-allelic, there are SNPs that can take on more than two nucleotides. While these cases may be rare, it is still desirable to impose as few restrictions as possible on tag SNP selection [136].

Second, the performance of current prediction-based methods is limited by certain restrictions, such as the *small-bounded location* or the *fixed number* of tag SNPs that can be used for predicting each tagged SNP. Although SNPs within close physical proximity are assumed to be in a state of high linkage disequilibrium (LD), recent studies have reported that the levels of LD vary across chromosomal regions [38, 151]. Therefore, as noted by Bafna *et al*. [7], *"... it is neither efficient nor desirable to fix the neighborhood in which tag SNPs are selected"*. Moreover, it is realistic to assume that a different number of tag SNPs

may be needed for predicting each tagged SNP.

Finally, most of current prediction-based methods require the *haplotype* – rather than the genotype – information of *tag* SNPs in order to predict the allele information of tagged SNPs for newly-genotyped samples. However, reliable haplotype information of tag SNPs may not be available. As noted in Section 2.3, obtaining haplotype information requires a separate *haplotype phasing* procedure, which refers to a computational/statistical procedure that deduces haplotype information from genotype information. However, as pointed by Halperin *et al*. [67], the accuracy of haplotype phasing based only on tag SNPs is limited due to the reduced linkage disequilibrium (LD) among tag SNPs. Therefore, it is reasonable to assume that reliable haplotype data are not available in the case of newly-genotyped samples.

We aim to address some of these restrictions and to improve the prediction performance of currently available predictive tag SNP selection methods. In the next section, we introduce the basic notations used throughout this chapter, and formally define the problem of predictive tag SNP selection.

4.2 Problem Definition

Suppose that we are given p SNPs on the target genomic region. Our goal is selecting a set of at most k tag SNPs ($k < p$) on the genomic region that maximizes the prediction accuracy for the remaining, unselected tagged SNPs. To formally define this problem of predictive tag SNP selection, we first introduce basic notation. We represent each SNP as a discrete random variable, $X_i\,(i = 1, ..., p)$, whose possible values are the 4 nucleotides, $\{\, a,\ g,\ c,\ t\,\}$. Let D be a dataset consisting of n haplotypes, $h_1, ..., h_n$, each containing the allele information for the p candidate SNPs, $X_1, ..., X_p$. The set D can be viewed as

an n by p matrix. Each row, D_{i-}, in D corresponds to haplotype h_i, while each column, D_{-j}, corresponds to the allele information for SNP X_j in the n haplotypes. D_{ij} denotes the allele information for the j^{th} SNP in the i^{th} haplotype.

As introduced in Section 2.1, a haplotype represents the allele information of contiguous SNPs on a *single* chromosome, while a genotype represents the *combined* allele information of the SNPs on a *pair* of chromosomes. Thus, the allele information of haplotypes takes on values from $\{\, a,\, g,\, c,\, t\,\}$, while that of genotypes takes on values from $\{\, a/a,\ a/g,\ a/c,\ a/t, ...,\ t/c,\ t/t\,\}$. When the combined allele information of a pair of haplotypes, h_j and h_k, comprises the genotype g_i, we say that h_j and h_k *resolve* g_i. For example, the two haplotypes $h_j{=}(\, a,\, g,\, a,\, c\,)$ and $h_k{=}(\, a,\, c,\, c,\, a\,)$ resolve the genotype $g_i{=}(\, a/a,\ c/g,\ a/c,\ a/c\,)$. We also refer to haplotypes h_j and h_k as the *complementary mates* of each other to resolve g_i, and consider them to be *compatible* with g_i.

We now define a prediction function within a probabilistic framework. Given the set V of random variables corresponding to the p SNPs, $V{=}\{X_1, ..., X_p\}$, let the set T consist of q selected tag SNPs, $T{=}\{X_{t_1}, ...X_{t_q}\}$ $(T \subset V)$. To predict the allele of a SNP $X_j \in V$ given the alleles of the tag SNPs in T, we use the posterior probability of X_j conditioned on the set T, $Pr(X_j \mid X_{t_1}, ...X_{t_q})$. That is, the allele whose conditional probability is the highest given the alleles of the predictive tag SNPs is taken to be the allele of the tagged SNP. When multiple maximum probability solutions exist, the most common allele of X_j is selected. To capture the idea that the predicted allele value can be either correct or incorrect, we introduce the following prediction indicator function I_p :

Definition 4.1. Prediction Indicator Function : *Given a set of p candidate SNPs,* $V = \{X_1, ..., X_p\}$, *a predictive tag SNP set,* $T = \{X_{t_1}, ...,\ X_{t_q}\}$, *a SNP,* $X_j \in V$, *and a haplotype,* D_{i-}, *a prediction indicator function* $I_p(X_j,\ T,\ D_{i-})$ *is defined as*:

$$I_p(X_j,\ T,\ D_{i-}) = \begin{cases} 1\ :\ if\ \ X_j \in T\ \ or \\ \qquad\qquad D_{ij} == \underset{x \in \{a,g,c,t\}}{argmax}\ Pr(X_j = x\ | X_{t_1} = D_{it_1}, ..., X_{t_q} = D_{it_q})\,; \\ 0\ :\ otherwise\,. \end{cases}$$

For a tag SNP $X_j \in T$, the function $I_p(X_{t_l}, T, D_{i-})$ is defined to be 1, because the allele information of a tag SNP is already correctly known. In other cases, $I_p(X_{t_l}, T, D_{i-})$ obtains the value 1 only if the predicted allele value is the same as the one in the given haplotype dataset D . We note that the prediction of each tagged SNP depends on the values of the tag SNPs, but not on other predicted tagged SNPs. Hence, prediction can be applied in any order. Using this prediction indicator function, we next formally define our objective.

Definition 4.2. Maximally Predictive Tag SNP Set: *Given a set of p SNPs, $V = \{X_1, ..., X_p\}$, a constant k ($k < p$), and a prediction indicator function I_p, a maximally predictive tag SNP set, $T = \{X_{t_1}, X_{t_2}, ..., X_{t_q}\}$, for a set of haplotypes D is defined as a subset T of V, ($T \subset V$), satisfying the following two criteria*:

$$1)\ |T| \leq k,\ \ and$$

$$2)\ T = \underset{T' \subset V}{argmax} \sum_{j=1}^{p} \sum_{i=1}^{n} I_p(X_j,\ T',\ D_{i-}).$$

That is, given the set V of random variables corresponding to the p SNPs, V=$\{X_1, ..., X_p\}$, we need to find a subset $T \subset V$, such that the size of T (i.e., $|T|$) is not greater than some pre-specified constant k, and the allele values of the SNPs in T can best predict the values of the remaining unselected ones, $V - T$, based on the function I_p. Our tag SNP selection method utilizes the framework of Bayesian networks to effectively compute the posterior probability term in the function I_p and to select a set of predictive tag SNPs. In the next

section, we briefly introduce the necessary background on Bayesian networks.

4.3 Bayesian Networks: Preliminaries

A Bayesian network (BN) is a graphical model describing joint probability distributions based on conditional independencies among its variables [82]. Given a finite set of random variables, $V = \{X_1, ..., X_p\}$, a Bayesian network has two components: a directed acyclic graph, G, and a set of conditional probability parameters, $\Theta = \{\theta_1, ..., \theta_p\}$. Each node in the graph G corresponds to a random variable X_j. An edge between two nodes represents a direct dependence between the two random variables, and the absence of an edge represents *conditional independence* between the variables. Using the conditional independence encoded in the structure of the BN [82], the joint probability distribution of the random variables in V can be computed as the product of their conditional probability parameters:

$$Pr(V) = \prod_{j=1}^{p} \theta_j = \prod_{j=1}^{p} Pr(X_j | parent(X_j)),$$

where *parent*(X_j) denotes the *parent* nodes of X_j.

The BN formalism enables the computation of the posterior probability of a target variable based on the observed value of other variables. This computation process is typically referred to as *BN inference*. Suppose that we have observed the discrete values of *q* variables, $X_{t_1} = e_1, ..., X_{t_q} = e_q$, in a BN. Based on this information, the conditional probability of X_j can be computed from the joint probability of V by marginalizing over *all unobserved variables except* X_j, denoted by $M = V - \{X_j, X_{t_1}, ..., X_{t_q}\}$ [82]. Let m denote any of the possible instantiation of the random variables in M. The posterior

probability of X_j can thus be calculated as:

$$\begin{aligned} & Pr(\, X_j = x \mid X_{t_1} = e_1, ..., \, X_{t_q} = e_q \,) \\ & = \frac{\sum\limits_{m} Pr(M = m, \, X_j = x, \, X_{t_1} = e_1, ..., \, X_{t_q} = e_q)}{Pr(X_{t_1} = e_1, ..., \, X_{t_q} = e_q \,)} \\ & = \frac{\sum\limits_{m} \prod\limits_{X_k \in V} Pr(X_k | parent(X_k))^*}{Pr(X_{t_1} = e_1, ..., \, X_{t_q} = e_q \,)}, \end{aligned} \tag{4.1}$$

where the summation is over all possible combinations of values m assigned to all the unobserved variables in M, while the value of every observed variable, X_{t_i}, is set to e_i and the value of X_j is set to x in $Pr(X_k|parent(X_k))^*$.

The *Markov blanket* is another central concept involved in Bayesian networks. The Markov blanket of a variable X_j includes the *parents* of X_j, the *children* of X_j, and the *other parents* of X_j's *children* [82]. In a BN, X_j is conditionally independent of all other variables given its Markov blanket. This conditional independence typically speeds up the calculation of the posterior probability, $Pr(\, X_j \mid X_{t_1} = e_1, ..., \, X_{t_q} = e_q \,)$, since when the Markov blanket of X_j is observed, only this information needs to be taken into account for computing the conditional distribution of X_j.

Numerous BN inference algorithms have been developed to compute this posterior probability exactly or approximately. We use the *Generalized Variable Elimination* algorithm implemented in JavaBayes [35] to compute the posterior probability used in our prediction indicator function I_p.

To use the BN inference algorithm, we must first identify the structure (G) and parameters (Θ) of the BN representing the haplotype data D. This process is referred to as *BN learning*. *Structure* learning aims to find the graph structure G which maximizes the

conditional probability of G given the data D, as follows:

$$G = \underset{G'}{argmax}\ Pr(G'|D) = \underset{G'}{argmax}\ \frac{Pr(D|G') \cdot Pr(G')}{Pr(D)}$$
$$= \underset{G'}{argmax}\ Pr(D|G') \cdot Pr(G').$$

We use the Minimum Description Length (MDL) score [103] to reflect the above probabilistic scoring. In the same vein, *parameter* learning in a BN aims to find Θ that maximizes the conditional probability of Θ given the data D, $Pr(\Theta|D)$. We use a maximum-likelihood approach to estimate Θ. We refer to the work by Jensen [82] or by Neapolitan [130] for more details about structure and parameter learning for Bayesian networks.

4.4 Methods for Tag SNP Selection

We present a new tag SNP selection method that selects a set of tag SNPs based on their accuracy in predicting tagged SNPs. To identify a predictor-predictee relationship among SNPs, the proposed method utilizes conditional independencies among SNPs in the framework of Bayesian networks (BNs). We thus call our tag SNP selection method BNTagger. In the following sections, we provide the details of the proposed method.

4.4.1 Overview

BNTagger aims to select a set of tag SNPs that can best predict the unselected tagged SNPs. However, finding this set of tag SNPs in the general case has been proven to be NP-complete [7]. To effectively identify a set of highly predictive SNPs, T, we use several heuristics, utilizing the framework of Bayesian networks (BNs) and the conditional independence captured in it. Bayesian networks have been previously used for haplotype block partitioning [64] and haplotype phasing [183], but to the best of our knowledge, this is the

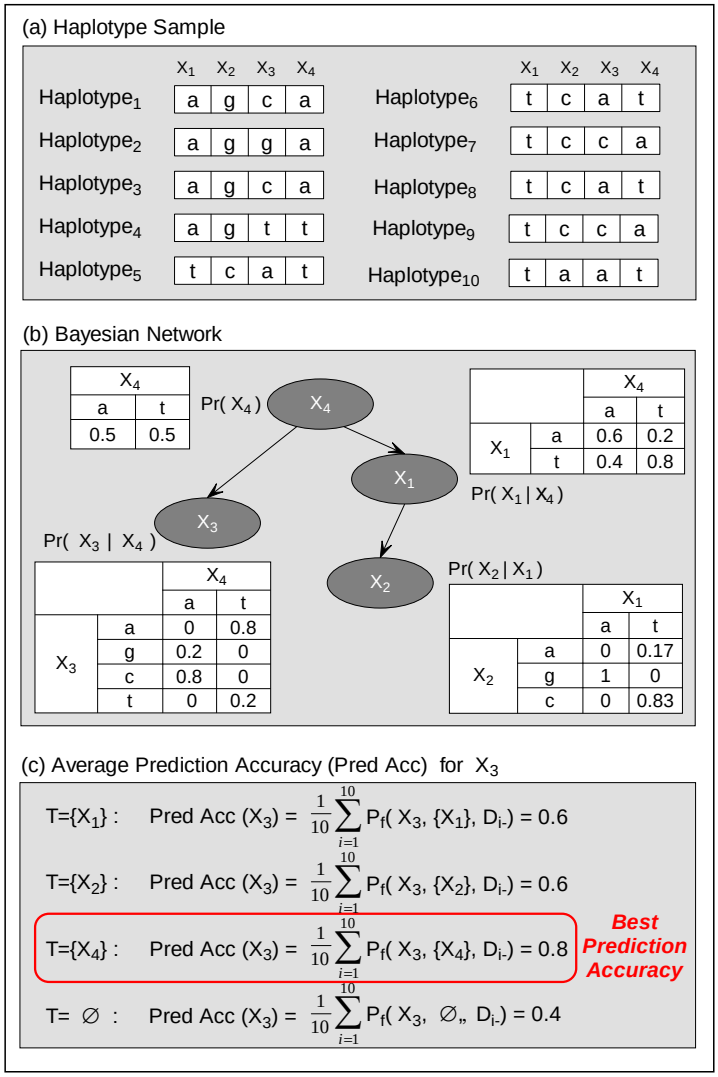


Figure 4.1: **A Bayesian network of SNPs and examples of prediction accuracy values.**

first time that they are applied to address the problem of tag SNP selection.

Figure 4.1 provides a simple example for how BNTagger utilizes the conditional independencies among SNPs to select tag SNPs. The sample here consists of ten haplotypes with four SNPs each (Figure 4.1.a); the BN structure that represents conditional independencies among the four SNPs, along with the probability parameters, is found via BN *structure* and *parameter* learning, respectively, and shown in Figure 4.1.b. For simplicity, the conditional probabilities are shown only for alleles occurring in the sample. The other probabilities are considered here to be zero.

To select tag SNPs given this Bayesian network, BNTagger uses a sequential greedy search. It first starts with an empty tag SNP set T, and sequentially examines the prediction accuracy for each SNP (node), X_j, averaged over the n haplotypes, based on the current set, T. If the prediction accuracy for SNP X_j, is smaller than a pre-specified threshold, BNTagger adds X_j into T as a new tag SNP, because X_j is not well-predicted by the current tag SNPs in T. Clearly, the order in which SNPs are evaluated is very important, since it can directly affect the selected set of tag SNPs and their prediction performance.

Unlike other methods that sequentially examine SNPs in the order of their *chromosomal location*, BNTagger examines the SNPs according to the *topological* order (from parents to children) in the BN. For example, in Figure 4.1.b, BNTagger first examines the root X_4, then its children X_3, X_1, and so on. Thus, when the prediction accuracy for each SNP X_j is evaluated given T, selected tag SNPs in the current set T are always non-descendants of X_j. In particular, when the parents of X_j are selected as tag SNPs in T, there are two advantages in following the topological order of BNs:

First, the parent-child relation in the BN encodes the direct dependence between the parents and the child node; that is, the state of a child node depends primarily on the

information of its parents. For example, Figure 4.1.c shows the prediction accuracy[1] for SNP X_3 assuming that each of the other SNPs, X_1, X_2, or X_4 is used as a single tag SNP, as well as when no tag SNPs are used. All the prediction accuracies are higher when tag SNP information is used than when it is not. Moreover, the best prediction accuracy is achieved when the parent of X_3, namely X_4, is used as a predictor.

Second, as shown in Definition 4.1, BNTagger calculates the prediction accuracy for each SNP X_j using the posterior probability of X_j *given* the allele information of the tag SNPs. To calculate this posterior, the product of the conditional probabilities in the BN must be computed as is shown in Equation (4.1). However, if the set of tag SNPs contains no descendants of X_j – which is always true due to our SNP evaluation order following the topological order of BNs – and the parents of X_j are already in the set of tag SNPs, the posterior probability, $Pr(\, X_j \mid X_{t_1}, ..., \, X_{t_q}\,)$, is the same as the conditional probability *parameter* associated with X_j, $Pr(\, X_j \,|\, parent(X_j)\,)$, due to the conditional independence encoded in the BN structure.

The simplification is derived as follows: Suppose that the current tag SNP set, $T = \{\, X_{t_1}, ..., \, X_{t_q}\,\}$, is divided into two mutually exclusive subsets; one is referred to as $parent{-}Tag(X_j)$, which consists of the parents of the SNP X_j in T, and the other is referred to as $remainingTag(X_j)$, which consists of the remaining tag SNPs in T. That is,

$$T = parentTag(X_j) \, \cup \, remainingTag(X_j)\,.$$

When all parents of the SNP X_j are selected as tag SNPs (i.e., $parent(X_j) = parentTag(X_j)$),

[1]The prediction indicator function I_p (Definition 4.1) is used in the equations in Figure 4.1.c.

$$
\begin{aligned}
& Pr(\, X_j \mid X_{t_1}, ..., \, X_{t_q} \,) \\
& = Pr(\, X_j \mid parentTag(X_j),\, remainingTag(X_j) \,) \\
& = Pr(\, X_j \mid parentTag(X_j) \,) \\
& = Pr(\, X_j \mid parent(X_j) \,).
\end{aligned}
$$

The value of this conditional probability parameter of X_j is a pre-calculated parameter, stored as a component of the BN. Therefore, the computation procedure of the prediction accuracy using I_p is much simplified. For instance, in Figure 4.1.c, the best prediction accuracy for the SNP X_3 is simply the maximum of its conditional probability parameters, $Pr(X_3|X_4)$, shown in Figure 4.1.b.

To summarize, the conditional independence structure and the conditional probability parameters in the BN guide BNTagger to find a set of highly predictive tag SNPs, and expedite the evaluation procedure. We note though that in order to use the BN components, BNTagger must first *learn* the structure and the parameters of the BN. Once the BN is constructed and the tag SNPs are selected, we also provide a prediction framework for newly-genotyped samples; as mentioned earlier, the main purpose of prediction-based tag SNP selection is to *predict* the allele information of unselected tagged SNPs based on that of the selected tag SNPs.

In conclusion, BNTagger consists of three stages:

1. Identification of the conditional independence relations among SNPs;

2. Selection of tag SNPs using two heuristics; and

3. Reconstruction of haplotype information for newly-genotyped samples.

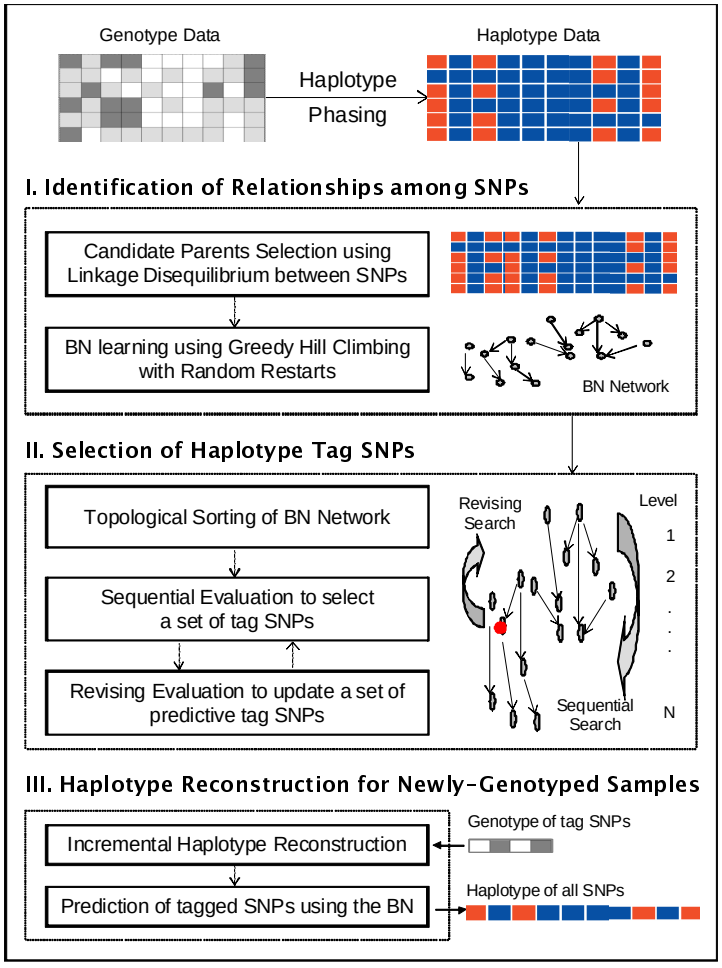


Figure 4.2: **Outline of tag SNP selection and reconstruction in BNTagger.**

In the first stage, BN learning is used to identify a graph structure, G, and a set of conditional probability parameters, Θ, that best explain the given haplotype data, D. In the second stage, a heuristic search is applied to the identified BN model to find a set of tag SNPs. The third stage provides the haplotype reconstruction framework for subsequent association studies. These three stages are depicted in Figure 4.2, and are further described in the following sections.

4.4.2 Identification of Conditional Independence Relations among SNPs

To use a Bayesian network as described above, its structure and parameters must first be *learned*. We implemented the *Sparse Candidate* algorithm [57], which accelerates BN learning by restricting the parents of each node to a small subset of candidates. To select candidate parents for each node, we use the non-random association among SNPs, known as *linkage disequilibrium* (LD). Disease-gene association studies are typically based on the assumption that there is high LD between a disease allele and adjacent SNPs [36], thus LD is widely used for quantifying relationships between SNPs in population genetics. Numerous LD measures have been proposed and been widely used. Among them, we use the *multi*-allelic[2] extension of Lewontin's linkage disequilibrium (LD) measure, D' [72], which is one of the most commonly used measures for multi-allelic SNPs [5]. We explain it here in detail.

Let X_1 be an m-allelic SNP, and X_2 be an n-allelic SNP. Let f_i^1 be the relative frequency of the i^{th} allele for SNP X_1, while f_j^2 be the relative frequency of the j^{th} allele for SNP X_2. Let f_{ij}^{12} be the relative joint frequency of the i^{th} allele occurring for SNP X_1 *and* the j^{th} allele occurring for SNP X_2 (where $i = 1, .., m$ and $j = 1, .., n$). Formally, the multi-allelic

[2]Most LD measures assume SNPs to have only two different alleles. Multi-allelic LD measures extend these bi-allelic LD measures, by allowing SNPs to have more than two different alleles.

extension of Lewontin's LD, D', is defined as:

$$D' = \sum_{i=1}^{m} \sum_{j=1}^{n} f_i^1 \cdot f_j^2 \, |\frac{f_{ij}^{12} - f_i^1 f_j^2}{D_{max}}|,$$

where D_{max} is the maximum value among the products of two relative frequencies of SNP X_1 and X_2, $f_i^1 \cdot f_j^2$ $(i = 1, .., 4;\ j = 1, .., 4)$. In principle, D' measures the difference between the observed (f_{ij}^{12}) and the expected frequency of haplotypes under independence assumption ($f_i^1 \cdot f_j^2$), normalizes the difference ($f_{ij}^{12} - f_i^1 \cdot f_j^2$) by the maximum LD ($D_{max}$), and weighs the result by the expected joint frequency under independence ($f_i^1 \cdot f_j^2$).

Using the measure D', BNTagger first selects candidate parents for SNP X_j from the set $V - \{X_j\}$, whose pairwise linkage disequilibrium with X_j, as measured by D', is in the top γ percent (here, $\gamma = 10$). The search for the optimal graph structure is performed using greedy hill climbing with random restarts. The learned network is then used for selecting better candidate parents for the next iteration, as proposed by Friedman *et al*. [57]. After N iterations (N=25,000), we select the graph structure with the best MDL score [103]. The conditional probability parameters $\Theta = \{\theta_1, ..., \theta_p\}$ are computed using maximum-likelihood estimation given the identified structure and the data.

4.4.3 Selection of Predictive Tag SNPs

Given the SNP-independence structure and the parameters of the BN constructed in the previous stage, we now identify a set of tag SNPs, T, for the haplotype data, D. We let a different combination of tag SNPs in T can be used to predict each tagged SNP. We thus identify a subset of predictive tag SNPs for each tagged SNP X_j, and denote this set by $T_{X_j} \subset T$.

As was demonstrated earlier, given the haplotype data, D, and the current set of tag

SNPs, T, we sequentially examine the prediction accuracy for each SNP, X_j. If the prediction accuracy for the SNP X_j is smaller than a pre-specified threshold, α, X_j is added to the set of tag SNPs, T. Otherwise, X_j is considered as a tagged SNP, and the current tag SNP set, T, is kept as its *candidate* set of predictive tag SNPs, T_{X_j}. We refer to this procedure as *sequential search*. When a new tag SNP is added to T during the sequential search, we re-evaluate the prediction accuracy for previously examined tagged SNPs using the updated T. If the prediction accuracy for the re-examined tagged SNP is increased by using the new set T, the candidate set of predictive tag SNPs associated with X_j is updated to the new T. We refer to this procedure as *revising search*.

To summarize, BNTagger sequentially identifies a global set of tag SNPs, T, based on their prediction accuracy, and iteratively updates the predictive set of tag SNPs, T_{X_j}, for each tagged SNP, X_j. To efficiently conduct these procedures, BNTagger uses two heuristics, as described below. The first heuristic is a SNP evaluation order that follows the topological order captured in the BN structure. BNTagger topologically sorts the nodes in the BN, which yields the *levels* of nodes as defined below, and conduct sequential search following this topological order.

Definition 4.3. Level *A level of a node X_j in a Bayesian network is recursively defined as*:

$$level(X_j) = \begin{cases} 1 & : \ if \ parent(X_j) = \phi \ ; \\ \max\limits_{X_k \in parent(X_j)} (level(X_k)) + 1 & : \ otherwise \ . \end{cases}$$

The sequential search is conducted in the order of the levels from low to high. This way, the level of tag SNPs in T is never greater than that of the currently examined node. As mentioned before, there are two advantages to this ordering when parents are tag SNPs: (1) The value of child nodes depends primarily on the information of their parents, and as such

parents tag SNPs are good predictors of their children SNPs; and (2) the child's posterior probability can be obtained directly from the network's conditional probability parameters.

The second heuristic is used for expediting the identification of predictive tag SNPs for each tagged SNP. That is, if the current set of tag SNPs, T, shows a prediction accuracy greater than a pre-specified threshold, β, for SNP X_j, we do not re-evaluate it any more. We formally define the current tag SNP set, T, as the *prediction blanket* of X_j, and use it as the final set of predictive tag SNPs for X_j. This second heuristic stems from an empirical observation that typically when the prediction accuracy for a tagged SNP, X_j, given the current set T, is sufficiently high, new tag SNPs do not significantly improve the accuracy. This phenomenon was also observed by Ackerman *et al*. [1]. Thus, we assume that it is unnecessary to examine the effect of every new tag SNP on the tagged SNPs that are already well-predicted. The loss in accuracy is typically negligible. Moreover, the potential overfitting of predictive tag SNP selection to the training data D is also reduced. Formally, we define the *prediction blanket* as follows:

Definition 4.4. Prediction Blanket *Given a prediction indicator function, I_p, and a constant β, the current set of tag SNPs, $T = \{X_{t_1}, ..., X_{t_q}\}$, is defined as the prediction blanket of X_j if the average prediction accuracy for X_j, over all haplotypes D_{i-}, given T is greater than β, that is*:

$$[\ \frac{1}{n}\sum_{i=1}^{n} I_p(\ X_j,\ T,\ D_{i-})\ ]\ >\ \beta.$$

As a matter of fact, in a Bayesian network, re-evaluation can be avoided whenever T_{X_j} is the Markov blanket of X_j, as information about newly-added tag SNPs does not affect the posterior probability of X_j given its Markov blanket. However, it is unlikely that *all* parents, *all* children, and *all* spouses of X_j (i.e., the complete Markov Blanket of X_j) will be included in the current tag SNP set T, unless T is very large. Thus, our prediction blanket

can be viewed as a relaxed version of the Markov blanket in the context of prediction. Our selection algorithm is summarized in Tables 4.1 and 4.2.

4.4.4 Reconstruction of Newly-Genotyped SNP Information

The ultimate purpose of prediction-based tag SNP selection is to reconstruct the information for all SNPs on the haplotype in newly-genotyped samples (for instance, in new association studies), using only the selected tag SNPs. We propose a practical framework for this reconstruction. Our reconstruction algorithm takes *genotype* data of tag SNPs as input, infers their resolving haplotypes (as defined in the first paragraph of Section 4.2), predicts the alleles of tagged SNPs using the Bayesian network model built in stage I, and outputs the *haplotype* information of *all* SNPs.

Let us elaborate the reconstruction procedure. Suppose that our tag SNP set T, which is identified in stage II, consists of q SNPs, that is, $T = \{X_{t_1}, ..., X_{t_q}\}$. Let $g = (x_{t_{11}}/x_{t_{12}}, ..., x_{t_{q1}}/x_{t_{q2}})$ be a *new genotype*, consisting of the combined allele information of the q tag SNPs. To deduce the complete haplotype information corresponding to the genotype g, we first select the most common haplotype in D, whose tag SNP information is *compatible* with g. The *complementary mate* of the haplotype can then be automatically constructed. If we cannot find any haplotype compatible with g in D, we create a new haplotype whose alleles are assigned as the major allele for each heterozygous tag SNP. Let h'_n be the new haplotype, and h'_{n_i} be its i^{th} element (where $i = 1, ..., q$). Given $g = (x_{t_{11}}/x_{t_{12}}, ..., x_{t_{q1}}/x_{t_{q2}})$, h'_{n_i} is defined as:

$$h'_{n_i} = \begin{cases} x_{t_{i1}} & : \ if \ \ x_{t_{i1}} = x_{t_{i2}} \ ; \\ \underset{x \in \{x_{t_{i1}} \, x_{t_{i2}}\}}{argmax} \ Pr(X_{t_i} = x) & : \ otherwise \ . \end{cases}$$

Table 4.1: **BNTagger: haplotype tagging SNP selection algorithm - sequential search**

D: training data (n haplotypes with p SNPs)
I_p: a prediction indicator function
V: a set of p SNPs $\{X_1, X_2, .., X_p\}$
T: a set of tag SNPs $\{T_{t_1}, ..., T_{t_q}\}$

// predefined constants
α: accuracy threshold for tag SNPs
β: accuracy threshold for prediction blanket

level[X_j]: the *level* of X_j in the BN
status[X_j]: the status of X_j
accuracy[X_j]: the prediction accuracy for X_j
prediction_blanket[X_j]: the prediction blanket for X_j

Main *SequentialSearch* (D, I_p) {
  T = ϕ;
  $\forall_j$ status[X_j]='unchecked';
  $\forall_j$ accuracy[X_j]=0;

  L = $\underset{j}{max}$ level[X_j];
  for each (level $1 \leq l \leq L$)
    for each (node X_j whose level is *l*)
      accuracy=$\frac{1}{n}\sum_{i=1}^{n} I_p(\ X_j,\ T,\ D_{i-}\ )$;
      if (accuracy $<\ \alpha$) // add this node as an tag SNP
        status[X_j] = 'tag SNP';
        T = $T \cup \{X_j\}$;
        call RevisingSearch($level[X_j]$);
      else if (accuracy $>\ \beta$) // the *prediction blanket* of X_j is found
        status[X_j] = 'blanket_found';
        prediction_blanket[X_j] = T;
      else // store T as a candidate set of predictive tag SNPs
        status[X_j] = 'tagged';
        prediction_blanket[X_j] = T;
        accuracy[X_j] = accuracy;
}

Table 4.2: **BNTagger: haplotype tagging SNP selection algorithm - revising search**

D: training data (n haplotypes with p SNPs)
I_p: a prediction indicator function
V: a set of p SNPs $\{X_1, X_2, .., X_p\}$
T: a set of tag SNPs $\{T_{t_1}, ..., T_{t_q}\}$

// predefined constants
α: accuracy threshold for tag SNPs
β: accuracy threshold for prediction blanket

level[X_j]: the *level* of X_j in the BN
status[X_j]: the status of X_j
accuracy[X_j]: the prediction accuracy for X_j
prediction_blanket[X_j]: the prediction blanket for X_j

Function $RevisingSearch\ (L)\ \{$
  for each (node X_k whose level[X_k] $\leq L$ and status[X_k]==‘tagged’)
    accuracy=$\frac{1}{n}\sum_{i=1}^{n} I_p(\ X_k,\ T,\ D_{i-}\ )$;
    if(accuracy $>\ \beta$)
      status[X_j] = ‘blanket_found’;
      prediction_blanket[X_k] = T;
    else if (accuracy $>$ accuracy[X_k])
      prediction_blanket[X_k] = T;
      accuracy[X_k] = accuracy;
}

The prior probability, $Pr(X_{t_i})$, can be computed using our Bayesian network model. Again, the complementary mate of the new created haplotype can then be automatically constructed. In either case, the inferred two haplotypes for g are separately used for predicting the alleles of each tagged SNP. We refer to this procedure as *incremental* haplotype reconstruction.

The principle of incremental haplotype reconstruction is based on Clark's parsimony approach [28]. That is, it tries to resolve an ambiguous genotype using one of the *already identified* haplotypes. Moreover, rather than picking any compatible haplotype, it selects the most common one, since common haplotypes are the most likely candidates under the random mating[3] assumption. Our haplotype reconstruction for the tag SNP genotype thus follows the widely-used maximum parsimony approach. However, it differs from conventional algorithms in utilizing the *existing* haplotype information of *all* previously known SNPs, rather than directly phasing the genotype information of tag SNPs. We believe that utilizing this *prior* haplotype information is necessary. As noted by Halperin *et al*. [67], haplotype phasing based only on the set of tag SNPs is not as reliable as haplotype phasing based on the original set of SNPs, due to the reduced linkage disequilibrium between tag SNPs.

Once the haplotype information of tag SNPs is deduced, we use the same prediction rule introduced in Section 4.2 to predict the tagged SNPs. That is, the allele whose conditional probability is the highest given the alleles of the tag SNPs is taken to be the allele for each tagged SNP. When multiple solutions exist, the most common allele of the tagged SNP is selected.

[3]Random mating involves individuals pairing by chance, not according to their genotypes or phenotypes [70] .

4.5 Experiments and Results

We conduct a comparative study to evaluate the performance of BNTagger compared to three state-of-the-art prediction-based methods: Eigen2htSNP [112], Block-free method [65, 7] and STAMPA [67]. In the following sections, we first summarize the experimental setting of the comparative study, and report the experimental results.

4.5.1 Evaluation Methods

We compare the performance of our method with that of three state-of-the-art predictive tag SNP selection methods: 1) The Eigen2htSNP method based on principal component analysis (PCA) [112]; 2) The Block-free method based on dynamic programming [65, 7]; and 3) The STAMPA method based on dynamic programming [67]. The problem of SNP selection has features similar to the problem of data reduction or compression, which has been widely addressed in computer science and engineering. One such method is principal component analysis (PCA). Lin and Altman [112] applied PCA for selecting tag SNPs with two heuristic options: *varimax* and *greedy*, and predicted each tagged SNP using *one* tag SNP whose correlation coefficient with the tagged one is the highest. Bafna *et al*. [7] and Halldórsson *et al*. [65] tested the Block-free method with two window sizes: 21 and 13, and used the majority vote of tag SNPs to predict each tagged SNP. Halperin *et al*. [67] also relied on the majority vote of tag SNPs for prediction, but unlike the previous two methods, they used the *genotype* data of tag SNPs rather than haplotype data.

All these methods aim to select a set of highly predictive tag SNPs for unselected, tagged SNPs. Therefore, they have all been evaluated using prediction accuracy. Accordingly, we use here prediction accuracy as the evaluation measure. We note that previously

the published results [67, 65, 7, 112] were all based on different datasets. To compare BNTagger with each of these methods, we obtained the dataset used to test each method, preprocessed it as described in the respective publication, and applied our algorithm to it. For evaluation, we use the same evaluation procedure used by each of the compared methods, utilizing *leave-one-out* cross validation for the Block-free and the STAMPA methods [67, 65, 7] and *10-fold* cross validation for Eigen2htSNP [112], as described in the respective publications. As Lin *et al*. [112] did not provide their 10-fold split, we ran the complete 10-fold cross validation procedure 10 times, each using a randomized 10-way split, to ensure robustness. In all cases, the average prediction accuracy is used as the ultimate evaluation measure. The prediction performance values of the compared methods for each dataset were directly taken from the respective publications [67, 65, 7, 112].

4.5.2 Test Data

Three public datasets, ACE (angiotensin converting enzyme) [154, 112], LPL (human lipoprotein lipase) [133, 7, 65], and IBD5 (inflammatory bowel disease 5) [38, 112, 67] were used for evaluation. These datasets were previously used to test the three compared methods, as reported in their respective publications. We first analyzed the genetic characteristics of each dataset based on: *gene diversity*, *linkage disequilibrium*, and *recombination rate*. *Gene diversity* refers to the probability that two haplotypes chosen at random from the sample are different [131], and it is is calculated as:

$$gene\_diversity\,(D) \;=\; (n/(n-1)) \cdot (1 - \sum_{i=1}^{h} p_i^2),$$

where D is a haplotype dataset, n is the total number of haplotypes in D, h is the number of

distinct haplotypes in D, and p_i is the relative frequency of the i^{th} distinct haplotype in D. Linkage disequilibrium (LD) between SNPs is estimated by the multi-allelic extension of Lewontin's LD, D' [72] (as introduced in Section 4.4.2), where the statistical significance of the standardized LD parameter is calculated using the χ^2 test with one degree of freedom. The recombination rate of each dataset is measured by the method proposed by Hudson *et al*. [79].

The first dataset ACE [154] contains 78 SNPs within a genomic region of $24Kb$ on chromosome $17q23$. Genotyping was done from 11 individuals. This dataset was used by Lin and Altman to test Eigen2htSNP [112]. Following their procedure, among the 78 original SNPs only 52 bi-allelic nonsingletons are analyzed. Partially due to the small number of SNPs and small sample size, this dataset shows high average LD (0.78) and relatively low gene diversity (0.876). The recombination rate is also relatively low (19.38%).

The second dataset LPL [133], which was used by Bafna *et al*. [7] and by Halldorsson *et al*. [65] to test the Block-free method, contains 88 SNPs spanning $5.5Kb$ on chromosome $19q13.22$. Genotyping was performed over 71 individuals. Following the analysis performed by Bafna *et al*. [7], we analyze only 87 bi-allelic SNPs. This dataset has high gene diversity (0.99) and low average LD (0.55), because it consists of haplotypes from three different populations. The four-gamete test shows 55.95% recombination or recurrent mutation.

The third dataset, IBD5 [38] contains 103 SNPs on chromosome $5q31$, spanning $500Kb$. Genotyping was performed over 129 father-mother-child trios from a European population. This dataset was used by Halperin *et al*. and by Lin and Altman to test the STAMPA [67] and the Eigen2htSNP [112] methods, respectively. Lin and Altman. [112] analyzed data from all 387 individuals using PHASE [171] for haplotype phasing. Halperin *et al*. [67]

Table 4.3: **Summary of test datasets**

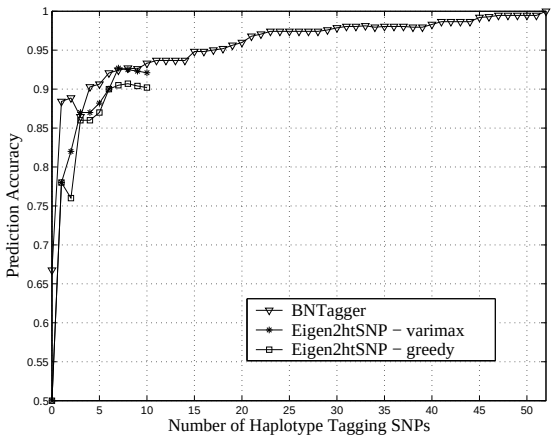
Data	SNP No	Haplotype No	Haplotype Phasing	Gene Diversity	Average LD (Std)	Average Recombination
ACE [112]	52	22	PHASE	0.876	0.78 (0.34)	19.38%
LPL [133]	87	142	known	0.991	0.55 (0.35)	55.95%
IBD5-1 [112]	103	774	PHASE	0.981	0.53 (0.27)	94.3%
IBD5-2 [38]	103	258	GERBIL	0.724	0.41 (0.23)	99.6%

analyzed data of only 129 individuals using GERBIL [96] for haplotype phasing. Thus, following both of these two procedures, we created two separate datasets from IBD5, denoted as IBD5-1 (corresponding to Lin and Altman's) and IBD5-2 (corresponding to Halperin's). Both these sets have low linkage disequilibrium and high recombination rates. The summary statistics of all datasets is given in Table 4.3.

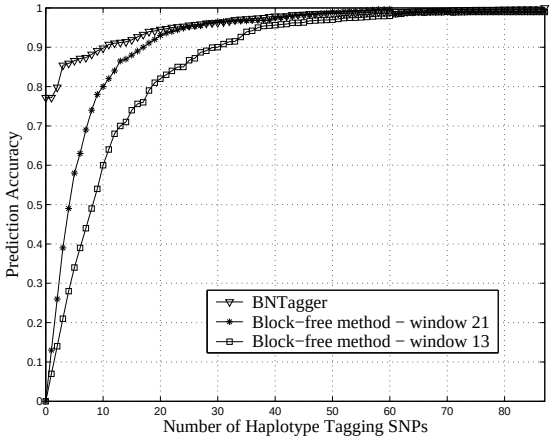
4.5.3 Test Results

We summarize the performance of BNTagger compared with that of the three state-of-the-art tag SNP selection methods in Figure 4.3. We also compute the p-value of the difference in performance, using the Wilcoxon-ranksum test with 5% significance level. Overall, BNTagger consistently outperforms other methods on all datasets. Most importantly, improvement in prediction performance is most notable when the number of selected tag SNPs is small, the average linkage disequilibrium in a dataset is relatively low, and the gene diversity is high. This is a major advantage of BNTagger, since most tag SNP selection methods have been known to suffer in those cases [36, 84, 6, 4, 21]. In other words, BNTagger retains its good performance even in what are considered to be hard cases.

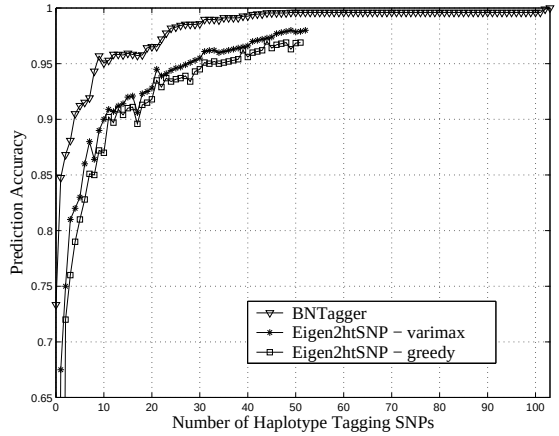
The prediction performance of Eigen2htSNP [112] is compared with ours using two



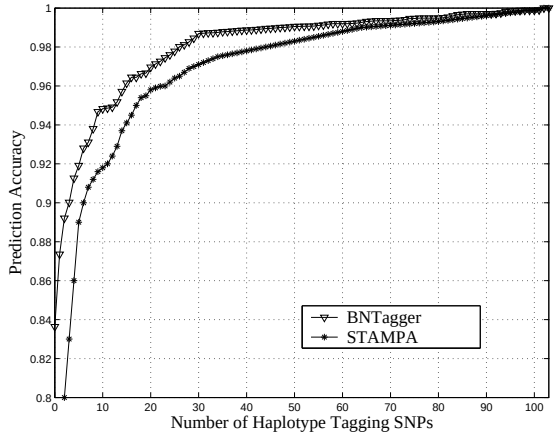
(a) ACE (Angiotensin Converting Enzyme)



(b) LPL (Human LipoProtein Lipase)



(c) IBD5-1 (Inflammatory Bowel Disease 5)



(d) IBD5-2 (Inflammatory Bowel Disease 5)

Figure 4.3: **Prediction performance of BNTagger and the compared methods for test datasets.**

datasets: ACE and IBD5-1. For the first dataset, ACE, Eigen2tag SNP-varimax shows performance comparable to ours (see Figure 4.3(a); p-values are 0.2933 for varimax and 4.88x 10^{-2} for greedy), but in the case of IBD5-1, its performance is considerably lower than ours, as shown in Figure 4.3(c) (p-values are 1.9489 x 10^{-6} for varimax and 1.5707x 10^{-8} for greedy). The prediction performance of the Block-free method [7, 65] is compared with ours using the LPL dataset. Their prediction accuracy substantially increases with the number of selected tag SNPs, as shown in Figure 4.3(b), but the difference in performance between our method and the Block-free method is significant when the number of tag SNPs is smaller than 30 (p-values are 4.2 x 10^{-3} for window 21, and 1.2552 x 10^{-9} for window 13). The prediction accuracy of STAMPA is compared with ours using the dataset that Halperin *et al*. [67] used, IBD5-2, as shown in Figure 4.3(d). Again, BNTagger outperforms STAMPA (p-value=0.7 x 10^{-2}), and the difference is significant as the number of tag SNPs gets smaller (below 60).

Overall, as shown in Figure 4.3, our method uses a small fraction of SNPs as tag SNPs (2.9%-11.5%) to achieve 90% prediction accuracy for all datasets: 4 tag SNPs among 52 SNPs (7.7%) for dataset ACE, 10 among 87 (11.5%) for LPL, 4 among 103 (3.9%) for IBD5-1, and 3 among 103 (2.9%) for IBD5-2. To achieve 95% prediction accuracy, we need 8.7%-32.7% of the target SNPs: 17 tag SNPs among 52 SNPs (32.7%) for dataset ACE, 22 among 87 (25.2%) for LPL, 9 among 103 (8.7%) for IBD5-1, and 13 among 103 (12.6%) for dataset IBD5-2. Table 4.4 summarizes the prediction accuracy of BNTagger with respect to the percentage of the selected tag SNPs.

As can be seen from Table 4.4, BNTagger can be reliably used even when the maximum number of tag SNPs is very small. This is a major advantage of BNTagger. The explicit goal of tag SNP selection is to save genotyping overhead, typically aiming at a *10-50 fold*

Table 4.4: **Prediction accuracy (in %) of BNTagger**

	Percentage of Selected tag SNPs				
dataset	0%	5%	10%	25%	50%
ACE	66.7	86.5	92.1	93.7	97.4
LPL	77.2	86.6	89.0	95.0	98.3
IBD5-1	73.3	91.2	95.3	98.4	99.6
IBD5-2	83.6	91.9	94.9	98.0	99.0

reduction in the number of target SNPs [136]. Thus, it is especially important to guarantee good prediction performance when the number of tag SNPs is a small fraction of the total number of SNPs. We note that, unlike other methods, BNTagger can predict the allele information of all SNPs even without any tag SNPs. In this case, the posterior probability of the predicted SNP X_j is the same as the prior probability of X_j. Thus, the prediction used by the function I_p, as shown in Definition 1, is still applicable even without selecting any tag SNPs.

4.6 Discussion

We presented BNTagger, a heuristic algorithm that uses the probabilistic framework of Bayesian networks to effectively identify a set of predictive tag SNPs. BNTagger outperforms other state-of-the-art predictive methods when compared over their own datasets and prediction measure. Moreover, its improved performance is especially notable when a small number of tag SNPs are selected. We believe that two main factors contribute to this improved performance:

1. We do not restrict the predictive tag SNPs for each tagged SNP to any bounded location.

2. We do not fix the number of the predictive tag SNPs for each tagged SNP.

In addition, heuristics based on the conditional independencies among SNPs guide BNTagger to effectively and efficiently find an improved set of tag SNPs in terms of prediction accuracy.

Another major advantage of BNTagger is that, after the tag SNPs are selected, it can directly reconstruct the *haplotype* information of newly-*genotyped* samples. BNTagger does not require prior haplotype phasing of tag SNPs, which might not be reliable [67]. Instead, it deduces the haplotype information of the new sample based on the haplotype training data that was originally used for tag SNP selection. In addition, BNTagger neither requires SNPs to be bi-allelic, nor does it require prior block-partitioning. Nevertheless, it shows significant improvement in prediction performance for datasets with high gene diversity and relatively low linkage disequilibrium. Thus, we believe that BNTagger provides a practical and comprehensive framework for tag SNP selection, and can form a reliable basis for subsequent disease-gene association studies.

Nevertheless, BNTagger has a number of drawbacks. First, the improved performance of BNTagger comes at the cost of compromised running time. Currently, its running time varies from dozens of minutes (when the number of SNPs is 52) to 2-4 hours (when the number is 103). Most of this time is spent on stage I, namely, learning the Bayesian network, rather than on tag SNP selection or on haplotype reconstruction. As BNTagger does not partition the haplotype data (neither through blocks nor through a sliding-window[4]), it considers all SNPs at once. That is, the conditional independence structure among all SNPs is learned simultaneously, which substantially increases its running time as the number of SNPs increases. In practice, we argue that based on the clinical importance of disease-gene

[4]Sliding-window-based algorithms confine the predictive tag SNPs for each tagged SNP to the ones in the pre-defined neighborhood (i.e., sliding-window) of the tagged SNP [123].

association studies [36], improved prediction performance takes priority over running time – when the time is not prohibitively long. Nevertheless, our future research will focus on improving the speed of BNTagger, while minimizing loss in prediction performance. This will most likely involve the evaluation of alternative heuristics and optimization criteria.

Second, BNTagger is not applicable to genome-scale genetics studies. As mentioned previously, BNTagger is build on the framework of Bayesian networks, for which learning and inference is not yet scalable to the size of a whole genome (such as several hundreds of thousands of SNPs per each chromosome). To address this scalability issue, we plan to apply a hierarchical selection approach. For example, SNPs can be first divided into subgroups with high pairwise linkage disequilibrium using clustering. Tentative tag SNPs are then selected from each subgroup, and BNTagger can be applied to the SNPs for further reducing the number of predictive tag SNPs.

Third, BNTagger does not directly set the number of selected tag SNPs. Rather, it selects tag SNPs based on their prediction accuracy with respect to a predefined threshold (α). Thus, by adjusting this threshold, the number of selected tag SNPs can be controlled. We intend to revise our selection algorithm so that the number of tag SNPs can be explicitly set, if needed.

Finally, we used the multi-allelic extension of Lewontin's linkage disequilibrium (LD), D' [72], to expedite the learning procedure in stage I. We plan to apply other multi-allelic LD measures, and examine whether different measures affect the learned networks, the selected set of tag SNPs, and their prediction performance.

Chapter 5

A Classification System for Selecting Functionally Significant SNPs

In the previous chapter, we have introduced a Bayesian network-based heuristic method for selecting a set of informative tag SNPs. In this chapter, we describe the web-based public database service, F-SNP (Functional Single Nucleotide Polymorphism), for supporting another major SNP selection approach, called *functional SNP selection*. F-SNP integrates functional information about SNPs from a variety of bioinformatics tools and databases, and classifies a subset of the assessed SNPs as *functional*. These functional SNPs are likely to have deleterious effects on major bio-molecular functions, and as such, more likely to underlie the etiology of disease. We provide the motivation for the F-SNP service in Section 5.1, and describe the database construction procedure in Section 5.2. Section 5.3 provides statistics on the database contents and describes the web interface to the database. Finally, we discuss the impact of the proposed work in Section 5.4.

5.1 Motivation and Objectives

We have constructed the F-SNP (Functional Single Nucleotide Polymorphism) database to facilitate the selection of functionally significant SNP markers for genetic studies. As discussed in Section 3.2, a variety of web services and public databases have been recently introduced to prioritize SNPs by their putative deleterious functional effects. These tools examine the functional category of genomic regions where each SNP occurs (for example, exons, splice sites, or transcription regulatory sites), and predict the potential corresponding functional effects that the SNP may have, using a variety of machine-learning approaches and experimental data. These computational methods, along with other tools in molecular genetics and epidemiology, are expected to enhance the identification of disease-causing SNPs underling many common and complex human diseases [149, 174, 148].

Yet, most such tools and systems that prioritize functionally significant SNPs provide only partial information about the functional significance of SNPs. That is, they examine the putative deleterious effects of SNPs with respect to only a *single* biological function, such as either protein coding or splicing regulation (but not both). Thus, to comprehensively analyze the functional significance of SNPs, researchers must spend much time and effort to separately apply multiple tools, and interpret/integrate their often conflicting predictions.

The F-SNP database aims to address this limitation by providing a comprehensive collection of functional information about SNPs. Specifically, F-SNP provides information about potential deleterious effects of SNPs with respect to the four major bio-molecular functional categories, namely, *splicing regulation*, *transcriptional regulation*, *protein translation*, and *post-translational modification*. Researchers can thus identify SNPs that may have deleterious effects on protein structure or function, or interfere with proper regulation

of mRNA transcription or alternative splicing.

Moreover, the F-SNP database aims to provide easy-to-use web interface so that its users can efficiently retrieve functional information about SNPs via diverse exploration routes. The following sections describe the construction procedure of the F-SNP database, provide a brief description of its current contents, and explain the web-based interface.

5.2 Database Construction

The construction procedure for the F-SNP database involves three main steps: 1) integrating primary databases for SNPs, genes, and human diseases; 2) assessing the functional effects of SNPs using external function-assessment tools and databases; and 3) identifying functionally significant SNPs using a majority-vote classifier. We further describe the details of each step in the following sections.

5.2.1 Integrating Primary Databases

SNPs and Genes We downloaded the dataset of 11,811,594 human SNPs and their annotations from the dbSNP build 126 [167] and Ensembl release 42 [78] databases. We also downloaded the list of 38,550 human genes [1] along with their primary information such as gene symbol, alias names, chromosomal location, and gene type from NCBI Entrez Gene (downloaded Dec. 12, 2006) [117].

SNP to Gene Mapping To link SNPs with specific genes, SNPs that are located along the gene region (including 5kb upstream and 5kb downstream) were identified for each

[1] Entrez Gene is NCBI's database for gene-specific information, encompassing annotated genomic regions for tRNA, rRNA, snRNA, scRNA, snoRNA, miscRNA, protein-coding and pseudo genes. As of May 1, 2009, Entrez Gene provides information about 40,590 genes for human.

gene. As a result, a total of 4,043,147 SNPs are mapped to 23,630 human genes.

Gene to Disease Mapping From NCBI's Genes and Disease site (http://www.ncbi.nlm.nih.gov/disease/), we retrieved the list of 85 human genetic disorders, categorized by the 16 body parts that they affect (downloaded Jan. 29, 2007). To link candidate genes with the 85 diseases, we downloaded the dataset of a gene-disease map from NCBI's OMIM database (downloaded Jan. 30, 2007) [69]. Accordingly, 2,374 genes were mapped to 85 human genetic disorders.

5.2.2 Assessing the Functional Effects of SNP

We assess the deleterious effects of SNPs using a variety of existing, publicly available bioinformatics tools and databases for function assessment. In particular, we focus on the potential deleterious effects of SNPs with respect to the following four major categories of biological function:

- *Protein Coding*: SNPs in protein coding regions may cause a deleterious amino acid substitution (called *non-synonymous* or *missense* SNPs) or interfere with protein translation by creating a new start/stop codon or frameshift (called *nonsense* SNPs);

- *Splicing Regulation*: SNPs in exonic or intronic splicing regulatory sites may interfere with splicing regulation for pre-mRNAs, resulting in detrimental exon skipping or intron retention;

- *Transcriptional Regulation*: SNPs in transcription regulatory regions (such as transcription factor binding sites, CpG islands, microRNAs, etc.) can alter the affinity of transcription regulators to their binding sites, and thus disrupt proper gene regulation;

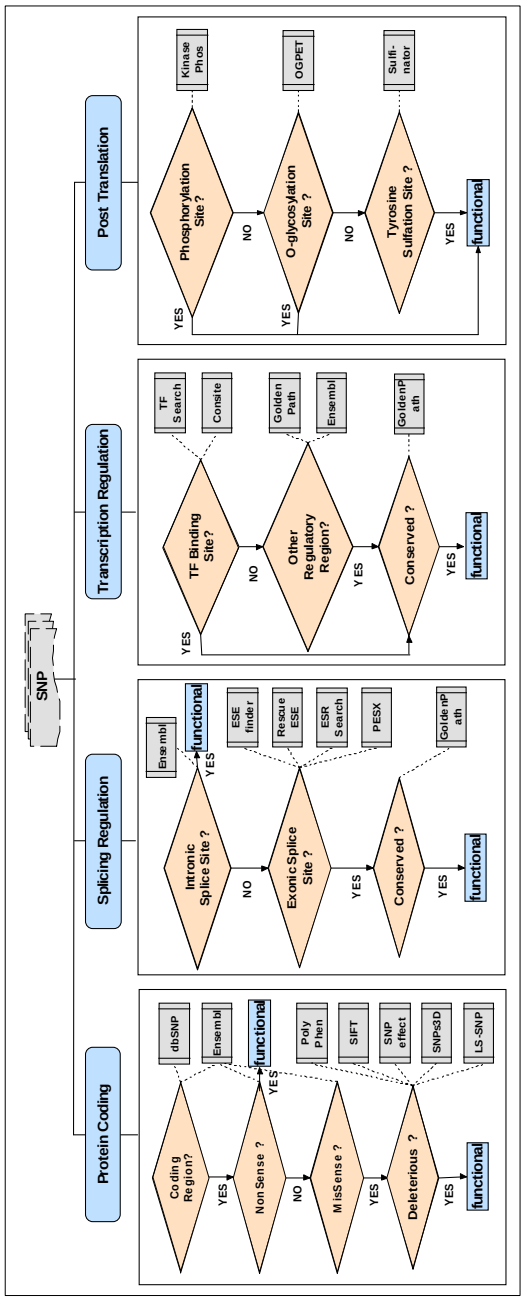


Figure 5.1: **The prediction flow-chart for four major bio-molecular functional categories. To clarify the decision procedure, the paths leading to a non-functional decision are not shown in the flow chart. Each SNP is examined for deleterious effects with respect to each functional category, namely protein coding, splicing regulation, transcriptional regulation, and post-translation modification – as shown in the top part of the figure. For each category a series of tests is executed to determine whether the examined SNP has a functional impact. First the type (coding, intronic etc.) of the genomic region is identified, using data from dbSNP [167] and Ensembl [78]. Once this is determined, other tests are performed as shown in the flow-chart within each rectangle. When multiple tools are used to predict the deleterious functional effects of SNPs, the decision is made based on a majority vote of the predicted results.**

- *Post-Translational Modification*: SNPs in protein coding regions may alter post-translational modification sites (such as phosphorylation, o-glycosylation, or tyrosine sulfation sites), interfering with proper post-translational modification.

Figure 5.1 illustrates the detailed function-assessment process: Each SNP is examined for deleterious effects with respect to each functional category (that is, protein coding, splicing regulation, transcriptional regulation, and post-translational modification – as shown at the top part of the figure). For each category a series of tests is executed to determine whether the examined SNP has a deleterious functional impact. First the type of the genomic region (such as *exon*, *intron*, *splice site*, 5'/3' *un-translated regions* of a gene (UTR), or *upstream* or *downstream* from a gene) is identified, using data from dbSNP [167] and Ensembl [78]. Each SNP is then examined based on its genomic location for its possible deleterious effects along each bio-molecular functional category.

For example, to assess if a SNP has a deleterious effect on protein coding, it first must be located on a coding region. Ensembl [78] is used to examine if this is a *nonsense* mutation, in which case the SNP is considered to be *deleterious* [2]. Otherwise - if the SNP is a *missense* mutation, it is further tested by five different tools, PolyPhen [147], SIFT [132], SNPeffect [153], SNPs3D [189] and LS-SNP [88] to check if the missense mutation is deleterious. A majority vote among these tools concludes the process, and identifies the SNP as either having a potentially deleterious functional impact (denoted '*functional*' in the figure) or not. We further describe the decision procedure in Section 5.2.3.

When a SNP is located on genomic regions of which the function is currently unspecified, it is examined by all the tools; As we do not know the function of the region, we need to examine the putative effects of SNPs with respect to all four bio-molecular functions.

[2]Nonsense SNPs are often considered to have most deleterious effects; They lead to a premature termination of amino acid peptides, often resulting in direct loss of protein function [185].

We note that, to make a robust functional assessment for SNPs, we use multiple, independent bioinformatics tools that are based on different data, algorithms, or theory for examining each functional category. The tools, PolyPhen [147], SIFT [132], SNPeffect ver. 2.0 [153], SNPs3D [189], and LS-SNP [88] are used to identify non-synonymous deleterious SNPs; ESEfinder release 3.0 [22], RescueESE [187], ESRSearch [63], and PESX [200] are used to identify SNPs in exonic splice regions; The Ensembl database release 42 [78] is used to identify nonsense SNPs and SNPs in intronic splice sites; TFSearch ver. 1.3 [3] and Consite [159] are used to identify transcriptional regulatory SNPs in promoter regions; The Ensembl release 42 [78] and GoldenPath [101] databases are used to identify SNPs in other transcriptional regulatory regions (such as microRNA, CpGIslands); KinasePhos [77], OGPET ver. 1.0 [59], and Sulfinator [124] are used to examine post-translation modification sites.

In addition to the function assessment tools for SNPs, GoldenPath (downloaded Dec. 2006) [101] is used to identify genomic regions that are conserved across multiple species (currently: chimpanzee, dog, mouse, rat, chicken, zebrafish and fugu). We use this information to filter out possible false-positive predictions of regulatory regions [11], such as 'transcription factor binding sites' or 'exonic splicing sites' – as shown in the two middle boxes in Figure 5.1. SNPs occurring in non-conserved regulatory regions are not selected as functional. This strategy is used because there is a high rate of false positive findings of regulatory sequences by *in silico* prediction tools due to the short length of such sequences, typically 6- to 8-mers [22]. Table 5.1 summarizes the list of the 16 integrated tools and databases in detail.

Table 5.1: **Integrated bioinformatics tools and databases. For each possible functional category into which a SNP may be classified, the table provides the tools that examine this function, and the URL from which the respective tool is available (as of Feb. 2009). The category *Conservedness* in the last row is not a functional category in-and-of itself, but is informative in determining the effects of SNP on splicing and transcriptional regulation.**

Function	**Tool**	**URL**
Protein Coding	PolyPhen [147]	http://genetics.bwh.harvard.edu/pph/data/index.html
	SIFT [132]	http://blocks.fhcrc.org/sift/SIFT.html
	SNPeffect [153]	http://snpeffect.vib.be/index.php
	SNPs3D [189]	http://www.snps3d.org/modules.php?name=SNPtargets
	LS-SNP [88]	http://alto.compbio.ucsf.edu/LS-SNP/Queries.html
	Ensembl [78]	http://www.ensembl.org/index.html
Splicing Regulation	ESEfinder [22]	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi
	RescueESE [187]	http://genes.mit.edu/burgelab/rescue-ese/
	ESRSearch [63]	http://ast.bioinfo.tau.ac.il/
	PESX [200]	http://cubweb.biology.columbia.edu/pesx/
	Ensembl [78]	http://www.ensembl.org/index.html
Transcriptional Regulation	TFSearch [3]	http://www.cbrc.jp/research/db/TFSEARCH.html
	Consite [159]	http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/
	GoldenPath [101]	http://genome.ucsc.edu/
	Ensembl [78]	http://www.ensembl.org/index.html
Post-Translation Modification	KinasePhos [77]	http://kinasephos.mbc.nctu.edu.tw/
	OGPET [59]	http://ogpet.utep.edu/
	Sulfinator [124]	http://www.expasy.ch/tools/sulfinator/
Conservedness	GoldenPath [101]	http://genome.ucsc.edu/

5.2.3 Summarizing the Functional Importance of SNPs

In addition to providing the raw output from the 16 integrated tools stating the functional effects of SNPs, F-SNP also denotes a subset of the assessed SNPs as *functional* SNPs; these are SNPs that are predicted by a *majority* of the integrated tools to be *deleterious* with respect to *at least* one biological function of a gene or a gene product. To identify the functional SNPs, we employ a classification method, to which we refer as F-SNP-C (F-SNP Classification).

The F-SNP-C method works as follows: For each of the four bio-molecular functional categories, a SNP is assigned into one of three classes: Class 1 indicates *irrelevance* to the corresponding biological function; Class 2 indicates that the SNP is *relevant* to the biological function, but predicted to be *benign* or has *no evidence of deleterious effects*; Class 3 indicates that the SNP is likely to be *deleterious* to the function.

For example, SNPs *outside* a protein coding region are considered to be irrelevant to protein coding, and as such are assigned to Class 1 with respect to 'protein coding'. Among the SNPs *within* a protein coding region, nonsense SNPs and some missense SNPs are predicted to have deleterious effects to protein coding (by a majority of the used assessment tools), and are thus assigned to Class 3; the remaining SNPs within the protein coding region are assigned to Class 2. Similarly, the SNPs within a highly conserved splice regulatory region or transcriptional regulatory region are assumed to be deleterious with respect to the corresponding regulatory function [11], and are thus assigned to Class 3, while the SNPs within non-conserved regulatory regions are only relevant to the respective function, and are thus assigned to Class 2.

We examine the deleterious functional effects of SNPs with respect to four major bio-molecular functional categories. As a result, four class labels are assigned to each SNP,

one for each of the four categories of biological function. To assign a single functional significance class to each SNP, we follow Bhatti *et al*. [11], and assign the *highest* class tag among all four categories as the functional significance class of the SNP.

For example, SNP rs4963 on gene ADD1 is assigned to *Class 3* with respect to 'protein coding' and *Class 1* with respect to the other functional categories, 'splicing regulation', 'transcriptional regulation', and 'post-translational modification'. The functional significance class of SNP rs4963 is set to *3* because it is highly significant for the protein coding function.

Currently, F-SNP denotes the SNPs assigned to Class 3 as *functional* SNPs. These SNPs need further investigation in disease-gene studies due to two reasons; i) They are likely to have deleterious effects with respect to at least one of the major bio-molecular functions; and ii) The prediction is supported by a majority of the used assessment tools.

5.3 Database Contents and Web Interface

The F-SNP database, release 1.0 (as of May 2009), contains the assessed functional information for 559,322 SNPs within 18,282 candidate genes for 85 major human diseases. We will continuously update F-SNP to provide functional information about additional SNPs. Detailed statistics of the current F-SNP database are provided in Table 5.2.

For each functional category, the number of SNPs for which the function has been assessed by the 16 integrated tools is shown in the middle column. The number of SNPs that F-SNP indicates to be potentially deleterious is shown on the right. Among the 154,140 SNPs examined for the protein coding category, about 43% SNPs were predicted to be potentially deleterious (66,899 among 154,140 SNPs). The ratio drops to 11%, 17%, and 7% for the splicing regulation, transcriptional regulation, and post-translation categories,

Table 5.2: **Statistics of functionally assessed SNPs in F-SNP, Release 1.0 (as of Feb. 2009). For each functional category, the number of SNPs for which the function has been assessed using the 16 tools and databases integrated into F-SNP is shown in the middle column. The number of SNPs that F-SNP indicates to be potentially deleterious is shown on the right.**

Functional Category	# of Assessed SNPs	# of potentially deleterious SNPs
Protein Coding	154,140	66,899
Splicing Regulation	73,051	8,075
Transcriptional Regulation	453,710	78,296
Post-Translation	64,736	4,477
Total	559,322	115,356

respectively. We also note that more than 80% of the examined SNPs resides in non-coding regions, and were examined for their impact on transcriptional regulation (453,710 among 559,322 SNPs).

The F-SNP database is available at *http://compbio.cs.queensu.ca/F-SNP/*. The user can search the database by SNP identifier, gene, disease, or chromosomal regions. Figure 5.2 shows an example of results obtained from an interactive search concerned with breast cancer.

Search by SNP Identifier To obtain information about a single SNP the database can be searched by providing the SNP's *rs*-identifier from dbSNP build 126 [167]. The resulting page provides the primary information about the SNP along with its assessed functional information. The primary information includes the chromosomal location of the SNP, alleles, ancestral allele, validation status, type of genomic region, links to external databases,

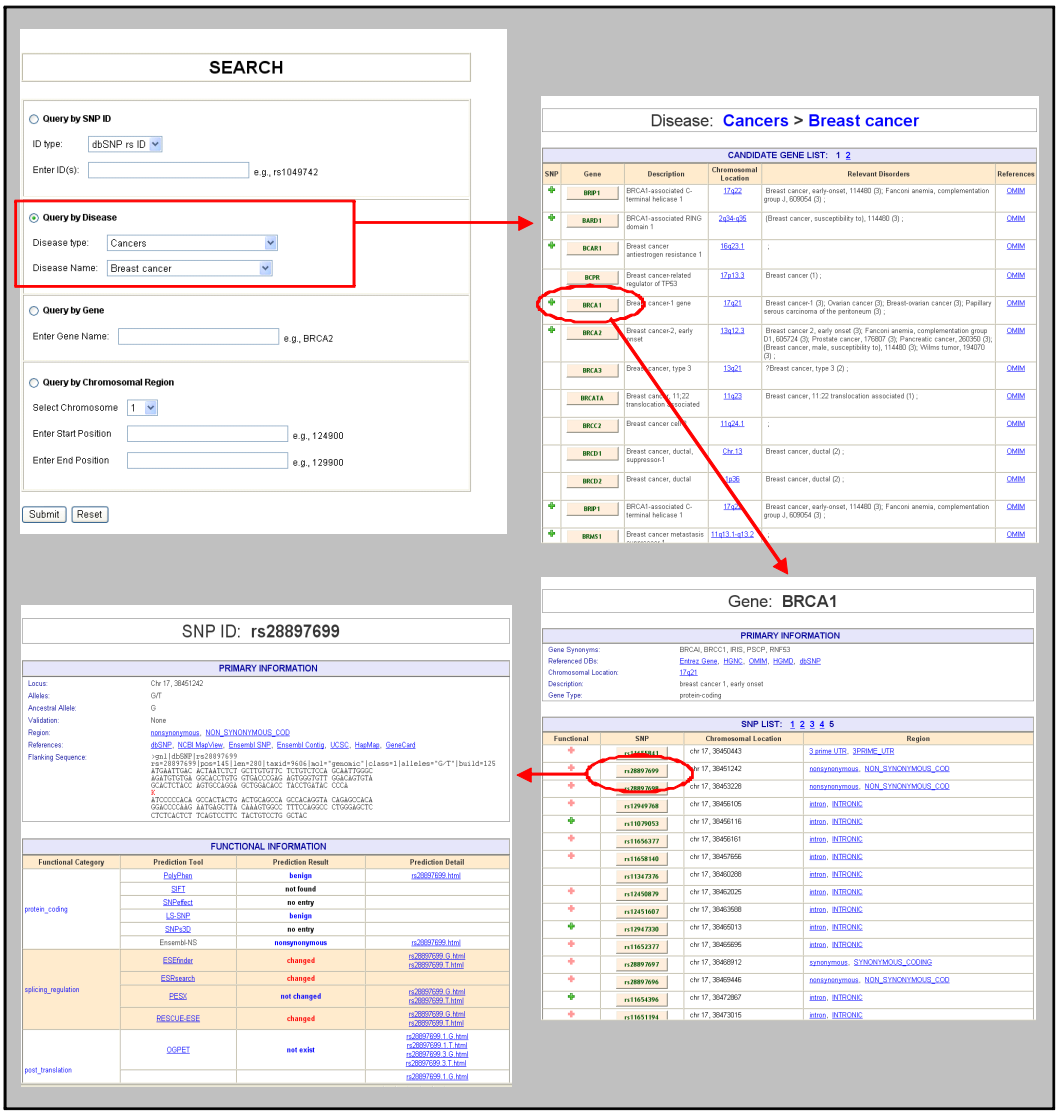


Figure 5.2: **Example of an F-SNP search session. (a) The initial search page is displayed, where the user selected the disease type to be Cancers, and the specific disease to be Breast cancer (Search by disease). (b) Results obtained after clicking the Submit button in panel (a), namely a list of genes associated with Breast cancer along with their associated chromosome location, known related disorders, and links to OMIM. The BRCA1 link (circled) is selected and clicked. (c) A detailed description of SNPs associated with BRCA1 is produced (demonstrating results of Search by gene). The SNP whose identifier is rs28897699 (circled) - indicated by a '+' mark to have associated functional information - is selected and clicked. (d) Information about the SNP rs28897699 is presented. (Demonstrates results of Search by SNP ID).**

namely dbSNP build 126 [167], NCBI Map Viewer homo sapiens build 125 [3], Ensembl release 42 [78], UCSC Genome Browser Mar. 2006 assembly [101], HapMap Rel 21a / phase II [33], and GeneCards ver. 2.37 [150], and the flanking gene sequence around the SNP. The functional information provided for each SNP includes functional category, integrated tools used, prediction results, and the detailed output from each predictive tool.

Search by Gene To find the SNPs located within a specific gene region, the database can be searched by providing the HUGO [4] name of the gene or of its protein. If no official HUGO name matches the input keyword, alias gene names, registered in NCBI Entrez Gene [117], are examined for the search. A table with all the SNPs linked to the gene is then produced, where a green '+' mark is shown next to each SNP for which the functional effects have been assessed, and a red '+' mark further indicates that the SNP was determined to have a potentially deleterious functional effect. The user can then click on each SNP to obtain the detailed functional information about it.

Search by Disease To identify SNPs that may be related to a specific disease the user can select the disease category and name. A table with all the genes relevant to the disease is produced. The user can then click on each gene to go to the gene information page. As described above, the gene information page lists all the SNPs linked to the gene, for which the user can retrieve further information.

[3]NCBI Map Viewer is the public web-based browsing service provided by NCBI. It allows users to view and search genomic regions using a graphical interface. The service is currently available at http://www.ncbi.nlm.nih.gov/mapview/.

[4]The Human Genome Organisation (HUGO) is the international organisation of scientists involved in human genetic and genomic research. For each known human gene, the HUGO Gene Nomenclature Committee assigns a gene name and symbol (short-form abbreviation) to ensure that each gene is only given one approved gene symbol.

Search by Chromosomal Region To study SNPs along a chromosomal region the user can provide the chromosome number, along with start/end positions. A table with all the SNPs within the region is produced and, as explained above, a '+' mark indicates the SNPs for which functional effects have been assessed. Again, the user can click on each SNP to obtain further information.

5.4 Discussion

The F-SNP (Functional Single Nucleotide Polymorphism) database aims to provide a comprehensive collection of functional information about SNPs, thus helping to expedite the functional SNP prioritization process. The current version of the F-SNP database provides functional information for 559,322 SNPs in 18,282 genes relevant to 85 major human diseases. Functional assessment of SNPs is done using 16 bioinformatics tools and databases.

There are two distinguishing features of the F-SNP database. For assessing the deleterious effect of SNPs along each functional category, F-SNP integrates multiple, independent bioinformatics tools that are based on different algorithms, data, and resources. No single tool can yet capture all the possible effects of SNPs on even one biological function [11]. F-SNP thus aims to complement possible shortcomings of an individual tool, by gathering information from multiple, independent resources. Researchers can also use the integrated functional information about SNPs, provided by F-SNP, to implement their own prediction tool for prioritizing functionally significant SNPs.

Another distinguishing feature of the F-SNP database is its integration of a human disease database – currently NCBI's Genes and Diseases – to facilitate identification of potential disease-causing SNPs. As shown, the F-SNP database provides a web interface that takes as input either a *disease*, a *gene*, a *chromosomal region*, or a *SNP identifier*. If the

input is a specific disease, its candidate genes, obtained from the integrated human disease database, are provided with their SNP information. Thus, researchers who are interested in a specific disease can retrieve a list of all the *known* candidate genes relevant to the disease, along with functional information for the SNPs within the selected genes with just a few mouse clicks.

The functional information provided for SNPs will be regularly updated as other prediction tools and bio-molecular experiments become available. We also plan to integrate additional human disease databases to include a broader spectrum of common and complex diseases.

We note, though, that the F-SNP database has a few limitations. First, while F-SNP denotes a subset of SNPs as *functional*, it does not numerically score or rank SNPs according to their functional significance. Budget considerations often force researchers to select a limited number of SNPs on the target genomic region for conducting association studies. When the number of putatively deleterious SNPs presented by F-SNP is larger than a pre-specified limit, selecting only some of the SNPs is not straightforward for researchers without additional ranking information. As a result, researchers may need to rely on other resources, such as the published literature, to finalize their decision.

Second, the F-SNP Classification (F-SNP-C) system is based on a simple majority vote of the functional prediction results for SNPs (such as '*deleterious*' vs. '*neutral*'), but it does not exploit additional information obtained from the used tools. For example, many function-assessment tools make predictions with different levels of confidence, and as such, provide numeric scores designating certainty – or uncertainty – regarding their own predictions. The F-SNP database does not take into account the confidence scores produced by the used tool.

Finally, some function-assessment tools are more (or less) reliable than others in terms of their prediction accuracy, but the F-SNP-C method currently weighs each tool's prediction equally. If the reliability of the different tools can be measured quantitatively, this information can be used to weigh each tool's prediction results for enhancing the voting procedure.

In the next chapter, we present a scoring method for functional SNP prioritization that addresses some of these limitations.

Chapter 6

A Scoring Approach for Selecting Functionally Significant SNPs

In the previous chapter, we have introduced the F-SNP database service and its classification system, F-SNP-C, for supporting functional SNP selection. In this chapter, we describe an integrative scoring method – for assessing the putative deleterious effects of SNPs – that improves upon the initial method. We first provide the motivation for the proposed system in Section 6.1. We formulate a scoring function that quantifies the putative deleterious functional effects of SNPs in Section 6.2, and describe the implemented scoring system in Section 6.3. In Section 6.4, we report the evaluation results of the proposed system using 112,949 SNPs on 580 disease-susceptibility genes, obtained from the OMIM (Online Mendelian Inheritance in Man) database [69]. Finally, we discuss the impact of our work and possible directions for future research in Section 6.5.

6.1 Motivation and Objectives

We propose an integrative scoring system that assesses the potential deleterious functional effects of SNPs within a probabilistic framework. The scoring system is based on the F-SNP database service, as it uses the functional assessment results for SNPs, provided by F-SNP. We thus call the proposed system F-SNP-Score. We have developed the F-SNP-Score system to improve upon current state-of-the-art methods for functional SNP selection, including the F-SNP Classification (F-SNP-C) system. While we have presented the limitations of current functional SNP selection systems in Sections 5.1 and 5.4, respectively, we briefly restate here the major motivation for the F-SNP-Score system.

Current systems, including F-SNP-C, for prioritizing functionally significant SNPs, suffer from three major limitations. First, their prioritization schemes do not take into account the uncertainty of the function prediction process (inherent in the tools used). Second, no system that uses multiple, independent bioinformatics tools for examining the deleterious functional effects of SNPs [188, 184, 11, 73, 109] considers the reliability of the different tools. Finally, most of current function-assessment systems classify SNPs into qualitatively distinct groups (such as 'irrelevant' vs. 'relevant' vs. 'deleterious'), but do not score or rank SNPs within each group.

The F-SNP-Score system aims to address these limitations by quantitatively assessing the putative deleterious functional effects of SNPs. Within a probabilistic framework, it combines the assessment results from multiple independent prediction tools, while taking into account the *prediction confidence* as well as the *tool reliability* of different tools. It assigns a specific numerical score to each SNP, representing its putative deleterious effects. Using this score, a limited subset of the most functionally significant SNPs can be ranked and selected.

In the next section, we introduce the basic notation used throughout this chapter, and formally define a scoring function for SNPs within a probabilistic framework.

6.2 Problem Definition

We aim to *quantitatively* measure the potential deleterious effects of SNPs with respect to four major bio-molecular functions, namely, splicing, transcription, translation, and post-translational modification. For simplicity, we refer to the assessed score as the *functional significance* (FS) score of each SNP. To formally define the scoring function for assessing the FS score, we first introduce basic notations.

Suppose that we are given p candidate SNPs on the target genomic region. Each SNP can be represented as a discrete random variable, X_i $(i = 1, .., p)$, whose possible values are the 4 nucleotides, $\{a, c, g, t\}$. The true (and *unknown*) functional category of SNP X_i is represented by another discrete random variable Y_i, whose value is 1 when SNP X_i is deleterious and 0 otherwise. We note that in most cases we do not know the true functional category Y_i of SNP X_i. We thus estimate it using q bioinformatics tools that predict, for each SNP X_i, the functional label (i.e., 'deleterious' or 'neutral') along four major bio-molecular functions: *protein coding*, *splicing regulation*, *transcriptional regulation*, or *post-translation modification*.

For each of the p SNPs and the q tools, we define two random variables, δ_{ij} and S_{ij} $(i = 1, ..., p\,;\, j = 1, ..., q)$. The variable δ_{ij} denotes the *functional label* assigned to the i^{th} SNP by the j^{th} tool, that is, $\delta_{ij} = 1$ when the j^{th} tool predicts SNP X_i to be deleterious, and 0 otherwise. The variable S_{ij} represents the *tool's own confidence score* with respect to the assigned label. The higher the value of S_{ij}, the more strongly the tool supports its own prediction, δ_{ij}. As different tools use different confidence scales, we define another random

variable, $\bar{S}_{ij}$, representing a normalized confidence score. The normalization procedure is explained in Section 6.3.3.

We also define a random variable, F_{jk}, to indicate the *bio-molecular functions that each tool examines*. We first define the set $\mathbb{F} = \{$'*protein coding*', '*splicing regulation*', '*transcriptional regulation*', '*post-translational modification*'$\}$ consisting of the four bio-molecular functions with which we are concerned. For each of the q tools and the four bio-molecular functions in $\mathbb{F}$, a random variable F_{jk} $(j = 1, ..., q,\ k \in \mathbb{F})$ is defined such that its value is 1 when the j^{th} tool examines the deleterious effects of SNPs on function k, and 0 otherwise.

Last, for each tool, we define a continuous random variable TR_j $(j = 1, ..., q)$, corresponding to the *tool reliability* (TR) score of the j^{th} tool. This score represents *how likely the tool is to correctly categorize SNPs as deleterious*. The computation procedure of the TR score is explained in Section 6.3.2.

Based on the parameters TR_j, F_{jk}, δ_{ij}, and $\bar{S}_{ij}$, the functional significance score of SNP X_i, denoted by FS_i, is defined as follows:

Definition 6.1. ***Functional Significance (FS) score of SNP*** X_i

$$FS_i \stackrel{\text{def}}{=} \max_{k \in \mathbb{F}} \frac{\sum_{j=1}^{q} F_{jk} \cdot TR_j \cdot (\delta_{ij} \cdot \bar{S}_{ij})}{\sum_{j=1}^{q} F_{jk} \cdot TR_j}.$$

That is, for each bio-molecular functional category k, we compute the *weighted average* of the *confidence* of each prediction tool with respect to the *deleterious* effect of the SNP, X_i, where the weight is the reliability score of each tool. We note that by multiplying by δ_{ij}, the confidence score of each tool is counted only when the tool predicts the SNP to be deleterious (that is, $\delta_{ij} = 1$). We also note that although summation is done over

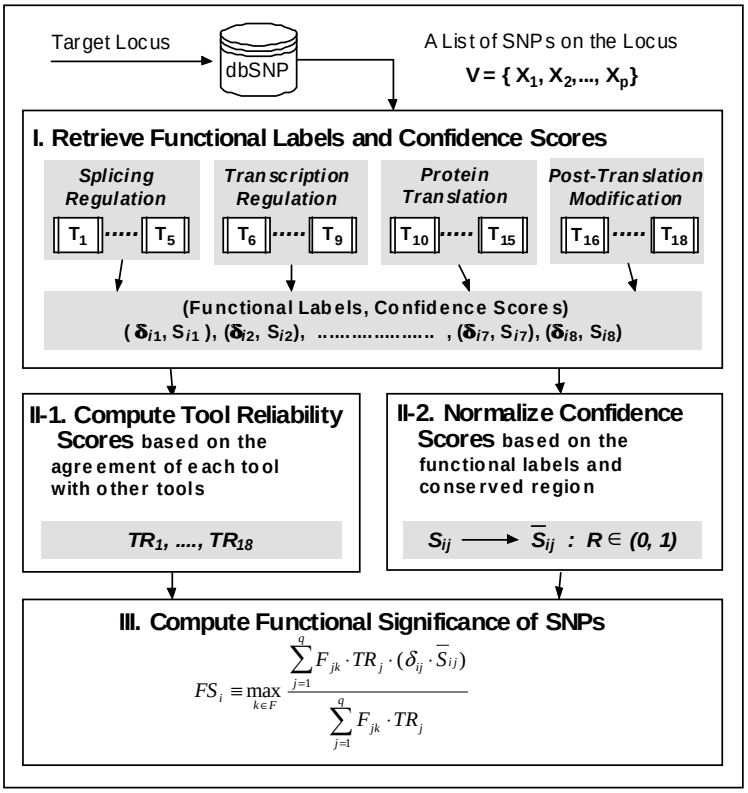


Figure 6.1: **Outline of the assessment process. In step I, we retrieve the predicted functional labels of SNPs from the different tools, along with their confidence scores. In step II, we compute tool reliability, and normalize the confidence scores. In step III, we compute the functional significance score of SNPs as shown in Definition 6.1.**

all the tools (that is, $j = 1$ to q, where q is the total number of tools, regardless of the bio-molecular functional category that each tool examines), F_{jk} allows only the tools that examine the specific bio-molecular functional category k to be considered. The maximum score, over all examined bio-molecular functions, is then assigned as the FS score for the SNP. In the next section, we describe the details of the implemented scoring system.

6.3 Methods for Assessing Functional Significance

Our system conducts three main steps to assess the functional significance (FS) score of SNPs. Figure 6.1 outlines the process. In step I, the functional labels (either, '*deleterious*' or '*neutral*') of SNPs, predicted by q bioinformatics tools, are retrieved. Confidence scores corresponding to the predictions are also retrieved when available. In step II-1, the reliability score of each tool is computed based on its tendency to agree with other tools' predictions. In step II-2, the confidence scores, obtained in step I, are normalized to a value between 0 and 1. In step III, the FS score of SNPs is computed as defined in Definition 6.1. We further describe each step in the following sections.

6.3.1 Retrieval of Predicted Labels and Confidence Scores

Given a set of p SNPs, $\{X_1, ..., X_p\}$, we first retrieve their predicted functional labels (that is, '*deleterious*' or '*neutral*') and corresponding confidence scores from *sixteen* publicly

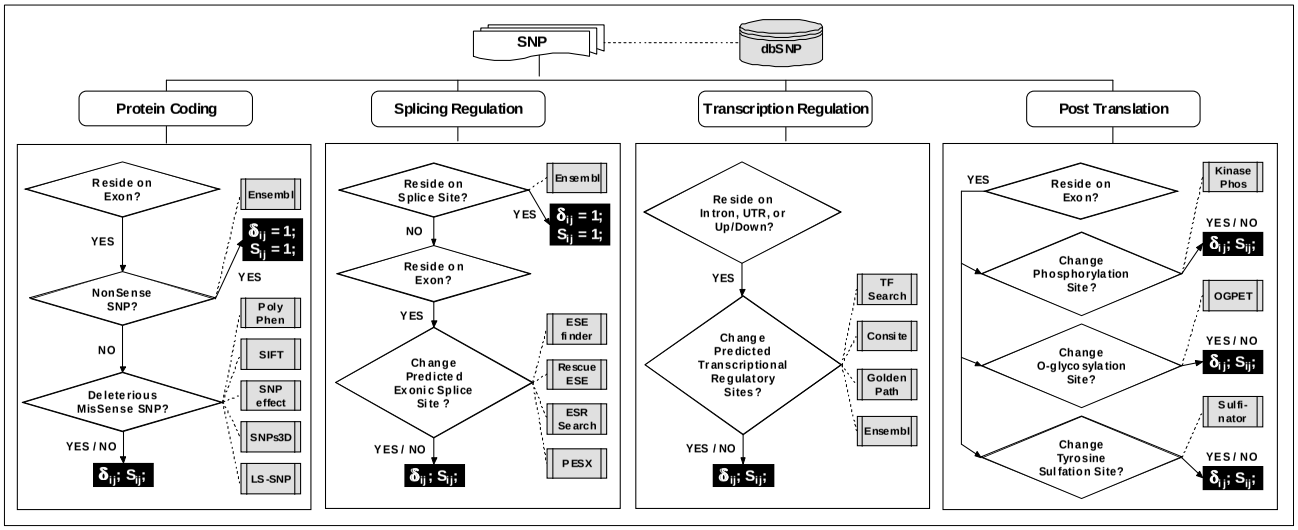


Figure 6.2: **The retrieval flow-chart for four major bio-molecular functional categories. Each SNP is examined for deleterious effects with respect to each functional category, namely, *protein coding*, *splicing regulation*, *transcriptional regulation*, and *post-translation modification* – as shown in the top part of the figure. For each category a series of tests is executed to determine whether the examined SNP has a functional impact. First the type (coding, intronic etc.) of the genomic region is identified, using data from db-SNP [167] and Ensembl [78]. Once this is determined, other tests are performed as shown in the flow-chart within each rectangle. Finally, the predicted functional label (either 'deleterious' or 'neutral') and the confidence score is retrieved from the used tools. For the Protein Coding category, the Ensembl [78] system is used to identify nonsense SNPs, and the web services, PolyPhen [147], SIFT [132], SNPeffect [153], SNPs3D [189], LS-SNP [88] are used to predict deleterious missense SNPs. For Splicing Regulation, Ensembl [78] is used to identify SNPs in canonical splice sites, and ESEfinder [22], RescueESE [187], ESRSearch [63], and PESX [200] are used to examine SNPs in exonic splice sites. For Transcriptional Regulation, TFSearch [3], ConSite [159], GoldenPath [101], Ensembl [78] are used to identify SNPs changing transcriptional regulatory sites. For Post-Translational Modification, KinasePhos [77], OGPET [59], and Sulfinator [124] are used.**

available web services and databases (that is, currently $q = 16$). Figure 6.2 illustrates the retrieval flow-chart. Most of the chart is the same as the prediction flow-chart for the F-SNP database service, shown in Figure 5.1. However, there are two aspects in which the flow-chart for the F-SNP-Score system differs from that associated with F-SNP.

First, at the end of each assessment process involving the j^{th} tool, the F-SNP-Score system obtains both the functional label, δ_{ij}, and the confidence score, S_{ij}, for SNP X_i $(i = 1, ..., p;\ j = 1, ..., q)$. We note, though, that as shown in Figure 6.2, in two cases the confidence score is always set to 1: (1) when the SNPs create a new start/stop codon or frameshift (that is, *nonsense* SNPs); or (2) when the SNPs occur in the first two or in the last two bases of intronic splice sites (that is, SNPs altering *canonical* splice sites). These SNPs are assigned the highest level of a confidence score, namely 1, because their deleterious effects to either '*protein coding*' or '*splicing regulation*' is unequivocal. Nonsense SNPs are often considered to have most deleterious effects, leading to a premature termination of amino acid peptides [185]. The change to the canonical splice sites is also known to be detrimental as suggested by the high selection pressure on the sites among mammalian genomes [18]. Other SNP prioritization studies [11, 188, 184] assign the highest rank or score of functional impact to these two kinds of SNPs as well.

Second, the F-SNP-Score system does not examine whether regulatory sites – predicted by the used tools – are conserved across multiple species to decide the functional significance of the SNPs in the region. As previously stated, this strategy is used by F-SNP's Classification system to filter out possible false positive predictions for regulatory sites due to their short DNA sequence. Instead, the F-SNP-Score system incorporates the information about conservative regions to normalize the confidence scores obtained from the used bioinformatics tools. We describe the details of this step in Section 6.3.3.

6.3.2 Computation of Tool Reliability

The tool reliability score, TR_j denotes how likely the j^{th} tool $(j = 1, ...q)$ is to correctly predict *deleterious* SNPs . We express the tool reliability score using the conditional probability as defined below:

$$TR_j \stackrel{\text{def}}{=} Pr(Y_i = 1 \mid \delta_{ij} = 1).$$

That is, the tool reliability score TR_j represents the likelihood that the true functional category of an arbitrary SNP X_i is '*deleterious*' when the j^{th} tool does predict so. If the true labels of SNPs, Y_i $(i = 1, ..., p)$ are known, this score can be estimated statistically. For example, using a maximum likelihood (ML) approach, TR_j can be estimated as the ratio between the number of correctly predicted deleterious SNPs and the total number of deleterious SNPs predicted by the tool, as follows:

$$\begin{aligned} TR_j &\stackrel{\text{def}}{=} Pr(Y_i = 1 \mid \delta_{ij} = 1) \\ &\approx \frac{\sum_{i=1}^{d} I_{Y\delta\, ij}}{\sum_{i=1}^{d} I_{\delta\, ij}}, \end{aligned}$$

where

$$I_{Y\delta\, ij} = \begin{cases} 1 \,:\, if \;\; Y_i = 1 \;\; and \;\; \delta_{ij} = 1\,; \\ 0 \,:\, otherwise\,, \end{cases}$$

$$I_{\delta\, ij} = \begin{cases} 1 \,:\, if \;\; \delta_{ij} = 1\,; \\ 0 \,:\, otherwise\,, \end{cases}$$

and d is the number of SNPs whose true functional label is known.

However, in most cases we do not know the true functional categories of SNPs. We thus estimate the probability $Pr(Y_i = 1 \mid \delta_{ij} = 1)$ using the theoretical work proposed by Long *et al.* on classification [113]. When class labels are unknown, they propose to estimate the prediction accuracy of a classifier based on the extent to which the classifier tends to agree with other classifiers. Long *et al.* have proven that the conditional probability $Pr(\delta_{ij} = 1 \mid Y_i = 1)$ can be calculated in this context as follows:

$$Pr(\delta_{ij} = 1 \mid Y_i = 1) = Pr(\delta_{ij} = 1) + \sqrt{\frac{(1 - Pr(Y_i = 1))}{Pr(Y_i = 1)} \cdot \frac{(u_{jm} - u_j \cdot u_m) \cdot (u_{jn} - u_j \cdot u_n)}{(u_{mn} - u_m \cdot u_n)}}\ , \tag{6.1}$$

where the variables m and n represent the indices of any two distinct tools $(m \neq n \neq j)$, $u_{jm} \stackrel{\text{def}}{=} Pr(\delta_{ij} = 1,\ \delta_{im} = 1)$, and $u_j \stackrel{\text{def}}{=} Pr(\delta_{ij} = 1)$. For the detailed proof of Equation (6.1), we refer the reader to the work by Long *et al.* [113].

Using Bayes' rule and Equation (6.1), we compute the tool reliability score of the j^{th} tool, TR_j , as follows:

$$\begin{aligned} TR_j &\stackrel{\text{def}}{=} Pr(Y_i = 1 \mid \delta_{ij} = 1) \ = \ (by\ Bayes'\ rule) \\ &= Pr(\delta_{ij} = 1 \mid Y_i = 1) \cdot \frac{Pr(Y_i = 1)}{Pr(\delta_{ij} = 1)} \ = \ (by\ substituting\ Eq.(6.1)\ ) \\ &= Pr(Y_i = 1) \ + \ \sqrt{\frac{Pr(Y_i = 1)}{(1 - Pr(Y_i = 1))^{-1}} \cdot \frac{(u_{jm} - u_j \cdot u_m) \cdot (u_{jn} - u_j \cdot u_n)}{(u_{mn} - u_m \cdot u_n)(u_j)^2}}\ . \end{aligned}$$

Note that we use the *same* prior probabilities, $Pr(Y_i = 1)$ and $Pr(\delta_{ij} = 1)$ for all SNPs X_i, and as such, the tool reliability score is independent of the SNP X_i. To estimate $Pr(Y_i = 1)$, which is the prior probability of any SNP X_i to be deleterious, we take a conservative

maximum likelihood approach. That is, for each tool assessing the effect of a SNP on a specific bio-molecular function k, the fraction of SNPs that are *unanimously* predicted to be deleterious by *all* the tools assessing the same function k is used as an estimate for $Pr(Y_i = 1),\ (1 \leq i \leq p)$. We estimate $Pr(\delta_{ij} = 1)$, which is the prior probability of the j^{th} tool to predict any SNP X_i to be deleterious, as the fraction of the examined SNPs that are predicted to be deleterious by the j^{th} tool.

6.3.3 Normalization of Confidence Scores

To account for the fact that different tools use different scales to report their confidence scores, we normalize the obtained confidence scores S_{ij} to be a value between 0 and 1. The normalization formula is as follows:

$$\bar{S}_{ij} = \frac{1}{2} \cdot \left(\delta_{ij} + (1 - C_{ij}) \cdot \frac{(S_{ij} - \min\limits_{i} S_{ij})}{(\max\limits_{i} S_{ij} - \min\limits_{i} S_{ij})} \right) ,$$

where C_{ij} is 1 if X_i resides on a *nonconserved* regulatory site, and 0 otherwise; $1 \leq i \leq p$ and $1 \leq j \leq q$.

Intuitively, when SNP X_i is predicted to be deleterious (i.e., $\delta_{ij} = 1$), the confidence score S_{ij} is converted to a value between 0.5 and 1. Otherwise (i.e., $\delta_{ij} = 0$), S_{ij} is converted to a value between 0 and 0.5. We note that for the SNPs occurring in regulatory regions, we examine whether the SNP position on the genome is *conserved* across multiple species (such as chimpanzee, dog, mouse, rat, chicken, zebrafish, and fugu) – information that is obtained from GoldenPath [101] – to reduce the effects of possible false-positive predictions. As previously mentioned in Sections 3.2.2 and 5.2.2, there is a high rate of false positive findings of regulatory sites by *in silico* prediction tools due to the relatively

short length of regulatory DNA sequences (typically 6- to 8-mers) [198]. Thus, when the predicted regulatory sites are not within a conserved region, the confidence score for the SNP in the region is set to 0.5, reflecting the uncertainty regarding the functionality of the region, and the consequent lack of confidence about potential deleterious effects of the SNPs on the regulatory function.

We also note that some prediction tools, such as SNPeffect [153] or LS-SNP [88], do not provide confidence scores. For these systems, we impute the confidence scores using the confidence scores for the same SNP obtained from other tools. Suppose that the j^{th} tool, which examines the possible effects of SNP X_i on the bio-molecular functional category k, does not provide a confidence score on its prediction. Among the other tools that provide the confidence scores for the same function k, we denote the index of the tool whose tool reliability score is highest as t. The imputed value is calculated as:

$$\bar{S}_{ij} = min\left(\frac{TR_j}{TR_t} \cdot S_{it},\ 1\right).$$

That is, when the j^{th} tool is more reliable than the t^{th} tool (i.e., $TR_j > TR_t$), its confidence score is imputed to be higher than that of the t^{th} tool, but not greater than one. Otherwise (i.e., $TR_j \leq TR_t$), the confidence score stays the same or becomes reduced proportionally to the ratio of the respective tool reliabilities.

Given the prediction results obtained in step I and the tool reliability and normalized confidence scores computed in step II, the functional significance (FS) score of SNP X_i can be computed as stated earlier in Definition 6.1.

6.4 Experiments and Results

We applied the F-SNP-Score system to 112,949 SNPs located on 580 disease-susceptible genes. The OMIM (Online Mendelian Inheritance in Man) database [69], which is one of the most widely-used databases of human genes and genetic disorders, provides the references to the biomedical literature that report the existence of SNPs – on these 580 genes – that are either *disease-causing* or *associated with common disorders*. The list of SNPs linked to the 580 genes, along with their primary information (such as genomic location), were downloaded from the dbSNP database, build 126 [167]. The number of known *disease-causing* or *disease-associated* SNPs [1] on these 580 genes is 1,399. The remaining 111,550 SNPs on the 580 genes are not yet identified to be disease-related. For simplicity, we refer to the former set of 1,399 SNPs as *disease-related* SNPs, and to the latter set of 111,550 SNPs as *neutral* SNPs. We note, however, that currently known disease-related SNPs can explain only a fraction of the genetic basis of human disease, and as such, the set of SNPs that is temporarily classified as *neutral* may still include functionally significant SNPs with deleterious effects that are *not yet* identified.

In Section 6.4.1, we summarize the scoring results by the F-SNP-Score system for all 112,949 SNPs, and show the distinguishing features of disease-related SNPs compared to neutral SNPs. In Section 6.4.2, we further validate that our integrative scoring system improves upon other state-of-the-art methods when applied to the same set of SNPs.

[1] The list of referenced disease-causing or disease-associated SNPs was obtained from the FTP service at ftp://ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/OmimVarLocusIdSNP.bcp.gz (downloaded April 14, 2007).

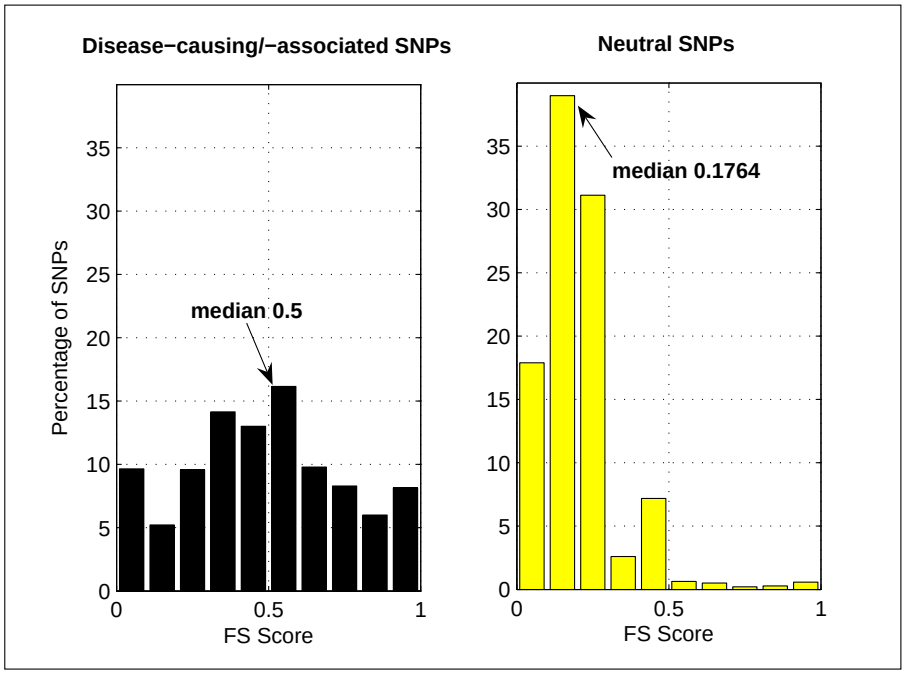


Figure 6.3: **The distribution of FS scores for disease-related SNPs and for neutral SNPs, assigned by F-SNP-Score. The X-axis represents the FS score for each group of SNPs, binned into 10 equal intervals, while the Y-axis represents the percentage of SNPs whose FS score is associated with that bin-score.**

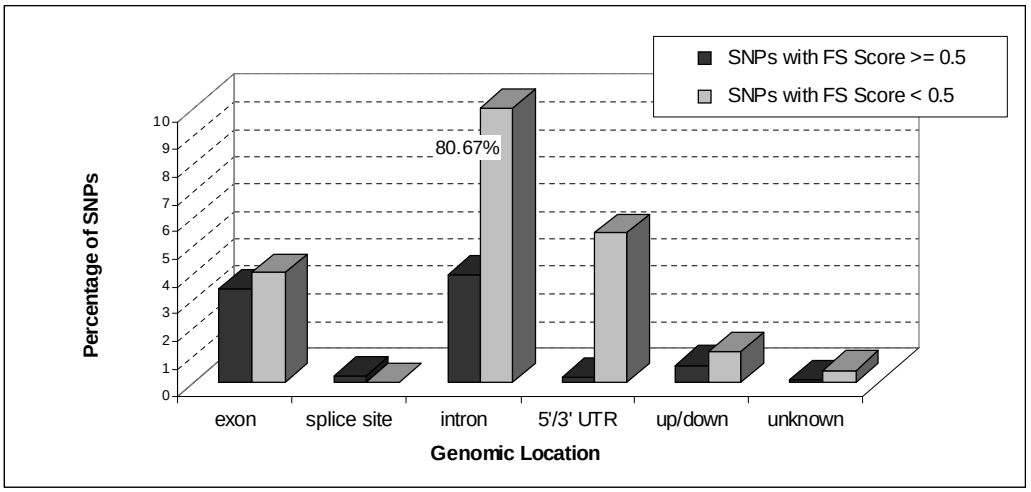
6.4.1 Review of the Scoring Results

First, we examine the scores that F-SNP-Score assigns to already known disease-related SNPs compared to neutral SNPs. Figure 6.3 shows the distribution of the FS score assigned to disease-related SNPs (shown on the left) along with that of SNPs currently assumed to be neutral (shown on the right). The figure clearly shows that the distribution of the FS scores for disease-related SNPs is significantly different from that of neutral SNPs on the same genes. The difference is also statistically significant, with a p-value of practically 0, according to the Kolmogorov-Smirnov two-side test with $\alpha = 0.05$. In particular, the median FS score for neutral SNPs is 0.1764, whereas, for disease-related SNPs, the median

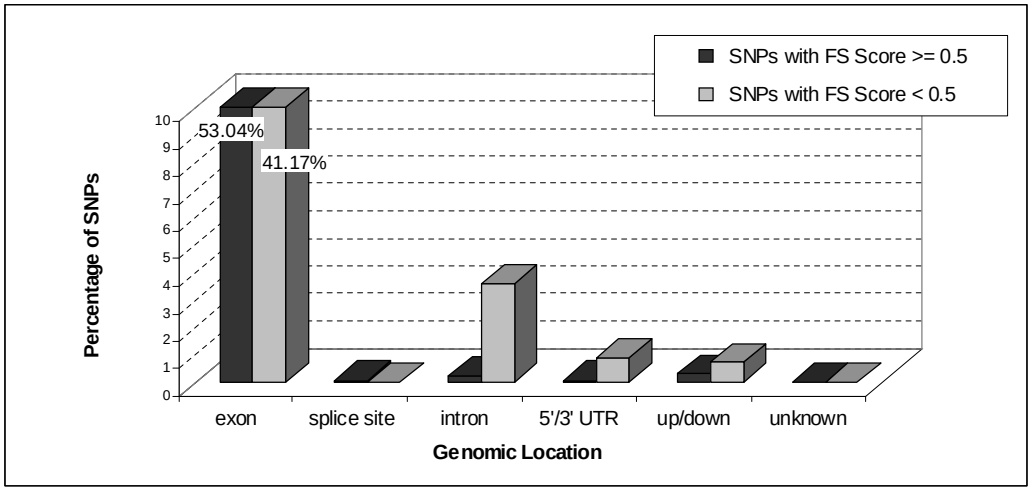
rises to 0.5. Moreover, 48.39% of disease-related SNPs are assigned an FS score greater than 0.5, whereas only 2.2% of neutral SNPs are assigned such a high score.

Some examples of known disease-related SNPs with a high functional significance score are as follow. Three SNPs on NAT2, namely, rs1799930, rs1801280, and rs1208 are included in the list of 1,399 disease-causing or disease-associated SNPs. The assessed FS scores for the three SNPs are 0.866, 0.584, and 0.858, respectively, designating high functional significance of the SNPs. In OMIM [69], all three SNPs are described to be relevant to slow acetylation activity [179], and rs1801280 has been reported to be strongly associated with susceptibility to bladder cancer and adverse drug reactions (OMIM ID: 243400) [138]. Other examples of high scoring, known disease-causing or disease-associated SNPs are rs7775 (FS score 1) and rs288326 (FS score 0.75), which have been reported as a strong risk factor for primary osteoarthritis of the hip in females (OMIM ID: 605083) [115].

Next, we examine the FS score distribution for SNPs based on their functional genomic regions. Figure 6.4-a shows the distribution for neutral SNPs, while Figure 6.4-b shows the same score distribution corresponding to disease-related SNPs. The X-axis denotes 6 types of genomic regions that are used in the decision procedure (shown in Figure 6.2). The Y-axis shows the percentage of SNPs occurring on each genomic region. To designate high FS scoring vs. low FS scoring SNPs on each region, we represent the percentage of SNPs whose FS score is at least 0.5 using black bars and the percentage of SNPs whose score is lower than 0.5 using gray bars. We note that the percentage of SNPs represented in each bar is calculated with respect to the whole set of neutral SNPs (in case of Figure 6.4-a) or to that of disease-related SNPs (in case of Figure 6.4-b). For clarity, the percentage is displayed only up to 10%.



(a) Neutral SNPs (111,550 SNPs)



(b) Known disease-related SNPs (1,399 SNPs)

Figure 6.4: **The distribution of low FS scoring vs. high FS scoring SNPs based on functional genomic locations. The X-axis denotes 6 types of genomic regions that are used in the decision procedure (shown in Figure 6.2), while the Y-axis shows the percentage of SNPs whose FS scores are at least 0.5 (black bars) vs. the percentage of SNPs whose scores are lower than 0.5 (gray bars) on each region type. We note that the percentage of SNPs represented in each bar is calculated with respect to the whole set of neutral SNPs (in case of Figure 6.4-a) or to that of disease-related SNPs (in case of Figure 6.4-b).**

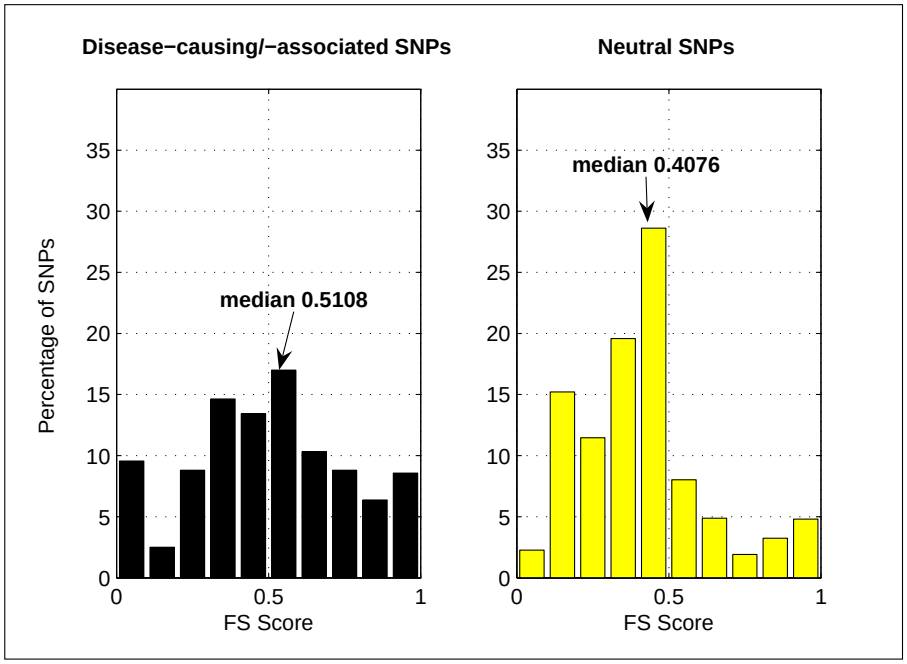


Figure 6.5: **The distribution of the assessed FS scores for *exonic* SNPs. The X-axis represents the FS score for each group of SNPs, binned into 10 equal intervals, while the Y-axis represents the percentage of SNPs whose FS score is associated with that bin-score.**

As shown in Figure 6.4-a, the majority of neutral SNPs are located within intronic regions, and the FS score for most intronic SNPs is lower than 0.5 (80.67%). A similar tendency is noted in 5'/3' untranslated regions (UTR), in regions that are upstream or downstream from genes, and in currently unspecified regions. In contrast, despite the relatively smaller number of SNPs on splice sites and on coding regions, these regions are enriched for high-scoring putatively deleterious SNPs. That is, an FS score of at least 0.5 is assigned to all SNPs in canonical splice sites and to 46.07% of the SNPs in coding regions. This scoring pattern is consistent with previous findings that mutations in splice sites and coding regions are likely to have direct impact on gene function [189, 22, 153]. It is thus highly likely that these SNPs with a high FS score may not be neutral at all, and future

disease-gene studies need to investigate them further.

In contrast, Figure 6.4-b shows the FS score distribution for disease-related SNPs as a function of their genomic regions. Unlike the case for neutral SNPs, most disease-related SNPs are located within exons (94.21%). This is indeed expected, as most association studies that validated these SNPs to be disease-related, have focused on protein coding SNPs, whose functional effects are relatively easy to pinpoint due to their direct impact on protein products. Aside for the outstanding proportion of exonic SNPs, disease-related SNPs show a similar scoring pattern to that of neutral SNPs. Most SNPs on intronic, 5'/3' UTR, and up/downstream regions are assigned an FS score lower than 0.5, but more than half of the SNPs on exonic regions (53.04%) and all SNPs on canonical splice sites are assigned an FS score of at least 0.5.

As is clear from the data shown above, most disease-related SNPs are located on exons, while most (currently assigned) neutral SNPs are located within introns. We thus need to examine whether the difference in FS-score distributions between the two sets of SNPs, shown in Figure 6.3, is merely an artifact of the difference in their genomic region. Figure 6.5 shows the distribution of assigned FS scores, this time only for 1,318 *exonic* SNPs that are already known to be disease-related (shown on the left) and for 8,228 *exonic* SNPs currently assumed to be neutral (shown on the right). As expected, the median score for exonic SNPs is higher than that of SNPs in all regions, both for disease-related SNPs and for neutral SNPs. Nevertheless, still only 22.86% of neutral exonic SNPs are assigned an FS score greater than 0.5, while the ratio rises to 56.30% for disease-related exonic SNPs. The Kolmogorov-Smirnov test with 5% significance level (that is, $\alpha = 0.05$) confirms that the two groups of exonic SNPs are unlikely to share a common score distribution (p-value 1.30e-079).

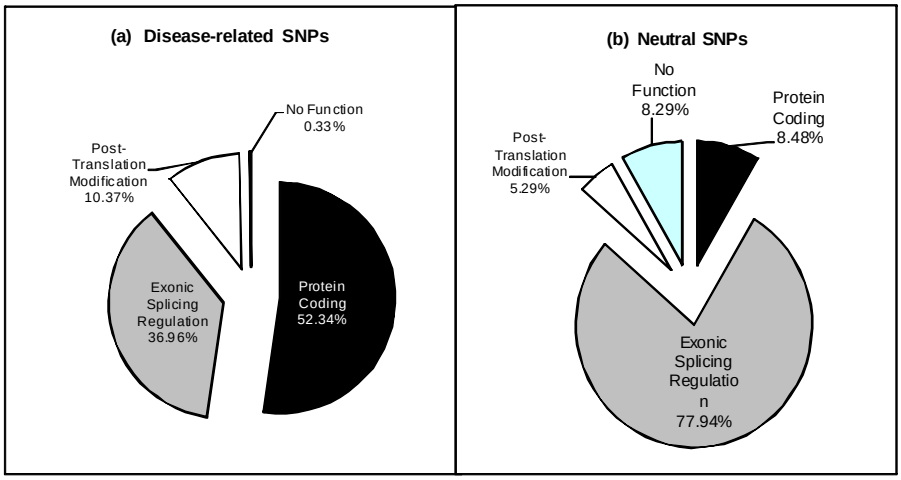


Figure 6.6: **The distribution of bio-molecular functions that the F-SNP-Score system predicts to be disrupted by disease-related exonic SNPs (shown on the left) and by neutral exonic SNPs (shown on the right).**

Last, we examine what kinds of bio-molecular functions that the F-SNP-Score system predicts the two groups of exonic SNPs mainly disrupt. Recall that SNPs in exonic regions may deleteriously affect either '*protein coding*', '*exonic splicing regulation*', or '*post-translational modification*' (as summarized in Figure 6.2), and F-SNP-Score assigns the maximum score over the three functional categories to each exonic SNP as its final FS score (as stated in Definition 6.1). We thus examine the proportion of the three predicted bio-molecular functions, which are used for assigning the final FS scores, among disease-related exonic SNPs and among neutral exonic SNPs.

Figure 6.6 summarizes the results. In the case of disease-related SNPs, more than half of the exonic SNPs are tagged by F-SNP-Score as affecting '*protein coding*', and about 37% of the SNPs are tagged as affecting '*exonic splicing regulation*'. Conversely, only 8.48% of neutral exonic SNPs are tagged as affecting '*protein coding*', while more than two thirds of them are tagged as affecting '*exonic splicing regulation*'. In either case, '*post-translational modification*' is rarely predicted as a source for potential deleterious

effects of SNPs.

6.4.2 Comparative Study

To validate that the F-SNP-Score system indeed improves upon state-of-the-art methods, we compare F-SNP-Score with three scoring methods for functional SNP selection that numerically assess putative deleterious effects of SNPs: SNPselector [184], FastSNP [188], and as a simple baseline, Simple Majority Vote. SNPselector and FastSNP are widely used public web-services for prioritizing functionally significant SNPs, while Simple Majority Vote is a baseline scoring scheme based on a majority vote. We briefly describe the three compared methods as follows.

SNPselector is a web-based SNP selection system. It prioritizes SNPs based on their tagging informativeness, SNP allele frequencies and source, function, regulatory potential and repeat status. In addition, SNPselector provides a numeric score for each SNP, called *function score*, which designates the possible effects of SNPs on gene transcript structure or on protein product. The score is a real number between 0.6 and 1.0; the higher the score is, the more deleterious the effects of the SNPs are expected to be.

FastSNP is another web-service for SNP function analysis and prioritization. It prioritizes high-risk SNPs according to their phenotypic risks and putative functional effects using 11 bioinformatics tools. FastSNP assigns to each SNP an integer score between 0 and 5, called *risk rank*, which quantifies how likely the SNP is to have functional effects leading to disease phenotypes.

Finally, as a baseline, we compute the functional significance score of SNPs using a simple majority vote. For example, when one third of the tools that examine the deleterious effects of SNPs on protein coding predict the SNP to be deleterious, a value of 1/3 is

assigned as its score with respect to the 'protein coding' category. This majority vote-based score is calculated for the other three functional categories, and the maximum value among the four scores is assigned as the FS score of the SNP. Our scoring scheme is distinguishable from this simple majority vote as it takes into account the certainty of each prediction (through normalized *confidence-scores*) as well as the reliability of each tool (through *tool-reliability-scores*). We refer to this baseline method as Simple Majority Vote.

To compare the three scoring schemes to ours, we generated test datasets using the following sampling procedure. For each disease-related SNP X_i, one neutral SNP is selected uniformly at random in the *same functional region* on the *same gene* as X_i. This selection is done for all disease-related SNPs. As a result, a dataset of 1,399 SNP-pairs, one disease-related and one randomly selected neutral, is generated. We repeat this procedure M times, generating M test datasets (here, M=100). We note that, by limiting the random selection to the same functional region on the same gene, we reduce any bias that may arise from sampling along different functional or chromosomal regions.

Using the test datasets, we examine how well each system distinguishes disease-related SNPs from neutral SNPs. Intuitively, a better scoring system would assign a higher functional score to disease-related SNPs than to neutral SNPs. First, we measure this tendency by directly computing the percentage of disease-related SNPs that are assigned a higher functional significance score than their paired, randomly selected neutral SNPs, averaged over M test datasets. We refer to this measure as *Higher Score* (%).

Second, instead of directly comparing each SNP pair, we compare the distribution of FS scores for disease-related SNPs with the score-distribution for neutral SNPs. The paired t-test can examine the hypothesis whether two score distributions share the same mean. We thus separately conduct the paired t-test on each of the M datasets, and count the number

Table 6.1: **The results of a comparative study based on two evaluation measures: Higher Score and Paired T-Test. Higher Score directly computes the percentage of disease-related SNPs that are assigned a higher functional significance (FS) score than their paired, randomly selected neutral SNPs, averaged over M test datasets. Paired T-Test computes the percentage of M t-tests that rejected the hypothesis whether the FS score distribution for disease-related SNPs is different from that for neutral SNPs.**

	EVALUATION MEASURE	
SYSTEM	Higher Score	Paired T-Test (avg. p-value)
F-SNP-Score	63.82 %	1.00 (0.00)
FastSNP	61.15 %	1.00 (3.61e-127)
SNPselector	55.39 %	1.00 (6.91e-125)
Simple Majority Vote	45.42 %	0.93 (0.01)

of times that the hypothesis is rejected by the t-test along with their average p-value. The rejection implies that the FS score distribution of disease-related SNPs is distinct from that of likely neutral SNPs. Therefore, scoring schemes with a high proportion of rejections are preferred. We refer to this second measure as *Paired T-Test*.

Table 6.1 summarizes the results of the comparative study. Overall, F-SNP-Score improves upon all the compared systems. According to the Higher Score measure, F-SNP-Score assigns higher functional significance scores to about 64% of known disease-related SNPs than to neutral SNPs. FastSNP comes second, and SNPselector and Simple Majority Vote follow. The score difference between F-SNP-Score and the compared systems is also statistically significant (p-values are 6.96e-038, 4.82e-105, and 5.26e-174 for FastSNP, SNPselector, and Simple Majority Vote, respectively, using the paired t-test, $\alpha = 0.05$). It

is notable that F-SNP-Score greatly improves upon Simple Majority Vote, which demonstrates the utility of the confidence and the tool reliability scores, integrated into our scoring scheme.

In the case of the Paired T-Test measure, the first three systems, namely, F-SNP-Score, FastSNP, and SNPselector perform the same; all of the paired t-tests rejected the hypothesis of the same mean for disease-related SNPs and neutral SNPs with a significance level of at least 0.05. However, the average p-value of the rejected hypotheses is smallest (that is, practically zero) for F-SNP-Score among the three, which means that the score distribution of disease-related SNPs and that of neutral SNPs are most disparate when their FS scores are assigned by F-SNP-Score. In the case of Simple Majority Vote, only 93% of the paired t-tests rejected the hypothesis of the same mean. The average p-value for the rejected hypotheses is also the largest among all the compared systems.

6.5 Discussion

We have presented a new integrative scoring system, F-SNP-Score, for assessing the putative deleterious effects of SNPs. The F-SNP-Score system combines assessments from multiple independent computational tools, using a probabilistic framework that takes into account the certainty of each prediction as well as the reliability of different tools. An empirical study over 580 disease-associated genes taken from the OMIM database shows that F-SNP-Score provides distinct scoring patterns that are consistent with well-established findings about functional SNPs. A comparative study based on two evaluation measures also shows that F-SNP-Score improves upon other SNP scoring systems in terms of distinguishing known disease-related SNPs from likely neutral SNPs.

Two main features distinguish F-SNP-Score from others. First, we integrate multiple

tools to overcome the incomplete or erroneous predictions of individual prediction tools. While a single tool may fail to capture the deleterious effects of many SNPs, a combination of multiple independent tools, which are based on different resources and algorithms, are less likely to all make the same error. Thus the tools are likely to complement each other, and typically compensate for each other's errors. As a result, the effect of possible false-negative or false-positive predictions in any single tool is reduced when computing the combined FS score. Our improved results suggest that this hypothesis indeed holds in practice.

Second, unlike other scoring systems, we take into account the reliability of different tools as well as the certainty of each prediction made by the tools. To the best of our knowledge, this is the first SNP prioritization approach to measure the reliability of individual tools and to use this information along with the confidence scores obtained from each tool.

We note, though, that the FS score assigned by F-SNP-Score to about 45% of disease-related SNPs is still below 0.5. There are two possible explanations for this seemingly inappropriate FS score. First, even though some SNPs, obtained from the OMIM database, show a positive statistical correlation with common disorders in some association studies, these SNPs may not all be actual disease-causing mutations. Some of these SNPs may represent false positive findings, or may simply be correlated with actual disease-causing mutations. Our future study will focus on investigating the actual disease-causing mutations that could be located near SNPs known to be disease-related with low FS scores.

Second, while the disease-related SNPs may indeed be disease-causing mutations, our current scoring scheme may not capture them properly. For example, in addition to the biomolecular functions that we currently examine, there could be other genetic mechanisms

that have a profound impact on human pathogenesis. We thus plan to update F-SNP-Score through combining other epidemiological resources, such as literature information, as well as integrating more prediction tools for each bio-molecular function.

Chapter 7

Two Birds, One Stone: Selecting Functionally Informative Tag SNPs

In the previous chapters, we have introduced three SNP selection systems, namely BN-Tagger, F-SNP (and its classification system, F-SNP-C) and F-SNP-Score: the first is developed for tag SNP selection, while the latter two are for functional SNP selection. In this chapter, we describe the first integrative SNP selection method for identifying SNPs that are *both* informative tagging and functionally significant. We provide the motivation for the proposed system in Section 7.1. In Section 7.2, we formally define the problem of functionally informative tag SNP selection as a multi-objective optimization problem, and introduce the basic notation used throughout this chapter. Section 7.3 describes a heuristic selection algorithm based on incremental, greedy search. Section 7.4 reports the evaluation results of the proposed system using a comparative study, and Section 7.5 concludes this chapter and outlines future directions.

7.1 Motivation and Objectives

We propose an integrative SNP selection method that supports both tag SNP selection and functional SNP selection within one selection process. Despite their distinct merits, neither tag SNP selection nor functional SNP selection shares the other's advantage. That is, methods for the selection of informative tag SNPs do not take into account the functional significance of SNPs; Similarly, methods for identifying functionally significant SNPs do not attempt to capture the allele information of the complete target locus.

As a result, there have been a few efforts to support these two SNP selection approaches within one selection framework [184, 30, 73]. Typically, these systems view the identification of informative tag SNPs and of functionally significant SNPs as two distinct optimization problems, and address each selection problem independently. That is, they separately conduct tag SNP selection and function-based SNP selection, and combine the two selected sets as a last step. A major shortcoming of such systems, in addition to the ad-hoc nature of the combination, is that the number of selected SNPs can be much larger than necessary.

To address this limitation, we propose an integrative SNP selection system that simultaneously identifies SNPs that are *both* informative tagging and carry a deleterious functional effect – which in turn means that they are likely to be disease-related. We formulate SNP selection as a multi-objective optimization problem, to which we refer as *functionally informative tag SNP selection*. We define a single objective function, incorporating both allelic information and functional significance of SNPs, and present a heuristic selection algorithm that we show, through a comparative study, to improve upon other state-of-the-art systems. To the best of our knowledge, the idea of combining the two notions of SNP selection – the function-based and the information-based – into a single optimized selection process is new, and was not attempted before.

In the next section, we formally define the problem of functionally informative tag SNP selection, and introduce the basic notation used throughout the chapter.

7.2 Problem Definition

We aim to select a subset of at most *k* SNPs on the target locus (where k is a pre-specified number) whose allele information is as informative as that of the whole set of SNPs, while including those SNPs that are most functionally significant. We refer to the problem as *functionally informative tag SNP selection*. Before we formulate and address this problem, we first introduce basic notation.

Suppose that our target locus contains p consecutive candidate SNPs. As previously stated, we represent each SNP as a discrete random variable, X_j $(j = 1, ..., p)$, whose possible values are the 4 nucleotides, $\{a, c, g, t\}$. For each value $x \in \{a, c, g, t\}$, there is a probability $Pr(X_j = x)$ that the nucleotide *x* is assigned to the genomic position of SNP X_j. Let V=$\{X_1, ..., X_p\}$ denote the set of random variables corresponding to the p SNPs. We are given a haplotype dataset, D, containing the allele information of n haplotypes, each of which consists of the p SNPs in V. As stated in Chapter 4, the set D can be viewed as an n by p matrix; each row, D_{i-}, in D corresponds to the allele information of the p SNPs comprising haplotype h_i, while each column, D_{-j}, corresponds to the allele information of SNP X_j in each of the n haplotypes. We denote by D_{ij} the allele information of the j^{th} SNP in the i^{th} haplotype. To formally address functional significance of SNPs, we denote by e_j the functional significance score for each SNP X_j in V, and define $E = \{e_1, ..., e_p\}$ to be the set of scores for all the SNPs on the target genomic locus. We explain how these values are obtained in Section 7.4.

For a subset of SNPs, $T \subset V$, we define an objective function, $f(T|D, E)$, to reflect

both the allele information carried by the SNPs in T about the remaining SNPs in $V - T$, and the functional significance of the SNPs in T. The problem of *functionally informative tag SNP selection* can then be stated as follows:

Problem	:	Functionally Informative Tag SNP Selection
Input	:	A set of SNPs, V; A maximum number of SNPs to select, k; A haplotype dataset, D; A set of functional significance scores, E;
Output	:	A set of SNPs $T = \underset{T \text{ such that } T \subset V \ \& |T| \leq k}{argmax} f(T|D, E)$.

That is, to select a subset of functionally informative tag SNPs, we need to find among all possible subsets of the original SNPs in the set V, an optimal subset of SNPs, T, of size $\leq k$, based on the objective function $f(T|D, E)$.

Our first task is to define the objective function, $f(T|D, E)$. To do so, we first introduce two simpler objective functions, denoted by $f_1(T|D)$ and $f_2(T|E)$; the former measures the allelic information, while the latter measures the functional significance of a SNP set T, based on the haplotype data D and the functional significance score set E, respectively.

Definition 7.1. Information-based Objective : *Given a set of p candidate SNPs,* $V = \{X_1, ..., X_p\}$, *a subset of k SNPs,* $T = \{X_{t_1}, ..., X_{t_k}\}$ $(T \subset V)$, *and a dataset* D *of* n *haplotypes, we define an information-based objective function,* $f_1(T|D)$, *as*:

$$f_1(T|D) = \frac{1}{np} \sum_{j=1}^{p} \sum_{i=1}^{n} I(X_j,\ T,\ D_{i-})$$

where

$$I(X_j,\ T,\ D_{i-}) = \begin{cases} 1\ :\ \text{if}\ X_j \in T \quad \text{or} \\ \qquad\qquad D_{ij} == \underset{x \in \{a,c,g,t\}}{argmax}\ Pr(X_j = x\ | X_{t_1} = D_{it_1}, ..., X_{t_q} = D_{it_k})\,; \\ 0\ :\ \text{otherwise}\,. \end{cases}$$

Recall that this prediction indicator function, $I(X_j,\ T,\ D_{i-})$, has been previously defined for our tag SNP selection method in Chapter 4, Eq. 4.1. The function I returns 1 if the SNP X_j is selected as a tag SNP (i.e., $X_j \in T$) or if its allele in the i^{th} haplotype (i.e., D_{ij}) is correctly predicted based on the allele information of the SNPs in T. We note that, by using the conditional probability expression, the allele value assigned to D_{ij} is the one that is most likely to occur given the allele information of the predictive tag SNPs in T. Otherwise, the function I returns 0. To summarize, the allelic information provided by a SNP set, T, with respect to a given haplotype dataset D, is measured by the average proportion of the correctly predicted alleles of each SNP, X_j, given the allele information of the SNPs in T.

This information-based objective function, $f_1(T|D)$ follows the *prediction-based tag SNP selection approach*, which aims to select a subset of SNPs (i.e., tag SNPs) that can best predict the alleles of the remaining, unselected SNPs (i.e., tagged SNPs) [7, 67, 164]. This approach is appealing since: (1) it does not require prior block partitioning [7]; (2) it tends to select a small number of SNPs [8]; and (3) it works well even for genomic regions with low linkage disequilibrium [107]. An in-depth discussion and survey of information-based tag SNP selection approaches was given in Section 3.1 and in other reviews [105, 66].

Definition 7.2. Function-based Objective : *Given a set of p candidate SNPs,* $V = \{X_1, ..., X_p\}$, *a set of k SNPs,* $T \subset V$, *and a set of functional significance scores,* $E =$

$\{e_1, ..., e_p\}$, *we define a function-based objective function*, $f_2(T|E)$ *as*:

$$f_2(T|E) = \frac{\sum_{j=1}^{p} e_j \cdot I_T(X_j)}{\sum_{j=1}^{p} e_j}$$

where

$$I_T(X_j) = \begin{cases} 1 \ : \ \text{if} \ \ X_j \in T \, ; \\ 0 \ : \ \text{otherwise} \, . \end{cases}$$

That is, the functional significance of a SNP set T is the normalized sum of the functional significance of SNPs in T.

Based on the two functions defined above, we next define a single objective function, $f(T|D, E)$, incorporating both allelic information and functional significance.

Definition 7.3. Functionally Informative Objective Function : *Given a set of k SNPs*, $T \subset V$, *a haplotype dataset*, D, *a functional significance score set*, $E = \{e_1, ..., e_p\}$, *and a parameter value*, α $(0 \leq \alpha \leq 1)$, *we define the functionally informative (FI) objective function*, $f(T|D, E)$ *as*:

$$f(T|D, E) = \alpha \cdot f_1(T|D) + (1 - \alpha) \cdot f_2(T|E) \, .$$

The parameter α is a weighting factor, which allows us to adjust the importance of information-based selection with respect to that of functional significance. In the work described here, we assign an equal weight to the two criteria, that is, $\alpha = 0.5$. We refer to the value assigned by this function to the subset of SNPs T, as the *FI-score* of T.

To summarize, we are looking for a subset of at most k SNPs, T, that is both functionally significant and likely to correctly predict the remaining SNPs in $V - T$. Bafna *et al.* [7]

have previously shown that finding k most informative tag SNPs is NP-complete. Based on this, we take it as a conjecture that the current problem is also NP-complete. In the next section, we thus introduce a heuristic algorithm to address the problem.

7.3 Algorithm for Selecting Functionally Informative Tag SNPs

Our selection algorithm takes an incremental, greedy approach. It starts with an empty tag SNP set, T, and iteratively adds one SNP to T until a maximum number, k, of SNPs are selected. Each greedy selection step identifies a SNP whose addition to T will result in the maximum increase in the value of the functionally informative objective function (FI-score) with respect to the current tag SNP set, T.

We first explain the basis for our greedy incremental selection process. Let $T^{(m)}$ denote the set of m selected SNPs after the m^{th} iteration, where $m = 0, ..., k$ and $T^{(0)} = \emptyset$. The FI-score of $T^{(m)}$ was defined in Definition 8.2 as follows:

$$\begin{aligned} f(T^{(m)}|D,E) &= \alpha \cdot f_1(T^{(m)}|D) + (1-\alpha) \cdot f_2(T^{(m)}|E) \\ &= \sum_{j=1}^{p} \left[\alpha \cdot \left(\frac{1}{np} \cdot \sum_{i=1}^{n} I(X_j, T^{(m)}, D_{i-}) \right) + (1-\alpha) \cdot \left(\frac{e_j}{\sum_{l=1}^{p} e_l} \cdot I_{T^{(m)}}(X_j) \right) \right]. \end{aligned}$$

Note that the FI-score of $T^{(m)}$ is the weighted sum of the allelic information of $T^{(m)}$ and the functional significance of $T^{(m)}$ for each SNP X_j $(j = 1, ..., p)$. For simplicity, we denote the contribution of each SNP X_j to the FI-score of $T^{(m)}$ as $f_j(T^{(m)}|D, E)$, and refer to it

as the FI-score of X_j with respect to $T^{(m)}$. That is,

$$f_j(T^{(m)}|D,E) \;=\; \left[\alpha \cdot \left(\tfrac{1}{np} \cdot \sum_{i=1}^{n} I(X_j, T^{(m)}, D_{i-})\right) + (1-\alpha) \cdot \left(\frac{e_j}{\sum_{l=1}^{p} e_l} \cdot I_{T^{(m)}}(X_j)\right)\right],$$

and

$$f(T^{(m)}|D,E) \;=\; \sum_{j=1}^{p} f_j(T^{(m)}|D,E)\,.$$

In the next iteration, $m+1$, we aim to select a SNP, $X^{(m+1)}$, whose addition to $T^{(m)}$ will maximally increase the FI-score. Using the FI-score of X_j with respect to $T^{(m)}$, $f_j(T^{(m)}|D,E)$, defined above, this goal can be stated as follows:

$$X^{(m+1)} \;=\; \underset{X \in\, (V - T^{(m)})}{argmax} \sum_{j=1}^{p} \left(\, f_j(\, T^{(m)} \cup \{X\} \mid D,E) - f_j(T^{(m)}|D,E) \,\right).$$

Our algorithm is outlined in Table 7.1. It starts with an empty set of tag SNPs, T, and computes the FI-score of each SNP with respect to the current set T. We note that although no SNP is currently selected, our algorithm can still predict the allele information of SNPs, and can thus lead to a different FI-score for each SNP. The reasoning is that in this initial case where T is empty, the posterior probability, $Pr(X_j|T)$, shown in the definition of the function I within Definition 7.2, is simply the prior probability, $Pr(X_j)$. That is, we always predict the alleles of X_j, $D_{ij}(i = 1, ..., n)$, as the major allele of the SNP. This approach is taken because it maximizes the expected prediction accuracy when no other information is given. At each subsequent iteration, the SNP that leads to the maximum increase in the FI-score is selected and added to T. The FI-score for each SNP is updated based on the augmented set T and used in the next iteration. This procedure is repeated until the set T

Input: A set of SNPs, V;
A maximum number of SNPs to select, k;
A haplotype dataset, D;
A set of functional significance scores, E;

Output: A set of tag SNPs, T;

$m \leftarrow 0$;
$T^{(m)} \leftarrow \emptyset$;
For each SNP $X_j \in V$
$FI_j \leftarrow f_j(T^{(m)}|D, E)$;
While $m < k$
For each t where $X_t \in V - T^{(m)}$
$\Delta_t^{(m)} \leftarrow \sum_{j=1}^{p} \left(f_j(T^{(m)} \cup X_t|D, E) - FI_j\right)$;
$X^{(m+1)} \leftarrow \underset{X_t \in V - T^{(m)}}{argmax}\ \Delta_t^{(m)}$;
$T^{(m+1)} \leftarrow T^{(m)} \cup X^{(m+1)}$;
For each $X_j \notin T^{(m)}$
$FI_j \leftarrow f_j(T^{(m+1)}|D, E)$;
$m \leftarrow m + 1$;
$T \leftarrow T^{(m)}$;

Table 7.1: **The incremental, greedy algorithm for selecting functionally informative tag SNPs.**

contains the pre-specified number of SNPs, k.

The time complexity of *each* incremental greedy selection is $O((p-m)^2 \cdot n)$, where $p-m$ is the number of SNPs that can be selected, and n is the number of haplotypes in a dataset D. As this iteration is repeated for $m = 0$ to $m = k-1$, the overall complexity of our algorithm is $O(k \cdot n \cdot p^2)$.

7.4 Experiments and Results

We compare the performance of the proposed integrative SNP selection method with that of two state-of-the-art selection systems supporting both tag SNP selection and functional SNP selection: TAMAL [73] and SNPselector [184]. For simplicity, we refer to the proposed method as FITS-Select (Functionally Informative Tag SNP Selector). In the following sections, we summarize the experimental setting of the comparative study, and report the evaluation results.

7.4.1 Experimental Setting

For evaluation, we have selected 14 genes that are involved in the etiology of common and complex diseases according to the OMIM database [69] and have disease-related SNPs identified and recorded by the HapMap Project [33]. To identify the candidate genes, we scanned the OMIM database for several major common and complex diseases, including diabetes, cancer, hypertension, and heart disease. The retrieved genes were then scanned to find those that have SNPs with possible deleterious functional effects reported in the biomedical literature and also have haplotype information available from the HapMap consortium [33]. From the genes satisfying these criteria, 14 were selected at random.

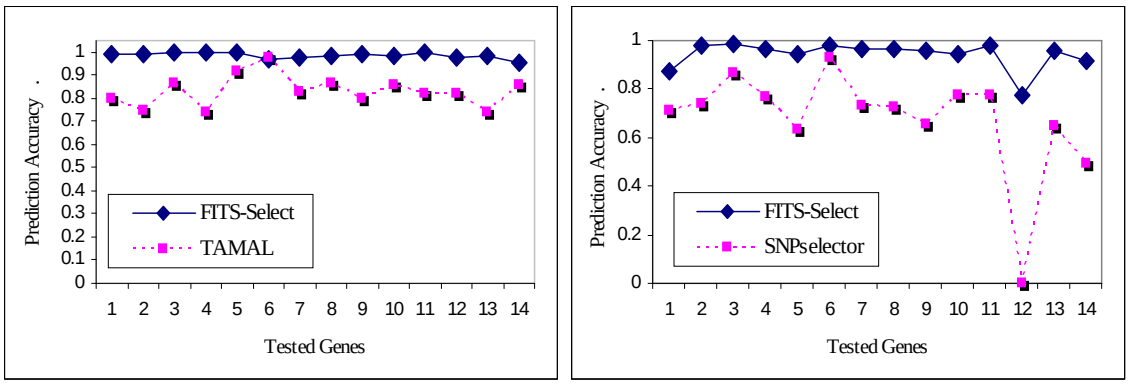
Table 7.2: **Summary of 14 test datasets. Linkage disequilibrium (LD) is estimated by the multi-allelic extension of Lewontin's LD, D' [72]. The number of SNPs selected by TAMAL [73] and by SNPSelector [184] are shown in the right column.**

Gene	Associated Disease	Locus	LD (D')	SNP #	Selected SNP #	
					TAMAL	SNPSel.
ADD1	Hypertension	4p16.3	0.7718	60	16	1
BRCA2	Breast Cancer	13q12.3	0.7657	106	28	13
CMA1	Hypertension	14q11.2	0.8361	20	6	4
ELAC2	Prostate Cancer	17p11	0.8336	35	13	2
ERBB2	Prostate Cancer	17q21.1	0.8104	8	6	1
F7	Heart Disease	13q34	0.8629	13	8	5
HEXB	Mental Retardation	5q13	0.7371	51	10	5
ITGB3	Heart Disease	17q21.32	0.6491	83	20	8
LEPR	Diabetes	1p31	0.7048	245	46	11
LTA	Heart Disease	6p21.3	0.7865	12	4	2
MSH2	Colon Cancer	2p22-p21	0.8413	51	18	4
NOS3	Alzheimer Disease	7q36	0.6183	16	7	0
PTPRJ	Colon Cancer	11p11.2	0.7863	115	32	7
TP53	Colon Cancer	17p13.1	0.7154	9	5	2

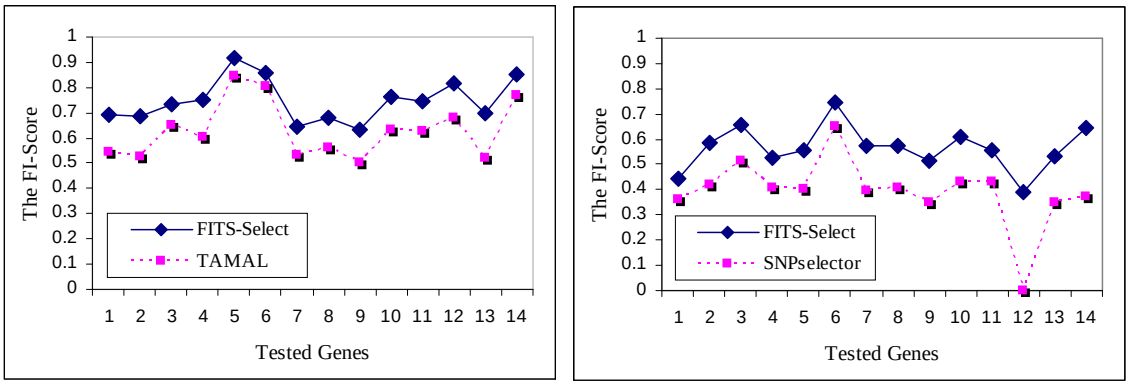
Table 7.2 provides the genetic characteristics of the 14 genes and their associated disease. Linkage disequilibrium (LD) is estimated by the multi-allelic extension of Lewontin's LD, D' [72]. The list of SNPs linked to the genomic location of each gene – including $10k$ upstream/downstream – was downloaded from the dbSNP database, build 126 [167]. The phased haplotype datasets for the SNPs were downloaded from the HapMap consortium website [33] for the CEU population (HapMap public release #20/phaseII). When no haplotype information exists for SNPs, the SNPs were excluded from the analysis. We have also downloaded the functional significance scores of the SNPs from the F-SNP database [109].

We compare our system with two state-of-the-art SNP selection systems that support both tag SNP selection and function-based SNP selection: TAMAL [73] and SNPselector [184]. The two systems share the same goal with our system, namely, selecting a set of tag SNPs, with significant functional effects on the molecular function of the genes, for association studies. However, they differ from our system in the assessment process for the functional significance of SNPs, the integrated bioinformatics tools, and the criteria used for selecting SNPs. Moreover, they conduct tag SNP selection and function-based SNP selection in two separate consecutive steps, while we address it as a single optimization problem.

As evaluation measures, we use Halperin's prediction accuracy [67] and the FI-score, introduced in Definition 8.2, (we note that the two systems to which we compare do not provide an evaluation measure). To compare the performance of the systems using the two measures, the SNP sets selected by each of the compared systems must include an equal number of SNPs. However, unlike our system, TAMAL and SNPselector do not allow the user to specify the number of selected SNPs, but rather calculate a subset of SNPs and provide it as their output. Thus, when they do not select the same number of SNPs for the same gene, they cannot be directly compared. Hence, for a fair comparison, we first apply each of the compared systems to each of the 14 test datasets, and then use our system on the same dataset to select the same number of SNPs as selected by the compared system. We then compute the two evaluation measures for the sets selected by each of the systems, and compare the resulting scores. The number of SNPs selected by TAMAL and SNPselector for the 14 tested genes is shown in Table 7.2. To ensure robustness of the results obtained from our system, we employ 10-fold cross validation 10 times, each using a randomized 10-way split of the n haplotypes. In all cases, the average performance is used in the



(a) The prediction accuracy of the selected tag SNPs for each gene



(b) The FI-score of the selected tag SNPs for each gene

Figure 7.1: **The performance of our system and the compared systems for 14 gene datasets.**

comparison.

7.4.2 Test Results

Figure 7.1 shows the performance of FITS-Select compared with TAMAL (left) and with SNPselector (right). The X-axis represents the 14 genes in an alphabetical order of their names, as listed in Table 7.2. In Figure 7.1-a (top), the Y-axis shows Halperin's prediction accuracy [67], and in Figure 7.1-b the Y-axis shows the FI-score for the selected SNP set

of the corresponding gene.

FITS-Select (upper solid line with diamonds) consistently outperforms the other two systems, TAMAL and SNPselector (lower dotted line with rectangles) on both evaluation measures. The performance difference in all cases is statistically significant, as confirmed by the Wilcoxon rank-sum test (p-values are 1.144e-005 and 4.7e-003 with respect to the TAMAL system and 1.7382e-005 and 5.6780e-004 with respect to the SNPselector system). We note that optimizing the FI-score when selecting SNPs does not compromise the predictive power of the SNPs selected by FITS-Select, that is, our selected SNPs still have a high prediction accuracy according to Halperin's original measure as demonstrated by Figure 7.1-a.

7.5 Discussion

We have presented a first integrative SNP selection system, FITS-Select, that simultaneously identifies SNPs that are both highly informative in terms of providing allele information for the target locus, and are of high functional significance. Our main contributions include the formulation of the problem of functionally informative tag SNP selection as a multi-objective optimization problem and presenting a heuristic selection algorithm to address the problem. An empirical study over a set of 14 disease-associated genes shows that our system improves upon current state-of-the-art systems.

While we have presented improved results, there are a number of limitations and possible extensions needed for this work. First, our SNP selection method, FITS-Select, assumes that both haplotype data and functional significance scores are given for *all* candidate SNPs to be examined. Therefore, when conducting the comparative study described in Section 7.4, we excluded SNPs that have no haplotype data or functional significance (FS) scores.

Computational methods, such as imputation, are needed to include these SNPs in the analysis.

Second, we presented a simple greedy search algorithm in which both tagging informativeness and functional significance of SNPs are incorporated into a single objective function expressed as a weighted sum [108]. Although the weighted sum approach has been widely used to solve multi-objective optimization problems, it is still limited by the fact that the selected SNP set depends on the predefined weighting factors. In the work described above, we have set an equal weight to the informative and the functional objective. It will be interesting to change the weight, and to see if and how the selection varies according to the given weight in the scoring function.

Third, we demonstrated the utility of our multi-objective SNP selection framework combining two objective functions, f_1 – based on tagging informativeness, and f_2 – based on function. However, our selection framework is general in a sense that other types of SNP selection criteria can be incorporated into it as well. For example, our information-based objective function, f_1, is currently defined following the tagged SNP prediction-based approach. It will be interesting to examine whether our selection framework works well with objective functions based on other tag SNP selection approaches, such as pairwise linkage disequilibrium (LD).

Finally, it is also interesting to apply other multi-objective optimization approaches, and to compare the selected SNP sets as well as their performance. In the next chapter, we present our second integrative SNP selection system that addresses some of these limitations.

Chapter 8

A Multi-objective Pareto Optimization Framework for Selecting Functionally Informative Tag SNPs

This chapter introduces our second integrative SNP selection method for identifying functionally informative tag SNPs. The proposed method is based on the notion of Pareto optimality, which is a well-established concept in game theory and engineering for addressing multi-objective optimization problems. We provide the motivation for the proposed method in Section 8.1. In Section 8.2, we formally define the problem of functionally informative tag SNP selection in the context of Pareto optimality. Section 8.3 describes an imputation algorithm for linkage disequilibrium and a heuristic algorithm for selecting functionally informative tag SNPs. Section 8.4 reports the evaluation results of the proposed method using a comparative study, and Section 8.5 concludes and outlines future directions.

8.1 Motivation and Objectives

In the previous chapter, we have introduced the first integrative SNP selection system that combines tag SNP selection and functional SNP selection into one unified selection process. We proposed a greedy selection algorithm in which both tagging informativeness and functional significance of SNPs are incorporated into a single objective function expressed as a weighted sum [108]. In that work, two objective functions, $f_1(x)$ and $f_2(x)$ are combined into one, by employing a linear combination of the form $\alpha_1 \cdot f_1(x) + \alpha_2 \cdot f_2(x)$. However, this formulation is still limited by the fact that the selected set of SNPs depends on the predefined weighting factors, α_1 and α_2, whose optimal value is unknown *a priori* in most cases.

In this chapter, we introduce a new multi-objective SNP selection system that, as one of its main contributions, overcomes this limitation by using the well-established, game-theoretic notion of *Pareto optimality* [98]. To the best of our knowledge, this idea was not applied before in any genetic variation study. The underlying theoretical principle is that when several, possibly competing, objectives are considered simultaneously, there may not exist a single global optimal solution that is superior with respect to *all* objectives; however, we can possibly find a *set* of *nondominated* solutions, formally called *Pareto optimal* solutions, to which no other solution is superior with respect to all objectives. Based on this notion of Pareto optimality, we propose a multi-objective simulated annealing algorithm that searches the space of Pareto optimal subsets of functionally informative tag SNPs. Our algorithm does not require a predefined weighting factor to combine different objectives, while it still selects only a small number of SNPs. We also present two heuristics to speed up the search process, and demonstrate the utility of the heuristics through a comparative study.

In addition, we present a new framework to calculate the tagging informativeness of SNPs with no allelic information. The HapMap consortium [33] provides haplotype information of a subset of SNPs on the human genome for three major populations. When SNPs in which we are interested are not members of the public haplotype, their tagging ability cannot be measured due to the lack of allele information over population samples. Conventional imputation algorithms for genome sequences are not helpful in this case, because the algorithms cannot estimate the missing allele information of the SNPs for the entire population samples. In order to address this problem, we propose to impute pairwise linkage disequilibrium (LD) of SNPs, rather than imputing the allele values themselves.

In the next section, we start by defining the problem of functionally significant tag SNP selection in the context of Pareto optimality.

8.2 Problem Definition

In Section 7.2, we have formally defined the problem of *functionally informative tag SNP selection*. As a basis, we use the same formulation of the problem, but redefine the multi-objective optimization function in the context of Pareto optimality. We also note that the information-based and the function-based objective functions introduced in this chapter extend our previous work. We redefine the two objective functions to incorporate the widely used concept of pairwise linkage disequilibrium (LD). This modification, along with our new imputation algorithm, enables to include SNPs with no haplotype information in our SNP selection process. We start by introducing the basic notation used throughout this chapter.

Suppose that the target locus contains p consecutive SNPs. As before, we represent each SNP as a discrete random variable, X_j $(j = 1, ..., p)$, whose possible values are

the 4 nucleotides, $\{a, c, g, t\}$. Let $V = \{X_1, ..., X_p\}$ denote a set of random variables corresponding to the p SNPs. A haplotype dataset, D, that contains the allele information of n haplotypes, each of which consists of the p SNPs in V, is provided as input. We are also given the set of functional significance (FS) scores for the p SNPs, which we denote by $E = \{e_1, ..., e_p\}$. We currently use the FS scores assessed by the F-SNP-Score system, introduced in Chapter 6. We note, though, that it is possible to use other functional scoring methods as well. Last, for a subset of SNPs, $T \subset V$, we define an objective function, $f(T|D, E)$, to reflect both the allele information carried by the SNPs in T about the remaining SNPs in $(V-T)$ and the functional significance represented by the SNPs in T.

The problem of *functionally informative tag SNP selection* is still formally stated as in Section 7.2:

Problem	:	Functionally Informative Tag SNP Selection
Input	:	A set of SNPs, V; A maximum number of SNPs to select, k; A haplotype dataset, D; A set of functional significance scores, E;
Output	:	A set of SNPs $T = \underset{T \ such \ that \ T \subset V \ \& |T| \leq k}{argmax} f(T|D, E)$.

However, alike in Chapter 7, we define the objective function, $f(T|D, E)$, as an *ordered pair* of two simpler objective functions, $f_1(T|D)$ and $f_2(T|D, E)$, where f_1 measures the allelic information of a SNP set T, while f_2 measures its functional significance. We also note that, as stated above, we redefine the two objective functions for incorporating the widely used concept of pairwise linkage disequilibrium (LD). Many tag SNP selection tools are based on the pairwise LD-based SNP selection approach (we provide the literature review in Section 3.1.2). Our new imputation algorithm, introduced in Section 8.3.1

expedites the computation procedure of the objective functions, as well.

The two objective functions, $f_1(T|D)$ and $f_2(T|D, E)$ are formally defined as follows:

Definition 8.1. Information-based Objective. *Given a set V of p SNPs, $V = \{X_1, ..., X_p\}$, a dataset D of n haplotypes, and a parameter value, α $(0 < \alpha < 1)$, we define the information-based objective $f_1(T|D)$ for a subset of SNPs, $T \subset V$, as*:

$$f_1(T|D) \stackrel{\text{def}}{=} \frac{1}{p} \sum_{j=1}^{p} I(X_j,\ T)$$

where

$$I(X_j,\ T) = \begin{cases} 1 \ : \text{ if } \exists X_s \in T \ \ such\ that \ \ \ LD(X_j, X_s|D) \geq \alpha; \\ 0 \ : \text{ otherwise.} \end{cases}$$

This objective function, $f_1(T|D)$, measures the allele information carried by the SNPs in T about the haplotype dataset D. It is based on the *pairwise LD-based tag SNP selection approach*, in which the smallest subset of SNPs is selected such that all unselected SNPs are in high LD with one of the selected tag SNPs [21]. Here, we express the objective as the number of SNPs in V whose maximum LD with selected tag SNPs in T is at least a pre-specified threshold α (here, α=0.8), based on the haplotype dataset D.

Definition 8.2. Function-based Objective. *Given a set V of p SNPs, $V = \{X_1, ..., X_p\}$, a dataset D of n haplotypes, a set of functional significance scores for the p SNPs, $E = \{e_1, ..., e_p\}$, and a parameter value, α $(0 < \alpha < 1)$, we define a function-based objective $f_2(T|D, E)$ for a subset of SNPs, $T \subset V$, as*:

$$f_2(T|E) \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{p} (e_j \cdot I_T(X_j))}{\sum_{j=1}^{p} e_j},$$

where I_T is a modified indicator function formally defined as:

$$I_T(X_j) = \begin{cases} 1 & : \textit{if } X_j \in T \text{ ;} \\ ld_j & : \textit{if } X_j \notin T \text{ and } \exists X_s \in T \textit{ such that } LD(X_j, X_s|D) \geq \alpha \text{ ;} \\ 0 & : \textit{otherwise ;} \end{cases}$$

and ld_j is the maximum LD between SNP X_j and the selected SNPs. That is, the functional significance (FS) of the subset T is computed as the normalized sum of the FS scores of the SNPs in T. This formulation is primarily based on a typical function-based SNP selection approach that aims to prioritize SNPs according to their functional significance scores [184, 188, 111].

However, by introducing ld_j, we modify the basic framework shown in Section 7.2, to account for the functional significance of unselected SNPs. That is, when a SNP X_j is not directly selected into T, but is in high LD with the SNPs in T (i.e., $LD(X_j, X_s|D) \geq \alpha$), we still allow a certain proportion (i.e., ld_j) of its functional significance score to be included in our computation. Specifically, although some functional SNPs are not directly selected, their association with disease is indirectly accounted for from the selected SNP, based on their LD. Therefore, the functional significance represented by the SNPs in T reflects not only the functional score of the SNPs in T, but also that of other SNPs which are in high LD with the selected SNPs. To the best of our knowledge, this is the first formulation of functional SNP selection that takes LD into account.

The functionally informative (FI) objective function, $f(T|D, E)$, is now defined as an

ordered pair $\langle f_1(T|D),\ f_2(T|D,E)\rangle$ as follows:

Definition 8.3. Functionally Informative Objective Function. *Given a set of k SNPs, $T \subset V$, a haplotype dataset, D, and a functional significance score set, $E = \{e_1, ..., e_p\}$, we define a functionally informative (FI) objective function, $f(T|D,E)$ as*:

$$f(T|D,E) = \langle\ f_1(T|D),\ \ f_2(T|D,E)\ \rangle\,.$$

Note that we aim to *simultaneously* optimize these two distinct and possibly competing objectives, f_1 and f_2. To achieve this goal, we adopt the notion of Pareto optimality defined as follows:

Definition 8.4. Pareto Optimality. Let T_i and T_j be two distinct subsets of V, of the same size, k.

1. T_i is said to *dominate* T_j if and only if
   $(f_1(T_i) \geq f_1(T_j)$ and $f_2(T_i) > f_2(T_j))$ or $(f_1(T_i) > f_1(T_j)$ and $f_2(T_i) \geq f_2(T_j))$.
   We denote this relationship by $T_i \succ T_j$.
2. T_i is called *Pareto optimal* if and only if no other subset of V dominates T_i.

Figure 8.1 shows an example of dominated and nondominated Pareto optimal solutions. Suppose that there are only seven subsets of the SNP set V, namely A, B, C, D, E, F, and G, chosen as candidate sets of functionally informative tag SNPs. The X-axis represents the information-based objective score, f_1 of each subset, and the Y-axis represents the corresponding function-based objective score, f_2. Based on our formulation, the higher the f_1 and f_2 scores are, the better the subset is. For clarity, in the figure we denote the f_1 and the f_2 scores of subset F as X_F and Y_F, and show dashed lines to the respective values, drawn perpendicular to the X- and Y-axis, respectively.

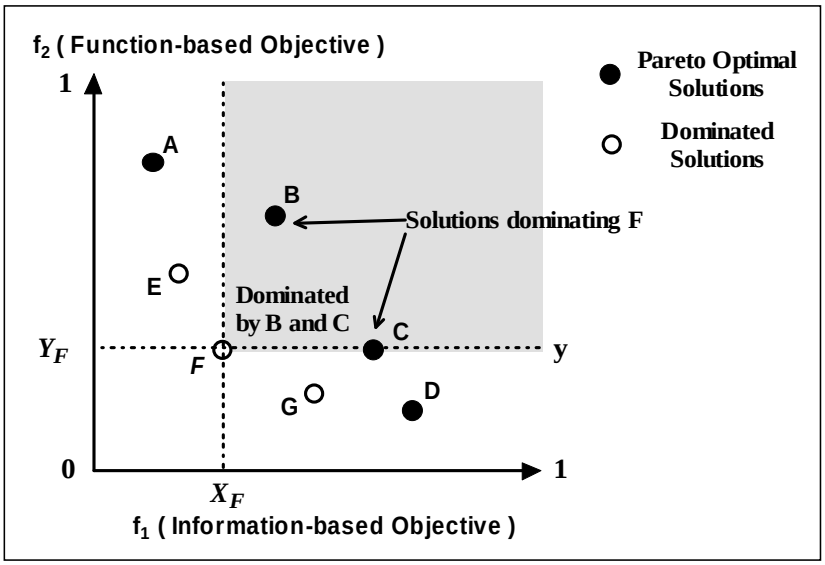


Figure 8.1: **Dominated and non-dominated Pareto optimal solutions. The X-axis represents the information-based objective score, f_1, and the Y-axis represents the function-based objective score, f_2. To illustrate an example of dominated solutions, we denote the f_1 and the f_2 scores of subset F as X_F and Y_F, and show dashed lines to the respective values on the X- and Y-axis, respectively. Solutions B and C dominate F as they show higher scores than F with respect to at least one objective. Among seven candidate solutions, nondominated Pareto optimal solutions are displayed as black dots, while dominated solutions are displayed as unfilled circles.**

Solution B *dominates* solution F as it shows higher scores than F with respect to both objectives; solution C also dominates F as it shows a higher score than F with respect to f_1 and the same score as F with respect to f_2. However, solutions B and C do not dominate each other, as each of them shows better performance than the other, with respect to one objective, but is doing worse than the other with respect to the other objective. Similarly, solutions $A, B, C,$ and D (shown as black dots in Figure 8.1) form a nondominated Pareto optimal set in this example. None of them are dominated by other solutions; each of these solutions shows higher score than all others on at least one objective, thus equally "optimal"

unless a specific preference toward one objective is stated.

In summary, among all possible SNP subsets of maximum size k, we aim to select all Pareto optimal subsets of functionally informative tag SNPs based on Definition 8.4. The problem of finding k most informative tag SNPs is proven to be NP-complete [7]. We compute our information-based objective independently of the function-based objective, which means that simultaneously considering the two selection objectives does not reduce the complexity of the problem. In the next section, we thus propose a heuristic framework, (which, like all heuristics, looks for a *locally* optimal solution), to address the problem of functionally informative tag SNP selection within the framework of Pareto optimality.

8.3 Methods for Pareto-based SNP Selection

Our SNP selection system consists of two main steps. First, we calculate the pairwise linkage disequilibrium (LD) among all candidate SNPs. When the allele information is not available, we impute the corresponding LD values. Second, we select the Pareto (locally) optimal sets of functionally informative tag SNPs using a multi-objective simulated annealing algorithm. We describe the details of each step in the following subsections.

8.3.1 Computing the Linkage Disequilibrium of SNPs

To efficiently compute the score of the information-based objective function, $f_1(T|D)$, we calculate the pairwise LD between all pairs of candidate SNPs in advance. As a measure of pairwise LD, we currently use the *multi*-allelic extension of Lewontin's linkage disequilibrium (LD) measure, D' [72].

As stated in Section 4.4.2, the LD computation procedure is as follows: Let X_i be an

m-allelic SNP, and X_j be an n-allelic SNP. Let f_k^i be the relative frequency of the k^{th} allele for SNP X_i, while f_l^j be the relative frequency of the l^{th} allele for SNP X_j, counted from the haplotype dataset D (where $k = 1, .., m$ and $l = 1, .., n$). We denote the relative joint frequency of the k^{th} allele occurring for SNP X_i *and* the l^{th} allele occurring for SNP X_j by f_{kl}^{ij}. The LD between the two SNPs, X_i and X_j, is computed as:

$$LD(X_i,\ X_j \mid D) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_i^1 \cdot f_j^2 \ |\frac{f_{ij}^{12} - f_i^1 f_j^2}{D_{max}}|,$$

where D_{max} is the maximum value among the products of two relative frequencies of SNP X_1 and X_2, $f_i^1 \cdot f_j^2$ $(i = 1, .., 4;\ j = 1, .., 4)$.

When the allele information of a SNP is not available in D, we impute its LD value according to the general characteristic of LD; the level of LD tends to decrease in proportion to the physical distance between SNPs [70]. Thus, for each SNP, X_i, with no allele information, we choose two nearest SNPs *with* allele information, one from the left and the other from the right side of the SNP, as shown in Figure 8.2. We call these two SNPs the *proxy* SNPs of X_i, and use their allele information to estimate the pairwise LD of X_i with others. The imputation algorithm is formulated as follows:

$$LD(X_i,\ X_j \mid D) \stackrel{\text{def}}{=} \frac{d(X_i, X_i^L) \cdot LD(X_i^R,\ X_j \mid D) + d(X_i, X_i^R) \cdot LD(X_i^L,\ X_j \mid D)}{d(X_i, X_i^L) + d(X_i, X_i^R)},$$

where X_i^L and X_i^R denote the left and the right proxy SNPs of X_i respectively, and $d(X,Y)$ denotes the distance in base-pairs between the location of any two SNPs, X and Y, on the genomic sequence. In short, we calculate the pairwise LD between X_j and X_i as a weighted average of the pairwise LD between X_j and the two proxy SNPs of X_i. We note that the closer the proxy SNP is to X_i on the genomic sequence (in terms of physical base pairs distance), the higher weight its LD has. This imputation procedure is illustrated in

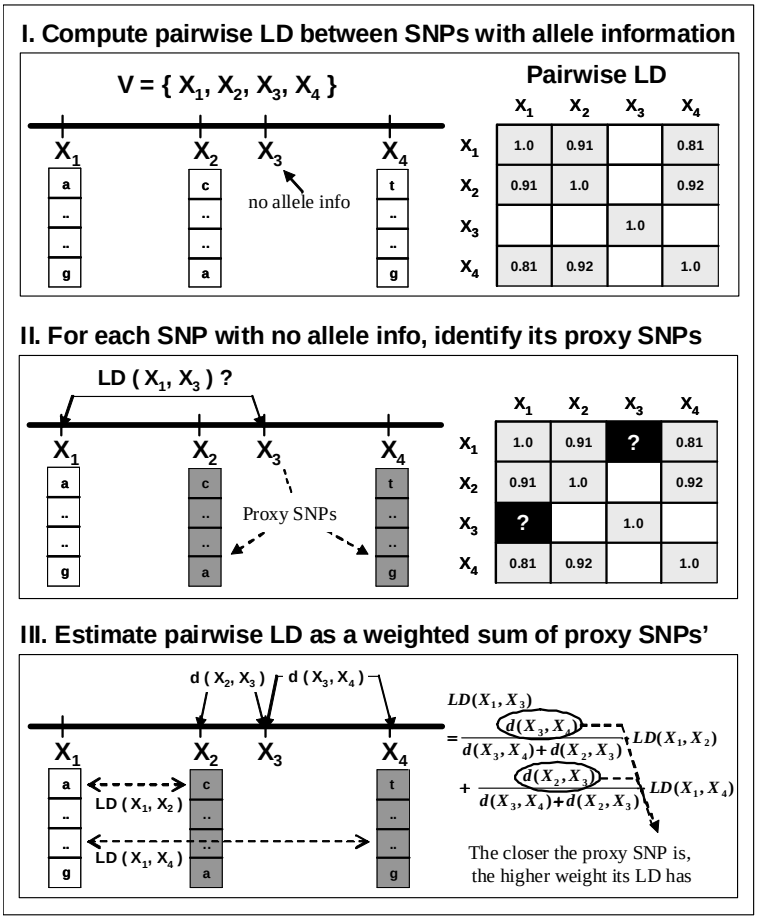


Figure 8.2: **The imputation procedure for inferring the linkage disequilibrium (LD) of SNPs with no allele frequency information. First, pairwise LD is computed between SNPs with allele information. Second, for each SNP with no allelic information, two neighboring SNPs with allelic information, one on each side, are selected. Third, the allele information of the selected neighboring SNPs are used to compute the pairwise LD of the SNP with no allelic information with others.**

Figure 8.2.

8.3.2 Selecting Functionally Informative Tag SNPs

Our selection algorithm is based on a multi-objective simulated annealing (SA) algorithm [97], which has been successfully used for addressing many combinatorial optimization problems [37, 48, 47]. Table 8.1 summarizes the proposed algorithm.

First, we choose a randomly generated subset of k SNPs as a current solution, T_c, and compute the score-pair $f(T_c|D, E)$. Second, while a temperature parameter t is greater than a minimum threshold t_{min}, the following three steps are repeated: 1) A neighbor set of the current solution T_c, referred to as T_n, is generated (as explained later in this section)[1];

2) If T_n is Pareto optimal among the sets we examined, T_n is added to the Pareto optimal solutions with respect to the examined sets, $\mathcal{PO}$, and replaces T_c for the next iteration; otherwise, it replaces T_c with a probability P_{accept}. The probability P_{accept} is updated as a function of $f(T_c|D, E), f(T_n|D, E)$ and t; 3) The temperature t is reduced by a rate of r_c. This whole procedure is repeated M times. In the experiments described here, we empirically set the SA parameters as follows: $t_0 = 1.3,\ r_c = 0.9999,\ t_{min} = 0.001$, and $M = 10^3$.

To guide an efficient SA search, we introduce two heuristics for generating a new neighbor solution. First, in order to find a neighbor SNP set that is likely to dominate the current set T_c, we utilize the score of each SNP with respect to the two selection objectives, f_1 and f_2. That is, for each SNP X_i, we compute the objective scores, $f_1(\{X_i\}|D)$ and $f_2(\{X_i\}|D, E)$ $(i = 1, ..., p)$, before starting the search. When generating a new neighbor

[1]We note that the size of a new neighbor set is fixed to k. It is straightforward to show that for each subset of size $k-1$, we can always find a subset of size k of which the selection objective $f_1(T|D)$ increases, while $f_2(T|D, E)$ does not decrease.

Table 8.1: **The multi-objective simulated annealing algorithm for searching the Pareto optimal sets of functionally informative tag SNPs.**

Input: A set of SNPs, $V = \{ X_1, ..., X_p \}$;
A set of functional significance scores, $E = \{ e_1, ..., e_p \}$;
A haplotype dataset D;
The maximum number of SNPs to select, k;

Output: Sets of Pareto optimal solutions, $\mathcal{PO} = \{ T_1, ... \}$;

Algorithm:
Compute $LD = \{ ld_{11}, ..., ld_{pp} \}$;

$\mathcal{PO} \leftarrow \emptyset$;
$m \leftarrow 0$;

While $(m < M)$
$t \leftarrow t_0$;
$T_c \leftarrow T_0$; // A set of randomly selected k SNPs from V;
Compute $f(T_c|D, E) = \langle f_1(T_c|D),\ f_2(T_c|E) \rangle$;

While $(t > t_{min})$
$T_n =$ neighbor (T_c);
Compute $f(T_n|D, E) = \langle f_1(T_n|D),\ f_2(T_n|E) \rangle$;

If ($\exists T_i \in \mathcal{PO},\ T_n \succ T_i$)
remove $\forall\, T_i \in \mathcal{PO}$ s.t. $T_n \succ T_i$;
$\mathcal{PO} \leftarrow \mathcal{PO} \cup \{T_n\}$;
Else if $\forall\, T_i \in \mathcal{PO},\ T_i \not\succ T_n$
$\mathcal{PO} \leftarrow \mathcal{PO} \cup \{T_n\}$;
EndIf

$P_{accept}(\, T_c,\, T_n,\, t\,) \leftarrow \min \left\{ 1, exp\left(\underset{j \in \{1,2\}}{max} \frac{\left(f_j(T_n) - f_j(T_c)\right)}{t} \right) \right\}$;

If ($T_n \succ T_c$ **or** $P_{accept} > random$)
$T_c \leftarrow T_n$;
EndIf

$t \leftarrow r_c \cdot t$;
EndWhile

$m \leftarrow m + 1$;

EndWhile

for the current set of functionally informative SNPs, T_c, we first determine whether to focus on the information-based objective, f_1, or on the function-based objective, f_2, by flipping an unbiased coin. Suppose that the information-based objective f_1 is selected. We now select a SNP X_r from T_c to be replaced by a SNP X_a from $\big(V - T_c\big)$. X_r is chosen with probability P_{remove}, which is *inversely* proportional to its f_1 score, while X_a is chosen with probability P_{add}, which is *directly* proportional to its f_1 score. That is,

$$\begin{aligned} P_{add} &= \frac{f_1(\{X_i\}|D)}{\sum\limits_{X_i \in (V - T_c)} f_1(\{X_i\}|D)}, \quad \text{and} \\ P_{remove} &= \frac{\Big(f_1(\{X_i\}|D)\Big)^{-1}}{\sum\limits_{X_i \in T_c} \Big(f_1(\{X_i\}|D)\Big)^{-1}}. \end{aligned}$$

A second heuristic is used to expedite and diversify the coverage of the search space. Instead of generating a new neighbor by replacing one SNP at a time, we simultaneously replace several SNPs in the initial search period, and gradually decrease the number of replaced SNPs as the search progresses. As a result, farther neighbors are examined in the initial stages of the search, diversifying the search area, while the later stages, which are expected to search closer to the optimum, focus on neighbors that are closer to the current solution. This strategy helps avoid local optima. In Section 8.4, we show the utility of these two heuristics by comparing the performance of our selection algorithm with and without them.

The time complexity of each iteration is $O(p)$, where p is the number of candidate SNPs. As this iteration is repeated for a maximum of $\big(M \cdot \log(t_{min}/t_0)/\log r_c\big)$ times, the overall complexity of our selection algorithm is $O\big(p \cdot M \cdot \log(t_{min}/t_0)/\log r_c\big)$. The computation procedure of the pairwise LD between p SNPs is $O(n \cdot p^2)$, where n is the number of haplotypes, and p is the number of SNPs in dataset D.

8.4 Experiments and Results

We conduct a comparative study to evaluate the performance of the proposed integrative SNP selection system compared to other state-of-the-art selection systems that support both tag SNP selection and functional SNP selection: TAMAL [73] and SNPselector [184]. In the following sections, we summarize the experimental setting of the comparative study, and report the evaluation results.

8.4.1 Experimental Setting

We applied our method to 34 disease-susceptibility genes for lung cancer, as summarized by Zhu *et al*. [204]. This dataset includes a larger number of genes compared to the dataset of 14 genes we used for evaluating FITS-Select in Section 7.4.1. The list of SNPs linked to the genes, including 10k upstream and downstream regions, was retrieved from the dbSNP database [167]. The haplotype datasets for the genes were downloaded from the HapMap consortium for the CEU population (public release #20/phaseII) [33]. We summarize the primary information about the 34 genes, such as gene symbol and the total number of linked SNPs, in the left-most part of Table 8.2. We note that we used a larger number of genes compared to the experiments that we did in Chapter 7 to ensure more generalized performance.

We compare the performance of our system with that of two state-of-the-art SNP selection systems that support both tag SNP selection and functional SNP selection: SNPselector [184] and TAMAL [73]. The compared systems share the same goal as ours, namely, selecting an informative set of tag SNPs with significant functional effects. However, they address tag SNP selection and functional SNP selection as two separate optimization problems, while we address it as a single multi-objective optimization problem.

Table 8.2: **Evaluation results of three Pareto optimal search algorithms, SA_1, SA_0, and RS against the two compared systems, SNPselector and TAMAL. Under the name of each compared system, the left-most column shows the number of SNPs, k, selected by the compared system, for the corresponding gene. The remaining three columns, SA_1, SA_0, and RS, typically show the e_1 score, that is, the percentage of the identified Pareto optimal solutions that *dominate* the compared system's solution, computed for each of the respective Pareto optimal search algorithms. In the few cases where the solutions are *dominated* by the compared system's solution, e_2 is shown (denoted by †). Cases where there is no dominating nor dominated solution are indicated by a dot.**

Gene	Total	SNPselector				TAMAL			
Symbol	SNP #	k	SA_1	SA_0	RS	k	SA_1	SA_0	RS
ADRB2	153	41	100.0	100.0	33.3	17	66.6	100.0	50.0†
APEX1	83	27	100.0	100.0	100.0	19	100.0	100.0	100.0
ATR	181	36	100.0	100.0	100.0	20	100.0	100.0	50.0
CDKN1A	116	34	100.0	100.0	100.0	20	100.0	100.0	100.0
CYP1A1	49	34	100.0	100.0	100.0	10	100.0	100.0	75.0
CYP1B1	172	51	100.0	100.0	100.0	28	100.0	100.0	100.0
NQO1	86	6	100.0	100.0	50.0	8	100.0	100.0	50.0
EPHX1	148	27	80.0	25.0	14.2†	23	25.0	·	·
ERCC2	210	27	100.0	100.0	100.0	30	50.0	50.0	·
ERCC4	289	41	100.0	100.0	44.4	49	50.0	50.0	20.0†
ERCC5	261	43	88.8	25.0	·	43	11.1	·	·
GSTP1	70	27	100.0	100.0	100.0	14	100.0	100.0	50.0
LIG4	107	27	100.0	100.0	100.0	27	20.0	100.0	25.0†
MBD1	65	24	100.0	100.0	100.0	19	100.0	100.0	50.0
MGMT	550	36	100.0	100.0	71.4	81	20.0	25.0†	·
MMP9	111	33	100.0	100.0	100.0	16	100.0	100.0	33.3
MTHFR	206	42	100.0	100.0	100.0	24	50.0	100.0	50.0
MTR	372	27	75.0	100.0	100.0	33	14.2	33.3	33.3
MTRR	212	31	100.0	100.0	100.0	32	33.3	50.0	75.0†
NBN	355	21	100.0	100.0	100.0	38	67.0	100.0	50.0†
POLB	143	25	100.0	100.0	100.0	18	100.0	100.0	100.0
RAD23B	197	12	100.0	100.0	100.0	29	20.0	16.6	·
SOD2	188	31	20.0	·	·	27	25.0	25.0†	20.0†
SULT1A1	180	39	100.0	100.0	100.0	6	33.3	100.0†	100.0†
TP53	307	46	50.0	100.0	100.0	11	50.0	33.3	66.6†
XPC	237	35	100.0	100.0	100.0	29	20.0	·	·
XRCC1	152	46	80.0	33.3	·	46	20.0	·	·
XRCC2	253	13	100.0	100.0	100.0	25	33.3	33.3	50.0†
XRCC3	158	11	100.0	100.0	100.0	37	50.0	100.0	33.3
EXO1	283	35	33.3	100.0	·	36	20.0	100.0	66.6†
HDAC5	111	21	100.0	100.0	100.0	13	100.0	100.0	100.0
POLI	239	23	75.0	100.0	100.0	24	25.0	100.0	·
REV1	307	53	100.0	100.0	100.0	32	50.0	100.0	50.0

† denotes the e_2 measure.

TAMAL [73] enables users to select haplotype tag SNPs (using Gabriel's method [58] or the Tagger method [39]), or functionally significant SNPs such as SNPs leading to non-synonymous or synonymous amino acid changes, or SNPs altering canonical splice sites, promoter regions, or transcriptional regulatory regions. To identify tag SNPs, we selected the Tagger method option as it is based on the same pairwise linkage disequilibrium (LD)-based objective, $f_1(T|D)$, as ours. SNPselector [184] prioritizes SNPs based on their tagging informativeness, SNP allele frequencies, functional significance, regulatory potential, and repeat scores. Same as TAMAL, it recognizes SNPs that are likely to alter protein function, splicing regulation, or transcriptional regulation as functionally significant. The tagging informativeness is calculated also based on the pairwise LD-based criterion.

As described in Section 7.4.1, TAMAL and SNPselector do not allow users to specify the maximum number of selected SNPs. Thus, for a fair comparison, we first apply TAMAL and SNPselector to the dataset of 34 genes, and apply our system on the same dataset to select the same number of SNPs as selected by each compared system. We denote our full-fledged multi-objective simulated annealing algorithm that employs the two heuristics by SA_1. In addition, we demonstrate the utility of our heuristics by examining the performance of two baseline search algorithms for identifying (locally) Pareto optimal solutions: 1) The same simulated annealing algorithm, described in Table 8.1, without the proposed two heuristics, which we denote by SA_0; and 2) A naïve selection algorithm that randomly generates M solutions (here, $M = 10^4$) and identifies (locally) Pareto optimal subsets within the M solutions. We refer to this naïve selection algorithm as RS.

As evaluation measures, we define two statistics based on Pareto optimality. First, for each Pareto optimal search algorithm, we compute the percentage of its SNP set results that *dominate* the solution found by the compared system (following Definition 8.4, one

solution *dominates* the other, if it is at least as good as the other according to one objective, and is strictly better than the other according to another objective). This measure examines whether our Pareto optimal search algorithm indeed performs *better* than the compared system. We refer to this first measure as e_1. If there is no dominating solution (i.e., e_1=0), we compute the percentage of the Pareto optimal solutions that *are dominated by* the solution found by the compared system. This second measure examines whether our search algorithm performs *worse* than the compared system. We refer to this measure as e_2. We note that some Pareto optimal solutions are neither dominant nor dominated by the compared solution[2]. Therefore, the sum of the two evaluation measures, e_1 and e_2 could be less than 100%.

8.4.2 Test Results

Table 8.2 summarizes the evaluation results of the three Pareto optimal search algorithms, SA_1, SA_0, and RS against the two compared systems, SNPselector and TAMAL. The two leftmost columns show gene symbols and the total number of SNPs linked to each gene. The remaining columns are divided into two parts, corresponding to the two compared systems. In each part, the leftmost column shows the number of SNPs, k, which is chosen by each compared system for the corresponding gene. The remaining three columns, SA_1, SA_0, and RS show the evaluation measure, e_1(the majority of the cases) or e_2 (denoted by †) computed for the corresponding search algorithm, respectively. When both of e_1 and e_2 are 0, which means that the compared solution is neither dominant nor dominated by our Pareto optimal solutions, we display it with a dot.

[2]This happens when the Pareto optimal solutions outperform the compared solution with respect to one objective, but are worse with respect to the other objective. In other word, the compared solution is also a Pareto optimal solution.

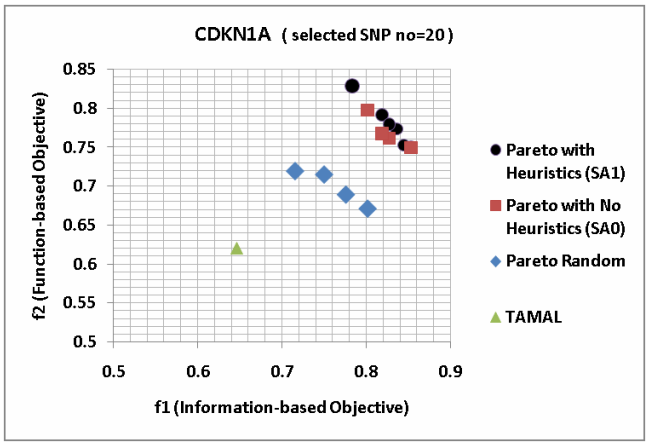


Figure 8.3: **The performance of Pareto optimal solutions identified by three search algorithms, SA_1, SA_0, and RS, and that of the solution selected by TAMAL for gene CDKN1A. The X-axis represents the information-based objective score, f_1, while the Y-axis represents the function-based objective score, f_2. The solutions identified by SA_1, SA_0, and RS are marked with black dots, red rectangles, and blue diamonds, respectively. TAMAL's solution is marked with a green triangle.**

Overall, the SA_1 algorithm that uses the two proposed heuristics always finds Pareto optimal subsets that *dominate* the compared solutions. The difference between our dominating solution and the compared system's solution is statistically significant with respect to both selection objectives. Using the paired t-test with 5% significance level ($\alpha = 0.05$), p-values are 8.29e-179 for $f_1(T|D)$ and 8.15e-157 for $f_2(T|D, E)$ in the case of SNPselector, and 7.02e-073 and 5.76e-005 for TAMAL. In contrast, the naïve SA_0 algorithm, which does not employ any heuristics, fails to find dominating solutions in 8 cases (shown as † or · in Table 8.2). In three cases, SA_0's solutions are dominated by the compared system's (shown as † in Table 8.2). The random search algorithm RS fails to find dominating solutions in 23 cases, while producing dominated solutions in 11 cases.

We further confirmed that the SA_1 algorithm performs better than the two naïve approaches, SA_0 and RS; SA_1's solutions *dominate* SA_0's solutions on 31 genes and RS's solutions on all 34 genes. The difference is statistically significant, as confirmed by the paired t-test (p-values are 1.37e-004 for f_1 and 3.11e-015 for f_2 with respect to SA_0, and 2.43e-149 and 3.89e-179 for RS).

Figure 8.3 shows an example of the identified solutions by SA_1, SA_0, RS, and the compared system, in this case, TAMAL, for the gene CDKN1A. The gene CDKN1A plays a critical role in the cellular response to DNA damage, and its allelic variants are known to be associated with lung cancer [169]. The number of SNPs selected by TAMAL is 20, which is approximately 17% of the SNPs linked to the gene. As is clearly shown in Figure 8.3, SA_1 identifies a set of solutions (that is, 5 different subsets of 20 SNPs each, shown as black dots in the figure), for which both information-based objective, f_1, and function-based objective, f_2, outperform the solutions found by SA_0, and greatly outperform RS and TAMAL.

8.5 Discussion

In this chapter, we presented a new multi-objective optimization framework for selecting functionally informative tag SNPs. It *simultaneously* identifies SNPs that are both highly informative as tag SNPs for all other SNPs on the target locus and are of high functional significance. For the first time, we applied the notion of Pareto optimality, which has been extensively used in other fields, to address the problem of SNP selection. A comparative study over a set of 34 disease-susceptibility genes for lung cancer shows that our system improves upon current state-of-the-art SNP selection systems that support both tag SNP

selection and functional SNP selection, as well as upon other general-purpose search algorithms for identifying Pareto optimal solutions.

It appears that two main factors contribute to this improved performance. First, we take into account both objectives at the same time, thus ensuring the optimization of both objectives given the limited number of selected SNPs. Second, instead of searching for a single optimum, which may not exist due to possibly competing selection objectives, we search for a collection of all Pareto locally-optimal subsets. Our comparative study shows that a broad range of Pareto optimal solutions exist for all 34 genes. Researchers can thus examine possible trade-offs between the obtained Pareto optimal solutions, without deciding *a priori* one best combination of distinct selection objectives.

In this work, we demonstrated the utility of our multi-objective SNP selection framework in the context of pairwise linkage disequilibrium. As discussed in Section 7.5, our selection framework is general in a sense that other types of SNP selection criteria can be incorporated into it as well. It is also straightforward to include additional SNP selection criteria or to use different functional significance (FS) score of SNPs, if preferred.

A disadvantage of exploring a set of Pareto optimal solutions is the increased running time. In our analysis, it takes from 17 minutes (in the case of CYP1A1, number of SNPs is 49) to 1174 minutes (in the case of MGMT, number of SNPs is 550) to reach convergence. Our future research will thus focus on improving the search speed by employing additional heuristics. Addressing this scalability issue is also critical for applying the method to genome-wide association studies.

In this work, we did not specify a criteria to select one solution from a set of Pareto optimal solutions. We believe that additional SNP selection criteria, such as fitness to SNP array design, can be used to finalize the decision. In the near future, we thus plan

to investigate other SNP selection objectives, and examine how the objectives can be used to prioritize the selected Pareto optimal solutions. We also plan to examine other search algorithms used for addressing multi-objective optimization problems.

Chapter 9

Conclusion

This chapter summarizes the major contributions of this dissertation work and outlines possible directions for future research. Section 9.1 presents a summary of the algorithms and systems introduced in this thesis, and describes the major contributions. Section 9.2 discusses the limitations of the proposed work and suggests a number of future research directions to enhance it.

9.1 Summary of Major Contributions

In this thesis, we addressed the problem of selecting a set of SNP markers for supporting effective disease-gene association studies. SNPs are the most common form of genetic variations on the human genome, and as such, they have been widely used as genetic markers for studying common and complex human diseases. However, the tremendous number of SNPs, which is estimated at more than eleven million [167], poses challenges to the genotyping and analysis procedure associated with such studies. Our goal is to support effective genetic association studies for common and complex human diseases, by providing

effective prioritization methods for SNP markers based on both their allele information and functional significance.

To achieve the goal, we have presented several novel algorithms and a database system based on the two major SNP selection approaches: tag SNP selection and functional SNP selection. In addition, we have proposed an innovative approach to combine both tag SNP selection and functional SNP selection into one unified selection process. Improved performance of all the proposed methods was demonstrated through comparative studies. The summary of the major contributions are as follows:

- We presented a new tag SNP selection method, BNTagger, to identify a subset of SNPs that can effectively predict the allele information of the complete SNP set on the target genomic region. By allowing the number or the location of predictive tag SNPs to vary, BNTagger improves prediction performance over that of state-of-the-art predictive methods. BNTagger is also more widely applicable than other tools, as it is neither limited to bi-allelic SNPs, nor requires an additional haplotype phasing procedure.

- We constructed a web-based public database service, F-SNP, to provide a comprehensive collection of functional information about SNPs. Using 16 external databases and function-assessment tools for SNPs, F-SNP provides users with information about putative deleterious effects of SNPs on protein structure, function, post-translational modification, splicing regulation, and transcriptional regulation. A web interface enables easy navigation for obtaining functional information about SNPs through multiple starting points and exploration routes, including relevant gene and disease information.

- We described a classification method, F-SNP-C (F-SNP Classification) that designates a subset of the SNPs assessed by the F-SNP system as functional. The functional SNPs are the ones that are predicted by a majority of the function-assessment tools to be deleterious with respect to major bio-molecular functions. Therefore, F-SNP-C enables users to identify functionally significant SNPs that are more likely to be associated with disease or with functional impairment.

- We presented a scoring scheme, F-SNP-Score, to quantitatively assess the deleterious functional effects of SNPs. Using a probabilistic framework, F-SNP-Score quantifies the functional assessment results obtained from multiple independent tools, while taking into account the certainty of each prediction as well as the reliability of different tools. An empirical study over 580 disease-associated genes shows that F-SNP-Score assigns much higher functional significance (FS) scores to known disease-related SNPs than to likely neutral SNPs. The calculated functional significance scores of SNPs are currently provided through our public web-based database service, F-SNP.

- We proposed a novel integrative approach, FITS-Select (Functionally Informative Tag SNP Selector), to identify a subset of SNPs that are both informative tagging and functionally significant. We formalized the problem of SNP selection as a multi-objective optimization problem and presented a heuristic selection algorithm based on a single objective function, incorporating both allele information and functional significance of SNPs. An empirical study over a set of 14 disease-associated genes shows that our system improves upon current state-of-the-art systems. This work is the first method that combines the two notions of SNP selection – the function-based and the information-based – into a single optimized selection process.

- We presented an additional integrative approach, based on the game-theoretic notion of Pareto-optimality, for selecting functionally informative tag SNPs. The presented method extends and improves on our own FITS-Select method. The information-based and the function-based objective functions were redefined to incorporate the widely used concept of pairwise linkage disequilibrium. Moreover, we employed the notion of Pareto-optimality in the search for functionally informative tag SNPs. A comparative study based on 34 genes shows that the proposed selection algorithm improves upon state-of-the-art methods that support both tag SNP selection and functional SNP selection, as well as upon other general Pareto-based search algorithms.

In addition, this thesis work provides the following contributions pertaining to computer science in general:

- The work, presented in Chapter 4, applied a domain-specific concept, namely, pairwise linkage disequilibrium, to guide the learning procedure of the Bayesian network topology. This integration shows that utilizing constraints from specific problem domains enhances the model learning procedure. We expect such integration of constraints to be applicable to other domains.

- The scoring scheme, presented in Chapter 6, proposes a novel approach to combine diverse information from multiple sources within probabilistic framework. The approach is general, and thus can be applicable to other problems that require analysis of data emerging from multiple sources, especially when the true class labels for many data instances are not available.

- The two multi-objective optimization methods, presented in Chapters 7 and 8, showed a new application of multi-objective optimization frameworks in human genetics and

medicine. Moreover, the work clearly demonstrated that combining distinct problem solving criteria into one unified process is possible, and indeed improves upon separate optimization approaches.

9.2 Future Work

The work described in this thesis comprises one step toward the goal of identifying disease-causal variants underlying common and complex human diseases. Specific issues and future research directions have already been discussed in each chapter. We thus provide here more general directions for extending the work.

Further Utilizing Bayesian Networks Topology The proposed tag SNP selection method, BNTagger, is based on the framework of Bayesian networks (BNs) to identify predictive tag SNPs. In particular, BNTagger's heuristic selection algorithm utilizes of a certain aspect of the topology of BNs, as the topology captures the dependence and conditional independence relationship among SNPs. Analyzing the topology of the BNs learned for SNPs may provide further insights about the relationships among SNPs. Some of the possible questions to which the analysis may provide answers are: Does the topology of the networks show patterns consistent with other features of SNPs? For example, are nodes corresponding to SNPs in close physical proximity on the genome located close to each other in the network graphs? Are the selected tag SNPs concentrated in a specific part of the network or scattered over the whole network? If so, what inferences can we draw from such an observation? It would be also interesting to examine whether the topology of BNs varies across different haplotype/genotype datasets and populations.

Evaluation through Simulation Studies In this thesis, comparative studies based on multiple datasets were used primarily to evaluate performance. While comparative studies can indeed demonstrate the improved performance of the proposed methods over the state-of-the-art, theoretical evaluation remains to be done. Specifically, we are interested in conducting simulation studies to examine the performance of the proposed methods under various genomic/evolutionary/experimental conditions. For example, as discussed in Section 3.1.4, several factors are known to affect the effectiveness of a tag SNP selection strategy. These include: the sample size, SNP densities, allele frequencies, the level of linkage disequilibrium on the genomic region, population structure, and inheritance modes of disease-causal variants. The influence of these conditions on the effectiveness of selected tag SNPs remains to be studied.

Extending F-SNP-Score for Scoring a Set of SNPs The F-SNP-Score system assesses the putative deleterious effects of an *individual* SNP with respect to the four major biomolecular functional categories: protein coding, splicing regulation, transcriptional regulation, and post-translational modification. To the best of our knowledge, all the current function-assessment tools and databases for SNPs take this single SNP-based assessment approach. Basic assumptions under the single SNP-based assessment are that 1) the functional effects of an individual SNP can be assessed based on its genomic properties (such as the chromosomal location and allele information); and 2) the effects are not affected by other SNPs. While these assumptions greatly simplify the function-assessment process, they do not reflect the epistatic [1] effects of genes and mutations. It is now widely accepted that most biological systems that underlie cellular, developmental and physiological

[1]Epistasis refers to the interaction between genes at two or more loci. When epistasis takes place, the phenotype of one locus is altered or masked by effects of another locus [141].

function are composed of many elements that interact with one another [141]. Similarly, mutations are assumed to interact such that their combined effect on fitness is reinforced (also known as *synergistic epistasis*), mitigated (known as *antagonistic epistasis*) or cumulated (which means *no epistasis*) [118]. Therefore, it is more realistic to assume that functional significance of each SNP depends on that of other SNPs. Assessing the putative deleterious effects of an individual SNP in the context of functionally relevant other SNPs remains a challenging and important task to address.

Applying F-SNP-Score to Association Studies Another interesting research direction is to incorporate the functional significance score, assessed by F-SNP-Score, into large-scale association studies. As discussed in Section 3.2.4, functional significance (FS) scores of SNPs can be used for prior selection of SNP markers as well as for post evaluation of SNP markers after association with disease is identified. Furthermore, the assessed FS scores can be directly applied to large-scale association studies to reduce the chance of missing true positive associations, also known as the *multiple testing problem*. In statistics, the multiple testing problem occurs when a large number of statistical inferences are conducted simultaneously. Due to the large number of hypothesis tests, there is a high chance of detecting false positive associations that occur by chance. Thus, to control the false positive error rate, more conservative p-values are used to examine association tests, which raises the multiple testing problem. We plan to develop association tests that incorporate prior information on the putative deleterious effects of SNP to deal with the multiple testing problem. A possible route (which we have started exploring) is to use the FS scores of SNPs as weights to adjust the p-value for individual hypothesis testing.

Developing Genome-wide SNP Selection Methods The SNP selection methods discussed in this thesis are not directly applicable to the whole genome due to their computational complexity. However, much current interest is focused on genome-wide association studies that examine hundreds of thousands of SNPs at the same time. Genome-wide studies are more promising than traditional candidate gene-based studies with respect to common and complex diseases, in which a combination of multiple genetic variations contributes to an individual's risk. Therefore, an important extension of the work will be to develop SNP selection methods that scale up to the whole genome. In particular, we are interested in developing a genome-wide integrative SNP selection method that takes into account both functional significance and tagging effectiveness of SNPs. As discussed in Section 8.5, the greatest obstacle to the full scale extension is the current computational complexity. There are also other difficulties that complicate the selection procedure. That is, the genomic structure of the human genome is much more complicated than that of a single gene; It consists of both gene and intergenic regions, more complex linkage disequilibrium structures, and different levels of functional structures. Leveraging the qualitative characteristics of the genomic structure as well as reducing the computational complexity of the selection algorithm will be the key issue to address in order to apply the work on a genome-wide scale.

Bibliography

[1] H. Ackerman, S. Usen, R. Mott, A. Richardson, F. Sisay-Joof, P. Katundu, T. Taylor, R. Ward, M. Molyneux, M. Pinder, and D. P. Kwiatkowski. Haplotype analysis of the TNF locus by association efficiency and entropy. *Genome Biology*, 4(4):R24. 1–R24. 13, 2003.

[2] J. Akey, L. Jin, and M. Xiong. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, 9:291–300, 2001.

[3] Y. Akiyama. TFSEARCH: searching transcription factor binding sites. *Web Service: http://www.cbrc.jp/research/db/TFSEARCH.html*, 1998. Last accessed date: May 5, 2009.

[4] S. I. Ao, K. Yip, M. Ng, D. Cheung, P. Fong, I. Melhado, and P. C. Sham. CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, 21:1735–1736, 2005.

[5] Y. Aulchenko, T. I. Axenovich, I. Mackay, and C. M. van Duijn. miLD and booLD programs for calculation and analysis of corrected linkage disequilibrium. *Annals of Human Genetics*, 67:372–375, 2003.

[6] H. I. Avi-Itzhak, X. Su, and F. M. De La Vega. Selection of minimum subsets of

single nucleotide polymorphism to capture haplotype block diversity. *In Proceedings of Pacific Symposium on Biocomputing (PSB)*, pages 466–477, 2003.

[7] V. Bafna, B. V. Halldórsson, R. Schwartz, A. G. Clark, and S. Istrail. Haplotypes and informative SNP selection algorithms: don't block out information. *In Proceedings of the 7th International Conference on Computational Molecular Biology (RECOMB)*, pages 19–27, 2003.

[8] P. I. W. De Bakker, R. R. Graham, D. Altshuler, B. E. Henderson, and C. A. Haiman. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple population. *In Proceedings of Pacific Symposium on Biocomputing (PSB)*, 11:478–486, 2006.

[9] L. Bao, M. Zhou, and Y. Cui. nsSNPanalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research*, 33(Webserver issue):W480–W482, 2005.

[10] T. Barzuza, J. S. Beckmann, R. Shamir, and I. Peer. Computational problems in perfect phylogeny haplotyping: xor-genotypes and tag SNPs. *Combinatorial Pattern Matching, 15th Annual Symposium, In Proceedings of Lecture Notes in Computer Science*, 3109:14–31, 2004.

[11] P. Bhatti, D. M. Church, J. L. Rutter, J. P. Struewing, and A. J. Sigurdson. Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. *American Journal of Epidemiology*, 164(8):794–804, 2006.

[12] P. Bonizzoni, G. D. Vedova, R. Dondi, and J. Li. The haplotyping problem: an

overview of computational models and solutions. *Journal of Computer Science and Technology*, 18(6):675–688, 2003.

[13] P. E. Bonnen, P. J. Wang, M. Kimmel, R. Chakraborty, and D. L. Nelson. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Research*, 12:1846–1853, 2002.

[14] S. R. Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*, 124(5):439–450, 2008.

[15] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220:49–65, 1991.

[16] L. R. Brunham, R. R. Singaraja, T. D. Pape, A. Kejariwai, P. D. Thomas, and M. R. Hayden. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. *PLOS Genetics*, 1(6):739–747, 2005.

[17] K. M. Burkett, M. Ghadessi, B. McNeney, J. Graham, and D. Daley. A comparision of five methods for selecting tagging single-nucleotide polymorphisms. *BMC Genetics*, 6 (Suppl 1):S71, 2005.

[18] M. Burset, I. A. Seledtsov, and V. V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, 28(21):4364–4375, 2000.

[19] 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*, 67:543–556, 2003.

[20] C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429:446–452, 2004.

[21] 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*, 74(1):106–120, 2004.

[22] L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, and A. R. Krainer. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, 31(13):3568–3571, 2003. Web service available at http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home. Last accessed date: May 5, 2009.

[23] H. Chang and T. Fujita. PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. *Biochemical and Biophysical Research Communications*, 287:288–291, 2001.

[24] C. Chelala, A. Khan, and N. R. Lemoine. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5):655–661, 2009.

[25] E. X. Chen and L. L. Siu. Development of molecular targeted anticancer agents: successes, failures and future directions. *Current Pharmaceutical Design*, 11(2):265–272, 2005.

[26] B. N. Chorley, X. Wang, M. R. Campbell, G. S. Pittman, M. A. Noureddine, and

D. A. Bell. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutation Research*, 659(1-2):147–157, 2008.

[27] R. H. Chung and D. Gusfield. Perfect phylogeny haplotyper: haplotype inferral using a tree model. *Bioinformatics*, 19(6):780–781, 2003.

[28] D. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology of Evolution*, 7:111–122, 1990.

[29] D. Clayton. SNPHAP: a program for estimating frequencies of large haplotypes of SNPs. *http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt*, 2002.

[30] L. Conde, J. M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Research*, 34(Web Server Issue):W621–W625, 2006.

[31] L. Conde, J. M. Vaquerizas, C. Ferrer-Costa, X. de la Cruz, M. Orozco, and J. Dopazo. PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Research*, 33:W501–W505, 2005.

[32] L. Conde, J. M. Vaquerizas, J. Santoyo, F. Al-Shahrour, S. Ruiz-Llorente1, M. Robledo, and J. Dopazo. PupaSNP finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Research*, 32(Web Server Issue):W242–W248, 2004.

[33] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[34] F. J. Couch and B. L. Weber. Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene, Breast Cancer Information Core. *Human Mutation*, 8:8–18, 1996.

[35] F. G. Cozman. Generalizing variable elimination in Bayesian networks. *Workshop on Probabilistic Reasoning in Artificial Intelligence*, pages 27–32, 2000.

[36] D. C. Crawford and D. A. Nickerson. Definition and clinical importance of haplotypes. *Annual Review of Medicine*, 56:303–320, 2005.

[37] P. Czyzak and A. Jaszkiewicz. Pareto simulated annealing - a metaheuristic technique for multiple objective combinatorial optimization. *Journal of Multi-Criteria Decision Analysis*, 7:34–47, 1998.

[38] 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*, 29(2):229–232, 2001.

[39] P. de Bakker, R. Yelensky, I. Peer, S. Gabriel, M. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature Genetics*, 37:1217–1223, 2005.

[40] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine scale mapping. *Genomics*, 29:311–322, 1995.

[41] K. Ding, J. Zhang, K. Zhou, Y. Shen, and X. Zhang. htSNPer1. 0: software for haplotype block partition and htSNPs selection. *BMC Bioinformatics*, 6(1):38, 2005.

[42] K. Ding, K. Zhou, J. Zhang, J. Knight, X. Zhang, and Y. Shen. The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. *Molecular Biology and Evolution*, 22(1):148–159, 2005.

[43] Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping problem. *In Proceedings of the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 585–600, 2005.

[44] P. Divina, A. Kvitkovicova, E. Buratti, and I. Vorechovsky. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *European Journal of Human Genetics*, [Epub ahead of print]:1–7, 2009.

[45] P. A. Doris. Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension*, 39:323–331, 2002.

[46] P. Duggal, E. M. Gillanders, R. A. Mathias, G. P. Ibay, A. P. Klein, A. B. Baffoe-Bonnie, L. Ou, I. P. Dusenberry, Y. Tsai, P. S. Chines, B. Q. Doan, and J. E. Bailey-Wilson. Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium. *BMC Genetics*, 6 (Suppl 1):S73, 2005.

[47] J-D. Duha and D. G. Brown. Knowledge-informed Pareto simulated annealing for multi-objective spatial allocation. *Computers, Environment and Urban Systems*, 31:253–281, 2007.

[48] M. Ehrgott and X. Gandibleux. A survey and annotated bibliography of multiobjective combinatorial optimization. *OR Spectrum*, 22(4):425–460, 2000.

[49] H. Ellegren. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proceedings of the Royal Society. Series B. Biological Sciences*, 274(1606):1–10, 2007.

[50] W. G. Fairbrother, R. F. Yeh, P. A. Sharp, and C. B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002. Web service available at http://genes.mit.edu/burgelab/rescue-ese/. Last accessed date: May 5, 2009.

[51] D. Fallin, A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, and N. J. Schork. Genetic analysis of case/control data using estimated haplotype frequencies: application to apoe locus variation and alzheimer's disease. *Genome Research*, 11:143–151, 2001.

[52] D. Fallin and N. J. Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, 67:947–959, 2000.

[53] C. Ferrer-Costa, J. L. Gelp, L. Zamakola, I. Parraga, X. de la Cruz, and M. Orozco. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, 21(14):3176–3178, 2005.

[54] K. A. Frazer, L. Elnitski, D. M. Church, I. Dubchak, and R. C. Hardison. Cross-species sequence comparisons: a review of methods and available resources. *Genome Research*, 13:1–12, 2003.

[55] D. Fredman, G. Munns, D. Rios, F. Sjöholm, M. Siegfried, B. Lenhard, H. Lehväslaiho, and A. J. Brookes. HGVbase: a curated resource describing human

DNA variation and phenotype relationships. *Nucleic Acids Research*, 32(Database Issue):D516–D519, 2004.

[56] R. R. Freimuth, G. D. Stormo, and H. L. McLeod. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. *Human Mutation*, 25(2):110–117, 2005.

[57] N. Friedman, I. Nachman, and D. Peer. Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 206–215, 1999.

[58] S. B. Gabriel, S. F. Scahffner, H. Nguyen, J. M. Moore, J. Roy, B. lumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altschuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[59] T. Gerken, C. Tep, and J. Rarick. The role of peptide sequence and neighboring residue glycosylation on the substrate specificity of the uridine 5'-diphosphate-alpha-n-acetylgalactosamine:polypeptide n-acetylgalactosaminyl transferases t1 and t2: kinetic modeling of the porcine and canine submaxillary gland mucin tandem repeats. *Biochemistry*, 43(30):9888–9900, 2004.

[60] W. Gibson. Resolution of the species problem in African trypanosomes. *International Journal of Parasitology*, 37(8-9):829–838, 2007.

[61] D. B. Goldstein. Islands of linkage disequilibrium. *Nature Genetics*, 29:109–211, 2001.

[62] D. B. Goldstein, K. R. Ahmadi, M. E. Weale, and N. W. Wood. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends in Genetics*, 19(11):615–622, 2003.

[63] A. Goren, O. Ram, M. Amit, H. Keren, G. Lev-Maor, I. Vig, T. Pupko, and G. Ast. Comparative analysis identifies exonic splicing regulatory sequences - the complex definition of enhancers and silencers. *Molecular Cell*, 21(6):769–781, 2006. Web service available at http://ast.bioinfo.tau.ac.il/. Last accessed date: May 5, 2009.

[64] G. Greenspan and D. Geiger. Model-based inference of haplotype block variation. *In Proceedings of the Annual International Conference on Research in Computational Molecular Biology (RECOME)*, pages 131–137, 2003.

[65] B. V. Halldórsson, V. Bafna, R. Lippert, R. Schwatz, F. M. De La Vega, A. G. Clark, and S. Istrail. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Research*, 14:1633–1640, 2004.

[66] B. V. Halldórsson, S. Istrail, and F. M. De La Vega. Optimal selection of SNP markers for disease association studies. *Human Heredity*, 58:190–202, 2004.

[67] E. Halperin, G. Kimmel, and R. Shamir. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, 21(Suppl 1):i195–i203, 2005.

[68] J. Hampe, S. Schreiber, and M. Krawczak. Entropy-based SNP selection for genetic association studies. *Human Genetics*, 114:36–43, 2003.

[69] A. Harmosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30:52–55, 2002.

[70] D. L. Hartl and A. G. Clark. *Principles of Population Genetics 3rd*. Sunderland MA: Sinauer Associates, 1997.

[71] J. He and A. Zelikovsky. MLR-tagging: informative SNP Selection for Unphased Genotypes. *Bioinformatcis*, 22(20):2558–2561, 2006.

[72] P. Hedrick. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117(8):331–341, 1987.

[73] B. M. Hemminger, B. Saelim, and P. F. Sullivan. TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics*, 22(5):626–627, 2006.

[74] D. Holste, G. Huo, V. Tung, and C. B. Burge. Hollywood: a comparative relational database of alternative splicing. *Nucleic Acids Research*, 34(Database Issue):D56–D62, 2006.

[75] B. Horne and N. J. Camp. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26:11–21, 2004.

[76] C. Houdayer, C. Dehainault, C. Mattler, D. Michaux, V. Caux-Moncoutier, S. Pags-Berhouet, C. D. d'Enghien, A. Laug, L. Castera, M. Gauthier-Villars, and D. Stoppa-Lyonnet. Evaluation of *in silico* splice tools for decision-making in molecular diagnosis. *Human Mutation*, 29(7):975–982, 2008.

[77] H. Huang, T. Lee, S. Tseng, and J. Horng. KinasePhos: a web tool for

identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Research*, 33(Web Server Issue):W226–W229, 2005. Web service available at http://kinasephos.mbc.nctu.edu.tw/. Last accessed date: May 5, 2009.

[78] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Research*, 35(Database Issue):D610–617, 2007.

[79] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.

[80] J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Cézard, J. Belaiche, S. Almer, C. Tysk, C.A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou, and G. Thomas. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411(6837):599–603, 2001.

[81] X. de la Cruz J. R. Goñi and M. Orozco. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Research*, 32:354–360, 2004.

[82] F. Jensen. *Bayesian networks and decision graphs*. Springer-Verlag, New York, 1997.

[83] H. Jeong, I. Herskowitz, D. L. Kroetz, and J. Rine. Function-altering SNPs in the human multidrug transporter gene ABCB1 identified using a Saccharomyces-based assay. *PLOS Genetics*, 3(3):e39, 2007.

[84] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. L. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29(2):233–237, 2001.

[85] L. B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10:1435–1444, 2000.

[86] R. Judson, B. Salisbury, and J. Schneider. How many SNPs does a genome-wide haplotype map require? *Parmacogenomics*, 3:379–391, 2002.

[87] O. V. Kalinina, R. B. Russell, A. B. Rakhmaninova, and M. S. Gelfand. Computational method for predicting protein functional sites with the use of specificity determinants. *Molecular Biology*, 41(1):137–147, 2006.

[88] R. Karchin, M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler, and A. Sali. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, 21(12):2814–2820, 2005. Web service available at http://modbase.compbio.ucsf.edu/LS-SNP/. Last accessed date: May 5, 2009.

[89] X. Ke and L. R. Cardon. Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, 19(2):287–288, 2003.

[90] X. Ke, C. Durrant, A. P. Morris, S. Hunt, D. R. Bentley, P. Deloukas, and L. R. Cardon. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Human Molecular Genetics*, 13(21):2557–2565, 2004.

[91] X. Ke, M. M. Miretti, J. Broxholme, S. Hunt, S. Beck, D. R. Bentley, P. Deloukas, and L. R. Cardon. A comparison of tagging methods and their tagging space. *Human Molecular Genetics*, 14(18):2757–2767, 2005.

[92] A. E. Kel, E. Gösling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579, 2003.

[93] D. S. Kim. Thyroid cancer: are molecular studies making any difference? *Journal of Laryngology and Otology*, 121(10):917–926, 2007.

[94] N. Kim, A. V. Alekseyenko, M. Roy, and C. Lee. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Research*, 35(Database Issue):D93–D98, 2007.

[95] C. Kimchi-Sarfaty, J. M. Oh, I-W Kim, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. A silent polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315:525–528, 2007.

[96] G. Kimmel and R. Shamir. GERBIL: genotype resolution and block identification using likelihood. *In Proceedings of the National Academy of Sciences (PNAS)*, 102(1):158–162, 2005.

[97] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 22:671–680, 1983.

[98] A. P. Kirman. Pareto as an economist. *The New Palgrave: A Dictionary of Economics*, 5:804–808, 1987.

[99] N. A. Kolchanov, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, M. A. Pozdnyakov, N. L. Podkolodny, A. N. Naumochkin, and A. G. Romashchenko. Transcription regulatory regions database (TRRD): its status in 2002. *Nucleic Acids Research*, 30:312–317, 2002.

[100] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genetics*, 27:234–236, 2001.

[101] R. Kuhn, D. Karolchik, A. Zweig, H. Trumbower, D. Thomas, A. Thakkapallayil, C. Sugnet, M. Stanke, K. Smith, et al. The UCSC genome browser database: update 2007. *Nucleic Acids Research*, 35(Database Issue):D668–673, 2007. Downloaded date: Dec. 12, 2006.

[102] P. Y. Kwok and M. Xiao. Single-molecule analysis for molecular haplotyping. *Human Mutation*, 23(5):442–446, 2004.

[103] W. Lam and F. Bacchus. Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence*, 10:269–293, 1994.

[104] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8:995–1005, 2007.

[105] P. H. Lee. Computational haplotype analysis: an overview of computational methods

in genetic variation study. Technical Report 2006-512, Queen's University, Queen's Unviersity, Kingston, ON, Canada, February 2006.

[106] P. H. Lee, J. Jung, and H. Shatkay. Functionally informative tag SNP selection using a Pareto-optimal approach: playing the game of life. *Submitted to WABI 2009*, 2009.

[107] P. H. Lee and H. Shatkay. BNTagger: improved tagging SNP selection using Bayesian networks. *In Proceedings of the 14th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Supplement of Bioinformatics*, 22(14):e211–219, 2006.

[108] P. H. Lee and H. Shatkay. Two birds, one stone: selecting functionally informative tag SNPs for disease association studies. *In the Proceedings of the Workshop of Algorithms in Bioinformatics (WABI)*, pages 61–72, 2007.

[109] P. H. Lee and H. Shatkay. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Research*, 36(Database Issue):D820 – D824, 2008.

[110] P. H. Lee and H. Shatkay. Ranking single nucleotide polymorphisms by potential deleterious effects. *In the Proceedings of American Medical Informatics Association Annual Symposium (AMIA)*, 6:667–671, 2008.

[111] P. H. Lee and H. Shatkay. An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics*, 25(8):1048–1055, 2009.

[112] Z. Lin and R. B. Altman. Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics*, 75:850–861, 2004.

[113] P. M. Long, V. Varadan, S. Gilman, M. Treshock, and R. A. Servedio. Unsupervised evidence integration. In *Proceedings of the 22nd international conference on Machine learning (ICML)*, volume 119, pages 521–528, 2005.

[114] G. G. Loots and I. Ovcharenko. rVISTA 2. 0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Research*, 32(Web Server Issue):W217–W221, 2004.

[115] J. Loughlin, B. Dowling, K. Chapman, L. Marcelline, Z. Mustafa, L. Southam, A. Ferreira, C. Ciesielski, D. A. Carson, and M. Corr. Functional variants within the secreted frizzled-related protein 3 gene are associated with hip osteoarthritis in females. *In Proceedings of the National Academy of Sciences (PNAS)*, 101(26):9757–9762, 2004.

[116] X. Lu, T. Niu, and J. S. Liu. Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. *Genome Research*, 13:2112–2117, 2003.

[117] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database Issue):D54–D58, 2005.

[118] S. Maisnier-Patin, J. R. Roth, A. Fredriksson, T. Nystrom, O. G. Berg, and D. I. Andersson. Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nature Genetics*, 37:1376–1379, 2005.

[119] V. D. Marinescu, I. S. Kohane, and A. Riva. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 6:79, 2005.

[120] A. Mas, E. Blanco, G. Monux, E. Urcelay, FJ Serrano, EG de la Concha, and A. Martínez. DRB1-TNF-α-TNF-β haplotype is strongly associated with severe aortoiliac occlusive disease, a clinical form of atherosclerosis. *Human Immunology*, 66(10):1062–1067, 2005.

[121] V. Matys, O. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database Issue):D108–D110, 2006.

[122] D. Melzer. Genetic polymorphisms and human aging: association studies deliver. *Rejuvenation Research*, 11(2):523–526, 2008.

[123] 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*, 73:115–130, 2003.

[124] F. Monigatti, E. Gasteiger, A. Bairoch, and E. Jung. The Sulfinator: predicting Tyrosine Sulfation sites in protein sequences. *Bioinformatics*, 18(5):769–770, 2002. Web service available at http://ca.expasy.org/tools/sulfinator/. Last accessed date: May 5, 2009.

[125] M. V. Monsalve., F. M Salzano, J. L. Rupert, M. H. Hutz, K. Hill, A. M. Hurtado, P. W. Hochachka, and D. V. Devine. Methylenetetrahydrofolate reductase (MTHFR) allele frequencies in Amerindians. *Annals of Human Genetics*, 67:367371, 2003.

[126] A. Montpetit, M. Nelis, P. Laflamme, R. Magi, X. Ke, M. Remm, L. Cardon, T. J.

Hudson, and A. Metspalu. An evaluation of the performance of tag SNPs derived from hapmap in a caucasian population. *PLoS Genetics*, 2(3):e27, 2006.

[127] S. Mooney. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, 6(1):44–56, 2005.

[128] S. D. Mooney and R. B. Altman. MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, 19(14):1858–1860, 2003.

[129] V. K. Nalla and P. K. Rogan. Automated splicing mutation analysis by information theory. *Human Mutation*, 25:334–342, 2005.

[130] R. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

[131] M. Nei. *Molecular evolutionary genetics*. Columbia University Press, New York, 1987.

[132] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, 11(5):863–874, 2001. Web service available at http://blocks.fhcrc.org/sift/SIFT.html. Last accessed date: May 5, 2009.

[133] D. A. Nickerson, S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengaard, V. Salomaa, E. Boerwinkle, and C. F. Sing. Sequence diversity and large-scale typing of SNPs in the human apolipoprotine E gene. *Genome Research*, 10:1532–1545, 2000.

[134] T. Niu. Algorithms for inferring haplotypes. *Genetic Epidemiology*, 27(4):334–347, 2004.

[135] M. Pagano and K. Gauvreau. *Principles of Biostatistics, Second Edition*. Duxbury Thomson Learning, 2000.

[136] L. Palmer and L. Cardon. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366:1223–1234, 2005.

[137] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, and D. P. McDonough. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1722, 2001.

[138] E. Patin, L. B. Barreiro, P. C. Sabeti, F. Austerlitz, F. Luca, A. Sajantila, D. M. Behar, O. Semino, A. Sakuntabhai, N. Guiso, B. Gicquel, K. McElreavey, R. M. Harding, E. Heyer, and L Quintana-Murci. Deciphering the ancient and complex evolutionary history of human arylamine n-acetyltransferase genes. *American Journal of Human Genetics*, 78:423–436, 2006.

[139] F. P. Perera and I. B. Weinstein. Molecular epidemiology and carcinogen-DNA adduct detection: New approaches to studies of human cancer causation. *Journal of Chronic Diseases*, 35(7):581–600, 1982.

[140] M. Pertea, X. Lin, and S. L. Salzberg. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research*, 29(5):1185–1190, 2001.

[141] P. C. Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Review Genetics*, 9(11):855–867, 2008.

[142] A. R. Pico, I. V. Smirnov, J. S. Chang, R. F. Yeh, J. L. Wiemels, J. K. Wiencke, T. Tihan, B. R. Conklin, and M. Wrensch. SNPlogic: an interactive single nucleotide

polymorphism selection, annotation, and prioritization system. *Nucleic Acids Research*, 32(Database Issue):D803–D809, 2009.

[143] J. V. Ponomarenko, G. V. Orlova, T. I. Merkulova, E. V. Gorshkova, O. N. Fokin, G. V. Vasiliev, A. S. Frolov, and M. P. Ponomarenko. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. *Human Mutation*, 20:239–248, 2002.

[144] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant... or not? *Human Molecular Genetics*, 11(20):2417–2423, 2002.

[145] L. Prokunina and M. E. Alarcon-Riquelme. Regulatory SNPs in complex diseases: their identification and functional validation. *Expert Reviews in Molecular Medicine*, 6:1–15, 2004.

[146] Z. Qin, S. Gopalakrishnan, and G. Abecasis. An efficient comprehensive search algorithm for tagSNP selection using linkage disequilbrium criteria. *Bioinformatics*, 22(2):220–225, 2005.

[147] V. B. P. Ramensky and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acid Research*, 30(17):3894–3900, 2002. Web service available at http://genetics.bwh.harvard.edu/pph/. Last accessed date: May 5, 2009.

[148] T. R. Rebbeck, C. B. Ambrosone, D. A. Bell, S. J. Chanock, R. B. Hayes, F. F. Kadlubar, and D. C. Thomas. SNPs, haplotypes, and cancer: applications in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention*, 13(5):681–687, 2004.

[149] T. R. Rebbeck, M. Spitz, and X. Wu. Assessing the function of genetic variants in candidate gene association studies. *Nature Review Genetics*, 5:589–597, 2004.

[150] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Israel, 1997.

[151] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411:199–204, 2001.

[152] A. Reif, S. Herterich, A. Strobel, et al. A neuronal nitric oxide synthase NOS-I haplotype associated with schizophrenia modifies prefrontal cortex function. *Molecular Psychiatry*, 11(3):286–300, 2006.

[153] J. Reumers, J. Schymkowitz, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, and F. Rousseau. SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Research*, 33(Database Issue):D527–D532, 2005.

[154] M. J. Rieder, S. L. Taylor, A. G. Clark, and D. A. Nicerson. Sequence variance in the human angiotensin converting enzyme. *Nature Genetics*, 22:59–62, 1999.

[155] A. Riva and I. S. Kohane. A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics*, 5:33, 2004.

[156] K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang, Y. Takeshima, M. Matsuo, G. Sobue, and K. Ohno. *In vitro* and *in silico* analysis reveals an efficient algorithm

to predict the splicing consequences of mutations at the 5' splice sites. *Nucleic Acids Research*, 35(18):5995–6003, 2007.

[157] R. M. Salem, J. Wessel, and N. J. Schork. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics*, 2(1):39–66, 2005.

[158] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database Issue):D91–D94, 2004.

[159] A. Sandelin, W. W. Wasserman, and B. Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research*, 32(Web Server Issue):W249–W252, 2004. Web service available at http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/. Last accessed date: May 5, 2009.

[160] M. Sato, T. Sato, T. Izumo, and T. Amagasa. Genetic polymorphism of drug-metabolizing enzymes and susceptibility to oral cancer. *Carcinogenesis*, 20:1927–1931, 1999.

[161] C. T. Saunders and D. Baker. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology*, 322:891–901, 2002.

[162] S. Savas, D. Y. Kim, M. F. Ahmad, M. Shariff, and H. Ozcelik. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 13(5):801–807, 2004.

[163] T. G. Schulze, K. Zhang, Y. Chen, N. Akula, F. Sun, and F. J. McMahon. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Human Molecular Genetics*, 13(3):335–342, 2004.

[164] P. Sebastiani, R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane, and M. F. Ramoni. Minimal haplotype tagging. *In Proceedings of the National Academy of Sciences (PNAS)*, 100(17):9900–9905, 2003.

[165] B. S. Shastry. SNPs and haplotypes: genetic markers for disease and drug response (review). *International Journal of Molecular Medicine*, 11:379–382, 2003.

[166] B. S. Shastry. SNPs in disease gene mapping, medical drug development and evolution. *Journal of Human Genetics*, 52:871–880, 2007.

[167] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[168] N. Shimizu, M. Ohtsubo, and S. Minoshima. MutationView/KMcancerDB: A database for cancer gene mutations. *Cancer Science*, 98(3):259–267, 2007.

[169] A. Sjalander, R. Birgander, A. Rannug, A. K. Alexandrie, G. Tornling, and G. Beckman. Association between the p21 codon 31A1 (arg) allele and lung cancer. *Human Heredity*, 46:221–225, 1996.

[170] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper. Human gene mutation database. *Human Mutation*, 21:577–581, 2003.

[171] J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J. H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, 2001.

[172] N. O. Stitziel, T. A. Binkowski, Y. Y. Tseng, S. Kasif, and J. Liang. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Research*, 32(Database Issue):D520–D522, 2004.

[173] K. E. Sullivan, C. Wooten, D. Goldman, and M. Petri. Mannose-binding protein genetic polymorphisms in black patients with systemic lupus erythematosus. *Arthritis & Rheumatism*, 39:20462051, 1996.

[174] H. K. Tabor, N. J. Risch, and R. M. Myers. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Review Genetics*, 3:391–397, 2002.

[175] N. E. Taylor and E. A. Greene. PARSESNP: a tool for the analysis of nucleotide polymorphisms. *Nucleic Acids Research*, 31(13):3808–3811, 2003.

[176] P. D. Thomas and A. Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. mendelian disease: evolutionary evidence for differences in molecular effects. *In Proceedings of the National Academy of Sciences (PNAS)*, 101(43):15398–15403, 2004.

[177] H. Ueda, J. M. M. Howson, L. Esposito, J. Heward, H. Snook, G. Chamberlain, D. B. Rainbow, K. M. D. Hunter, A. N. Smith, G. Di Genova, et al. Association of the T-cell regulatory gene CTLA4 with susceptiblility to autoimmune disease. *Nature*, 423:506–511, 2003.

[178] University of Utah Genome Center. GeneSNPs. Web service available at: http://www.genome.utah.edu/genesnps/, 2007.

[179] K. P. Vatsis, K. J. Martell, and W. W. Weber. Diverse point mutations in the human gene for polymorphic N-acetyltransferase. *In Proceedings of the National Academy of Sciences (PNAS)*, 88:6333–6337, 1991.

[180] I. S. Vizirianakis. Clinical translation of genotyping and haplotyping data: implementation of *in vivo* pharmacology experience leading drug prescription to pharmacotyping. *Clinical Pharmacokinetics*, 46(10):807–824, 2007.

[181] P. Wang, M. Dai, W. Xuan, Richard C. McEachin, A. U. Jackson, L. J. Scott, B. Athey, S. J. Watson, and F. Meng. SNP function portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, 22(14):e523–e529, 2006.

[182] X. Wu, A. Luke, M. Rieder, K. Lee, E. J. Toth, D. Nickerson, X. Zhu, D. Kan, and R. S. Cooper. An association study of angiotensiongen polymorphisms with serum level and hypertension in an African-American population. *Journal of Hypertension*, 21(10):1847–1852, 2003.

[183] E. P. Xing, R. Sharan, and M. I. Jordan. Bayesian haplotype inference via the dirichlet process. *In Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 879–886, 2004.

[184] H. Xu, S. G. Gregory, E. R. Hauser, and J. E. Stenger. SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics*, 21(22):4181–4186, 2005.

[185] Y. Yamaguchi-Kabata, M. K. Shimada, Y. Hayakawa, S. Minoshima, R. Chakraborty, T. Gojobori, and T. Imanishi1. Distribution and effects of nonsense polymorphisms in human genes. *PLoS ONE*, 3(10):e3393, 2008.

[186] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biololgy*, 11(2-3):377–394, 2004.

[187] G. Yeo and C. B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *In Proceedings of the National Academy of Sciences (PNAS)*, 101(44):15700–15705, 2004. Web service available at http://genes.mit.edu/burgelab/rescue-ese/. Last accessed date: May 5, 2009.

[188] H. Y. Yuan, J. J. Chiou, W. H. Tseng, C. H. Liu, C. K. Liu, Y. J. Lin, H. H. Wang, A. C. Yao, Y. T. Chen, and C. N. Hsu. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Research*, 34(Web Server Issue):W635–W641, 2006.

[189] P. Yue, E. Melamud, and J. Moult. SNPs3D: candidate gene and SNP selection for

association studies. *BMC Bioinformatics*, 7:166, 2006. Web service available at http://www.snps3d.org/. Last accessed date: May 5, 2009.

[190] P. Yue and J. Moult. Identification of deleterious human SNPs. *Journal of Molecular Biology*, 356:1263–1274, 2006.

[191] S. H. Zeisel. Gene response elements, genetic polymorphisms and epigenetics influence the human dietary requirement for choline. *IUBMB Life*, 59(6):380–387, 2007.

[192] K. Zhang. Dynamic programming algorithm for haplotype block partitioning: application to human chromosome 21 haplotype data. *In Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 332–340, 2003.

[193] K. Zhang, P. Calabrese, M. Nordborg, and F. Sun. Haplotype block structure and its application to association studies: power and study designs. *American Journal of Human Genetics*, 71:1386–1394, 2002.

[194] K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *In Proceedings of the National Academy of Sciences (PNAS)*, 99(11):7335–7339, 2002.

[195] K. Zhang and L. Jin. HaploBlockFinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1301, 2003.

[196] 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 progrmming algorithms. *Bioinformatics*, 21(1):131–134, 2005.

[197] K. Zhang, Z. S. Qin, J. S. Liu, T. Chen, M. S. Waterman, and F. Sun. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Research*, 14:908–916, 2004.

[198] L. H. Zhang, D. P. Liu, and C. C. Liang. Finding regulatory sequences. *International Journal of Biochemistry*, 35(1):95–103, 2003.

[199] P. Zhang, H. Sheng, and R. Uehara. A double classification tree search algorithm for index SNP selection. *BMC Bioinformatics*, 5(89), 2004.

[200] X. H-F. Zhang and L. A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes and Development*, 18:1241–1250, 2004. Web service available at http://cubweb.biology.columbia.edu/pesx/. Last accessed date: May 5, 2009.

[201] H. Zhao, R. Pfeiffer, and M. H. Gail. Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4(2):171–178, 2003.

[202] T. Zhao, L. W. Chang, H. L. McLeod, and G. D. Stormo. PromoLign: a database for upstream region analysis and SNPs. *Human Mutation*, 23(6):534–539, 2004.

[203] C. L. Zheng, Y. S. Kwon, H. R. Li, K. Zhang, G. Coutinho-Mansfield, C. Yang, T. M. Nair, M. Gribskov, and X. D. Fu. MAASE: an alternative splicing database designed for supporting splicing microarray applications. *RNA*, 11(12):1767–76, 2006.

[204] Y. Zhu, A. Hoffman, X. Wu, H. Zhang, Y. Zhang, D. Leaderer, and T. Zheng. Correlating observed odds ratios from lung cancer case-control studies to SNP functional scores predicted by bioinformatics tools. *Mutation Research*, 639:80–88, 2008.

Appendix A

Program Source Codes

A.1 BNTagger

Server: redtape.cs.queensu.ca

Home Directory: /fs/hs/projects/BNTagger

A.1.1 To run Bayesian networks of SNPs

COMMAND

To learn all SNP datasets in a dataset file

java phd.bn.learning.LearnerMain DATA_SET_NAME DATA_NAME CANDIDATE _PARENTS_MODE redtape;

COMMAND OPTIONS:

* DATA_SET_NAME: the name of datasets such as LOCV (for leave one out cross validation) or 10fold_10set (for 10 fold cross validation 10 times)

* DATA_NAME: gene symbol such as ACE, LPR

* CANDIDATE_PARENTS_MODE

VALUE	MEANING
0	Probability.CORRELATION
1	LD.D_PRIME
2	LD.LARGE_DELTA_SQUARE
3	InformationTheory.MI
4	LD.PROXIMITY
5	LD.Q
6	LD.SMALL_D
7	LD.SMALL_DELTA
8	NO_RESTRICTION

INPUT_FILE:

A metafile with the information about datasets should be saved as a text file at /fs/hs/projects/BNTagger/data/DATA_NAME/DATA_SET_NAME.txt, and its exemplary contents are shown for gene ACE and dataset name LOCV2 as follows:

DATA_NAME TRAINING_DATA TEST_DATA FREQUENCY

ACE/LOCV2/1.6 ACE/LOCV2/training.1.6.txt ACE/LOCV2/test.1.6.txt 1

ACE/LOCV2/1.1 ACE/LOCV2/training.1.1.txt ACE/LOCV2/test.1.1.txt 1

ACE/LOCV2/6.7 ACE/LOCV2/training.6.7.txt ACE/LOCV2/test.6.7.txt 2

ACE/LOCV2/6.9 ACE/LOCV2/training.6.9.txt ACE/LOCV2/test.6.9.txt 1

ACE/LOCV2/1.7 ACE/LOCV2/training.1.7.txt ACE/LOCV2/test.1.7.txt 1

ACE/LOCV2/2.3 ACE/LOCV2/training.2.3.txt ACE/LOCV2/test.2.3.txt 1

ACE/LOCV2/4.5 ACE/LOCV2/training.4.5.txt ACE/LOCV2/test.4.5.txt 1

ACE/LOCV2/11.10 ACE/LOCV2/training.11.10.txt ACE/LOCV2/test.11.10.txt 1

ACE/LOCV2/1.8 ACE/LOCV2/training.1.8.txt ACE/LOCV2/test.1.8.txt 1

ACE/LOCV2/12.13 ACE/LOCV2/training.12.13.txt ACE/LOCV2/test.12.13.txt 1

To learn a specific set of consecutive SNP datasets in a dataset file

java phd.bn.learning.LearnerMain DATA_SET_NAME DATA_NAME CANDIDATE

_PARENTS_MODE START_DATA_INDEX END_DATA_INDEX redtape;

A.1.2 To select tag SNPs and evaluate the accuracy

COMMAND

To learn all SNP datasets in a dataset file

java phd.bn.prediction.PredictorMain DATA_NAME CANDIDATE

_PARENTS_MODE START_DATA_INDEX END_DATA_INDEX MAX_SNP_NO MODE

DATA_SET_NAME redtape;

COMMAND OPTIONS:

* MAX_SNP_NO: the total number of SNPs

* MODE: P (for prediction) or S (for evaluation summary)

A.2 F-SNP-Score

Server: redtape.cs.queensu.ca

Home Directory: /fs/hs/projects/F-SNP/perl/batch

A.2.1 To prepare datasets

(1) As a primary dataset, save a list of gene symbols to the following path as a text file delimited by newlines.

PATH

/fs/hs/projects/F-SNP/data/gene/GENE_SYMBOL_FILE_NAME

(2) Suppose that the name of the gene symbol file, saved as stated above, is all_gene.txt. Then, its absolute path should be as follows: /fs/hs/projects/F-SNP/data/gene/all_gene.txt.
To prepare secondary gene datasets (indexed from zero to 4) with a specified upstream/downstream region (in this example, 10000), execute the following four commands sequentially.

COMMANDs

/opt/perl5/bin/perl MainDataPreparation.cgi list2gene all_gene.txt no yes

/opt/perl5/bin/perl -w MainDataPreparation.cgi gene2snp snp_list no all_gene.out.txt 10000 10000

/opt/perl5/bin/perl -w MainDataPreparation.cgi gene2snp hapmap no all_gene.out.txt 10000 10000

/opt/perl5/bin/perl MainDataPreparation.cgi gene2snp no_function yes all_gene.out.txt 10000 10000 no

A.2.2 To run F-SNP batch services

COMMAND

perl MainRunFSNP.cgi FILE_IX (or ”all”) PROGRAM (or ”all”) all_gene.out.txt 10000 10000

COMMAND OPTIONS:

* FILE_IX

VALUE	MEANING
zero or 3	Run java programs such as ns/NSMain, sr/SRMain, pt/PTMain
1 or 2	Run java program, tr/TRMain
4	Run java program, sr/SRMain

* PROGRAM:

type the name of a specific program to run, e.g., PolyPhen, Consite

COMMAND EXAMPLES

To run all integrated programs using all functional files indexed from 0 to 4

/opt/perl5/bin/perl -w MainRunFSNP.cgi all all all_gene.out.txt 10000 10000

To run all integrated programs using each functional file

/opt/perl5/bin/perl -w MainRunFSNP.cgi zero all all_gene.out.txt 10000 10000

/opt/perl5/bin/perl -w MainRunFSNP.cgi 1 all all_gene.out.txt 10000 10000

/opt/perl5/bin/perl -w MainRunFSNP.cgi 2 all all_gene.out.txt 10000 10000

/opt/perl5/bin/perl -w MainRunFSNP.cgi 3 all all_gene.out.txt 10000 10000

/opt/perl5/bin/perl -w MainRunFSNP.cgi 4 all all_gene.out.txt 10000 10000

To run an individual program

phd/snp_function_prediction/tools/$COMMAND\_LIST$[j] redtape PROGRAM SNP_LIST_FILE

- *Prerequisite*: setenv CLASSPATH .:/fs/hs/projects/F-SNP/perl/batch/phd/ext/mysql-connector-java-3.1.8-bin.jar
- *Input*: SNP_LIST_FILE CommonVar::FSNP_dir . GENE_LIST_FILE.up_pos.function.FILE_IX.txt
- *Output*: CommonVar::fsnp_prediction_dir . TOOL_CATEGORY . GENE_LIST_FILE.up_pos.function.FILE_IX . TOOL.txt

A.2.3 To update F-SNP db

COMMAND

perl MainUpdateFSNP.cgi all_gene.out.txt TOOL_CATEGORY MODE UP DOWN FUNC_IX

COMMAND OPTIONS:

* TOOL_CATEGORY

VALUE	MEANING
0	protein_coding
1	splicing_regulation
2	post_translation
3	transcriptional_regulation
-1	all categories

* MODE

either *upload_result, upload_file, upload_summary*, or *all*

COMMAND EXAMPLES

/opt/perl5/bin/perl -w MainUpdateFSNP.cgi all_gene.out.txt 3 all 10000 10000 1

/opt/perl5/bin/perl -w MainUpdateFSNP.cgi all_gene.out.txt 3 all 10000 10000 2

/opt/perl5/bin/perl -w MainUpdateFSNP.cgi all_gene.out.txt 1 all 10000 10000 4

/opt/perl5/bin/perl -w MainUpdateFSNP.cgi all_gene.out.txt -1 all 10000 10000 3

/opt/perl5/bin/perl -w MainUpdateFSNP.cgi all_gene.out.txt -1 all 10000 10000 0

A.3 FITS-Selector

Server: redtape.cs.queensu.ca

Home Directory: /fs/hs/projects/BNTagger

COMMAND

java phd.evaluation.FBNTagger.FBNTaggerEvaluator

A.4 SA1

Server: redtape.cs.queensu.ca

Home Directory: /fs/hs/projects/FITagger/perl/pareto

COMMAND

To select functionally informative tag SNPs

/fs/hs/projects/FITagger/perl/pareto/main.cgi GENE_LIST_FILE

To analyze the performance of the selected functionally informative tag SNPs

/fs/hs/projects/FITagger/perl/pareto/main.cgi GENE_IX