11-6-2007

Genotype/Haplotype Tagging Methods and their Validation

Jun Zhang

GENOTYPE/HAPLOTYPE TAGGING METHODS AND THEIR VALIDATION

by

Jun Zhang

Under the Direction of Alex Zelikovsky

ABSTRACT

This study focuses how the MLR-tagging for statistical covering, i.e. either maximizing average R^2 for certain number of requested tags or minimizing number of tags such that for any non-tag SNP there exists a highly correlated (squared correlation $R^2 > 0.8$) tag SNP. We compare with tagger, a software for selecting tags in hapMap project. MLR-tagging needs less number of tags than tagger in all 6 cases of the given test sets except 2. Meanwhile, Biologists can detect or collect data only from a small set. So, this will bring a problem for scientists that the estimates accuracy of tag SNPs when constructing the complete human haplotype map. This study investigates how the MLR-tagging for statistically coverage performs under unbias study. The experiment results shows MLR-tagging still select small amount of SNPs very well even without observing the entire SNP in the sample.

INDEX WORDS: Tagging, Validation, Haplotype, Genotype, SNP,

GENOTYPE/HAPLOTYPE TAGGING METHODS AND THEIR VALIDATION

by

Jun Zhang

A Dissertation Submitted in Partial Fulfillment of Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2007

GENOTYPE/HAPLOTYPE TAGGING METHODS AND THEIR VALIDATION

by

Jun Zhang

Major Professor: Alex Zelikovsky
Committee: Raj Sunderraman
XiaoLin Hu

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
December 2007

DEDICATON

To my dear daughter, Jennifer, my husband, Jingwu and my parents

ACKNOWLEDGMENTS

The thesis would not have been possible without the help of many people. I would like to take this opportunity to express my deep appreciation to all those who helped me during my study in Georgia State University. First and foremost, I am deeply grateful to my major professor Dr. Alex Zelikovsky, for his help, guidance, encouragement, and the time that he has spent on directing my thesis. I could not finish the thesis work without his insightful direction. I also wish to thank my thesis committee members Dr. XiaoLin Hu and Dr. Raj Sunderraman for their helpful comments and useful suggestions on my thesis. I would like to thank all my friends, in particular Altun Gulsah for their help. I also appreciate support and assistance from our research group: Dumitru Brinza, Diana Mohan, Kelly Westbrooks, Stefan Gremalschi, Irina Astrovskaya and Qiong Cheng. Last but not least, I wish to thank my parents Xianjiao Zhang and Guizhi Dai, my husband Jingwu He and my daughter Jennifer for their love, constant support, and motivation.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Biology Background: SNPs, Haplotypes, Genotypes, and Notations

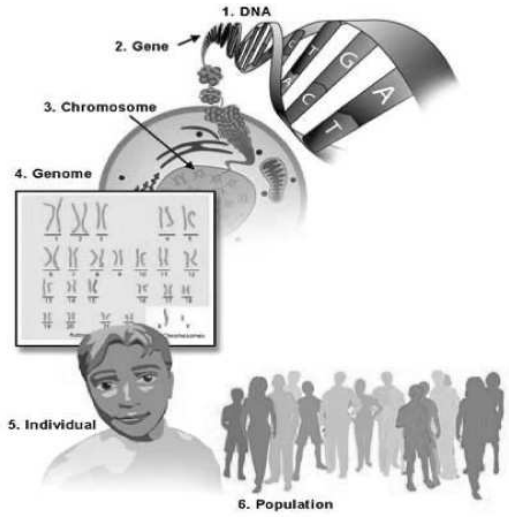

Figure 1.1. DNA, gene, chromosome, genome

Usually all living organisms are organized in 4 levels: Genome, chromosomes, genes, and DNA (see Figure 1.1). DNA is a double helical molecule with specific base pairing rules. Each of the two strands of the double helical structure serves as a template for synthesis of a new DNA strand during replication. Before a cell divides, the DNA within the cell nucleus is copied with exceptional fidelity. Information in DNA is organized into Genes, which is the second level. Genes make up Chromosomes, and all chromosomes taken together form an organism's Genome. Every cell in an Individual contains the genome. Cells are the fundamental working units of every living

organism. Each cell contains a complete copy of an organism's genome. The genome is distributed along chromosomes, which are made of compressed and entwined DNA. A gene is a segment of chromosomal DNA that directs the synthesis of a protein. DNA is made of two complimentary strands of nucleotides. A's complement is T and G's complement is C. Usually the more the living organism has evolved, the longer genome they have. The length of DNA is measured by the number of base pairs (bp).

Humans have 46 total chromosomes, two copies of each of 23 different types. Chromosomes 1 through 22 are the same in both males and females. The sex (X and Y) chromosomes differ between the sexes. Males have one X and one Y chromosome, whereas females have two X and no Y chromosomes. One copy of each chromosome type is inherited from the mother and one from the father. A father contributes an X chromosome to each of his daughters and a Y chromosome to each of his sons.

In diploid organisms each chromosome has two "copies" which are not completely identical. Each of two single copies is called a haplotype, while a description of the data consisting of mixture of the two *haplotypes* is called a *genotype*. For complex diseases caused by more than a single gene it is important to obtain haplotype data which identify a set of gene alleles inherited together. Genome difference between any two people is about 0.1% of genome. These differences are Single Nucleotide Polymorphisms (SNPs). Both substitutions have to be observed in the general population at a frequency greater than 1%. SNP's occur as frequently as every 100-300 bases. This implies that in an entire human genome there are approximately 10 to 30 million potential SNP's. More than 4 million SNP's have been identified and the information has been made publicly available. SNPs may occur in both coding (gene) and non-coding regions of the genome. Many SNPs have no effect on cell function, but they could predispose people to disease or influence their response to a drug.

The differences between any two human individuals are produced by mutation, crossing over and genetic recombination during fertilization (union of egg and sperm).

Mutation is the change in DNA of an organism which may result in that organism being different than its parents. While there are many causes of mutations, some factors are known which rapidly increase the incidence of mutation. In crossing over which occurs in the production of sex cells or gametes in meiosis, there is an exchange of chromosome pieces between the chromosome pairs associated with each other in this process

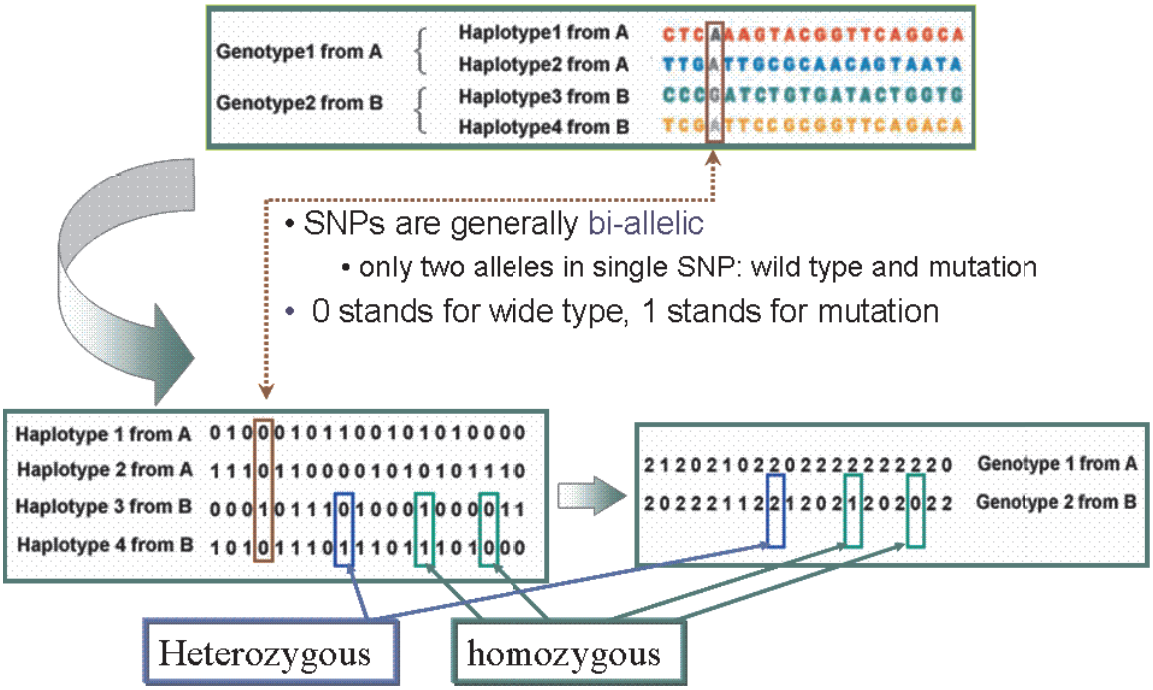

Figure 1.2. Encode

SNP's are bi-allelic and can be referred as 0 if it's a majority and 1, otherwise. If both haplotypes are the same allele, then the corresponding genotype is homogeneous, can be represented as 0 or 1. If the two haplotypes are different, then the genotype is represented as 2 (See Figure 1.2). Usually the major allele is expected to be the wild type and the minor allele is expected to be a mutation. It is important to study SNPs because they represent genetic differences among humans. Therefore biologists are searching for risk factors for genetic diseases among SNPs.

The Human Genome Project [2] is the organized, international effort to map and sequence the entire human genome. Much information about the human genome including maps and sequences are available through the internet. The great majority of the human DNA sequence has now been determined.

1.2 Tagging Motivation

Recent research found that it is essential to find a small subset of informative SNPs (tag SNPs) that may be used as good representatives of the rest of SNPs. Informative SNPs can be used for compaction of unphased genotype data. Indeed, recent successes in high throughput genotyping technologies (e.g., Affimetrix Map Arrays) drastically increase the length of available SNP sequences and they should be compacted to be feasible for fine genotype analysis. Traditionally, such SNPs are called tags and the selection procedure is referred as *tagging*. The decision which SNPs should be typed (also referred as *tag SNPs*) and which should be inferred is based on how well non-typed SNPs can be predicted from typed SNPs.

Informative SNP selection (Tagging) methods have been initially explored in statistical and pattern recognition community as well as following optimization community. In statistics, tags are required to *statistically cover* individual (non-tagged) SNPs or haplotypes (sets of SNPs), where the quality of statistical covering is usually measured by correlation, e.g., find minimum number of tags such that for any non-tag SNP there exists a highly correlated (squared correlation $R^2 > .8$) tag SNP [6, 7]. In the optimization community, the number of tags is usually minimized subject to upper bounds on *prediction error* measured as how non-tag SNPs can be predicted from the tag SNPs.

The generic informative SNP selection problem can be formulated as follows (see Figure 2.1:

Given a sample S of a population P of *individuals* (either haplotypes or genotypes) on m SNPs, find positions of k ($k < m$) tag SNPs such that one can predict (or statistically cover) an entire *individual* (haplotype or genotype) from its restriction onto the k tag SNPs.

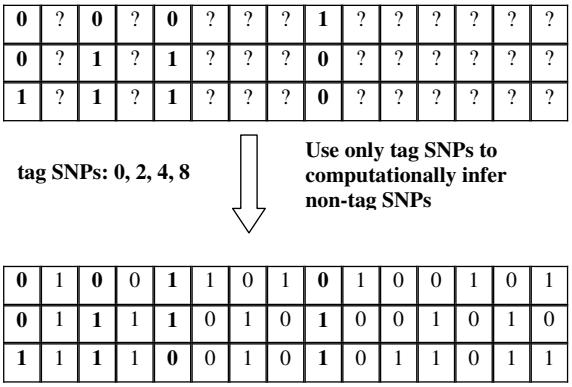

Figure 1.3. Problem formulation of Informative SNP Selection

The use of tag SNPs as a cost-effective means of capturing genetic diversity is widespread. However, the quality of the tag SNPs selected depends on the initial sample in which they are characterized. If the initial marker set is too sparse the tag SNPs chosen will capture less information than analysis suggests. Tag SNPs are commonly used as a means of capturing the genetic diversity in a region while minimizing the amount of genotyping to be performed. It is usual to select the tSNPs and judge their efficacy simultaneously. However, such an approach leads to biased estimates of tag SNP performance [28]. Thus it is unclear when using standard tag SNP procedures whether the initial marker coverage is sufficiently dense to select tag SNPs, and whether the tag SNPs selected from these markers capture the required proportion of the underlying variation in the region being studied. The density of markers required will vary from one region to another depending on factors such as recombination rate, marker frequency, mutation rate and population history. Recently a procedure was proposed for assessing the sufficiency of the marker density when selecting tag SNPs [42]. The procedure for estimating tag SNP unbiased performance

[42] is straightforward. If it is assumed that the k genotyped SNPs are drawn from the same distribution as the unobserved SNPs the 'true' performance of the tag SNPs can be estimated. Each of the k SNPs is excluded in turn, the tagging procedure performed on the remaining k-1 SNPs, and the proportion of the variance (R^2) at the excluded SNP explained by the haplotypes formed from the tag SNPs calculated. Averaging these k values should give an unbiased estimate of the performance of the tag SNPs selected from a set of k-1 SNPs. This 'leave-SNP-out' approach [42], as SNP-dropping) assumes both that the sample size is big enough that the haplotype frequencies are representative of the whole population, and that the observed SNPs have the same distribution as the unobserved SNPs. My work is to investigate MLR-tagging for statistical covering to see how it performs in the unbias tag selection.

1.3 Tagging Validation

The use of tag SNPs as a cost-effective means of capturing genetic diversity is widespread. However, the quality of the tag SNPs selected depends on the initial sample in which they are characterized. If the initial marker set is too sparse the tag SNPs chosen will capture less information than analysis suggests. Tag SNPs are commonly used as a means of capturing the genetic diversity in a region while minimizing the amount of genotyping to be performed. It is usual to select the tSNPs and judge their efficacy simultaneously. However, such an approach leads to biased estimates of tag SNP performance [28]. Thus it is unclear when using standard tag SNP procedures whether the initial marker coverage is sufficiently dense to select tag SNPs, and whether the tag SNPs selected from these markers capture the required proportion of the underlying variation in the region being studied. The density of markers required will vary from one region to another depending on factors such as recombination rate, marker frequency, mutation rate and population history. Recently a procedure was proposed for assessing the sufficiency of the marker density when

selecting tag SNPs [42]. The procedure for estimating tag SNP unbiased performance [42] is straightforward. If it is assumed that the k genotyped SNPs are drawn from the same distribution as the unobserved SNPs the 'true' performance of the tag SNPs can be estimated. Each of the k SNPs is excluded in turn, the tagging procedure performed on the remaining k-1 SNPs, and the proportion of the variance (R^2) at the excluded SNP explained by the haplotypes formed from the tag SNPs calculated. Averaging these k values should give an unbiased estimate of the performance of the tag SNPs selected from a set of k-1 SNPs. This 'leave-SNP-out' approach [42], as SNP-dropping) assumes both that the sample size is big enough that the haplotype frequencies are representative of the whole population, and that the observed SNPs have the same distribution as the unobserved SNPs. My work is to investigate MLR-tagging for statistical covering to see how it performs in the unbias tag selection.

1.4 Contribution

This thesis provides the following contributions.

Proposes a new SNP prediction using a robust tool for classification – Support Vector Machine (SVM). An extensive experimental study on various datasets including three regions from HapMap shows that the tag selection based on SVM SNP prediction can reach the same prediction accuracy as the methods of Halldorson et al. [17] on the LPL using significantly fewer tags. For example, our method reaches 90% non-tag SNP prediction accuracy using only three tags for Daly et al. [9] dataset with 103 SNPs. The proposed tagging method is also more accurate (but considerably slower) than multiple linear regression method of He et al. [20].

The corresponding software for haplotype tagging based on SVM is available for use.

Finalizes and implements tagging for statistical covering based on multiple linear regression. An experimental result shows the method are as good as the state-of-the-arts method for statical covering.

Unbiased validation of MLR-tagging for statistically coverage performs shows MLR-tagging still selects small amount of SNPs very well even without observing the entire SNP in the sample.

In the chapter 1, we introduce the biological background of my research, tagging problem formulation, and Motivation of Validation of Tagging Methods. Next we summarize the previous tagging methods and describe my previous work on hapltoye tag SNP selection by using support vector machine[21]. Chapter 3 describes the method for statistic tagging based on multiple linear regression and its corresponding software. Chapter 4 describes haplotype SNP selection by using support vector machine and the corresponding software. Finally, we discuss the unbias validation for MLR-tagging and its experiment results.

CHAPTER 2

TAGGING METHODS

The search for the association between complex diseases and single nucleotide polymorphisms (SNPs) has been recently received great attention. For these studies, it is essential to use a small subset of informative SNPs, named *tags*, accurately representing the rest of the SNPs. Firstly, informative SNPs can be used for selective SNP typing and computationally inferring all non-typed SNPs thus achieving considerable budget savings. Secondly, informative SNPs can be used for compaction of SNP data. Indeed, recent successes in high throughput genotyping technologies (e.g., Affimetrix Map Arrays) drastically increase the length of available SNP sequences and they should be compacted to be feasible for fine genotype analysis. This chapter summarizes the stat-of-the-art informative SNP seleciton tools.

2.1 Problem Formulation

The generic informative SNP selection problem can be formulated as follows (see Figure 2.1:

Given a sample S of a population P of *individuals* (either haplotypes or genotypes) on m SNPs, find positions of k ($k < m$) tag SNPs such that one can predict (or statistically cover) an entire *individual* (haplotype or genotype) from its restriction onto the k tag SNPs.

The use of tag SNPs as a cost-effective means of capturing genetic diversity is widespread. However, the quality of the tag SNPs selected depends on the initial sample in which they are characterized. If the initial marker set is too sparse the

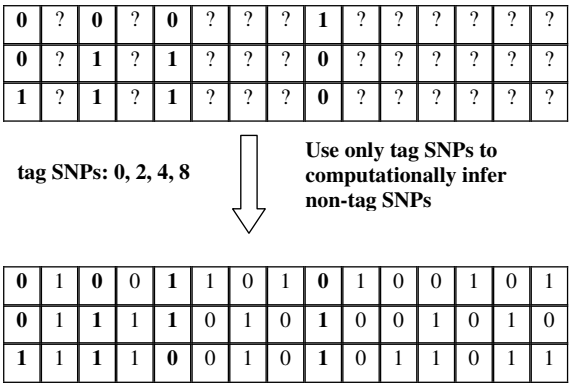

Figure 2.1. Problem formulation of Informative SNP Selection

tag SNPs chosen will capture less information than analysis suggests. Tag SNPs are commonly used as a means of capturing the genetic diversity in a region while minimizing the amount of genotyping to be performed. It is usual to select the tSNPs and judge their efficacy simultaneously. However, such an approach leads to biased estimates of tag SNP performance [28]. Thus it is unclear when using standard tag SNP procedures whether the initial marker coverage is sufficiently dense to select tag SNPs, and whether the tag SNPs selected from these markers capture the required proportion of the underlying variation in the region being studied. The density of markers required will vary from one region to another depending on factors such as recombination rate, marker frequency, mutation rate and population history. Recently a procedure was proposed for assessing the sufficiency of the marker density when selecting tag SNPs [42]. The procedure for estimating tag SNP unbiased performance [42] is straightforward. If it is assumed that the k genotyped SNPs are drawn from the same distribution as the unobserved SNPs the 'true' performance of the tag SNPs can be estimated. Each of the k SNPs is excluded in turn, the tagging procedure performed on the remaining k-1 SNPs, and the proportion of the variance (R^2) at the excluded SNP explained by the haplotypes formed from the tag SNPs calculated. Averaging these k values should give an unbiased estimate of the performance of the tag SNPs selected from a set of k-1 SNPs. This 'leave-SNP-out' approach [42], as

SNP-dropping) assumes both that the sample size is big enough that the haplotype frequencies are representative of the whole population, and that the observed SNPs have the same distribution as the unobserved SNPs. My work is to investigate MLR-tagging for statistical covering to see how it performs in the unbias tag selection.

2.2 Overview of Previous Work

Informative SNP selection (Tagging) methods have been previously explored in statistical and pattern recognition community as well as optimization community. In statistics, tags are required to *statistically cover* individual (non-tagged) SNPs or haplotypes (sets of SNPs), where the quality of statistical covering is usually measured by correlation, e.g., find minimum number of tags such that for any non-tag SNP there exists a highly correlated (squared correlation $R^2 > .8$) tag SNP [6, 7]. In the optimization community, the number of tags is usually minimized subject to upper bounds on *prediction error* measured as how non-tag SNPs can be predicted from the tag SNPs.

Previous research on tag SNP selection has explored both lossless and lossy methods. Lossless methods select a set of tag SNPs that capture 100% of the haplotypic variation in the sample population. Lossy methods typically select fewer tags than lossless methods, but with some tolerated amount of information loss.

Aviitzhak et al. [4] presented a method for selecting tags which can be used in both a lossless and a lossy manner. The central idea behind both their lossless and lossy methods is to eliminate tags that contribute the least to the Shannon entropy for the haplotype set. First, identical columns and complimentary columns are eliminated, then they eliminate columns that do not reduce the number of unique rows. They note that selecting a maximal linearly independent set of column vectors would miss opportunities to eliminate complimentary SNPs and illustrate that by the 2-by-2 identity matrix. Their lossless method reduces by 25% and 36% the number of

SNPs describing the haplotype diversity within an African-American and Caucasian population, respectively.

Zhang et al. [43] introduced a block-based, dynamic programming algorithm for haplotype inference that is capable of reconstructing 90% of the original data using only 35% of SNPs as tags. They used the partition-ligation expectation maximization algorithm for haplotype inference, and as a result, provided a method of performing association studies directly on genotype data.

Sebastiani et al. [35] described a lossless method called BEST (Best Enumeration of SNP Tags) for identifying a minimal set of tag SNPs from haplotype data. BEST selects tags by determining if a candidate tag is a boolean function of SNPs already chosen as tags. The BEST method selected 14% of SNPs as tags from an African-American population and 10% from an European-American population by considering individual genes each ranging from 5 to 229 SNPs in length. However, its effectiveness on a genome-wide scale is still unproven. According to their method, 95% of tags selected from the European-American population were also selected from the African-American population, which provides evidence for the a genetic bottleneck event that occurred long ago as hominids migrated out of Africa to settle Europe and Asia.

Halldorson et al. [17] defined the *informativeness* measure of how well a set of tags describes a haplotype sample. Both the informativeness measure, as well as their tag SNP selection method consider a graph whose vertices are SNPs; an edge is placed between to SNPs if one SNP can be used to reliably predict the other. Their method seeks the set of SNPs that maximizes the informativeness measure on the haplotype data. The method can achieve prediction rates of 90% based on only 20% of SNPs. Halldorsson's method differs from the others in that it is a *block-free* method. Block-based methods are restricted to identifying tags only within local contiguous sequences of SNPs where the haplotype diversity is low. Block-free methods have the

capability to identify tags across an entire genome. Like Halldorsson's method, the linear reduction method we propose is a block-free method.

Lee et al. [26] introduce BNTagger, a new method for tagging SNP selection, based on conditional independence among SNPs. Using the formalism of Bayesian networks (BNs), their system aims to select a subset of independent and highly predictive SNPs. For example, BNTagger uses 10% tags to reach 90% prediction accuracy. However, BNTagger comes at the cost of compromised running time. Its running time varies from several minutes (when the number of SNPs is 52) to 2-4 hours (when the number is 103).

Our tagging problem formulations and above approaches do not take into account haplotype frequency when selecting a tag SNPs. For a discussion of how haplotype frequency affects tag SNP selection, see [7, 11, 37].

2.3 idSelect

IdSelect, developed by Carlson et al.[6], used a greedy approach for tag SNP selection. They developed a greedy algorithm to identify subsets of tagSNPs for genotyping, selected from all SNPs exceeding a specified MAF threshold. Starting with all SNPs above the MAF threshold, the single site exceeding the threshold with the maximum number of other sites above the MAF threshold is identified. This maximally informative site and all associated sites are grouped as a bin of associated sites. Not all SNPs within the bin are interchangeable, because pairwise association is not an associative property: if R^2 exceeds the threshold for SNP pairs A/B and B/C, R^2 for SNP pair A/C might not exceed the threshold. Thus, because the bin is initially ascertained using a single SNP, all pairwise R^2 within bin are re-evaluated, and any SNP exceeding threshold R^2 with all other sites in the bin is specified as a tag SNP for the bin. Thus, one or more SNPs within a bin are specified as tagSNPs, and only one tag SNP would need to be genotyped per bin. The tag SNP can be

selected for assay on the basis of genomic context (coding vs. noncoding or repeat vs. unique), ease of assay design, or other user-specified criteria.

The binning process is iterated, analyzing all as-yet-unbinned SNPs at each round, until all sites exceeding the MAF threshold are binned. Each bin is reported as a set of all SNPs in the bin as well as the subset of tag SNPs within the bin, each of which is above the r2 threshold with all other SNPs in the bin. If an SNP does not exceed the r2 threshold with any other SNP in the region, it is placed in a singleton bin.

2.4 STAMPA

Halperin et al. [16] describes a new method STAMPA for SNP prediction and tag selection. A SNP is predicted by inspecting the two closest tag SNPs from both sides; the value of the unknown SNP is given by a majority vote over the two tag SNPs. They use dynamic programming to select tags to reach best prediction score. Their methods are compared with idSelect and HapBlock on a variety of data sets, and could predict with 80% accuracy the SNPs in the daly dataset[9] using only 2 SNPs as tags. In general, this problem is computationally difficult and the runtime of an exact algorithm may become prohibitively slow. Therefore, one can use heuristics for the selection of k tags following Halperin et al.[16] who compare relatively slow STAMPA with a fast random tag selection.

2.5 MLR

2.5.1 Introduction to Multiple Linear Regression

The general purpose of multiple linear regression is to learn the relationship between several independent variables and a response variable. The multiple linear regression model is given by

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x_2} + ... + \beta_k \mathbf{x}_k + \epsilon = \mathbf{X}\beta + \epsilon \tag{2.1}$$

where $\mathbf{y}$ is the response variable (represented by a column with n coordinates ($k \leq n-1$)), $\mathbf{x_i}, i = 1, \ldots, k$ are independent variables (columns), $\beta_i, i = 1, \ldots, k$ are regression coefficients, and ϵ (a column) is the model error. The regression coefficient β_i represents the independent contribution of the independent variable x_i to the prediction of $\mathbf{y}$.

2.5.2 Multiple Linear Regression Tagging

A. Zelikovsky and J. He [20] proposed a new SNP prediction method based on rounding of multivariate linear regression (MLR) analysis in sigma-restricted coding. When predicting a non-tag SNP, the MLR method accumulates information about all tag SNPs resulting in significantly higher prediction accuracy with the same number of tags than for the previously known tagging methods. They also showed that the tag selection strongly depends on how the chosen tags will be used – advantage of one tag set over another can only be considered with respect to a certain prediction method. Two simple universal tag selection methods have been applied: a (faster) stepwise and a (slower) local-minimization tag selection algorithms. An extensive experimental study on various datasets including 6 regions from HapMap shows that the MLR prediction combined with stepwise tag selection uses significantly fewer tags (e.g., up to two times less tags to reach 90% prediction accuracy) than the state-of-art methods of Halperin et al. [16] for genotypes and Halldorsson et al. [17] for haplotypes, respectively. Our stepwise tagging matches the quality of while being faster than STAMPA [16].

The MLR-tagging method computes $b_i, i = 1, \ldots, k$ estimating unknown *true coefficients* $\beta_i, i = 1, \ldots, k$ minimizing the error $||\epsilon||$ using the least squares method. Geometrically speaking, in the *estimation space span*(X), which is the linear closure of vectors $x_i, i = 1, \ldots, k$, we find the vector $\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k = Xb$ estimating y. The vector $\hat{y}$ minimizing distance (error) $||\epsilon|| = ||\hat{y} - y||$ is the projection

of y on $span(X)$ and equals

$$\hat{y} = X(X^tX)^{-1}X^ty \tag{2.2}$$

Given the values of independent variables $x^* = (x_1^*, \ldots, x_k^*)$, the MLR method can predict (estimate) the corresponding response variable y^* with

$$\hat{y}^* = x^*(X^tX)^{-1}X^ty \tag{2.3}$$

Formally, let T be the $(n) \times k$ matrix consisting of $n1$ rows corresponding to n sample genotypes $x_i, i = \overline{1,n}$, from X, $g_i = \{x_{i,1}, \ldots, x_{i,k}\}$, whose k coordinates correspond to k tag SNPs. The SNP s is represented by a (n)-column with values $y_i, i = \overline{1,n}$. The multiple linear regression gives the R^2 bewteen T and s

$$T = \begin{bmatrix} x_{1,1} & \ldots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,k} \end{bmatrix} \qquad s = \begin{bmatrix} y_{1,k+1} \\ \vdots \\ y_{n,k+1} \end{bmatrix}$$

CHAPTER 3

STATISTICAL COVERING

This chapter describes two methods I proposed for tag SNP selection: MLR-tagging for statistically covering and haplotype SNP selection by using support vector machine and the corresponding software.

3.1 Statistical Covering

A. Zelikovsky and J. He [20] proposed a new tag SNP selection method based on multiple linear regression (MLR) analysis, i.e., MLR-tagging. When predicting a non-tag SNP, the MLR-tagging method accumulates information about all tag SNPs resulting in significantly higher prediction accuracy with the same number of tags than for the previously known tagging methods. An extensive experimental study on various datasets including 10 regions from HapMap shows that the MLR-tagging for prediction matches the quality of while being faster than STAMPA [16]. Here, we introduce MLR-tagging for statistical covering e.g., find minimum number of tags such that for any non-tag SNP there exists a highly correlated (squared correlation $R^2 > .8$) tag SNP [6, 7] (see Figure 3.1).

3.1.1 Tag SNP Selection Based on SNP Statistical Covering

A. Zelikovsky and J. He [20] showed how to separate the tag selection from SNP prediction. Following their work, we fist define SNP statistical covering algorithm as follows:

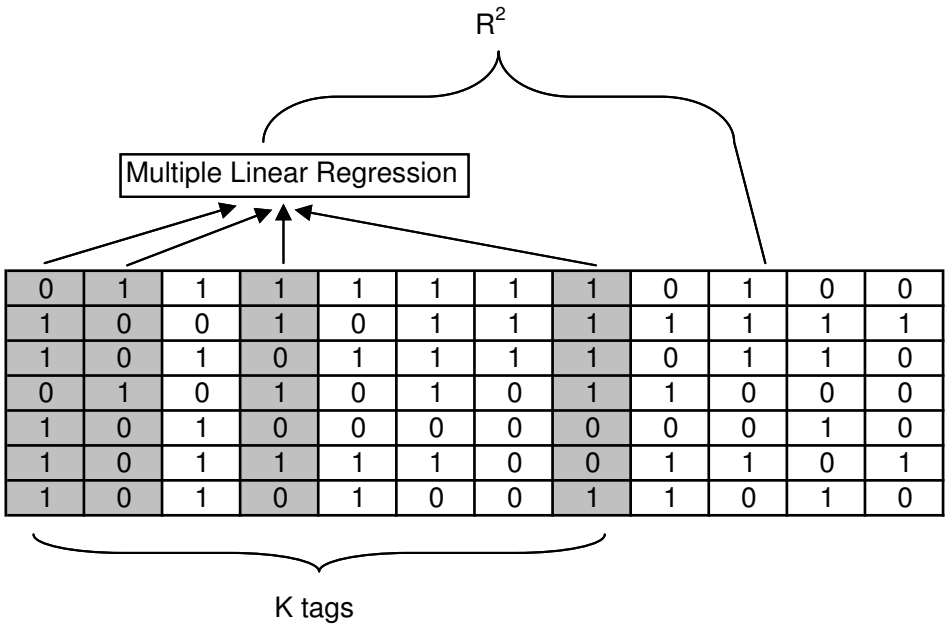

Figure 3.1. MLR-tagging for statistical covering. The shaded columns correspond to k tag SNPs and the clear columns correspond to $m - k$ non-tag SNPs. MLR-tagging method ensures that non-tag SNPs can be statistically covered by tags

A *SNP statistical covering algorithm* A_k accepts as its input the values of k tags $(t_1, \cdots, t_k)$ of a sample S. The output of A_k is R^2, that is, R^2 is correlation coefficient between the non-tag SNPs and k tags.

We select tags by using SNP statistical covering algorithm as follows: We can check each k-tuple of tags and choose the k-tuple either maximizing average R^2 for all SNPs or number of statistical covered SNPs. This manner of exhaustive search is very expensive in terms of running time. We introduced a greedy manner of selection. It starts with the auxiliary tag t_0, finds such tag t_1 which would be the best extension of $\{t_0\}$ and continue adding best tags until reaching the set of tags of the given size k. This produces *hereditary* set of tags, i.e., the chosen k tags contain the chosen $k - 1$ tags. This hereditary property may be useful in case if the set of tags can be extended. The runtime of greedy manner is $O(knmT)$, where T is the runtime of the SNP statistical covering algorithm.

3.1.2 Experimental Results

We apply our haplotype tagging algorithm (SVM/STA) to 4 dataset related a genetic disease. Dataset of ADIPOQ-EA-1 are collect from 90 individuals of 30 SNPs. Dataset of ADIPOQ-EA-2 are collect from 90 individuals of 65 SNPs. We apply the statistical tagging software on both genotype and haplotype data of ADIPOQ-EA-1 and ADIPOQ-EA-2. Table 3.1 and 3.2 shows how entire sample can be statical covered by number of tags.

Table 3.1. Statistically covered number (/percentage) of SNPs to number (/percentage) of tags in ADIPOQ-EA-1

ADIPOQ-EA(Hap) (180 x 30)				ADIPOQ-EA(Geno) (90 x 30)			
tags		Covered SNPs		tags		Covered SNPs	
tags	covered tags	tags	covered tags	tags	covered tags	tags	covered tags
1	5.26	5	26.31	1	5.26	5	26.31
2	10.52	9	47.36	2	10.52	10	52.63
3	15.78	11	57.89	3	15.78	11	57.89
4	21.05	12	63.15	4	21.05	12	63.15
5	26.31	13	68.42	5	26.31	13	68.42
6	31.57	14	73.68	6	31.57	14	73.68
7	36.84	15	78.94	7	36.84	5	78.94
8	42.15	16	84.21	8	42.15	16	84.21
9	47.36	17	89.47	9	47.36	17	89.47
10	52.63	18	94.74	10	52.63	18	94.74
11	57.84	19	100	11	57.84	19	100

Tagger is a tool for the selection and evaluation of tag SNPs from genotype data such as that from the International HapMap Project. It combines the simplicity of pairwise tagging methods with the efficiency benefits of multimarker haplotype approaches. Tagger produces a list of tag SNPs and corresponding statistical tests to capture all variants of interest, and a summary coverage report of the selected tag SNPs. We compare MLR statis-tagging with tagger. The results shows MLR-tagging has almost same quality as tagger as in 3.3. We do better than tagger in all cases of the given test sets except 2 cases.

3.1.3 Software

Here, we describe our tagging software base on multiple linear regression for statistical covering. statis-tagging software package implements a novel genotype tagging method based on multiple linear regression (MLR) analysis for statistical covering. The software selects stepwise tags based on a haplotype/genotype sample data and R2.

Downloading and Installing All relevant files including this pdf file are included in the tar files: available at http://alla.cs.gsu.edu/ software/stat-tagging. Download this tar file to your machine then extract the files from the archive.

tar -xvf statis-tagging.tar Currently, there is only Linux version available.

The package contains the following files:

1. taggingStatReadme.pdf: Readme file

2. statis-tagging: Binary code for tag selection

3. genoInput.txt: Sample input of a genotype population sample: 129 offspring genotypes each with 103 SNPs from Daly et al.[9]

4. tagFile.txt: Sample input of tag positions

For running MLRsta:

type ./MLR-tagging-stat genoInput.txt 0.8 tagFile.txt G"

First parameter = the file name of a genotype sample population "

Second parameter = Threshold R^2

Third parameter = the name of output tag file (it contains selected k tag positions)

Fourth parameter = H for haplotype file input and G for genotype file input

File Formats:

genoInput.txt contain the following lines:

The number of genotypes

The number N of SNPs in each genotype

Description of data (can be empty)

The first genotype represented by a sequence of 0/1/2's without gaps, 0 stands for homozygous major allele, 1 stands for homozygous minor allele, and 2 stands for heterozygous SNP.

.....

The last genotype

tagFile.txt consists of k+3 lines:

The number of tags

Description of data (can be empty)

Description of data (can be empty)

The position of the first tag (a number in the range from 0 to N-1, where N is the number of SNPs.)

.....

The last tag

Table 3.2. Statistically covered number (/percentage) of SNPs to number (/percentage) of tags in ADIPOQ-EA-2

ADIPOQ-EA(Hap) (180 x 65)				ADIPOQ-EA(Geno) (90 x 65)			
tags		Covered SNPs		tags		Covered SNPs	
tags	covered tags	tags	covered tags	tags	covered tags	tags	covered tags
1	1.53	17	26.15	1	1.53	15	23.08
2	3.07	43	66.15	2	3.07	26	40
3	4.61	51	78.46	3	4.61	33	50.76
4	6.15	56	86.15	4	6.15	38	58.46
5	7.69	58	89.23	5	7.69	43	66.15
6	9.23	60	92.31	6	9.23	45	69.23
7	10.76	61	93.84	7	10.76	47	72.30
8	12.31	63	96.92	8	12.31	48	73.84
9	13.84	64	98.46	9	13.84	49	75.38
10	15.38	65	100	10	15.38	50	76.92
-	-	-	-	11	16.92	51	78.46
-	-	-	-	12	18.46	52	80
-	-	-	-	13	20	53	81.23
-	-	-	-	14	21.03	54	83.26
-	-	-	-	15	23.07	55	84.87
-	-	-	-	16	24.61	56	86.15
-	-	-	-	17	26.15	57	87.69
-	-	-	-	18	27.69	58	89.23
-	-	-	-	19	29.23	59	90.76
-	-	-	-	20	30.76	60	92.37
-	-	-	-	21	32.30	61	93.84
-	-	-	-	22	32.84	62	95.38
-	-	-	-	23	35.84	63	96.92
-	-	-	-	24	36.92	64	98.51
-	-	-	-	25	38.46	65	100

Table 3.3. Number of tags needed to statistically cover entire dataset

DataSet (SNPs)	tags needed (MLR-tagging)	tags needed (tagger)
ADIPOQ-AA (15)	10	12
ADIPOQ-EA (19)	11	14
ADIPOR1-AA (16)	12	10
ADIPOR1-EA (12)	5	8
ADIPOR1-AA (71)	27	32
ADIPOR1-EA (65)	25	17

CHAPTER 4

HAPLOTYPE TAG SELECTION BASED ON SUPPORT VECTOR MACHINE

We propose a new SNP prediction using a robust tool for classification – Support Vector Machine (SVM). For tag selection we use a fast stepwise tag selection algorithm. An extensive experimental study on various datasets including three regions from HapMap shows that the tag selection based on SVM SNP prediction can reach the same prediction accuracy as the methods of Halldorson et al. [17] on the LPL using significantly fewer tags. For example, our method reaches 90% non-tag SNP prediction accuracy using only three tags for Daly et al. [9] dataset with 103 SNPs. The proposed tagging method is also more accurate (but considerably slower) than multiple linear regression method of He et al. [20].

4.1 SVM Overview

SVM has recently attracted a lot of attention in bioinformatics research (see, e.g. [38]). This is because SVM produces very accurate results comparatively with other data mining approaches such as Neural Networks. The SVM method is a learning system which is developed by Vapnik and Cortes [41]. SVM is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation. The basic principle behind SVM is to find an optimal maximal margin separating hyperplane between two classes. The goal is to maximize the margin between the solid planes separating the two classes and at the same time permit the least amount of errors as possible. SVM can also be used in the case

when the data is not linearly separable. In this case, the data is mapped to a high dimensional future space using a nonlinear function. When using SVM, the dot products (x,y) in the future space must be fed to the SVM, which can be computed through a positive definite kernel in the input space.

After given a training set (a set of pairs, input vector: features and target), SVM builds a model. This model is later applied to unknown test set where the model maps an input vector to $+1$ (positive class) or -1 (negative class) output target value.

SVMlight is an implementation of Vapnik's Support Vector Machine [41]. In this project, we have used SVM^{light} software as a black box to do the prediction. The SVM^{light} software has many features such as changing the kernel function and other parameters. We have used the Radial Basis Function (RBF) kernel in our project it is the default and recommended kernel function.

$$\exp(-\gamma * |u - v|^2)$$

For the trade-off between training error and margin, 0.05 is chosen (c value). Parameter gamma in RBF kernel was chosen as 0.1. These parameters were found by testing different values in our experiments. We used the same for all the experiments.

4.2 SVM Haplotype Tagging

This problem can be formulated as **Haplotype Tagging Problem** (see Figure 4.1). Given the full pattern of all haplotypes in a small population sample, find the minimum number of tag SNPs and the method for reconstructing each haplotype in the entire population from these tags.

This tagging problem formulation implicitly relies on a certain SNP prediction method. The corresponding **SNP prediction problem** is formulated as follows: Given the values of k tags of the individual x with unknown SNP s and n individuals with k tag SNP and known value of SNP s, find the value of s in x.

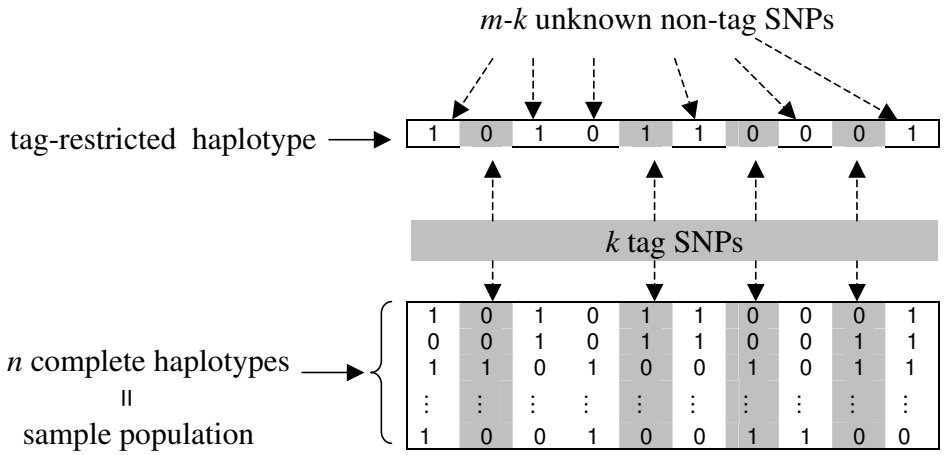

Figure 4.1. Haplotype Tagging Problem. The shaded columns correspond to k tag SNPs and the clear columns correspond to $m - k$ non-tag SNPs. The unknown $m - k$ non-tag SNP values in tag-restricted haplotype (top) are predicted based on the known k tag values and the sample population of n complete haplotypes.

In the SNP Prediction Problem, SVM builds a model after given n complete haplotypes as training set. Then when an unknown haplotype is given to SVM as a test sample, SVM is asked to predict the unknown SNP value (see Figure 4.2).

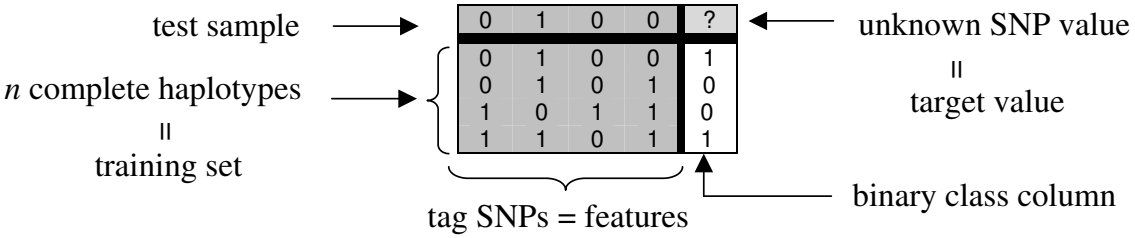

Figure 4.2. The SNP Prediction Problem. Each haplotype with k tags in the training set belongs to the 0- or 1- class. These binary class values are given in the last column. For a given k tag-restricted haplotype (test sample), the unknown non-tag SNP in the right corner should be classified based on the known tag SNP values and training set.

4.3 Experimental Results

We apply our haplotype tagging algorithm (SVM/STA) to very well known haplotype datasets. These datasets are original genotype datasets, but we phased them to obtain haplotypes using GERBIL algorithms [12].

Two gene Regions form HapMap. Two gene regions STEAP and TRPM8 from 30 CEPH family trios are obtained from HapMap [2]. We took the HapMap SNPs that are spanned by the gene plus 10KB upstream and downstream. The number of SNPs genotyped in each gene region is 23 and 102 SNPs. We only use 60 haplotypes of parents.

Chromosome 5q31. The data set collected by Daly et al. [9] is derived from the 616 kilobase region of human Chromosome $5q31$ that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. We only use 258 haplotypes of offsprings.

LPL The Clark et al. [6] data set consists of the haplotypes of 71 individuals typed over 88 SNPs in the human lipoprotein lipase (LPL) gene.

We apply leave-one-out cross-validation to evaluate the quality of the solution given by the tag SNP selection and prediction methods. One by one, each individual is removed from the sample. Then, tag SNPs are selected using only the remaining individuals. The "left out" individual is reconstructed based on its tag SNPs and the remaining individuals in the sample. The average number of errors in the reconstruction of all individuals is used as a measure of the overall prediction accuracy.

Table 4.1 presents the results of STA combined with SVM (SVM/STA) on leave-one-out experiments on the 3 haplotype datasets. Table 4.2 compares SVM/STA with multiple linear regression method (MLR) of He et al. [22] on the 3 haplotype datasets. The proposed tagging method is more accurate than multiple linear regression method of He et al. [20]. For example, for small number of tag SNPs, SVM/STA can obtain (up to 8%) better prediction accuracy than MLR with same number of tag SNPs. But SVM/STA is considerably slower. Indeed, for 5q31 dataset, SVM/STA needs 3 hours to select 1 tag SNPs while MLR only needs 0.77 seconds[1].

[1]All experiments are performed on a computer with Intel Pentium 4, 3.06Ghz processor and 2 GB of RAM.

Table 4.1. Leave-one-out tests are performed on 3 real haplotype datasets. The minimum number of tag SNPs needed to reach from 80% to 99% prediction accuracy is listed. The bold numbers indicate cases when the SVM/STA needs fewer tags than the MLR method of He et al. [20] for reaching same prediction accuracy.

	prediction accuracy %											
datasets (# of SNPs)	80	85	90	91	92	93	94	95	96	97	98	99
5q31 (103)	1	**1**	**3**	**3**	**4**	**5**	**6**	**8**	**10**	**22**	**42**	**51**
TRPM8 (101)	1	**1**	**2**	5	5	6	7	8	10	15	15	24
STEAP (22)	1	1	1	1	1	1	1	2	**2**	**2**	**2**	2

Table 4.2. The comparison of our proposed SVM/STA method and the MLR method of He et al. [20] over different number of tag SNPs.

datasets (# of SNPs)		methods	number of tag SNPs					
			1	2	4	6	8	10
5q31 (103)	prediction accuracy %	SVM/STA	86.81	89.32	92.24	94.09	95.28	96.09
		MLR	81.15	83.84	88.15	90.91	92.66	93.49
	running time	SVM/STA	3 hour	5 hour	11 hour	16 hour	18 hour	1 day
		MLR	0.77 sec	1.16 sec	4.07 sec	7.27 sec	11.26 sec	15.92 sec
TRPM8 (101)	prediction accuracy %	SVM/STA	88.89	90.50	90.67	93.67	95.56	96.74
		MLR	80.68	85.32	90.75	93.74	95.16	96.38
	running time	SVM/STA	1 hour	2 hour	5 hour	9 hour	16 hour	23 hour
		MLR	0.357 sec	0.787 sec	1.895 sec	3.376 sec	5.181 sec	7.373 sec
STEAP (22)	prediction accuracy %	SVM/STA	94.02	98.18	99.68	99.73	99.79	99.80
		MLR	90.79	96.16	99.13	99.71	99.78	99.78
	running time	SVM/STA	14 min	27 min	1 hour	2 hour	3 hour	4 hour
		MLR	0.034 sec	0.052 sec	0.118 sec	0.203 sec	0.304 sec	0.413 sec

We also compare SVM/STA with the methods of Halldorson et al. [17] and the method of He et al. [20] in leave-one-out tests on the LPL data set (see Figure 4.3). Note that the method of Halldorson et al. imputes a SNP based on the tag SNPs in the same neighborhood and in fact can be classified as a method for statistical coverage. If there is no tag SNPs in the neighborhood, then their method does not make any prediction. It is not surprising that it performs poorly for SNP prediction. The SVM/STA method reconstructs each SNP based on the values of *all* tag SNPs which may potentially be far away. On the LPL dataset, SVM/STA reaches, e.g., 90% accuracy using only one tag.

4.4 SVM-tagging Software

Here, we describe our tagging software base on multiple linear regression. SVM-tagging selects haplotype tag SNP using support vector machines. We first describe

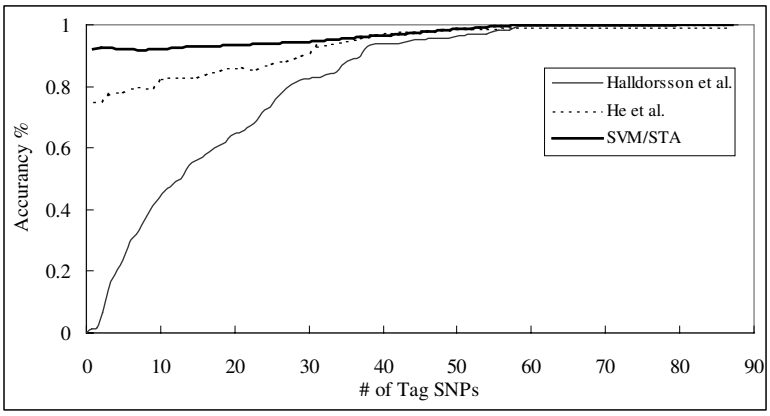

Figure 4.3. Comparison among three haplotype tagging method on LPL data: SVM/STA, Halldorson et al. [17], and He et al. [20] in a leave-one-out experiment. The x-axis shows the number of SNPs typed, and the y-axis shows the fraction of SNPs correctly imputed.

how to download, compile, and run SVMtagging package. Then we describe input and output formats.

Downloading and Installing All relevant files including this pdf file are included in the tar files: SVMtagging.tar - Linux version available at http://alla.cs.gsu.edu/ software.

Download this tar file to your machine then extract the files from the archive tar -xvf SVMtagging.tar

Compile leaveOneOutSVM.cpp: g++ -o leaveOneOutSVMleave OneOutSVM.cpp And checkExist.cpp: g++ -o checkExist checkExist.cpp

Download the SVMlight from http://svmlight.joachims.org/

Make sure perl is in $\sharp!/usr/bin/perl$

Make sure SVMlight and SVM-tagging package all are in same directory

Running the Program For running SVM-tagging:

type perl SVMtagging.perl hap.txt -g 10

- First parameter = the file name of a haplotype population sample

- Third parameter = desired number of tags K

The result tags will store in tag.txt.

File Formats hap.txt contain the following lines:

- The number of haplotypes

- The number N of SNPs in each haplotype

- Description of data (can be empty)

- The first haplotype represented by a sequence of 0/1's without gaps, 0 stands for major allele, 1 stands for minor allele.

.....

-The last haplotype

CHAPTER 5

UNBIASED VALIDATION OF TAGGING METHODS

Meanwhile, Biologists can detect or collect data only from a small part of population due to the reasons of technology and expense. For some data, we still have no way to detect them and they are unobserved to us. So, this will bring a problem for scientists that the estimates accuracy of tag SNPs when constructing the complete human haplotype map. Iles [28] and Weale et al. [42] proposed a procedure of 'dropping SNP' to investigate unbias performance of tagging method. Following their ideas, this study investigates how the MLR-tagging for statistically coverage performs under unbias study. The experiment results shows MLR-tagging still select small amount of SNPs very well even without observing the entire SNP in the sample.

5.1 Validation of Tagging Methods

5.1.1 Leave-one-out and Leave-many-out

The standard way to validate tagging method is to apply leave-one-out cross-validation to evaluate the quality of the (1) one by one, each genotype vector is removed from the sample, (2) tag SNPs are selected using only the remaining genotypes, and (3) the "left out" genotype is reconstructed based on its tag SNPs and the values of tag and non-tag SNPs in the remaining genotypes.

Instead of each time leave one out, leave-many-out cross validation method randomly removes certain percentage of sample, and then the remaining works as training sample.

5.1.2 Illis's validation procedure: Leave-SNP-out

Recently a procedure was proposed for assessing the sufficiency of the marker density when selecting tag SNPs [42]. The procedure for estimating tag SNP unbiased performance [42] is straightforward. If it is assumed that the k genotyped SNPs are drawn from the same distribution as the unobserved SNPs the 'true' performance of the tag SNPs can be estimated. Each of the k SNPs is excluded in turn, the tagging procedure performed on the remaining k-1 SNPs, and the proportion of the variance (R2) at the excluded SNP explained by the haplotypes formed from the tag SNPs calculated. Averaging these k values should give an unbiased estimate of the performance of the tag SNPs selected from a set of k-1 SNPs. This 'leave-one-out' approach [42], as SNP-dropping) assumes both that the sample size is big enough that the haplotype frequencies are representative of the whole population, and that the observed SNPs have the same distribution as the unobserved SNPs.

5.2 Experimental Results

The performance of the new approach was to test MLR method in iles' manner. We leave column out as a small sample. We selected percentage of 10%, 15% and 20% of entire dataset as our 'observed' data separately and the remainder classed as 'unobserved'. Then we randomly generated tag $k = 1, 2, .., 8$ and 10 or 30 were selected from 'observed' to calculate R^2 to find the maximal average R^2 in 'observed' region and 'unobserved' region. In such way, the average R^2 between tags and nonTag SNPs is maximum.Further, we use the tags selected from the leave-many-out sample to test how good these tags can statically cover the entire sample. As result, in Daly data, we use one tag can reach R^2 27% in its selected sample data, this one tag can cover the 15% of entire data. T represents tags, V represents visible SNPs and C represents the entire chrosome.In future, long haplotypes will be our aim. We will take a large dataset, such as $80K$ to test.

The length of the region simulated makes little difference to the accuracy of the results-what is important is the number of observed SNPs from which the tSNPs are selected. As the number of observed SNPs increases, so the estimates become more accurate.

5.2.1 Experimental Datasets

The following datasets are used to measure the quality of our algorithms. Currently, our algorithms cannot tolerate missing data. Following Halperin et al.[16], we use GERBIL [12] to phase the genotypes and then combine the resulting two haplotypes to recover any missing data.

Chromosome 5q31. The data set collected by Daly et al. [9] is derived from the 616 kilobase region of human Chromosome $5q31$ that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. We only use 258 haplotypes of offsprings.

Three gene regions. Three regions (ENm013, ENr112, ENr113) from 30 CEPH family trios obtained from HapMap ENCODE Project [2]. The number of SNPs genotyped in each region is 361, 412 and 515 respectively. Regions ENr123 and ENm010 from 2 population: 45 singles Han Chinese (HCB) and 44 singles Japanese(JPT). The number of SNPs genotyped in each region is 63 and 105.

5.2.2 The example of how we resoleved

The performance of the new approach was to test MLR method in iles' manner. We leave column out as a small sample. We selected percentage of 10%, 15% and 20% of entire dataset as our 'observed' data separately and the remainder classed as 'unobserved'. Then we randomly generated tag $k = 1, 2, .., 8$ and 10 or 30 were selected from 'observed' to calculate R^2 to find the maximal average R^2 in 'observed' region and 'unobserved' region. In such way, the average R^2 between tags and nonTag SNPs is maximum.Further, we use the tags selected from the leave-many-out sample

to test how good these tags can statically cover the entire sample. As result, in Daly data, we use one tag can reach R^2 27% in its selected sample data, this one tag can cover the 15% of entire data. T represents tags, V represents visible SNPs and C represents the entire chrosome.In future, long haplotypes will be our aim. We will take a large dataset, such as $80K$ to test.

The length of the region simulated makes little difference to the accuracy of the results-what is important is the number of observed SNPs from which the tSNPs are selected. As the number of observed SNPs increases, so the estimates become more accurate.

We used four datasets to test our result.We leave-many-out as a small sample,then we select Tag 1..K , in such way, the average R2 between tags and nonTag SNPs is maximum. Further, we use the tags selected from the leave-many-out sample to test how good these tags can statically cover the entire sample. As the number of observed SNPs increases, so the estimates become more accurate. When there are more than 10 tag SNPs is observed, the relationship between two SNPs is more closed. In ENr113 dataset, we selected percentage as 10get R2 as below Fig1. From testing on different dataset, we found the average R2 in observed small region is bigger than that in entire region. Using 10 tags with average R2 0.75 in visible sample can cover average R2 0.63 in the entire sample, 20 tags with average R2 0.95 can reach 0.85 in invisible region, 30 tags is for 0.91.

5.3 Discussion

An example of this for a region of different length with 8 SNPs 'observed' is shown in figure1. Here the average R2 captured at the unobserved SNPs by the tSNPs selected in the leave-many-out process is 0.949, the average estimated by the leave-many-out method is 0.833 while the average estimated by using only the observed SNPs is 0.950. It can be seen from the line chart of figure 2 that as the number of

Table 5.1. Tags for Daly data

tags	prec = 10		prec = 15		prec = 20	
	S Set	E Set	S Set	E Set	S Set	E Set
1	0.203	0.194	0.182	0.173	0.195	0.192
2	0.387	0.368	0.322	0.304	0.376	0.349
3	0.500	0.446	0.394	0.366	0.451	0.415
4	0.560	0.499	0.472	0.502	0.515	0.501
5	0.599	0.522	0.534	0.548	0.569	0.543
6	0.632	0.540	0.597	0.592	0.608	0.592
7	0.663	0.567	0.634	0.623	0.643	0.618
8	0.701	0.590	0.668	0.657	0.674	0.653

unobserved markers increases, so the apparent accuracy of the estimate increases. We provide the result using the other dataset as well. See Fig3, Fig4, Fig5.

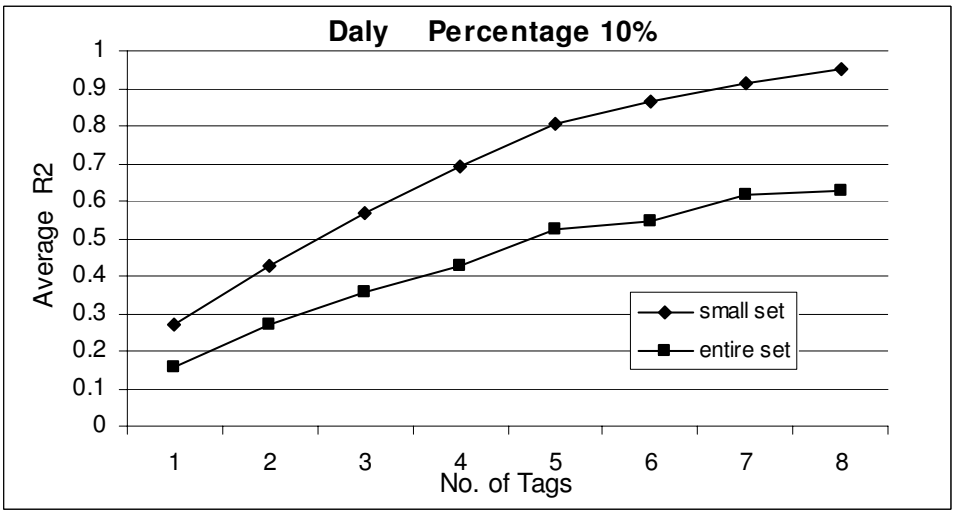

Figure 5.1. Tag are selected from 10% of entire data. Average R2 over 1..to 8 tags on Daly Data.

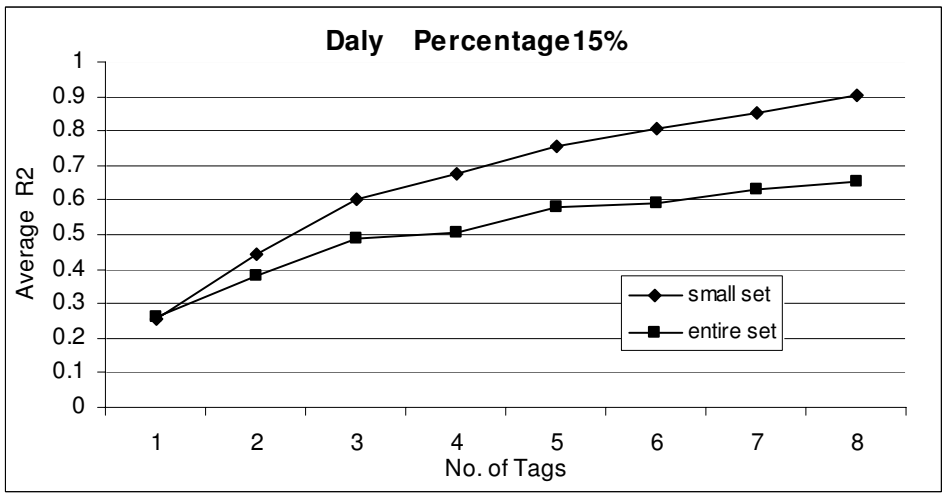

Figure 5.2. Tag are selected from 15% of entire data. Average R2 over 1..to 8 tags on Daly Data.

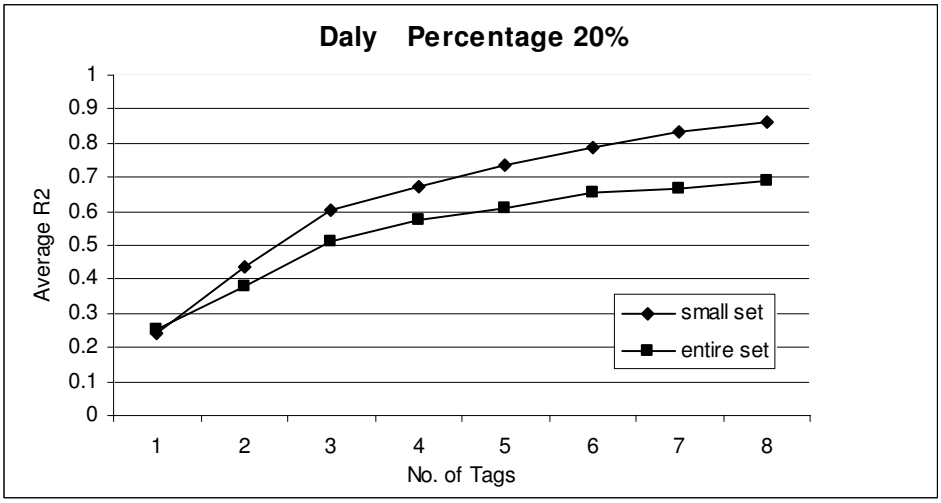

Figure 5.3. Tag are selected from 20% of entire data. Average R2 over 1..to 8 tags on Daly Data.

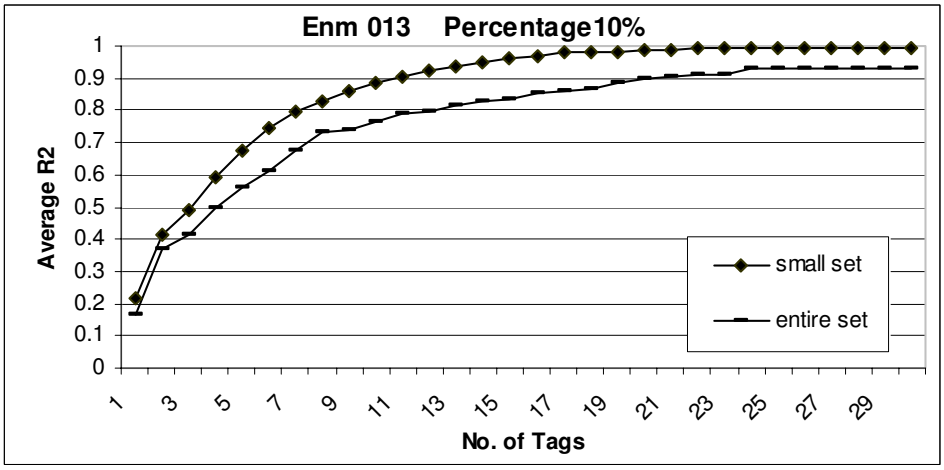

Figure 5.4. Tag are selected from 10% of entire data. Average R2 over 1..to 8 tags on Enm 013.

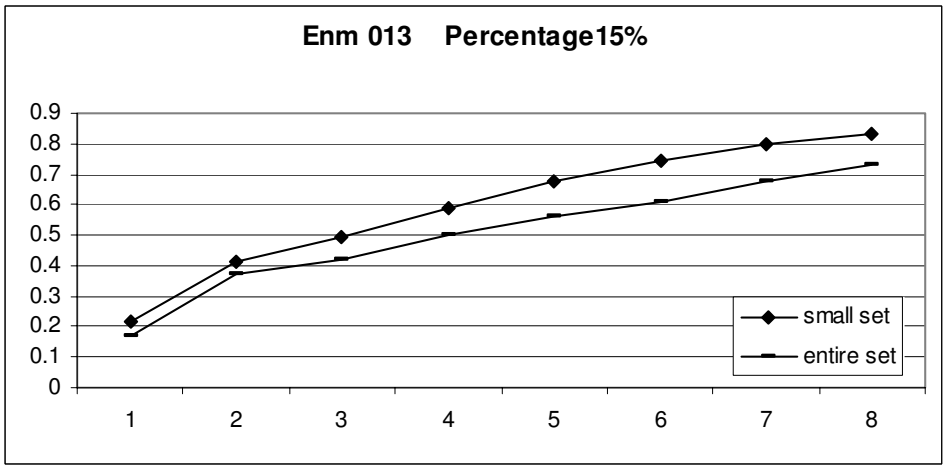

Figure 5.5. Tag are selected from 15% of entire data. Average R2 over 1..to 8 tags on Enm 013

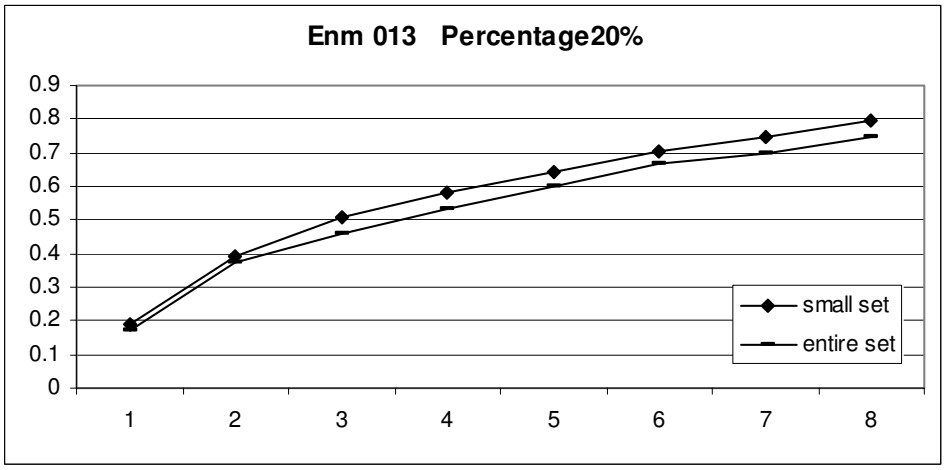

Figure 5.6. Tag are selected from 20% of entire data. Average R2 over 1..to 8 tags on Enm 013

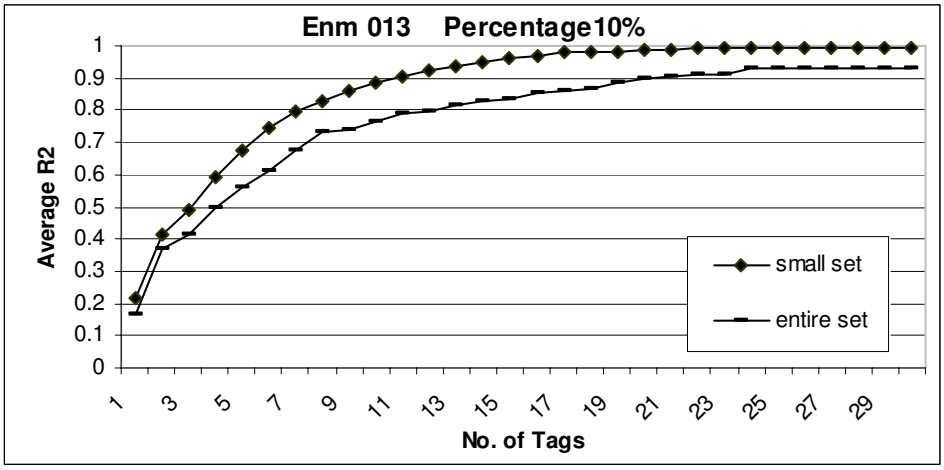

Figure 5.7. Tag are selected from 10% of entire data. Average R2 over 1..to 8 tags on ENr 112

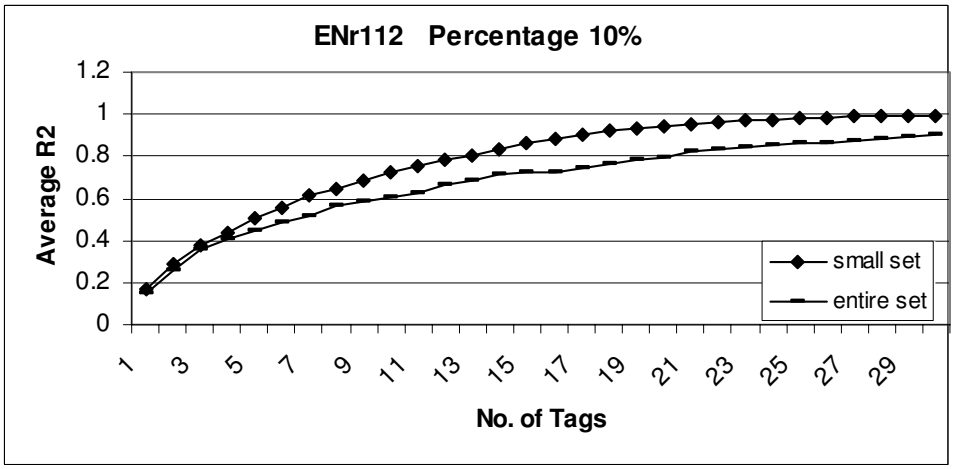

Figure 5.8. Tag are selected from 15% of entire data. Average R2 over 1..to 8 tags on ENr 112

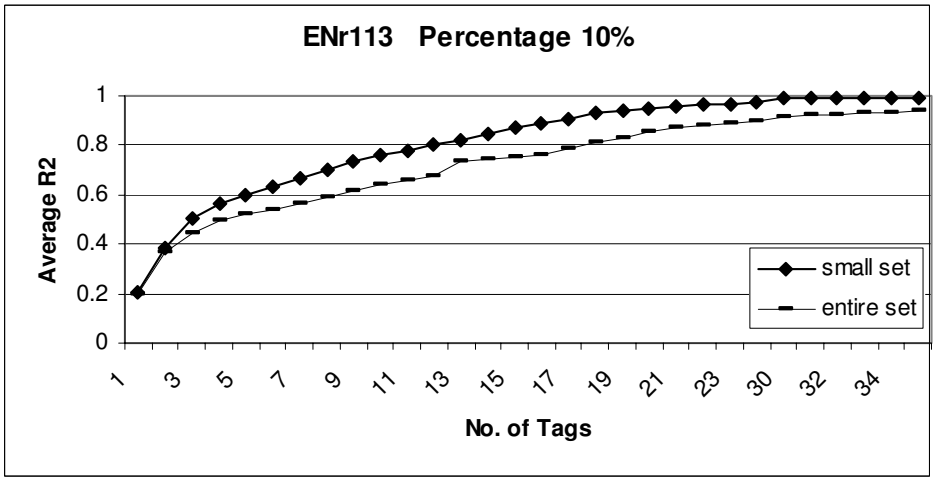

Figure 5.9. Tag are selected from 20% of entire data. Average R2 over 1..to 8 tags on ENr 112

BIBLIOGRAPHY

[1] Affymetrix (2005) *http://www.affymetrix.com/products/arrays/.*

[2] International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796, *http://www.hapmap.org.*

[3] Ackerman, H., Usen, S., Mott, R., Richardson, A., Sisay-Joof, F., Katundu, P., Taylor, T., Ward, R., Molyneux, M., Pinder, M., Kwiatkowski, D.P. (2003) 'Haplotypic analysis of the TNF locus by association efficiency and entropy', *Genome Biology*, Vol.4, pp. 24

[4] Avi-Itzhak, H.I., Su, X. and de la Vega, F.M. (2003) 'Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity', *Proceedings of Pacific Symposium on Biocomputing*, Vol. 8, pp. 466–477.

[5] Bafna, V., Halldorsson, B.V., Schwartz, R.S., Clark, A.G. and Istrail, S. (2003) 'Haplotypes and informative SNP selection algorithms: don't block out information', *Proceedings of the Seventh International Conference on Research in Computational Molecular Biology*, pp. 19–27.

[6] Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) 'Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium', *American Journal of Human Genetics*, Vol. 74, No. 1, pp. 106–120.

[7] Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). 'Detecting disease associations due to linkage disequilibrium using haplotype tags: a class

of tests and the determinants of statistical power', Human Heredity, Vol. 56, pp. 18–31.

[8] Clark, A., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. vol. 63 pp. 595–612, 1998.

[9] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) 'High resolution haplotype structure in the human genome', *Nature Genetics*, Vol. 29, pp. 229–232.

[10] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. and Altshuler, D. (2002) 'The structure of haplotype blocks in the human genome', *Science*, Vol. 296, pp. 2225–2229.

[11] Forton, J., Kwiatkowshi, D., Rockett, K., Luoni, G., Kimber, M. and Hull, J. (2005) 'Accuracy of Haplotype Reconstruction from Haplotype-Tagging Single-Nucleotide Polymorphisms', *American Journal of Human Genetics* Vol. 76, pp 438–448.

[12] Kimmel, G., and Shamir R.(2004). 'GERBIL: Genotype resolution and block identification using likelihood', *PNAS*, Vol. 102, pp 158–162.

[13] Halperin, E. and Eskin, E. Haplotype reconstruction from genotype data using imperfect phylogeny. Bioinformatics. Advance Access published on February 26, 2004.

[14] Halperin, E. and Karp, R. M. On the greedy set cover algorithm. In preperation, 2003.

[15] Halperin, E. and Karp, R. M. Perfect phylogeny and haplotype assignment. RECOMB, 2004.

[16] Halperin, E., Kimmel, G. and Shamir, R. (2005) 'Tag SNP Selection in Genotype Data for Maximizing SNP Prediciton Accuracy', *Bioinformatics*, Vol. 21, pp. 195-203.

[17] Halldorsson, B.V., Bafna, V., Lippert, R., Schwartz, R., de la Vega, F.M., Clark, A.G. and Istrail, S. (2004) 'Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies', *Genome Research* Vol. 14, pp. 1633–1640.

[18] He, J. and Zelikovsky, A. (2004) 'Linear Reduction Methods for Tag SNP Selection', *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology (EMBC'04)*, pp. 2840–2843.

[19] He, J. and Zelikovsky, A. (2004) 'Linear Reduction for Haplotype Inference', , *Proceedings of the Workshop on Algorithms in Bioinformatics (WABI'04)*, Vol. 3240, pp. 242–253.

[20] He, J. and Zelikovsky, A. (2005) 'Linear Reduction Method for Predictive and Informative Tag SNP Selection', , *International Journal Bioinformtics Research and Applications*, Vol 3, pp. 249–260.

[21] J. He, J. Zhang, G. Altun, A. Zelikovsky and Y. Zhang, Haplotype Tagging using Support Vector Machines, Proc. IEEE Intl Conf on Granular Computing (GRC 2006), May 2006, pp. 758-761.

[22] He, J. and Zelikovsky, A. (2006) "Tag SNP Selection Based on multiple Linear Regression," Proc. of Intl Conf on Computational Science (ICCS 2006), May 2006, LNCS 3992, pp. 750-757

[23] He, J. and Zelikovsky, A. (2006) "Haplotype Tagging based on SVM SNP Prediction," Proc. IEEE Intl Conf on Granular Computing (GRC 2006), May 2006, pp. 758-761

[24] He, J. and Zelikovsky, A. (2006) "MLS-tagging: Genotype Tagging Based on Multivariate Linear Regression," Application notes, Bioinformatics, to appear

[25] He, J. and Zelikovsky, A. (2006) "Multiple Linear Regression for Index SNP Selection on Unphased Genotypes," Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'06), September 2006, to appear.

[26] Lee, P.H. and Shatkay, H (2006) 'BNTagger: Improved Tagging SNP Selection using Bayesian Networks', *Proceeding of ISMB2006, in manuscript.*

[27] Brinza, D., He, J. and Zelikovsky, A. "Combinatorial Search Methods for Multi-SNP Disease Association," Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'06), September 2006, to appear.

[28] The effect of SNP marker density on the efficacy of haplotype tagging SNPs–a warning. (2005) Ann Hum Genet. 2005 Mar;69(Pt 2):209-15

[29] Obtaining Unbiased Estimates of Tagging SNP Performance, (2005) Annals of Human Genetics vol. 69, page 1–8.

[30] Judson, R., Salisbury, B., Schneider, J., Windemuth, A. and Stephens, J.C. (2002) 'How many SNPs does a genome-wide haplotype map require?', *Pharmacogenomics*, Vol. 3, pp. 379–391.

[31] Ke, X. and Cardon, LR. (2003) 'Efficient selective screening of haplotype tag SNPs', *Bioinformatics*, Vol. 170, pp. 287-288.

[32] Merikangas, KR., Risch, N. (2003) 'Will the genomics revolution revolutionize psychiatry', *The American Journal of Psychiatry*, 160:625-635.

[33] Pasaniuc, B. and Mandoiu, I. Highly Scalable Genotype Phasing by Entropy Minimization, submitted to EMBC06.

[34] Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S. and Cox, D. (2001) 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome', *Science*, Vol. 294, pp. 1719–1723.

[35] Sebastiani, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., and Ramoni, M. (2003) 'Minimal haplotype tagging', *Proceedings of the National Academy of Sciences*, Vol. 100, pp. 9900–9905.

[36] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag.

[37] Stram, D., Haiman, C., Hirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B. and Pike, M. (2003). 'Choosing haplotype-tagging SNPs based on unphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study', *Human Heredity*, Vol. 55, pp. 27–36.

[38] Y.C. Tang, Y.-Q. Zhang and Z. Huang, Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007.

[39] Taylor, B. and Zhulin, I. "In search of higher energy: metabolism -dependent behavior in bacteria," Molecular Microbiology, vol. 28, pp. 683-690, 1998.

[40] Thornberry, N. A., Rano, T. A., Peterson, E. P., Rasper, D. M., Timkey, T., Garcia-Calvo, M., Houtzager, V. M., Nordstrom, P. A., Roy, S., Vaillancourt,

J. P., Chapman, K. T. and Nicholson, D. W. (1997). A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. J Biol Chem 272, 17907-11.

[41] Vapnik, V. and Cortes, C. , "Support Vector Networks",*Machine Learning*, vol. 20, pp. 273–293, 1995.

[42] Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551C565

[43] Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., and Sun, F. (2004) 'Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies', *Genome Research*, Vol. 14, pp. 908–916.

[44] Zhao, H., Pfiffer, R. and Gail, MH. (2003) 'Haplotype analysis in population genetics and association studies', *Phamacogenomics*, 4:171-178.

[Zhang P.*et al.*, 2004] Zhang P., Sheng H. and Uehara R. (2004) A double classification tree search algorithm for index SNP selection, *BMC Bioinformatics*, Vol. 5, pp. 89–95