Computational Haplotype Analysis: An overview of computational methods in genetic variation study

Phil Hyoun Lee

Advisor: Dr. Hagit Shatkay

A depth paper submitted to the
School of Computing
conforming to the requirements for
the degree of Doctor of Philosophy

School of Computing
Queen's University
Kingston, Ontario, Canada, K7L 3N6
February 6, 2006

Contents

List of Figures

Abstract

One of the major interests of current genomics research is *disease-gene association*, that is, identifying which DNA variation or a set of DNA variations is highly associated with a specific disease. In particular, single nucleotide polymorphisms (SNPs), which are the most common form of DNA variation on the human genome, and a set of SNPs on one chromosome, referred to as a *haplotype*, are at the forefront of the disease-gene association studies. In general, when haplotype information is used for studying disease-gene association, it is called *haplotype analysis*. Numerous studies have shown that haplotype analysis can successfully identify the DNA variations relevant to several common and complex human diseases. However, despite its advantages over other approaches, the use of haplotype analysis has been limited due to the high cost and long operation time of bio-molecular methods for obtaining the haplotype information. To address this limitation, two computational procedures, namely, *Haplotype Phasing* and *Tag SNP Selection* have been incorporated in haplotype analysis, and now provide the most practical framework for conducting large-scale association studies. In this depth paper, we introduce an overview of computational haplotype analysis, survey the existing approaches for Haplotype Phasing and Tag SNP Selection, and discuss their open problems. Given the current state of the field, as presented in this survey, we plan to conduct further research in the area of Tag SNP Selection.

Chapter 1

Introduction

Understanding the genomic differences in the human population is one of the primary challenges of current genomics research [65]. The human genome can be viewed as a sequence of three billion letters from the nucleotide-alphabet $\{A,C,G,T\}$, and this sheer amount of data requires massive computational analysis. In more than 99 percent of the positions on the genome, the same nucleotide is shared across the population. However, one percent of the genome includes numerous genetic variations such as different nucleotide occurrences, deletion/insertion of a nucleotide, 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.

Early research [41, 55, 91] has focused on identifying which positions of the human genome are commonly variant and which are typically invariant. Generally, when a variation occurs in at least a certain percentage of a population (typically around 5-10%), it is considered a *common* variation [65]. To date, millions of the common DNA variations have been identified and are accessible in public databases [41, 55, 91]. These identified common variations, usually involve the substitution of a single nucleotide, and are called *single nucleotide polymorphisms* (SNPs - pronounced *snips*). The nucleotide at a position in which a SNP occurred is called an *allele*. The one with the dominant occurrence within a population is called the *major allele*, while the others are called the *minor alleles*. For example, if 80 percent of a population has the nucleotide A at a certain position of the genome while 20 percent of the population has the nucleotide T at the same position, then A is the major allele of the SNP, and T is the minor allele for it.

As a next step of genetic variation study, current interest is focused on *disease-gene association*, that is, identifying which DNA variation or a set of DNA variations is highly associated with a specific disease. Simple Mendelian diseases (e.g., Huntington disease, Sickle Cell Anemia, tuberous sclerosis, cystic fibrosis, etc.) are caused by an abnormal alteration of a single gene. However, most current common diseases (e.g., cancer, heart disease, obesity, diabetes, hypertension, asthma, etc.) are known to be affected by a combination of

two or more mutated genes along with certain environmental factors; thus, they are often called *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-scale population. Thus, traditional family-based analysis methods that were useful for a simple Mendelian disease [50], do not perform well for complex and common disease studies [33].

Recently, *haplotype*[1] *analysis* has been successfully applied to the identification of the DNA variations relevant to several common and complex diseases [11, 30, 54, 75, 87], and is now considered the most promising method for studying complex disease-gene association [58, 77, 90, 111]. In this depth report, I survey the existing approaches for performing two main computational procedures in haplotype analysis: *Haplotype Phasing* and *Tag SNP Selection*, and provide an overview of computational haplotype analysis.

The rest of the paper is organized as follows: Chapter 2 provides an overview of computational haplotype analysis; Chapters 3 and 4 introduce and discuss Haplotype Phasing and Tag SNP Selection, respectively; Chapter 5 concludes and outlines future research; and Appendix A contains statistical tests for Haplotype-Disease Association.

[1] A haplotype is a set of SNPs present on one chromosome. All definitions and terms pertaining to computational haplotype analysis are going to be introduced and defined in the next chapter.

Chapter 2

Computational Haplotype Analysis

This chapter starts by defining the basic genetic concepts in computational haplotype analysis. It then provides an overview of computational haplotype analysis, including its general objective, distinguishing features from previous approaches and essential computational procedures.

2.1 Basic Concepts in Computational Genetic Analysis

Population genetics studies genetic change in populations in order to understand the evolutionary significance of genetic variations, both within and between species [50]. Thus, it provides the basis for common and complex disease-gene association, that is, identifying a set of DNA variations that is common enough to be prevalent in the human population and has a causal connection to the elevated risk of a complex disease [111]. Since the ultimate aim of computational haplotype analysis is disease-gene association, we first need to define some basic concepts in population genetics to understand computational haplotype analysis.

2.1.1 Haplotypes, Genotypes, and Phenotypes

Suppose that we have chromosome samples from six individuals. Three of them have lung cancer and the others do not. We aim to identify a set of DNA variations associated with lung cancer using the chromosome samples. Due to experimental cost and time, only a limited region of the chromosome that 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 *locus*. A locus can be as large as a whole chromosome or as small as a part of a gene.

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

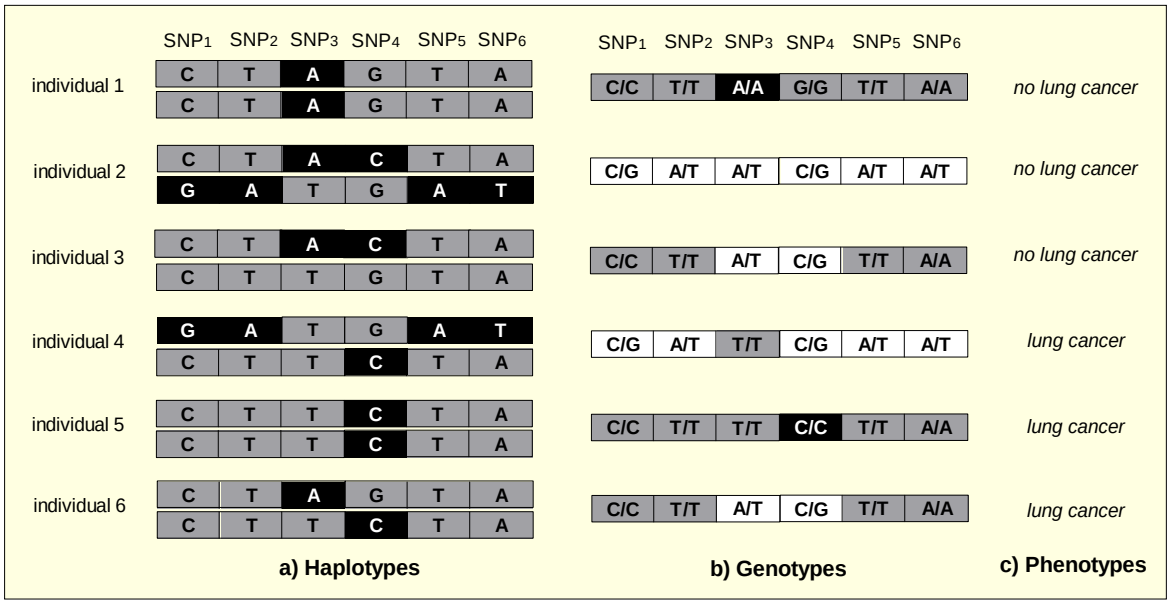

Figure 2.1: Haplotypes, Genotypes, and Phenotypes

the mother. Thus, every individual in our sample also has two alleles for each SNP, one on the paternal chromosome and the other on the maternal chromosome. For each SNP, the allele on one chromosome and the allele on the other can be either identical or different. When they are the same, the SNP is called *homozygous*. When they are different, the SNP is called *heterozygous*.

Suppose that our target locus contains six SNPs, and each SNP has only two different alleles (i.e., SNPs are assumed to be *bi-allelic*). The allele information is as shown in Figure 2.1-a). The major allele of the SNP is colored gray, and the minor is colored black. Each individual has *two* sets of six SNPs constructed from his/her two chromosomes. A set of SNPs present on one chromosome is referred to as a *haplotype* [19]. Notice that 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 long operation time, they are mainly used for small to moderate-size samples (typically from several to tens of individuals) [19]. For large-scale samples (typically from hundreds to thousands of individuals), high-throughput bio-molecular methods are used to identify the alleles of the target locus for each individual. The main limitation of the high-throughput methods lies in their lack of ability to distinguish the source chromosomes of each allele. Typically, such methods simply associate the two alleles with the SNP position, but do not determine their source chromosomes. This 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 sample. 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 with disease are referred to as *cases*, while the ones with no disease are referred to as *controls*. Figure 2.1-c) displays the phenotype information for our sample.

2.1.2 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) [33]. 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' two copies of chromosomes, and is passed by inheritance to a descendant. Figure 2.2 illustrates this 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* between SNPs is called *linkage equilibrium*.

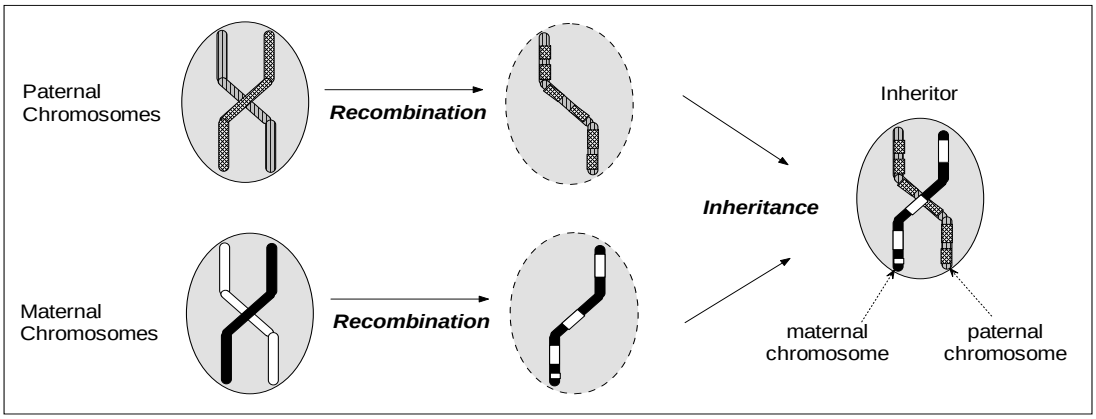

Figure 2.2: Recombination and Inheritance

Suppose that we have two SNPs s_1 and s_2. Let $|s_1|$ and $|s_2|$ denote the number of alleles that the SNPs s_1 and s_2 have, respectively. Let s_{1i} denote the i^{th} allele of the first SNP s_1, and s_{2j} denote the j^{th} allele of the second SNP s_2, where $i = 1, ..., |s_1|$ and $j = 1, ..., |s_2|$.

Under *linkage equilibrium*, the joint probability of two alleles s_{1i} and s_{2j} is expected to be equal to the product of the alleles' individual probabilities since s_1 and s_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 is not satisfied by two SNPs, that is, when their alleles are not independent, we consider them to be in a state of *linkage disequilibrium* (LD). In principle, when their allele dependence is large [1], 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. The probability of recombination increases with the distance between two SNPs [19]. 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 consisting of the SNPs is much smaller than expected under linkage equilibrium.

Recently, large-scale LD studies [20,32,84] 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 distinct number of 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 [23, 88]. 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 and the limited number of haplotypes due to the block structure of the human genome have provided the basis of computational haplotype analysis for disease-gene association. We introduce the detail of computational haplotype analysis in the following sections.

2.2 Computational Haplotype Analysis

Our ultimate goal is to identify a set of DNA variations that is highly associated with a specific disease. Haplotype, genotype, or even single-SNP information can be used to examine the association of genetic variation with the target disease. When haplotype information is used for studying disease-gene association, it is called *haplotype analysis*. *Single-SNP analysis* and *Genotype analysis* refer to the studies that use single-SNP information and genotype information, respectively.

[1] The absolute threshold differs in each LD measure. For details, refer to LD review articles [21, 57]

Haplotype analysis has several advantages compared to single-SNP analysis and genotype analysis. Single-SNP analysis cannot identify the association where a combination of several SNPs on one chromosome (i.e., a haplotype) is required to affect the phenotype of an individual [3,20,104]. 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 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.

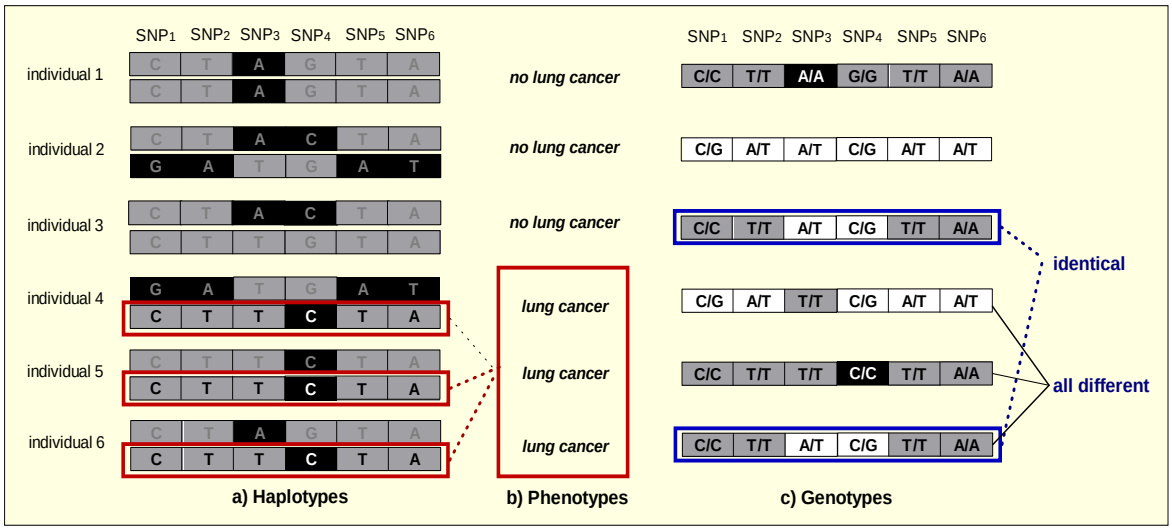

Figure 2.3: Difference between Haplotype Analysis and Genotype Analysis

Genotypes do not contain the source chromosome information, known as *phase*, thus they often hide the obvious association existing 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*, that is the one associated with the lung cancer phenotype, and the other one is unique for each case. Although all cases share the exact same haplotype *CTTCTA*, their genotypes, in Figure 2.3-c), all look different due to their unique haplotype. Worse, the genotype of individual 6, who has lung cancer, is identical to that of individual 3, who has no 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 by the high cost and long operation time of bio-molecular methods for obtaining the haplotype information. However, two computational procedures, *Haplotype Phasing* and *Tag SNP Selection* address

this problem, and greatly promote the use of haplotype analysis for disease-gene association. *Haplotype Phasing* deduces haplotype information from genotype data. *Tag SNP Selection* selects a subset of SNPs on a haplotype that is sufficiently informative to study disease-gene association but still small enough to reduce the genotyping overhead. When these computational procedures are used for haplotype analysis, the whole procedure is referred to as *computational haplotype analysis*.

Figure 2.4 summarizes the general procedures of computational haplotype analysis and of traditional haplotype analysis. Bio-molecular experiments are displayed in white boxes, and computational and statistical procedures are displayed in black boxes. Computational haplotype analysis consists of *Haplotype Phasing*, *Tag SNP Selection*, and *Haplotype-Disease Association* along with two genotyping experiments. Initially, a relatively small number of individuals are genotyped from a target population, and their haplotypes are inferred using *Haplotype Phasing* algorithms. Then, *Tag SNP Selection* algorithms select a small subset of SNPs on the haplotypes, which can represent the identified haplotypes with little loss of information. Using the selected small number of SNPs, second genotyping is done for a large number of individuals. Again, *Haplotype Phasing* algorithms are used to infer the haplotypes from these genotype data. Finally, *Haplotype-Disease Association*, that is identifying the association of a haplotype or a set of haplotypes with a target disease, is performed on the haplotypes.

In contrast to computational haplotype analysis, traditional haplotype analysis relies on bio-molecular experiments to directly obtain haplotype information. Thus, it can provide more accurate haplotype information than computational procedures, and, in the near future, the bio-molecular methods might become a standard technique for haplotype analysis [80]. However, until then, the two computational procedures, Haplotype Phasing and Tag SNP Selection, are expected to be of much use for large-scale association studies.

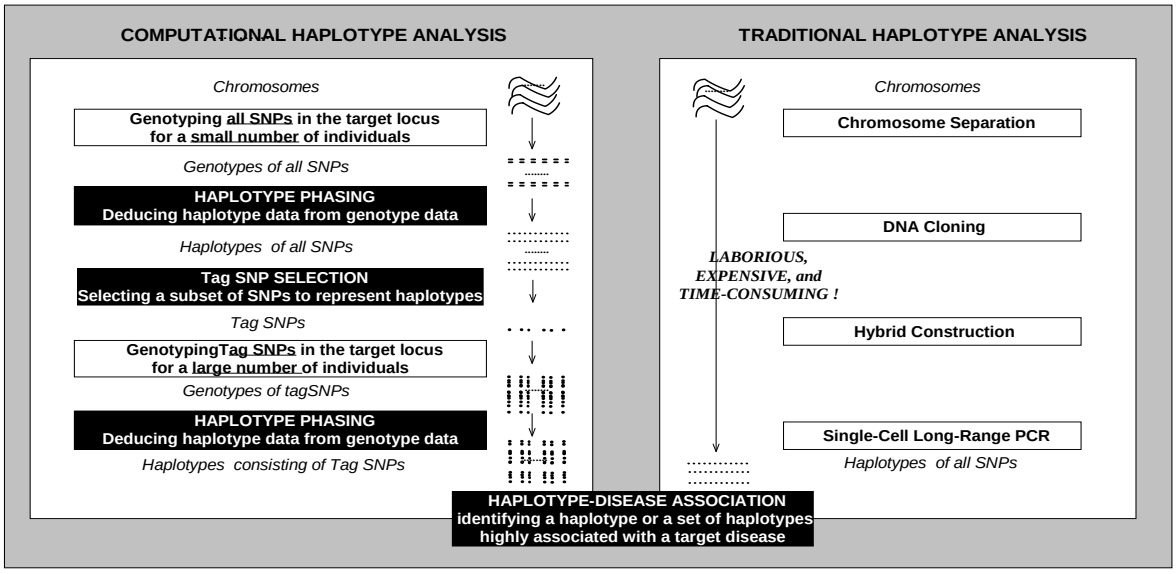

Figure 2.4: Computational Haplotype Analysis and Traditional Haplotype Analysis

Chapter 3

Haplotype Phasing

This chapter introduces Haplotype Phasing, that is, the computational process of deducing haplotypes from genotypes. The main concepts are introduced and formally defined in section 3.1. Haplotype Phasing algorithms are categorized by the four major approaches on which they are based: (1) parsimony; (2) phylogeny; (3) maximum-likelihood; and (4) Bayesian inference. Sections 3.2 to 3.5 introduce these approaches. We conclude with a discussion of open problems and future directions in Haplotype Phasing research.

3.1 Overview

Haplotype Phasing refers to the computational procedure of identifying haplotype information from genotype data. Formally, we define the Haplotype Phasing problem as follows: Let $G = \{g_1, ..., g_n\}$ be a set of n genotypes, where each genotype g_i consists of the combined allele information of m SNPs, $s_1, ..., s_m$. For simplicity, we represent $g_i \in G$ as a vector of size m whose j^{th} element g_{ij} $(i = 1, ..., n$ and $j = 1, ..., m)$ is defined as:

$$g_{ij} = \begin{cases} 0 & : \ when\ the\ two\ alleles\ of\ SNP\ s_j\ are\ major\ homozygous, \\ 1 & : \ when\ the\ two\ alleles\ of\ SNP\ s_j\ are\ minor\ homozygous, \\ 2 & : \ when\ the\ two\ alleles\ of\ SNP\ s_j\ are\ heterozygous. \end{cases}$$

Let H be the set of all[1] haplotypes consisting of the same m SNPs, $s_1, ..., s_m$. Like the genotype, each haplotype $h_i \in H$ is also a vector of size m. However, as introduced in Section 2.1.1, haplotypes represent the allele information of SNPs on *one* chromosome, while genotypes represent the *combined* allele information of SNPs on *two* chromosomes. Thus, the j^{th}

[1]Since we represent the allele of a SNP as either major or minor, the possible number of haplotypes consisting of m SNPs is 2^m.

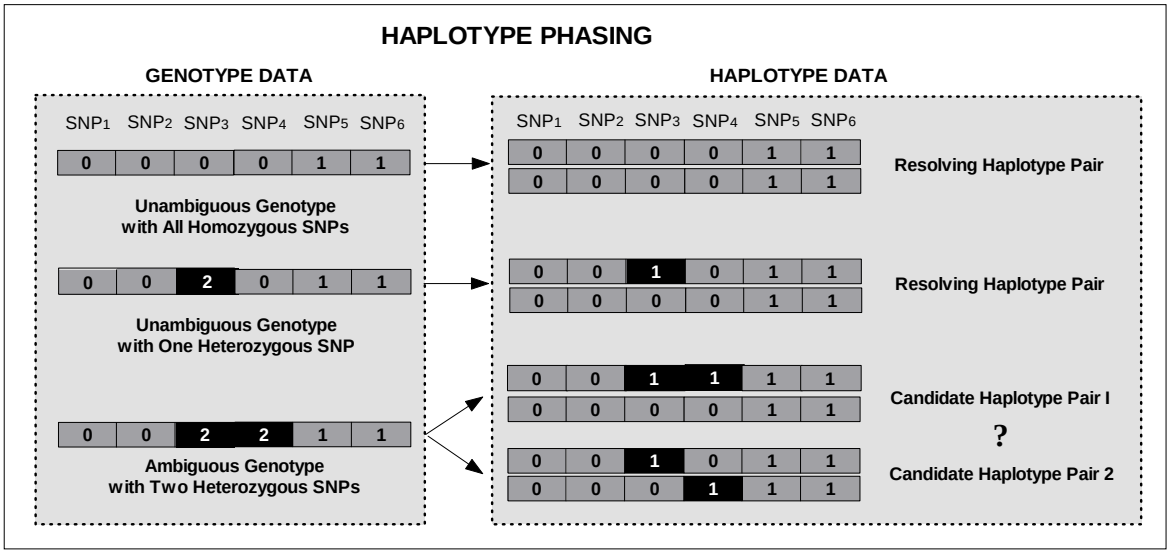

Figure 3.1: Haplotype Phasing and Ambiguous Genotypes

element, h_{ij}, of the haplotype h_i ($i = 1, ..., 2^m$ and $j = 1, ..., m$) is defined as:

$$h_{ij} = \begin{cases} 0 & : \quad when\ the\ allele\ of\ SNP\ s_j\ is\ major, \\ 1 & : \quad when\ the\ allele\ of\ SNP\ s_j\ is\ minor. \end{cases}$$

When the combined allele information of two haplotypes, $h_j \in H$ and $h_k \in H$, comprises the genotype g_i, we say that h_j and h_k *resolve* g_i and denote the relationship[2] as $h_j \oplus h_k = g_i$. The haplotypes h_j and h_k are referred to as the *complementary mates* of each other to resolve g_i, and each of them is considered to be *compatible* with g_i. The Haplotype Phasing problem can thus be defined as follows:

Problem	:	Haplotype Phasing
Input	:	A set of genotypes $G = \{ g_1, ..., g_n \}$
Output	:	A set of n haplotype-pairs
		$O = \{ < h_{i1}, h_{i2} > \mid h_{i1} \oplus h_{i2} = g_i,\ h_{i1}, h_{i2} \in H,\ 1 \leq i \leq n \}$.

In brief, to solve the Haplotype Phasing problem, one needs to find a set of haplotype-pairs that can resolve all genotypes in G.

However, the solution to the Haplotype Phasing problem is not straightforward due to *resolution ambiguity*. Figure 3.1 illustrates the problem. The genotype data of three individuals are displayed on the left. Each genotype consists of six SNPs. When two alleles of a SNP in a genotype are homozygous (i.e., either 0 or 1), the SNP is colored gray. When two

[2]The order of the haplotype-pair does not matter, that is, $h_j \oplus h_k = g_i$ is the same as $h_k \oplus h_j = g_i$.

alleles of a SNP in a genotype are heterozygous (i.e., 2), the SNP is colored black. The first genotype consists of all homozygous SNPs, while the second genotype contains one heterozygous SNP. For both of these genotypes, the resolving haplotype-pairs can be identified unambiguously as shown on the top right. However, in the case when there are c heterozygous SNPs in the genotype ($c > 1$) such as the third one in Figure 3.1, there are 2^{c-1} pairs of haplotypes that can resolve the genotype. Thus, the genotype cannot be uniquely resolved without additional biological insight or constraints. In this case, the genotype is considered *ambiguous*.

Numerous computational and/or statistical algorithms have been developed for addressing this ambiguity in Haplotype Phasing. The methods are typically grouped based on one of four principles: (1) parsimony; (2) phylogeny; (3) maximum-likelihood; (4) Bayesian inference. The former two solve the Haplotype Phasing problem as a combinatorial problem. They define an explicit objective function to resolve all genotypes, and aim to find a solution that optimizes the function [40]. The latter two are based on statistical inference methods. In addition to resolving all genotypes, they also estimate population haplotype frequencies.

Regardless of the approach, the performance of all Haplotype Phasing algorithms can be measured by *phasing accuracy*, that is, the proportion of the genotypes that are correctly resolved by the algorithm. When a simulation data set is used, the correct haplotype pair resolving each genotype is already known. In the case of a real data set, the one whose haplotype information has been directly obtained by bio-molecular experiments is used for evaluation. In the following sections, we introduce each of the four major Haplotype Phasing approaches.

3.2 Parsimony-based Methods

All parsimony-based approaches assume that a target population shares a relatively small number of common haplotypes due to linkage disequilibrium. Thus, they try to resolve an *ambiguous* genotype using one of *already identified* haplotypes.

The principle was first proposed by Clark [17]. Clark's algorithm begins by finding *unambiguous* genotypes, which contain only homozygous alleles or at most a single heterozygous allele. These genotypes can be uniquely resolved, so that their corresponding haplotype pairs are stored in the set of identified haplotypes, which is denoted by I. For each remaining *ambiguous* genotype, the set I is examined to see if it contains a haplotype that is *compatible* with the target genotype. When such a haplotype is found, the genotype is labeled *resolved*, and the haplotype's *complementary mate* is added to I. This process is iterated until all ambiguous genotypes are resolved or no new haplotype is found.

Clark's algorithm is simple, intuitive, and has been known to work well in practice [16]. However, it has several limitations: (1) it requires at least one unambiguous genotype; (2) genotypes may remain unresolved at the end of the procedure; and (3) a different order of

iteration may yield a different set of haplotypes. Simulation studies [17] showed that the first two limitations can be overcome if the size of the sample is large enough. Therefore, sampling enough individuals is practically important to apply Clark's algorithm. To address the last limitation, Clark proposed to repeat the whole procedure multiple times with different orderings of the data, and select the solution that resolves the largest number of genotypes. This criterion is referred to as *maximum-resolution*.

Gusfield [37] empirically verified the maximum-resolution criterion. Furthermore, he studied *what is the maximum number of genotypes that Clark's algorithm can resolve*, and defined it as the maximum-resolution (MR) problem. By reducing the satisfiability problem to the MR problem, he proved that the MR problem is NP-hard. In addition, an approximation algorithm based on linear programming was proposed to solve the MR problem. Although the experiments [37] show that this approach works well in practice, it may fail to find a solution.

In contrast to the *maximum-resolution* criterion, several groups [12, 38, 43, 53, 66, 67, 96] aimed to find *a minimum set of haplotypes that can resolve all genotypes in a data set*. This problem is referred to as the *maximum-parsimony* (MP) or *pure-parsimony* (PP) problem. As this problem is proven NP-hard [66], approximation algorithms and heuristics [12, 38, 43, 53, 67, 96] were proposed.

The early algorithms are based on integer linear programming [38], a greedy method [96], or a branch-and-bound rule [96]. However, their memory requirement increases exponentially with the problem size, limiting their applicability to small-size studies. Recently, several polynomial-space algorithms were suggested using integer linear programming [12, 42, 67].

All parsimony-based methods assume that the observed number of distinct haplotypes in a population is much smaller than the possible number of distinct haplotypes under linkage equilibrium. Therefore, when the data set does not satisfy this condition, the performance of parsimony-based methods becomes poor [10, 40].

3.3 Phylogeny-based Methods

Phylogeny-based approaches assume that the haplotypes in a population evolve along the *coalescent*, a popular genetic model which denotes a rooted tree describing the evolutionary history of a set of DNA sequences [50]. Thus, such approaches aim to find a set of haplotypes that resolves the target genotype data and follows the coalescent model as well.

In general, the coalescent is built using two assumptions: *infinite site mutation* and *no recombination*. The infinite-site-mutation assumption states that, at each SNP site, *a mutation only occurs once in the evolutionary history*. Therefore, a chromosome with mutation at one SNP site must be a descendant of the ancestral chromosome in which the mutation originally occurred. Moreover, any chromosome without this mutation cannot be a descendant of

a chromosome that has the mutation. The no-recombination assumption states that *the target region of DNA sequence was not recombined from a parent's two copies of chromosomes*, thus it can be considered to be inherited from just a single ancestor.

A *perfect phylogeny* [36] is a computational terminology corresponding to a coalescent tree of haplotypes. Let H be a set of $2n$ haplotypes $H = \{h_1, ..., h_{2n}\}$, where each haplotype h_i consists of m SNPs. A perfect phylogeny is defined as a rooted tree T with $2n$ leaves that satisfies the following properties:

1. Each of the $2n$ haplotypes labels exactly *one leaf* of T.
2. Each of m SNPs labels exactly *one edge* of T.
3. Every *internal edge* (i.e., one not connected to a leaf) is labeled by *at least one SNP*.
4. For any haplotype h_i, SNPs labeled on the path from the root to the leaf labeled by it, specify the SNPs whose allele is mutated (i.e., minor) in h_i.

Figure 3.2-a) shows a perfect phylogeny for a set of 4 haplotypes. In general, the root of a phylogeny is always assumed to be a haplotype whose alleles are all major (i.e., all 0's). A set of haplotypes has a perfect phylogeny *if and only if* for each pair of SNPs, there are no three haplotypes with values $(0, 1)$, $(1, 0)$, and $(1, 1)$ [36]. Figure 3.2-b) illustrates a violation to this condition. Haplotype 1, $(1, 0)$, has a mutation at the first SNP site, while haplotype 2, $(0, 1)$, has a mutation at the second SNP site. Thus, they cannot be descendants of each other, and two internal edges that denote the mutations at the first SNP and at the second are drawn. Haplotype 3, $(1, 1)$, has mutations at both SNP sites, thus it should be the descendant of the subtree that either haplotype $(1, 0)$ or $(0, 1)$ belongs to. However, to make haplotype 3 belong to either subtree, another edge denoting the mutation at either the first SNP or at the second should be added to the respective subtree. This violates the infinite-site-mutation assumption, that is, at each SNP site, a mutation can occur only once.

Gusfield [39] first proposed to use the perfect phylogeny to identify a set of haplotypes that evolves along a coalescent. He defined this problem as the *perfect phylogeny haplotype* (PPH) problem. Theorems and algorithms from graph and matroid theory were used to find a solution to the problem. The complexity of the presented algorithm is $O(nm \cdot \alpha(n, m))$, where n is the number of genotypes, m is the number of SNPs, and α is the inverse Ackerman function. Although the performance of this algorithm is nearly linear in the size of the input, the proposed approach is considered very difficult to understand and challenging to implement [6, 27, 40]. Thus, simpler but slower algorithms [6, 15, 27] were proposed subsequently. All of them have an $O(nm^2)$ time complexity. Recently, a linear time algorithm was developed by Ding et al. [24].

Although the performance of the perfect phylogeny-based methods have improved, all of them suffer from their strict conformity to the coalescent model; it is possible that *no perfect phylogeny solution* exists for a given genotype data set. In practice, real data often does not

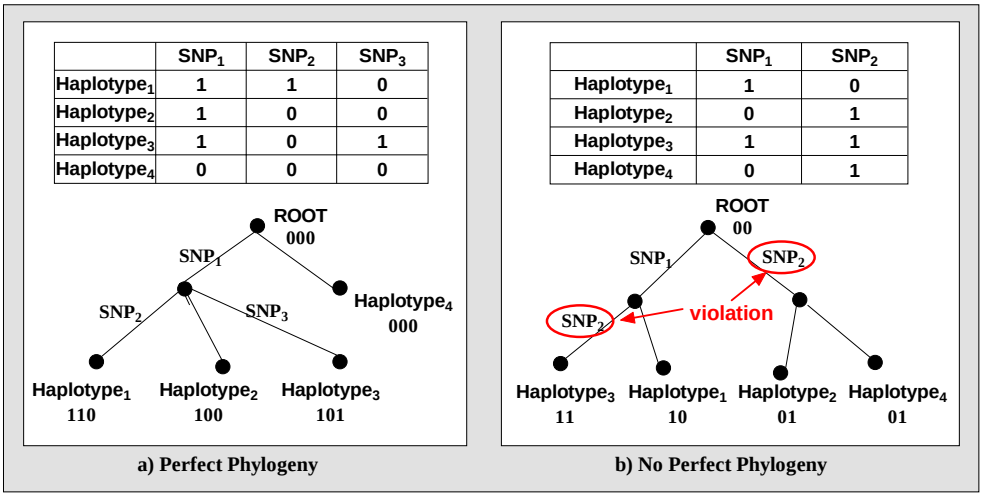

Figure 3.2: Perfect Phylogeny and Imperfect Phylogeny

perfectly fit the coalescent model. This may occur due to errors in genotyping or a violation of the infinite-site-mutation and/or no-recombination assumption during evolution [63].

Eskin et al. [27] tried to remove a minimum number of genotypes from the original data so that the remaining ones can be resolved by a set of haplotypes with a perfect phylogeny. The problem was proven NP-hard, but no heuristic solution was presented. Halperin et al. [45] tried to assign the values of missing alleles so that the resulting haplotypes can have a perfect phylogeny solution. Both approaches assume that a perfect phylogeny solution exists for a given genotype set, but it cannot be identified due to genotyping errors or missing alleles.

Imperfect phylogeny-based methods [28, 44] take a more realistic approach. In principle, the methods assume that *most but not all* haplotypes will fit the perfect phylogeny model. Thus, they consider a relaxed model that allows for a certain number of recurrent mutations and recombinations. Among multiple candidate solutions satisfying the relaxed model, the one with the maximum-likelihood given a genotype data set is chosen as the solution. However, handling the exponential number of candidate solutions remains an unsolved problem.

3.4 Maximum-Likelihood-based Methods

The parsimony-based methods and the phylogeny-based methods introduced above, aim to directly resolve each genotype with a pair of haplotypes. In contrast, *maximum-likelihood* (ML) methods are based on a rather indirect approach, *haplotype frequency estimation* (HFE). They aim to estimate *the haplotype distribution in a population, maximizing the likelihood of the genotype data*.

Let D be the genotype data of n individuals, where each genotype consists of m SNPs,

and the number of *distinct* genotypes in D is n'. Let g_i denote the i^{th} distinct genotype, and f_i denote the frequency of g_i in the data set D, where $i = 1, ..., n'$. Let H be the set of all haplotypes consisting of the same m SNPs. As explained in Section 3.1, the number of haplotypes in H is 2^m. Let h_j denote the j^{th} distinct haplotype in H, and p_j be the *population frequency* of haplotype h_j, where $j = 1, ..., 2^m$. Unlike the genotype sample frequencies, f_i, which we can directly calculate from the data set, the haplotype population frequencies, p_j, are unknown, and we need to estimate them.

Maximum-likelihood (ML) methods estimate the population haplotype frequencies, $\lambda = \{p_1, \ p_2, ..., p_{2^m}\}$ based on their likelihood, L, given the genotype data D. Initially, the likelihood, L, can be stated as the probability of genotypes comprising D as:

$$L(\lambda) = Pr(D \mid \lambda) \approx \prod_{i=1}^{n'} Pr_\lambda(g_i)^{f_i} = \prod_{i=1}^{n'} \Big(\sum_{\{\forall <h_k,h_l> \mid h_k \oplus h_l = g_i\}} Pr_\lambda(h_k, h_l) \Big)^{f_i}. \quad (3.1)$$

In brief, the likelihood of the data D is the product of the probabilities of all genotypes in D. Each genotype g_i occurs f_i times in D, and its probability $Pr_\lambda(g_i)$ can be computed by summing the joint probability of each haplotype pair that can resolve the genotype. Under the assumption of random mating, known as the *Hardy-Weinberg Equilibrium* (HWE) assumption, the joint probability $Pr_\lambda(h_k, h_l)$ of two haplotypes can be computed as the product of the two population haplotype frequencies, p_k and p_l. When $k = l$, $Pr_\lambda(h_k, h_l) = (p_k)^2$. Otherwise[3], $Pr_\lambda(h_k, h_l) = 2p_k p_l$. Thus, the joint probability $Pr_\lambda(h_k, h_l)$ in Equation 3.1 can be substituted with the product of the two population haplotype frequencies accordingly, and the population frequencies that maximize Equation 3.1 are computed. Using the estimated population frequencies, each genotype can be resolved by the haplotype pair with the maximum population frequency among all pairs compatible with the genotype.

Several groups [29, 48, 49, 72] independently proposed the expectation maximization (EM) algorithm to estimate the maximum-likelihood haplotype frequencies. The EM procedure is defined as follows: Initially, arbitrary values are assigned to the target haplotype frequencies $p_1, ..., p_{2^m}$, which we refer to as $p_1^{(0)}, ..., p_{2^m}^{(0)}$. In the Expectation step, the haplotype frequencies are used to estimate the expected genotype frequency $\hat{Pr}_\lambda(h_k, h_l)^{(t)}$ where (t) denotes the t^{th} iteration. In the Maximization step, the expected genotype frequency $\hat{Pr}_\lambda(h_k, h_l)^{(t)}$, computed in the previous step, is used to re-estimate the haplotype frequencies $p_1^{(t+1)}, ..., p_q^{(t+1)}$. The expectation and maximization steps are repeated until the change in the haplotype frequency in consecutive iterations is less than some predefined value. The time complexity for one iteration of the EM algorithm is $O(n2^k)$ where n is the number of genotypes, and k is the maximum number of heterozygous SNPs in the genotypes.

The main limitation of the EM algorithm lies in the exponential increase in the number of possible haplotypes as the number of heterozygous SNPs in a genotype grows. Thus, the

[3]Since we do not know the phase information of the given genotype, two different haplotypes can have two phases: p_k on the maternal chromosome and p_l on the paternal, and vice versa.

number of SNPs that can practically be handled by the EM algorithms is often limited to about 12 [86]. To address this problem, a *partition-ligation* (PL) strategy [18, 69, 86] and a *block-partitioning* strategy [86] were proposed. Both of these solutions take a divide-and-conquer approach. They divide a set of SNPs into a small number of contiguous subgroups, identify the set of most probable haplotypes for each subgroup, and combine the selected haplotypes from all subgroups through a bottom-up approach.

The partition-ligation methods partition the set of SNPs into subsets of equal size (e.g., 8 contiguous SNPs), while the block-partitioning methods partition it into subsets of different sizes, where each subset satisfies a given block definition[4]. There are several issues that need further research. For example, when using the divide-and-conquer approach, the solutions are often only locally optimal with respect to the whole region [34, 86]. It is also unknown whether the fixed size of subgroups in the partition-ligation methods or the different block boundaries in the block-partitioning methods affect the overall accuracy of the proposed methods. Nevertheless, both approaches are used in practice.

Once the EM algorithms were successfully applied to Haplotype Phasing, their performance under various conditions was examined [31, 62, 64, 94]. All EM-based methods assume the *Hardy-Weinberg Equilibrium (HWE)*, that is, they assume that each genotype is composed of two haplotypes randomly mated. Initially, Excoffier et al. [29] reported that the HWE condition is more likely to be satisfied when the number of individuals in a sample is large, so the EM algorithm is appropriate for analyzing a large-size sample. However, subsequent simulation studies [31, 94] demonstrated that the EM algorithm is reliable and robust even when the HWE assumption is violated.

The effect of *genotyping errors* and of *missing alleles* on the performance of the EM algorithm was studied by Kirk et al. [64] and by Kelly et al. [62], respectively. Under moderate to strong levels of linkage disequilibrium (LD) among SNPs, the absence of up to 30% of data was reported not to affect the overall accuracy of the EM algorithm [62]. However, genotyping errors were reported to substantially reduce the estimation accuracy of the EM algorithm, particularly under low LD [64]. Thus, Kelly et al. [62] concluded that ambiguous data are better treated as unknown. More extensive studies are needed to confirm this conclusion.

In general, the performance of EM-based methods for the Haplotype Phasing problem was demonstrated to be accurate and robust under a wide range of parameter settings [31, 62, 94]. However, several shortcomings still exist. First, the EM algorithm strongly depends on its initial condition, and does not guarantee a global optimum. To overcome this, EM-based methods should be run multiple times with different initial conditions. Second, the variance of the haplotype frequency estimation is not accurately known [29, 48]. Last, calculating confidence intervals and conducting statistical tests under the EM algorithm typically involve

[4]Blocks are typically defined based on limited haplotype diversity [84], linkage disequilibrium [32], and recombination [97]

approximations, which require a large sample to be most accurate [9].

3.5 Bayesian Inference-based Methods

Like the maximum-likelihood (ML) methods introduced above, Bayesian inference methods take a statistical approach. However, while ML methods aim to find a set of exact model parameters Θ that maximize the probability of genotype data $G = \{g_1, ..., g_n\}$ given the model, that is, $Arg \underset{\Theta}{max}\ Pr(G|\Theta)$, Bayesian inference methods aim to find the *posterior distribution* of the model parameters given the genotype data G, which is $Pr(\Theta|G)$. Moreover, in ML methods, Θ denotes a set of unknown haplotype frequencies in a population, while in Bayesian inference methods, Θ denotes a set of each genotype's resolved haplotype pairs. Thus, where H is a set of haplotype pairs resolving the given genotypes, Bayesian inference methods aim to find the posterior probability $Pr(H|G)$. However, computing $Pr(H|G)$ exactly is not feasible in the general case [9]. Thus, Markov Chain Monte-Carlo (MCMC) techniques are used to obtain approximate samples from $Pr(H|G)$, and their expectation is presented as the final solution.

One popular MCMC technique is Gibbs sampling. Its essential application to haplotypes is as follows: Let $H^{(t)}$ denote the set of haplotype pairs resolving all genotypes at the t^{th} iteration, $H^{(t)}_{-i}$ denote the set of haplotype pairs resolving all genotypes *except* g_i at the t^{th} iteration, and $H^{(t)}_{i}$ denote the set including only the haplotype pair resolving the genotype g_i at the t^{th} iteration. Gibbs sampling starts with an initial guess $H^{(0)}$. An ambiguous genotype g_i is then randomly selected. Under the assumption that a current estimation of $H^{(t)}$ is correct for all genotypes except g_i, the new haplotype pair resolving g_i, that is, $H^{(t+1)}_{i}$, is sampled from the distribution $Pr(H_i|G, H^{(t)}_{-i})$. This random selection and update is iterated, until we get approximate samples from the haplotype distribution for the genotype set G, $Pr(H|G)$.

The first Bayesian inference method in the context of Haplotype Phasing was proposed by Stephens et al. [93]. Their algorithm exploits ideas from *coalescent* theory to guide their Gibbs sampling procedure. As defined in section 3.3, a coalescent is a rooted tree representing the evolutionary relationships among haplotypes. Along the coalescent, haplotypes evolve one SNP at a time. Thus, whenever a genotype cannot be resolved using existing haplotypes, new haplotypes that are most similar to the existing common[5] ones are generated to resolve the genotype.

Other Bayesian inference methods [70,79,101] use similar MCMC sampling techniques, but employ priors of different forms. The priors include: simple Dirichlet [70,79] and Dirichlet Process [101]. In addition, Lin et al. [70] use neighboring information of heterozygous SNPs to resolve each genotype.

[5]Due to its preference to common haplotypes, some [70] interpreted this approach as a kind of a parsimony approach.

Bayesian inference methods using MCMC techniques are often compared with maximum-likelihood (ML) methods that use the EM algorithm. However, the performance of both approaches varies under different genotype compositions and with different accuracy measures [2,68,92,102,110], making it difficult to decide whether one approach is superior to the other. Most importantly, the two approaches both have their own merits and shortcomings. Unlike ML methods, Bayesian inference methods can be applied to samples consisting of a large number of SNPs or to samples in which a substantial portion of haplotypes occur only once [9, 42]. In addition, MCMC techniques can explore the whole state space, thus they avoid local maxima given sufficient running time [81]. Last, Bayesian inference methods can incorporate prior knowledge to guide their estimation procedure. In contrast, ML methods require less computing time [9], and are easier to check for convergence than Bayesian inference methods [86]. Furthermore, the performance of ML methods is robust even under the violation of their basic assumption, Hardy-Weinberg Equilibrium, while the performance of Bayesian inference methods is reported to be affected by a deviation of data from their basic assumption, coalescent theory [92].

3.6 Discussion

Arguably, two statistical approaches, namely maximum-likelihood (ML) using the EM algorithm and Bayesian inference using MCMC techniques, are most popular for Haplotype Phasing [80]. First and most importantly, empirical comparison studies [98, 102] show that the phasing accuracy of the two statistical approaches is somewhat better than that of the combinatorial approaches. In addition, parsimony-based methods and phylogeny-based methods often present *multiple* solutions, making it difficult to compare their performance with other methods. Last, statistical approaches are applicable even when only ambiguous genotypes are in the data and when no perfect phylogeny solution exists, that is, these approaches can be used even when parsimony-based methods and perfect-phylogeny methods cannot be applied.

Although the performance of the two statistical approaches is quite promising [2,94,102], there are several difficulties that none of the current Haplotype Phasing methods can address well. First, the phasing accuracy of all methods decreases as linkage disequilibrium (LD) drops [2, 31]. This poor accuracy occurs more often when a large number of SNPs are examined [43], since LD tends to decrease as the distance between SNPs increases.

Second, most algorithms work well for data sets with few or no genotyping errors or missing alleles [62, 64]. However, very often, allele information is incorrect or missing due to imperfection of current genotyping technology. Missing allele information increases the combinatorial complexity of the Haplotype Phasing problem. The genotyping error problem is even more difficult to solve, since in general, we do not know which alleles are incorrect.

Last, most Haplotype Phasing algorithms show a poor phasing accuracy for rare haplotypes (i.e., ones with a population frequency $<$ 1-5%) [2, 37, 40, 94]. This problem occurs since most Haplotype Phasing algorithms are based on population genetic assumptions which prefer common haplotypes (i.e., occurring in more than 5%-20% of the population). However, it is not clear yet whether rare haplotypes or rather common ones are important for the etiology of disease [19].

In conclusion, future research of Haplotype Phasing should focus on improving the performance of algorithms for data sets with low LD, genotyping errors, missing alleles, and rare haplotypes.

Chapter 4

Tag SNP Selection

This chapter introduces Tag SNP Selection. An overview of the problem is given in section 4.1. Tag SNP Selection algorithms are categorized into four major approaches based on: (1) haplotype diversity; (2) pairwise association; (3) tagged SNP prediction; and (4) phenotype association. We give a brief introduction of each approach and conclude with a discussion of open problems and future directions.

4.1 Overview

In most large-scale disease studies, genotyping all SNPs in a candidate region for a large number of individuals is still costly and time-consuming. Thus, selecting a subset of SNPs that is sufficiently informative to conduct disease-gene association but small enough to reduce the genotyping overhead, a process known as *Tag SNP Selection*, is a critical problem to solve. In general, the selected SNPs on a haplotype are referred to as *haplotype tag SNPs* (htSNPs), and the unselected SNPs are referred to as *tagged SNPs*.

Formally, we define the Tag SNP Selection problem as follows: Let $S = \{s_1, ..., s_m\}$ be a set of m SNPs in a candidate region, and $D = \{h_1, ..., h_n\}$ be a data set of n haplotypes consisting of the m SNPs. As defined in Section 3.1, $h_i \in D$ is a vector of size m whose vector element is 0 when the allele of a SNP is *major* and 1 when it is *minor*. Suppose that the maximum number of htSNPs is k, and a function $f(T', D)$ evaluates how well the subset $T' \subset S$ represents the original data D. Then, the Tag SNP Selection problem can be stated as follows:

Problem	:	Tag SNP Selection
Input	:	A set of SNPs S, A set of haplotypes D, A maximum number of htSNPs k
Output	:	A set of htSNPs $T = \underset{T' \ s.t. \ T' \subset S \ \& \ \lvert T' \rvert \leq k}{argmax} f(T', D)$.

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, among all possible subsets of the original SNPs.

Initially, Tag SNP Selection was motivated by *linkage disequilibrium* (LD) introduced in Section 2.1.2 [33]. When high LD exists between SNPs, their allele information might be almost the same. Thus, we can select one from those redundant SNPs so that, even with only a subset of original SNPs, most information in a haplotype is retained. However, what comprises the best htSNP selection strategy is still an open problem [99].

Researchers proposed *a variety of measures* to represent the information of haplotypes, and tried to identify the subset of SNPs that optimizes these measures. The relations among these measures and their effect on the selection of htSNPs are still the subject of ongoing research. Most importantly, unlike Haplotype Phasing, there is no gold standard to evaluate the performance of different approaches [23]. Thus, the performance of Tag SNP Selection algorithms is often evaluated based on their own information measure, which makes comparison among different approaches difficult.

We group here the algorithms for Tag SNP Selection into four categories based on the approach they take to measure the *information of haplotypes*: (1) haplotype diversity; (2) pairwise association among SNPs; (3) tagged SNP prediction; and (4) phenotype association. In the following sections, we introduce each of them.

4.2 Haplotype Diversity-based Methods

Recent observation of *the block structure of the human genome* [20,32,84] demonstrates that the human genome can be partitioned into discrete blocks such that within each block, most of the population (i.e., 80-90%) shares a very small number of common haplotypes (i.e., 3-5 haplotypes). Based on this assumption, early htSNP selection research aimed to find *a subset of SNPs that can capture most of the limited haplotype diversity in the original data*.

Figure 4.1 illustrates how a set of htSNPs 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 4.1-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 4.1-b). Therefore, the information about 2 SNPs might 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 htSNPs.

A variety of haplotype diversity measures were proposed. Some [59, 84] use *the number of haplotypes that are uniquely distinguishable by the candidate subset* T' as a measure of

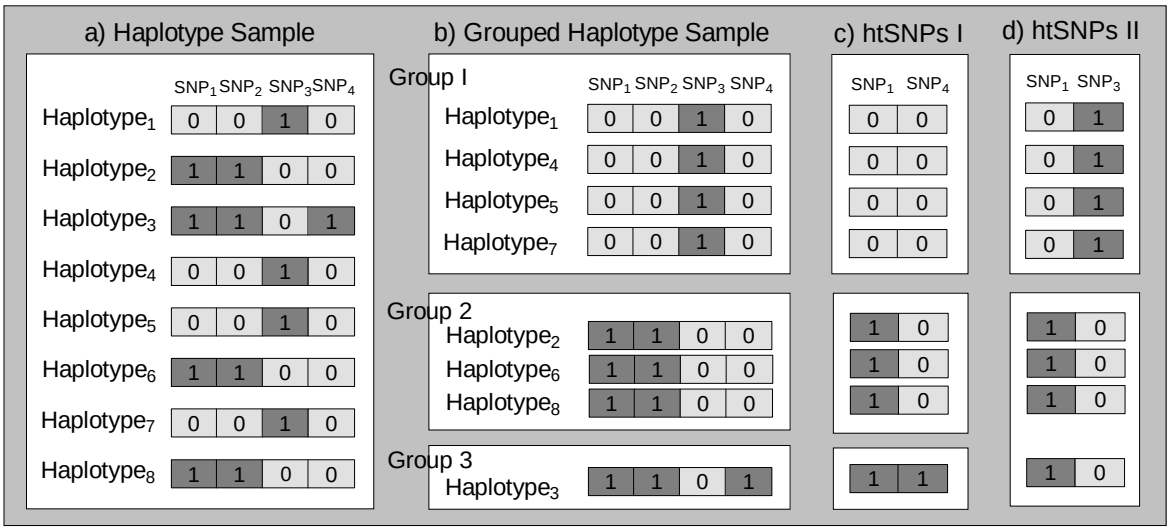

Figure 4.1: Tag SNP Selection based on limited haplotype diversity

the haplotype diversity captured by T'. For example, in Figures 4.1-c) and 4.1-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. [56] define the haplotype diversity *not* captured by the candidate subset T' (i.e., 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 4.1-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 4.1-d), which makes its *residual* haplotype diversity greater than 0. Thus, T' with the *smallest residual* haplotype diversity is selected as the set of htSNPs.

Another popular haplotype diversity measure is *Shannon's Entropy* (H) [1, 5, 47, 58, 78]. 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 htSNP set T', haplotypes are partitioned into groups so that the ones in the same group share the same alleles at the SNPs $\in T'$. 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 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 [1, 5, 13, 18, 47, 56, 58, 59, 78, 84] exhaustively examine all subsets of the original SNP set S, 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 [109], a branch-and-bound rule [22], dynamic programming [103–108], and principal component analysis (PCA) [52, 71, 76].

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 htSNP 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 htSNPs from each block might not be the optimal set of htSNPs for a whole region [34]. Furthermore, as introduced in 2.1.2, regions of low linkage disequilibrium exist *between* blocks [19]. 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 htSNPs depends on the block-partitioning method used [23, 82, 88].

4.3 Pairwise Association-based Methods

Pairwise association-based approaches rely on the idea that a set of htSNPs should be *the smallest subset of available SNPs that are capable of predicting a disease locus* on a haplotype. However, the disease locus is generally the one we are looking for, and is not known ahead of time. Instead, pairwise association between SNPs is used as an estimate for the predictive power with respect to the the disease locus. In principle, a set of htSNPs is selected such that *all SNPs on the haplotype are highly associated with one of the htSNPs*. This way, although the SNP that is relevant to the disease may not be selected as an htSNP, the association of the target disease with that SNP can be indirectly deduced from the htSNP that is highly associated with it. In most studies, non-random association of SNPs (i.e., linkage disequilibrium (LD)) introduced in Section 2.1.2, is used to estimate the pairwise association.

Byng et al. [13] first proposed to use cluster analysis for pairwise association-based htSNP selection. The original set of SNPs is partitioned into hierarchical clusters, where SNPs within the same cluster have at least 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 [4, 14, 100] proposed that a htSNP 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 ensure the htSNP selection property, *minimax clustering* [4] and *greedy binning algorithm* [14, 100] were proposed.

In minimax clustering, the *minimax* distance between two clusters C_i and C_j is defined as $D_{minimax}(C_i, C_j) = \min_{\forall s \in (C_i \cup C_j)} (D_{max}(s))$, 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 based on their 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 of each merged cluster is selected as the cluster representative.

The greedy binning algorithm works as follows: First, it examines all pairwise LD relationship between SNPs, and for each SNP, counts the number of other SNPs whose pairwise LD with the SNP is greater than pre-specified level σ. The SNP that has the largest counting number is then clustered together with its associated SNPs, and becomes the htSNP 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 $O(cnm^2)$, where the number of clusters is c, the number of haplotypes is n, and the number of SNPs is m. Thus, in general, they run faster than the methods based on haplotype diversity, and do not require a prior block-partitioning procedure. 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 htSNPs than other methods [34, 61, 74, 88].

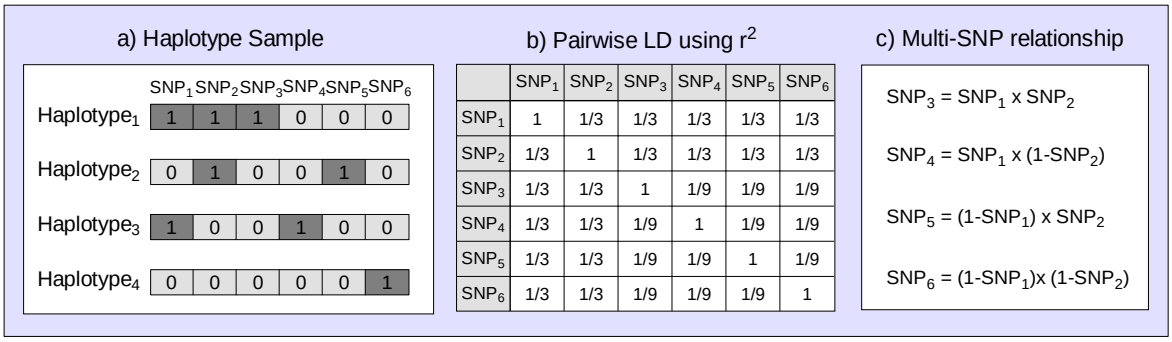

Figure 4.2: Pairwise linkage disequilibrium (LD) among SNPs and multi-SNP dependencies

Figure 4.2 illustrates this weakness of pairwise association-based methods. Suppose that our sample consists of four haplotypes with six SNPs, as shown in Figure 4.2-a). If we

measure pairwise LD between the SNPs using the most common LD measure, correlation coefficient r^2 [34], no two SNPs have pairwise LD greater than 0.5, as shown in Figure 4.2-b). Thus, pairwise association-based methods will select all six SNPs as htSNPs. However, as shown in Figure 4.2-c), the allele of SNP 3, 4, 5, and 6 can be perfectly represented by the alleles of SNP 1 and 2. Thus, if we consider multi-SNP dependencies, only two SNPs, namely SNP 1 and 2, are sufficient to represent all the six SNPs.

4.4 Tagged SNP Prediction-based Methods

Tagged SNP prediction-based approaches consider htSNP selection as a reconstruction problem of the original haplotype data using only a subset of SNPs. Thus, they aim to select *a set of SNPs that can predict the unselected (i.e., tagged) SNPs with little error*. In general, after the selected htSNPs are genotyped, the alleles of the tagged SNPs are predicted using the alleles of the htSNPs, and disease-gene association is conducted based on the reconstructed full haplotype data. Therefore, these methods present a prediction rule for tagged SNPs along with the selected set of htSNPs.

Bafna et al. [7, 43] first proposed to select htSNPs 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. To measure how well a set of SNPs, $S = \{s_1, ..., s_k\}$, can predict the SNP, t, Bafna et al. define a measure called *informativeness* as:

$$I(S,t) = Pr_{i \neq j}(\bigcup_{l=1}^{k} E^{s_l}_{i,j} | E^t_{i,j}).$$

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 htSNPs of each tagged SNP to those that are within a relatively close physical proximity w to the predicted. However, the exponential time complexity $O(nk2^w)$ of dynamic programming needs to be reduced. Recently, Halperin et al. [46] proposed a polynomial time dynamic programming algorithm, but, in principle, their improvement results from limiting the number of htSNPs for each tagged SNP to 2.

Both methods proposed a *majority vote* as a reconstruction rule for tagged SNP alleles. Suppose that our sample consists of six haplotypes with five SNPs, and SNP 1 and 2 are selected as htSNPs as shown in Figure 4.3-a). We call this sample the htSNP selection sample. As introduced in Section 2.2, second genotyping is conducted to obtain the alleles of the *selected htSNPs* for a large number of individuals. To reconstruct the ungenotyped alleles (i.e., the tagged SNPs) for each haplotype in this new sample, first, the haplotypes whose htSNP alleles are the same as those of the new haplotype are identified in the htSNP selection sample. In Figure 4.3-b-1), these haplotypes are marked by a solid box. Each tagged SNP

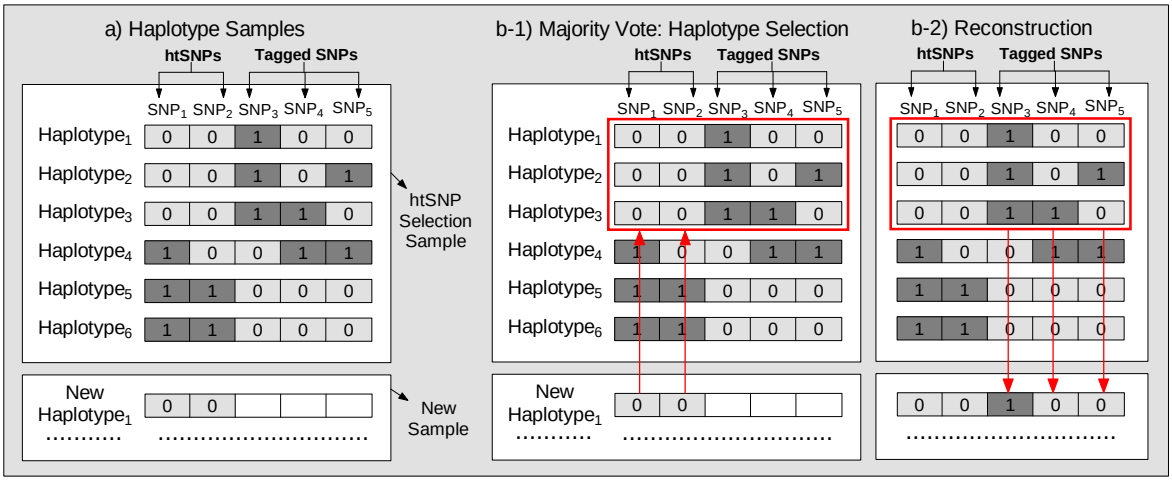

Figure 4.3: Majority Vote in Tagged SNP Prediction-based Methods

in the new haplotype is assigned the allele that occurs most often in the haplotypes identified above, as shown in Figure 4.3-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 htSNPs. As a result, the number of selected htSNPs is often smaller than that of pairwise association-based [8]. In addition, all dynamic programming methods [7, 43, 46] guarantee to find a global optimum with respect to the given measure. However, their prediction efficiency is still limited by some restrictions such as the small-bounded location or the fixed number of htSNPs. Further research is needed to effectively address this problem.

4.5 Phenotype Association-based Methods

Phenotype association-based approaches assume the availability of phenotype information, and try to find *a set of SNPs that can distinguish individuals carrying the disease* (i.e., case) *from individuals with no disease* (i.e., control). These SNPs are then used as the set of htSNPs. Under this view, htSNP selection is a kind of feature selection, which aims to select a set of features that distinguishes between two classes (case/control) with little error. As a result, classification techniques and test statistics measuring association between features and class labels, are used in this context.

Despite its simplicity, one of the most popular classifiers is the naïve Bayes classifier [95]. It assumes that the allele of one SNP is conditionally independent from that of others given the individual phenotype, and classifies each haplotype as *case* or *control* based on its probability of belonging to each of these classes. The subset T' of SNPs with the best classification

accuracy is then selected as the set of htSNPs. The main limitation of this approach lies in the *conditional independence* assumption among SNPs used by the naïve Bayes classifier. In reality, non-random association (i.e., linkage disequilibrium) exists among SNPs [52]. Shah et al. [89] addressed this problem by using a feature selection method considering correlation among SNPs. Their algorithm selects a feature if it correlates with a target class label but not with any other features that have already been selected.

The above phenotype association-based methods focus on selecting a set of SNPs that accurately partitions the given data into case and control classes. Hoh et al. [51] proposed a method that not only classifies the given data well but also guarantees its performance at a statistically significant level. Their algorithm is based on a bootstrap technique [26]. Suppose that our original data consists of n haplotype-phenotype pairs. First, one replicate set A is made by sampling n haplotype-phenotype pairs from the original data with replacement. Second, additional 1000 replicate sets, $B_1, ...B_{1000}$, are made, where each set consists of n haplotype-phenotype pairs sampled from the replicate A with replacement. These latter 1000 replicates represent the samples in which no association exists between haplotypes and phenotypes, thus their phenotype and haplotype labels are randomly permuted. Last, the SNPs whose sum of association score is higher in A than in at least (1-α) x100 (typically α=0.05) percent of the random samples $B_1, ...B_{1000}$, are selected. This procedure is iterated a pre-specified number of times, and the SNPs selected at least 50% of the iterations are presented as the set of htSNPs. The decision on test statistic for measuring the association between a SNP and a phenotype was left for future research.

Phenotype association-based methods are directly related to the main goal of computational haplotype analysis, namely, disease-gene association. The main limitation of a phenotype classification-based approach lies in its need of phenotype information, which may not be available ahead of time. In addition, usually, the number of haplotypes used for Tag SNP Selection is relatively small. Thus, the selected htSNPs that classify the small sample very well, may not perform as well on a larger sample. This can directly affect the performance of subsequent disease-gene association.

4.6 Discussion

The feasibility of Tag SNP Selection has been empirically demonstrated by simulation studies [34, 60, 61, 74, 104]. The results suggest that Tag SNP Selection can yield about 2-5 fold savings in the genotyping efforts. Most importantly, Zhang et al. [104] 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 htSNPs whose size is $1/4$ of the original SNP set is only 4 percent.

[1]The power of association tests is the probability that the test rejects the *false* null hypotheses [83].

However, several pitfalls still exist:

1) Most Tag SNP Selection algorithms focus on covering common haplotypes or common SNPs rather than rare ones [111]. Common variations are of interest because many common human diseases have been explained by common DNA variations rather than by rare ones [25, 61, 85]. Furthermore, practically, a much larger sample size is needed to identify rare haplotypes [65]. However, as discussed in Section 3.6, it is still an open question whether common variations or rare ones influence the susceptibility to common and complex disease.

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 [73], or the distribution of haplotype pairs for each genotype rather than a single resolved pair [111]. Until now, no htSNP selection methods consider this uncertainty of inferred haplotype data.

3) All the algorithms described above assume that the set of htSNPs selected from a given sample will work well for another sample from the same population. However, to ensure the generalized performance, a sufficient number of individuals should be sampled for Tag SNP Selection. For example, Goldstein et al. [34] reported that at least 100 chromosomes, that is, 200 haplotypes, are needed when the number of SNPs 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.

Along with ways to address these limitations, more comparative studies are needed to understand the merits and shortcomings of different Tag SNP selection approaches. To date, little comparative study has been done, and existing studies report conflicting results. By comparing 5 selection methods (2 haplotype diversity-based, 2 pairwise LD-based, and 1 equal spaced), Burkett et al. [74] concluded that different approaches result in considerably *different number* of htSNPs, and even between two methods based on the same approach, the *proportion of commonly selected* htSNPs is strikingly small (i.e., 30%). In contrast, Duggal et al. [82] reported that despite the differences in the number of selected htSNPs, the *proportion of commonly selected* SNPs among 6 haplotype diversity-based methods is, in general, consistently high (about 50%-95%). In addition, Ke et al. [60, 61] reported that the *prediction ability* of htSNPs is highly concordant between haplotype diversity and pairwise association-based methods. It is difficult to generalize any of these conclusions since they are based on different data sets and also different evaluation measures. Further research should clarify these conflicting results as well as establish a common testbed to evaluate the performance of different selection approaches.

Chapter 5

Conclusion

Along with the completion of the human genome project, one of the major interests of current genomics research is disease-gene association, that is, identifying which DNA variation or a set of DNA variations is highly associated with a specific disease. Computational haplotype analysis, and specifically, its two procedures, *Haplotype Phasing* and *Tag SNP Selection*, provide the most practical framework for conducting large-scale association studies. They provide inexpensive, fast, and relatively accurate performance. Thus, until the overhead of low-throughput bio-molecular experiments becomes less formidable, computational haplotype analysis will be in demand. In this paper, we introduced the major computational and statistical approaches of Haplotype Phasing and Tag SNP Selection along with their biological motivations. Of course, certain limitations still exist in both, and future research should focus on improving them.

Missing alleles, genotyping errors, and low linkage disequilibrium among SNPs are the common difficulties with which all Haplotype Phasing algorithms are confronted. Further improvement of Tag SNP Selection requires the ability to handle rare haplotypes, uncertainty in haplotype data, and small sample size. In addition, thorough evaluation of different approaches and development of a common testbed are also open for more research.

As for our future research directions, we plan to work in the area of Tag SNP Selection. Compared to Haplotype Phasing, for which two statistical approaches have been used as an established tool, Tag SNP Selection research has been started very recently, and still faces many challenges. As discussed in Chapter 4, what comprises the best htSNP selection strategy is still an open problem, and no standard evaluation measure has been proposed. However, we believe that the tagged SNP-prediction-based approach, introduced in Section 4.4, has several advantages over others: (1) it does not rely on prior block-partitioning; (2) it utilizes multi-SNP relationship; and (3) it does not require phenotype information. Currently, in spite of these advantages, the performance of the tagged SNP-prediction-based methods is limited by the computational complexity of the dynamic programming procedures. We plan to apply other machine learning techniques in an attempt to improve the performance. In par-

ticular, we consider probabilistic approaches as an alternative to the dynamic programming procedures. Probabilistic methods have been successfully applied to Haplotype Phasing, and most importantly, nondeterministic characteristic of haplotype data may be better represented in a probabilistic framework. The development, examination and application of such probabilistic approaches is to become the topic for the rest of my PhD work.

Appendix A

Haplotype-Disease Association

Haplotype-Disease Association aims to identify which haplotype or a set of haplotypes is highly associated with a target disease, using haplotype samples from a group of individuals carrying the disease (i.e., cases) and a group of individuals not carrying the disease (i.e., controls). This appendix briefly introduces two of Haplotype-Disease Association tests: chi-square goodness-of-fit test and odds ratio test. We start by explaining their common basis in Section A.1, and introduce the two tests in Section A.2 and A.3, respectively.

A.1 Common basis of Haplotype-Disease Association

Let H be a set of haplotypes occurring either in cases or controls. Suppose that we want to examine whether association exists between a haplotype $h \in H$ and a target disease d. Let $Pr_h(d)$ be the probability of disease incidence among the individuals possessing the haplotype h, and $Pr_{\neg h}(d)$ be the probability of disease incidence among the individuals not possessing this haplotype. Our null hypothesis, H_0, can be stated as:

$$H_0 : \ Pr_h(d) = Pr_{\neg h}(d).$$

To test the null hypothesis, all haplotype-disease association tests use two contingency tables: one representing the *observed* frequency of haplotypes in cases and controls and the other representing their *expected* frequency under the null hypothesis. Table A.1 displays the first contingency table, O, where each cell, O_{ij}, represents the *observed* number of the target haplotype h or that of the other haplotypes, $H - \{h\}$, in the sets of cases and of controls, respectively.

Under the null hypothesis, the probability of disease incidence, $Pr(d)$, is not affected by possessing the haplotype h or not. Thus, it is simply the number of cases divided by the total number of individuals, $Pr(d) = \frac{a+b}{a+b+c+d}$. The *expected* number of individuals that have the disease d as well as the haplotype h is the product of the number of observed individuals with haplotypes, $(a + c)$, and the probability of disease incidence, $Pr(d)$,

Table A.1: The contingency table O of the observed haplotype frequencies in the sets of cases and of controls

	Haplotype h	Haplotypes $\in H - \{h\}$	Total
Case	a	b	$a+b$
Control	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Table A.2: The contingency table E of the expected haplotype frequencies in the sets of cases and of controls under the null hypothesis

	Haplotype h	Haplotypes $H - \{h\}$	Total
Case	$(a+c) \times (\frac{a+b}{a+b+c+d})$	$(b+d) \times (\frac{a+b}{a+b+c+d})$	$a+b$
Control	$(a+c) \times (\frac{c+d}{a+b+c+d})$	$(b+d) \times (\frac{c+d}{a+b+c+d})$	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

$$(a+c) \times Pr(d) = (a+c) \times (\tfrac{a+b}{a+b+c+d}).$$

Thus, the second contingency table, E, of the *expected* number of individuals with and without haplotype h, under the null hypothesis, can be calculated as Table A.2. We denote O_{ij} and E_{ij} as the data cell in the i^{th} row and the j^{th} column in tables O and E, respectively (where $i = 1, 2$ and $j = 1, 2$).

A.2 Chi-Square Goodness-of-Fit Test

The χ^2 goodness-of-fit test examines *how well the observed data agree with the expectation under the null hypothesis* H_0 [83]. The χ^2 statistic is defined as:

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

In principle, if H_0 is correct, the observed frequency O_{ij} should not deviate greatly from the expected frequency E_{ij} under H_0; thus, the value of the χ^2 statistic should be small.

When all the expected or the observed frequencies are large (> 5) [83], the distribution of the χ^2 statistic can be approximated by the χ^2 distribution. Thus, the null hypothesis is rejected if $\chi^2 > \chi^2_\alpha$, where χ^2_α can be obtained from the χ^2 distribution with degree one (typically, α=0.05).

When some of the expected or observed frequencies are small (< 5), permutation tests [35] are used instead of the χ^2 distribution. The case-control labels in the data set are randomly permuted, and two contingency tables, O and E, are constructed from the permuted

sample. Finally, the χ^2 statistic is calculated based on the two contingency tables. By performing this permutation step a large number of times, we can define an empirical distribution for the χ^2 statistic under the null hypothesis, which can be used instead of the χ^2 distribution.

A.3 Odd Ratio Test

One commonly used measure of the relative probability of disease is the *odds*. If an event takes place with probability p, the *odds* for the occurrence of this event is defined as $p/(1-p)$ to one [83]. For example, if the probability p of disease is $2/3$, the odds in favor of the disease is $\{(2/3)/(1-2/3)\} = 2$ to one. Thus, the probability that the disease occurs is twice as large as the probability that it does not. Based on this, the *odds ratio* (OR) can be defined to compare the probabilities for disease occurrence in two groups, one group possessing the haplotype h and the other does not.

Let $exposed$ be a group of individuals with the haplotype h and $unexposed$ be a group of individuals with haplotypes $\in H - \{h\}$. The odds ratio (OR) is defined as:

$$OR = \frac{P(disease|exposed)/\{1-P(disease|exposed)\}}{P(disease|unexposed)/\{1-P(disease|unexposed)\}}.$$

The value of the odds ratio is 1 if disease incidence is statistically independent of the haplotype h. Typically, if the 95% confidence interval of the odds ratio does not include 1 [83], we reject the null hypothesis.

The odds ratio test is useful when the difference in the disease incidence between two groups $exposed$ and $unexposed$ is small. For example, when the odds of $exposed$ is 0.05 and the odds of $unexposed$ is 0.02, their absolute difference is 0.03. However, their odds ratio is 2.5, which strongly indicates association between disease incidence and the $exposed$ group, as the exposed group is more than twice as likely to get the disease than the unexposed group.

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. Haplotypic analysis of the TNF locus by association efficiency and entropy. *Genome Biology*, 4(4):R24.1–13, 2003.

[2] R.M. Adkins. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genetics*, 5(22):1–7, 2004.

[3] 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.

[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] 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*, pages 466–477, 2003.

[6] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: a direct approach. Technical Report CSE-2002-21, UC Davis Computer Science, 2002.

[7] V. Bafna, B. V. Halldorsson, 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*, pages 19–26, 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*, 2006.

[9] M.A. Beaumont and Bruce Rannala. The bayesian revolution in genetics. *Nature Reviews - Genetics*, 5:251–260, 2004.

[10] 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.

[11] 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.

[12] D.G. Brown and I.M. Harrower. A new formulation for haplotype inference by pure parsimony. Technical Report CS-2005-03, University of Waterloo, 2005.

[13] M. C. Byng, J. C. Whittaker, A. Pl Cuthbert, C. G. Mathew, and C. M. Lewis. SNP subset selection for genetic association studies. *Annals of Human Genetics*, 67:543–556, 2003.

[14] 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 disequilirium. *American Journal of Human Genetics*, 74:106–120, 2004.

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

[16] A.G. Clark, K.M Weiss, D.A. Nickerson, et al. Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, 63:595–612, 1998.

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

[18] D. Clayton. SNPHAP: A program for estimating frequencies of large haplotypes of SNPs. http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt, exists in Jan. 2006.

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

[20] M.J. Daly. High-resolution haplotype strucutre in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

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

[22] K. Ding, J. Zhang, K. Zhou, Y. Shen, and X. Zhang. htSNPer1.0: software for haplotype block partition and htSNPs selection. *BMC Bioinformatics*, 6(38):1–7, 2005.

[23] 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.

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

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

[26] B. Efron. *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, 1982.

[27] E. Eskin, E. Halperin, and R.M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, 1(1):1–20, 2003.

[28] E. Eskin, E. Halperin, and R.M. Karp. Large scale reconstruction of haplotypes from genotype data. *In Proceddings of the Annual International Conference on Research in Computational Molecular Biology*, pages 104–113, 2003.

[29] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology of Evolution*, 12:921–927, 1995.

[30] 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.

[31] 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.

[32] S.B. Gabriel, S.F. Scahffner, H. Nguyen, J.M. Moore, et al. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

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

[34] 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.

[35] P. Good. *Permutation Tests: A practical guide to resamping methods for testing hypotheses*. Springer-Verlag, 1993.

[36] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[37] D. Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *Journal of Computational Biology*, 8(3):305–323, 2001.

[38] D. Gusfield. Haplotype inference by pure parsimony. *In Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching*, pages 144–155, 2002.

[39] D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. *In Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology*, pages 166–175, 2002.

[40] D. Gusfield and S.H. Orzack. *CRC Handbook in Bioinformatics*, chapter 1. Haplotype Inference, pages 1–25. CRC Press, 2005.

[41] H. Haga, R. Yamada, Y. Ohnishi, Y. Nakamura, and T. Tanaka. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: Identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. *Journal of Human Genetics*, 47(11):605–610, 2002.

[42] B. V. Halldorsson, V. Bafna, N. Edwards, and R. Lippert. *SNPs and Haplotype Inference*, chapter 2. A Survey of Computational Methods for Determining Haplotypes, pages 26–47. Springer-Verlag Berlin Heidelberg, 2004.

[43] B. V. Halldorsson, 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.

[44] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–1849, 2003.

[45] E. Halperin and R.M. Karp. Perfect phylogeny and haplotype assignment. *In Proceedings of the Annual International Conference on Research in Computational Molecular Biology*, pages 10–19, 2004.

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

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

[48] M. Hawley and K. Kidd. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86:409–411, 1995.

[49] M.E. Hawley, A.J. Pakstis, and K.K. Kidd. A computer program implementing the EM algorithm for haplotype frequency estimation. *American Journal of physiological Anthropololy*, 18:104, 1994.

[50] P.W. Hedrick. *Genetics of pouplation*. Jones and Bartlett Publishers, 3rd Edition, 2004.

[51] J. Hoh, A. Wille, R. Zee, S. Cheng, R. Reynolds, K. Lindpaintner, and J. Ott. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Annual Human Genetics*, 64:413–417, 2000.

[52] 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.

[53] Y.T. Huang, K.M. Chao, and T. Chen. An approximation algorithm for haplotype inference by maximum parsimony. *2005 ACM symposium on applied computing*, pages 146–150, 2005.

[54] J. Hugot et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411(6837):599–603, 2001.

[55] The international HapMap Consortium. The international HapMap Project. *Nature*, 426:789–796, 2003.

[56] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, et al. High-resolution haplotype strucutre in the human genome. *Nature Genetics*, 29(2):233–237, 2001.

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

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

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

[60] 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.

[61] 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.

[62] E.D. Kelly, F. Sievers, and R. McManus. Haplotype frequency estimation error analysis in the presence of missing genotype data. *BMC Bioinformatics*, 5(188):1–13, 2004.

[63] G. Kimmel and R. Shamir. Maximum likelihood resolution of multi-block genotypes. *In Proceedings of the 11th International Conference on Computational Molecular Biology*, pages 2–9, 2004.

[64] K.M. Kirk and L.R. Cardon. The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, 10:616–622, 2002.

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

[66] G. Lancia, C.M. Pinotti, and R. Rizzi. Haplotyping populations: Complexity and approximations. Technical Report DIT-02-080, University of Trento, 2002.

[67] G. Lancia, M.C. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony: complexity of exact and approximation algorithms. *Journal on Computing*, 16(4):348–359, 2004.

[68] H. Lee, H. Cho, and H. Song. Comparison of EM algorithm and PHASE for haplotype frequency estimation with diverse accuracy measures. *In Proceedings of the Spring Conference, Korea Statistical Society*, pages 229–234, 2004.

[69] S.S. Li, N. Khalid, C. Carlson, and L.P. Zhao. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, 4(4):513–522, 2003.

[70] S. Lin, D. culter, M. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *American Journal of Human Genetics*, 71:1129–1137, 2002.

[71] 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.

[72] J.C. Long, R.C Williams, and M. Urbanek. An EM algorithm and testing strategy for multiple locus haplotypes. *American Journal of Human Genetics*, 56(3):799–810, 1995.

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

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

[75] A. Mas, E. Blanco, G. Monux, et al. DRB1-TNF-alpha-TNF-beta haplotype is strongly associated with severe aortoiliac occlusive disease, a clinical form of atherosclerosis. *Human Immunology*, 66(10):1062–1067, 2005.

[76] 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.

[77] R. W. Morris and N. L. Kaplan. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology*, 23:221–233, 2002.

[78] R. Mott. Marker selection by maximum entropy. http://www.well.ox.ac.uk/ rmott/SNPS/, exists in Jan. 2006.

[79] T. Niu, Z.S. Qin, X. Xu, and J.S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.

[80] M. Nothnagel. *The definition of multilocus haplotype blocks and common diseases*. PhD thesis, University of Berlin, 2004.

[81] G. Olund. Analysis and implementation of statistical algorithms capable of estimating haplotypes in phase-unknown genotype data. Master's thesis, Kungliga Tekniska Hogskolan, 2004.

[82] Duggal P, Gillanders E M, Mathias R A, Ibay G P, et al. Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium. *BMC Genetics*, 6 (Suppl 1):S73, 2005.

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

[84] N. Patil, A.J. Berno, D.A. Hinds, et al. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1722, 2001.

[85] 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. common disease/common variant.

[86] Z.S. Qin, T. Niu, and J.S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, 71:1242–1247, 2002.

[87] 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*, Epub ahead of print, 2006.

[88] 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.

[89] S. C. Shah and A. Kusiak. Data mining and genetic algorithm based gene/SNP selection. *Artifical Intelligence in Medicine*, 31:183–196, 2004.

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

[91] 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*, 29:308–311, 2001.

[92] M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction. *American Journal of Human Genetics*, 73(5):1162–1169, 2003.

[93] M. Stephens, N.J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4):978–989, 2001.

[94] S.A. Tishkoff, A.J. Pakstis, G. Ruano, and K.K. Kidd. The accuarcy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *American Journal of Human Genetics*, 67:518–522, 2000.

[95] A. Tsalenko, A. Ben-Dor, N. Cox, and Z. Yakhini. Methods for analysis and visualization of SNP genotype data for complex diseases. *In Proceedings of Pacific Symposium on Biocomputing*, pages 548–561, 2003.

[96] L. Wang and Y. Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780, 2003.

[97] N. Wang, J.M. Akey, K. Zhang, K. Chakraborty, and L. Jin. Recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics*, 71:1227–1234, 2002.

[98] X. Wang. Hit: a haplotype inference testbed. Technical report, Department of Electrical and Computer Engineering - CAPSL, University of Delaware, 2003.

[99] M.E. Weale, C. Depondt, S.J. Macdonald, A. Smith, P.S. Lai, S.D. Shorvon, N.W. Wood, and D.B. Goldstein. Selection and evaluation of tagging SNPs in the nueronal-sodium-channel gene scn1a: implications for linkage-disequilibrium gene mapping. *American Journal of Human Genetics*, 73(3):551–565, 2003.

[100] X. Wu, A. Luke, M. Rieder, et al. An association study of angiotensiongen polymorphisms with serum level and hypertension in an african-american population. *Journal of Hypertension*, 21(10):1847–1852, 2003.

[101] 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*, pages 879–886, 2004.

[102] C.F. Xu, K. Lewis, K.L. Cantone, P. Khan, C. Donnelly, N. White, N. Crocker, P.R. Boyed, D.V. Zaykin, and I.J. Purvis. Effectiveness of computational methods in haplotype prediction. *Human Genetics*, 110:148–156, 2002.

[103] 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*, pages 332–340, 2003.

[104] 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.

[105] 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*, 99(11):7335–7339, 2002.

[106] K. Zhang and L. Jin. Haploblockfinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1301, 2003.

[107] 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.

[108] 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.

[109] P. Zhang, H. Sheng, and R. Uehara. A double classification tree search algorithm for index SNP selection. *BMC Bioinformatics*, 5(89):1–6, 2004.

[110] S. Zhang, A.J. Pakstis, K.K. Kidd, and H. Zhao. Comparison of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *American Journal of Human Genetics*, 69:906–914, 2001.

[111] H. Zhao, R. Pfeiffer, and M. H. Gail. Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4(2):171–178, 2003.