Two Birds, One Stone: Selecting Functionally Informative Tag SNPs for Disease Association Studies

Phil Hyoun Lee and Hagit Shatkay

Computational Biology and Machine Learning Lab
School of Computing, Queen's University
Kingston, ON, Canada
{lee, shatkay}@cs.queensu.ca

Abstract. Selecting an informative subset of SNPs, generally referred to as tag SNPs, to genotype and analyze is considered to be an essential step toward effective disease association studies. However, while the selected informative tag SNPs may characterize the allele information of a target genomic region, they are not necessarily the ones directly associated with disease or with functional impairment. To address this limitation, we present a first integrative SNP selection system that simultaneously identifies SNPs that are both informative and carry a deleterious functional effect – which in turn means that they are likely to be directly associated with disease. We formulate the problem of selecting functionally informative tag SNPs as a multi-objective optimization problem and present a heuristic algorithm for addressing it. We also present the system we developed for assessing the functional significance of SNPs. To evaluate our system, we compare it to other state-of-the-art SNP selection systems, which conduct both information-based tag SNP selection and function-based SNP selection, but do so in two separate consecutive steps. Using 14 datasets, based on disease-related genes curated by the OMIM database, we show that our system consistently improves upon current systems.

1 Introduction

Identifying *single nucleotide polymorphisms*[1] (SNPs) that are involved in complex common diseases, such as cancer, is a major challenge in current molecular epidemiology. Due to their genome-wide prevalence, knowledge of such SNPs is expected to be essential for unraveling the genetic etiology of human diseases, and thus, for enabling timely diagnosis, treatment, and, ultimately, prevention of disease. However, genotyping[2] and analyzing all the SNPs on the human genome [2] is practically infeasible as the number of SNPs is estimated at over ten million [3]. Thus, selecting a subset of SNPs that is sufficiently informative to conduct disease-gene association but still small enough to reduce the genotyping and analysis overhead, a process known as *tag SNP selection*, is a key step toward effective association studies.

[1] A single nucleotide polymorphism (SNP) is the substitution of a single nucleotide at a certain position on the genome [1].

[2] Genotyping is the biomolecular process of identifying the nucleotide of a genetic variation [1].

A variety of measures and algorithms have been proposed for tag SNP selection, and their utility has been empirically demonstrated by simulation studies or by association studies. Yet, while the selected informative tag SNPs may effectively characterize the allele information of a target genomic region, they are not necessarily the ones directly associated with disease or with functional impairment. Given this limitation, SNPs with deleterious functional effects have drawn recent attention [4,5]. Typically, SNPs occurring in functional genomic regions are more likely to cause functional distortion and, as such, more likely to underly disease-causing variations [2,6]. As of yet, methods for the selection of informative tag SNPs do not take into account the functional significance of SNPs; similarly, methods for identifying disease-related SNPs do not attempt to capture the allele information of the complete target locus[3].

The identification of informative tag SNPs and of functionally significant SNPs can be viewed as two distinct optimization problems with possibly conflicting objectives. Consequently, current systems that try to support both information-based tag SNP selection and function-based SNP selection [7, 8] address each selection problem independently. That is, they separately conduct tag SNP selection and function-based SNP selection, and combine the two selected sets as a last step. A major shortcoming of such systems is that the number of selected SNPs can be much larger than necessary. Moreover, the functional SNPs selected may not be predictive of the other SNPs in the locus, while the predictive SNPs selected may have no relation to disease.

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

In Sec. 2, we formulate the problem of functionally informative tag SNP selection, and introduce the basic notations that are used throughout the paper. Section 3 describes our functional-significance assessment process and our heuristic algorithm for selecting functionally informative SNPs. Section 4 reports the results from a comparative study. Section 5 summarizes our findings and outlines future directions.

2 Functionally Informative Tag SNP Selection

We are concerned with identifying a set of SNPs associated with a given disease. The relevant target locus on the genome can be as large as a whole chromosome or as small as a part of a gene. Disease association studies typically involve the following steps: 1) chromosome samples are obtained from *cases* bearing the disease and from *controls* (people not bearing the disease); 2) The allele information for all the SNPs on the target

[3] A locus is the chromosomal location of the target region for biomolecular experiments [1].

locus is obtained (*genotyped*) from the chromosome samples; 3) a subset of SNPs that is most associated with the disease phenotype[4] is identified. However, in practice, due to experimental cost and time, not all the SNPs on the target locus can be genotyped or analyzed. We thus need to select a subset of at most k SNPs on the target locus (where k is a pre-specified number) whose allele information is as informative as that of the whole set of SNPs, while including those SNPs that are most functionally significant. We refer to the problem as *functionally informative tag SNP selection*. Before we formulate and address this problem, we introduce here the basic notations that are used throughout this paper.

Suppose that our target locus contains p consecutive SNPs. Each SNP can be represented as a discrete random variable, X_j $(j = 1, ..., p)$, whose possible values are the 4 nucleotides, $\{a, g, c, t\}$. For each value $x \in \{a, g, c, t\}$, there is a probability $Pr(X_j = x)$ that X_j is assigned the nucleotide x. Let V=$\{X_1, ..., X_p\}$ denote the set of random variables corresponding to the p SNPs. We are given a haplotype[5] dataset, D, containing the allele information of n haplotypes, each of which consists of the p SNPs in V. The set D can be viewed as an n by p matrix; each row, D_{i+}, in D corresponds to the allele information of the p SNPs comprising haplotype h_i, while each column, D_{+j}, corresponds to the allele information of SNP X_j in each of the n haplotypes. We denote by D_{ij} the allele information of the j^{th} SNP in the i^{th} haplotype. To formally address functional significance of SNPs, we denote by e_j the functional significance score for each SNP X_j in V, and define $E = \{e_1, ..., e_p\}$ to be the set of scores for all the SNPs. We further discuss how these values can be obtained in Sec. 3.1.

For a subset of SNPs, $T \subset V$, we define an objective function, $f(T|D, E)$, to reflect both the allele information carried by the SNPs in T about the remaining SNPs in $V - T$, and the functional significance of the SNPs in T. The problem of *functionally informative tag SNP selection* can then be stated as follows:

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

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

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

Definition 1. ***Information-based Objective.*** *Given a set of k SNPs, $T = \{X_{t_1}, ..., X_{t_k}\}$, and a dataset D of n haplotypes, we define an information-based objective function, $f_1(T|D)$, as:*

[4] A phenotype is the physical, observed manifestation of a genetic trait [1].
[5] A haplotype is a set of consecutive SNPs present on one chromosome [1].

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

where

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

The function I returns 1 if the allele of the j^{th} SNP in the i^{th} haplotype (i.e., D_{ij}) is correctly predicted based on the allele information of the SNPs in T. We note that, by using the conditional probability expression, the allele of D_{ij} is predicted as the one that is most likely to occur given the allele information of predictive tag SNPs in T [6]. Otherwise, the function I returns 0. To summarize, the allelic information provided by a SNP set, T, with respect to a given haplotype dataset D, is measured by the average proportion of the correctly predicted alleles of each SNP, X_j, given the allele information of the SNPs in T.

This information-based objective function, $f_1(T|D)$, was introduced in our previous work [9], and is based on the *prediction-based tag SNP selection approach* [10, 11], which aims to select a subset of SNPs (i.e., tag SNPs) that can best predict the alleles of the remaining, unselected SNPs (i.e., tagged SNPs). This approach is appealing since: (1) it does not require prior block partitioning [12]; (2) it tends to select a small number of SNPs [13]; and (3)it works well even for genomic regions with low linkage disequilibrium[7] [9]. An in-depth discussion and survey of information-based tag SNP selection approaches is given elsewhere [14, 15].

Definition 2. ***Function-based Objective.*** *Given a set of k SNPs, $T \subset V$, and a set of functional significance scores, $E = \{e_1, ..., e_p\}$, we define a function-based objective function, $f_2(T|E)$ as:*

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

where

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

In other words, the functional significance of a SNP set T is the normalized sum of the functional significance of SNPs in T. We note that, for the vast majority of SNPs, no experimental evidence is yet available to substantiate their functional significance [2]. We thus define and evaluate the functional significance of SNPs using a large variety of bioinformatics tools for function-assessment. The details of our assessment procedure are described in Sec. 3.1.

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

[6] Note that for any SNP $X_{t_l} \in T$, $I(X_{t_l}, T, D_{i+})$ is by definition always 1.

[7] Linkage disequilibrium (LD) refers to the non-random association of SNPs [1].

Definition 3. ***Functionally Informative Objective Function.*** *Given a set of k SNPs, $T \subset V$, a haplotype dataset, D, a functional significance score set, $E = \{e_1, ..., e_p\}$, and a parameter value, α $(0 \leq \alpha \leq 1)$, we define a functionally informative (FI) objective function, $f(T|D, E)$ as:*

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

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

To summarize, we are looking for a subset of at most k SNPs, T, that is both functionally significant and likely to correctly predict the remaining SNPs in $V - T$. Bafna et al. [12] have previously shown that finding k most informative tag SNPs is NP-hard. Based on this, we take it as a conjecture that the current problem is also NP-hard, (the proof is beyond the scope of this paper). The next section introduces a function-assessment process and a heuristic algorithm to address the problem.

3 Models and Algorithms

Our SNP selection system involves two main steps: 1) assessing the functional significance, e_j, of SNPs and 2) selecting a set of functionally informative tag SNPs, T. These are described next.

3.1 Assessing the Functional Significance of SNPs

Using a variety of existing, publicly available bioinformatics tools, we examine the deleterious effects of SNPs on the molecular function of their genomic region. In particular, we focus on the following three major categories of biological function:

- *Protein Coding*: SNPs in protein coding regions may cause an amino acid substitution (i.e., a *missense* mutation) or interfere with protein translation (i.e., a *nonsense* mutation).
- *Splicing Regulation*: SNPs in splicing regulatory regions may affect alternative splicing or result in exon skipping or intron retention.
- *Transcriptional Regulation*: SNPs in transcription regulatory regions (e.g., transcription factor binding sites, CpG islands, regulatory RNAs) can alter the affinity of the binding sites, and disrupt proper gene regulation.

We assess the functional significance of SNPs based on their location and possible deleterious effects along these three functional categories. Figure 1 illustrates the following assessment process:

For each of the three categories, a SNP is separately assigned into one of three classes[8]: Class 1 indicates *irrelevance* to the biological function; Class 2 indicates that the SNP is *relevant* to the biological function, but predicted to be benign or has no evidence of deleterious effects; Class 3 indicates that the SNP is likely to be *deleterious*.

[8] Thus, a SNP is assigned three class labels; one label for each of the three functional categories.

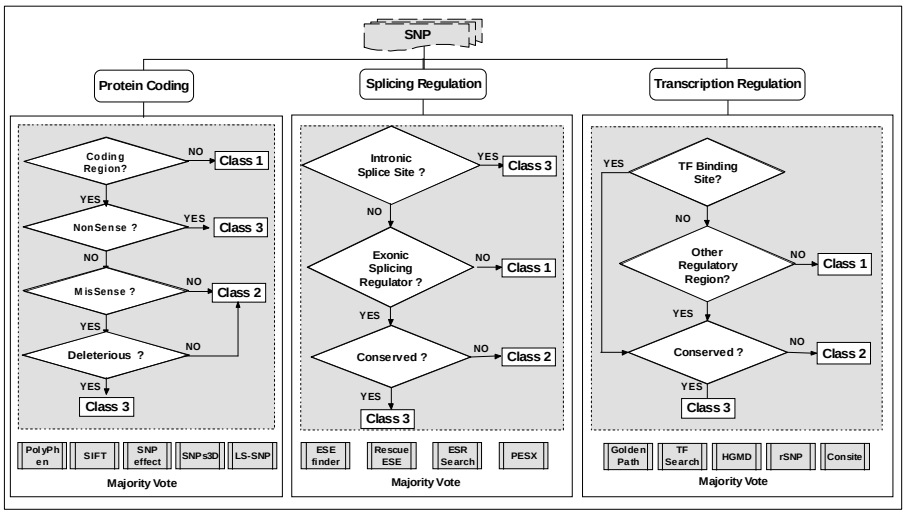


Fig. 1. Our functional significance assessment system.

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

To make a robust assessment, we use multiple bioinformatics tools that are based on different data, algorithms, or theory for examining each biological functional category. The tools, PolyPhen [16], SIFT [17], SNPeffect [18], SNPs3D [19], and LS-SNP [20] are used to examine missense SNPs; ESEfinder [21], RescueESE [22], ESRSearch [23], and PESX [24] are used to identify the SNPs in exonic splice regions; TFSearch [25] and Consite [26] are used to identify transcriptional regulatory SNPs in promoter regions; Ensembl [27], GoldenPath [28], and HGMD [29] databases are used to identify SNPs in other transcriptional regulatory regions (e.g., microRNA); and Ensembl [27] database is used to identify nonsense SNPs and the SNPs in intronic splicing sites.

The classes assigned to each SNP, with respect to each functional category are decided by a majority vote of the integrated tools in the category. As a result, three class labels are assigned to each SNP, one for each of the three categories of biological function. To assign a single functional significance value to each SNP, we follow Bhatti et al. [2], and assign the highest class tag along all three categories as the functional significance score, e_j, for the SNP X_j. For example, SNP rs4963 on gene ADD1 is assigned

to *Class 3* with respect to Protein Coding, *Class 1* with respect to Splicing Regulation, and *Class 1* with respect to Transcription Regulation. The functional significance score of SNP rs4963 is thus *3* because it is highly significant for the protein coding function.

3.2 Selecting Functionally Informative Tag SNPs

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

We first explain the basis for our greedy incremental selection process. Let $T^{(m)}$ denote the set of m selected SNPs after the m^{th} iteration, where $m = 0, ..., k$ and $T^{(0)} = \emptyset$. The FI-score of $T^{(m)}$ was defined in Def. 3 as follows:

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

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

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

and

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

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

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

Our algorithm is outlined in Fig. 2. It starts with an empty set of tag SNPs, T, and computes the FI-score of each SNP with respect to the current set T. We note that although no SNP is currently selected, our algorithm can still predict the allele information of SNPs, and can thus lead to a different FI-score for each SNP. The reasoning is that in this initial case where T is empty, the posterior probability, $Pr(X_j|T)$, shown in

Input: a set of SNPs, V; a maximum number of SNPs to select, k;
a haplotype dataset, D; a set of functional significance scores, E;
Output: a set of tag SNPs, T.

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

Fig. 2. The incremental, greedy algorithm for selecting functionally informative tag SNPs.

the definition of the function I within Def. 1, is simply the prior probability, $Pr(X_j)$. That is, we always predict the alleles of X_j, $D_{ij}(i = 1, ..., n)$, as the major allele of the SNP. This approach is taken because it maximizes the expected prediction accuracy when no other information is given. At each subsequent iteration, the SNP that leads to the maximum increase in the FI-score is selected and added to T. The FI-score for each SNP is updated based on the augmented set T and used in the next iteration. This procedure is repeated until the set T contains the pre-specified number of SNPs, k.

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

4 Experiments and Results

4.1 Experimental Setting

For evaluation, we have selected 14 genes that are involved in the etiology of common and complex diseases according to the OMIM database [30] and have disease-related SNPs identified and recorded by the HapMap Project [31]. To identify the candidate genes, we scanned the OMIM database for several major common and complex diseases, including diabetes, cancer, hypertension, and heart disease. The retrieved genes were then scanned to find those that have SNPs with possible deleterious functional effects reported in the biomedical literature and also have haplotype information available from the HapMap consortium [31]. From the genes satisfying these criteria, 14

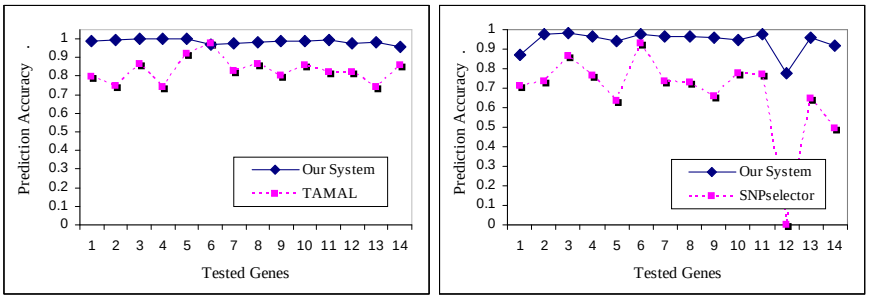
were selected at random. Table 1 provides the genetic characteristics of the 14 genes and their associated disease. The haplotype datasets of the 14 genes were downloaded from the HapMap project website [31]; The genomic location of each gene, including a $10k$ promoter region, was used to download the phased haplotype data (HapMap public release #20/phaseII) for the CEU population.

Table 1. Summary of 14 test datasets. Linkage disequilibrium (LD) is estimated by the multi-allelic extension of Lewontin's LD, D' [32]. The number of SNPs selected by TAMAL and by SNPSelector are shown in the right column.

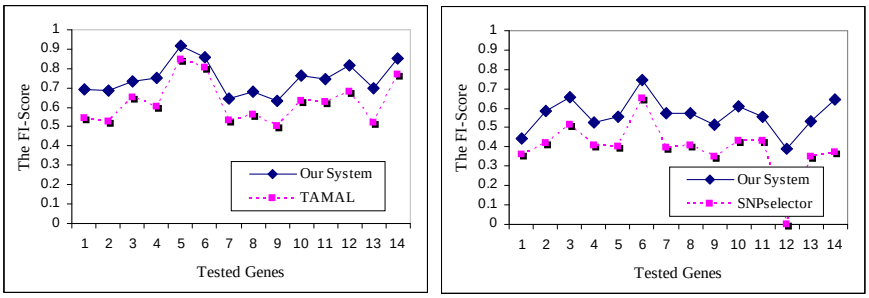
Gene	Target Disease	Locus	LD (D')	Total # of SNPs	# of Selected SNPs	
					TAMAL	SNPSelector
ADD1	Hypertension	4p16.3	0.7718	60	16	1
BRCA2	Breast Cancer	13q12.3	0.7657	106	28	13
CMA1	Hypertension	14q11.2	0.8361	20	6	4
ELAC2	Prostate Cancer	17p11	0.8336	35	13	2
ERBB2	Prostate Cancer	17q21.1	0.8104	8	6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

We compare our system with two state-of-the-art SNP selection systems that support both tag SNP selection and function-based SNP selection: TAMAL [7] and SNPselector [8]. The two systems share the same goal with our system, namely, selecting a set of tag SNPs, with significant functional effects on the molecular function of the genes, for association studies. However, they differ from our system in the assessment process for the functional significance of SNPs, the integrated bioinformatics tools, and the criteria used for selecting SNPs. Moreover, they conduct tag SNP selection and function-based SNP selection in two separate consecutive steps, while we address it as a single optimization problem.

As evaluation measures, we use Halperin's prediction accuracy [11] and the FI-score, introduced in Def. 3, (we note that the two systems to which we compare do not provide an evaluation measure). To compare the performance of the systems using the two measures, the SNP sets selected by each of the compared systems must include an equal number of SNPs. However, unlike our system, TAMAL and SNPselector do not allow the user to specify the number of selected SNPs, but rather calculate a subset of SNPs and provide it as their output. Thus, when they do not select the same number of SNPs for the same gene, they cannot be directly compared. Hence, for a fair compari-



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



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

Fig. 3. The performance of our system and the compared systems for 14 gene datasets.

son, we first apply each of the compared systems to each of 14 test datasets, and then use our system on the same dataset to select the same number of SNPs as selected by the compared system. We then compute the two evaluation measures for the sets selected by each of the systems, and compare the resulting scores. The number of SNPs selected by TAMAL and SNPselector for the 14 tested genes is shown in Table 1. To ensure robustness of the results obtained from our system, we employ 10-fold cross validation 10 times, each using a randomized 10-way split of the n haplotypes. In all cases, the average performance is used in the comparison.

4.2 Results

Figure 3 shows the performance of our system compared with TAMAL (left) and with SNPselector (right). The X-axis represents the 14 genes in an alphabetical order of their names, as listed in Table 1. In Fig. 3(a) (top), the Y-axis shows Halperin's prediction accuracy [11], and in Fig. 3(b) the Y-axis shows the FI-score for the selected SNP set of the corresponding gene. Our system (upper solid line with diamonds) consistently outperforms the other two systems, TAMAL and SNPselector (lower dotted line with rectangles) on both evaluation measures. The performance difference in all cases is statistically significant, as confirmed by the Wilcoxon rank-sum test (p-values are 1.144e-005 and 4.7e-003 with respect to the TAMAL system and 1.7382e-005 and 5.6780e-004

with respect to the SNPselector system.). We note that optimizing the FI-score when selecting SNPs does not compromise the predictive power of the SNPs selected by our system, that is, our selected SNPs still have a high prediction accuracy according to Halperin's original measure as demonstrated by Fig. 3(a).

5 Conclusions

We have presented a first integrative SNP selection system that simultaneously identifies SNPs that are both highly informative in terms of providing allele information for the target locus, and are of high functional significance. Our main contributions include the formulation of the problem of functionally informative tag SNP selection as a multi-objective optimization problem, presenting a heuristic selection algorithm to address the problem, and proposing an assessment process for scoring the functional significance of SNPs. An empirical study over a set of 14 disease-associated genes shows that our system indeed improves upon current state-of-the-art systems. In the near future we plan to apply a general computational approach, such as goal programming [33], for addressing the multi-objective optimization problem of selecting functionally informative tag SNPs. We also plan to apply a probabilistic approach to assess the functional significance of SNPs.

References

1. Hedrick, P.: Genetics of pouplation. Jones and Bartlett Publishers, 3rd Edition (2004)
2. Bhatti, P., Church, D., Rutter, J.L., Struewing, J.P., Sigurdson, A.J.: Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. American Journal of Epidemiology **164** (2006) 794–804
3. Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic Acids Research **29** (2001) 308–311
4. Brunham, L.R., Singaraja, R.R., Pape, T.D., Kejariwai, A., Thomas, P.D., Hayden, M.R.: Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLOS Genetics **1** (2005) 739–747
5. Rebbeck, T.R., Ambrosone, C.B., Bell, D.A., Chanock, S.J., Hayes, R.B., Kadlubar, F.F., Thomas, D.C.: SNPs, haplotypes, and cancer: applications in molecular epidemiology. Cancer Epidemiology, Biomarkers & Prevention **13** (2004) 681–687
6. Conde, L., Vaquerizas, J.M., Ferrer-Costa, C., de la Cruz, X., Orozco, M., Dopazo1, J.: PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. American Journal of Epidemiology **33** (2005) W501–W505
7. Hemminger, B.M., Saelim, B., Sullivan, P.F.: TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. Bioinformatics **22** (2006) 626–627
8. Xu, H., Gregory, S.G., Hauser, E.R., Stenger, J.E., Pericak-Vance, M.A., Vance, J.M., Zuchner, S., Hauser, M.A.: SNPselector: a web tool for selecting SNPs for genetic association studies . Bioinformatics **21** (2005) 4181–4186
9. Lee, P.H., Shatkay, H.: BNTagger: improved tagging SNP selection using Bayesian networks. Bioinformatics **22** (2006) e211–219
10. Sebastiani, P., Lazarus, R., Weiss, S.T., Kunkel, L.M., Kohane, I.S., Ramoni, M.F.: Minimal haplotype tagging. Proceedings of the National Academy of Sciences **100** (2003) 9900–9905

11. Halperin, E., Kimmel, G., Sharmir, R.: Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics **21** (2005) i195–i203
12. Bafna, V., Halldorsson, B.V., Schwartz, R., Clark, A.G., Istrail, S.: Haplotypes and Informative SNP Selection Algorithms: Don't Block Out Information. In Proceedings of the 7th International Conference on Computational Molecular Biology (2003) 19–26
13. Bakker, P.D., Graham, R.R., Altshuler, D., Henderson, B., Haiman, C.: Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple population. In Proceedings of Pacific Symposium on Biocomputing (2006)
14. Halldorsson, B.V., Istrail, S., Vega, F.D.L.: Optimal selection of SNP markers for disease association studies. American Journal of Epidemiology **58(3-4)** (2004) 190–202
15. Lee, P.H.: Computational haplotype analysis: An overview of computational methods in genetic variation study. Technical Report 2006-512, Queen's University, Kingston, ON, Canada (2006) WEB URL: http://www.cs.queensu.ca/TechReports/Reports/2006-512.pdf.
16. Ramensky, V., Sunyaev, S.: Human non-synonymous SNPs: surver and survey. Nucleic Acid Research **30** (2002) 3894–3900
17. Ng, P., Henikoff, S.: Predicting deleterious amino acid substitutions. Genome Research **11** (2001) 863–874
18. Reumers, J., Schymkowitz, J., Ferkinghoff-Borg, J., Stricher, F., Serrano, L., Rousseau, F.: SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acid Research. **33** (2005) D527–532
19. Yue, P., Melamud, E., Moult, J.: SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics **7** (2006) 166
20. Karchin, R., et al.: LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics **21** (2005) 2814–2820
21. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., Krainer, A.R.: ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Research **31** (2003) 3568–3571
22. Yeo, G., Burge, C.B.: Variation in sequence and organization of splicing regulatory elements in vertebrate genes. In the Proceeding of Proc. Natl. Acad. Sci. **101(44)** (2004) 15700–15705
23. Fairbrother, W.G., Yeh, R.F., Sharp, P.A., Burge, C.B.: Predictive identification of exonic splicing enhancers in human genes. Science **297** (2002) 1007–1013
24. Zhang, et al.: Exon inclusion is dependent on predictable exonic splicing enhancers. Molecular and Cellular Biology **25(16)** (2005) 7323–7332
25. Akiyama, Y.: TFSEARCH: Searching Transcription Factor Binding Sites. WEB URL: http://www.rwcp.or.jp/papia/ (1998)
26. Sandelin, A., Wasserman, W.W., Lenhard, B.: ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Research **32(Web Server issue)** (2004) W249–252
27. Hubbard, T.J.P., et al.: Ensembl 2007. Nucleic Acids Research (2007; Database issue)
28. Karolchik, D., et al.: The ucsc genome browser database. Nucl. Acids Res **31(1)** (2003) 51–54
29. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., Cooper, D.N.: Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mrna splicing. Human Mutation **28(2)** (2007) 150–158
30. McKusick-Nathans Institute of Genetic Medicine, J.H.U., National Center for Biotechnology Information, N.L.o.M.: Online Mendelian Inheritance in Man, OMIM (TM). (WEB URL: http://www.ncbi.nlm.nih.gov/omim/)
31. The International HapMap Consortium, .: The International HapMap Project. Nature **426** (2003) 789–796
32. Hedrick, P.: Gametic disequilibrium measures: proceed with caution. Genetics **117** (1987) 331–341
33. Lee, S.M.: Goal programming for decision analysis. Auerback, Philadelphia (1972)