BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 8 2009, pages 1048–1055
doi:10.1093/bioinformatics/btp103

Genetics and population analysis

An integrative scoring system for ranking SNPs by their potential deleterious effects

Phil Hyoun Lee* and Hagit Shatkay

Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON, K7L 3N6, Canada

Received on December 8, 2008; revised on January 21, 2009; accepted on February 17, 2009

Advance Access publication February 19, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Identifying single nucleotide polymorphisms (SNPs) that underlie common and complex human diseases, such as cancer, is of major interest in current molecular epidemiology. Nevertheless, the tremendous number of SNPs on the human genome requires computational methods for prioritizing SNPs according to their potentially deleterious effects to human health, and as such, for expediting genotyping and analysis. As of yet, little has been done to quantitatively assess the possible deleterious effects of SNPs for effective association studies.

Results: We propose a new integrative scoring system for prioritizing SNPs based on their possible deleterious effects within a probabilistic framework. We applied our system to 580 disease-susceptibility genes obtained from the OMIM (Online Mendelian Inheritance in Man) database, which is one of the most widely used databases of human genes and genetic disorders. The scoring results clearly show that the distribution of the functional significance (FS) scores for already known disease-related SNPs is significantly different from that of neutral SNPs. In addition, we summarize distinct features of potentially deleterious SNPs based on their FS score, such as functional genomic regions where they occur or bio-molecular functions that they mainly affect. We also demonstrate, through a comparative study, that our system improves upon other function-assessment systems for SNPs, by assigning significantly higher FS scores to already known disease-related SNPs than to neutral SNPs.

Availability: http://compbio.cs.queensu.ca/F-SNP and http://compbio.cs.queensu.ca/RankingSNPs/default.html.

Contact: lee@cs.queensu.ca

1 INTRODUCTION

Much effort in current epidemiology, medicine and phamarco-genomics is focused on the identification of genetic variations that are involved in common and complex diseases. In particular, single nucleotide polymorphisms (SNPs), which are substitutions of single nucleotides at specific positions on the genome occurring in >1% of the human population, are at the forefront of such studies, as they are the most common form of genetic variations on the genome. Knowledge of disease-causing SNPs is expected to enable early diagnosis, effective treatment and ultimately, prevention of target disease.

Typically, the first step toward identifying causal SNPs for common and complex human diseases, involves association studies. However, due to the tremendous number of SNPs on the human genome, estimated at over 10 million (Sherry *et al.*, 2001), there is a clear need to prioritize SNPs based on their potential deleterious functional effects (Bhatti *et al.*, 2006). For instance, SNPs occurring in functional genomic regions such as protein-coding or regulatory regions are more likely to have deleterious effects, and, as such, more likely to underlie disease. By focusing on a small number of these functionally significant SNPs that are likely to be involved in disease, a substantial amount of genotyping and analysis overhead can be reduced.

To pursue this aim, a variety of web services and public databases have been recently introduced to prioritize SNPs by their putative deleterious effects on major bio-molecular functions [for review, see Rebbeck *et al.* (2004)]. These tools examine the functional category of genomic regions where each SNP occurs, such as exons, splice sites or transcription regulatory sites, and predict the potential corresponding functional effects that the SNP may have, using a variety of machine learning approaches. These computational methods, along with other tools in molecular genetics and epidemiology, are expected to enhance the identification of SNPs underling human diseases (Rebbeck *et al.*, 2004).

Yet, such tools and systems, which prioritize functionally significant SNPs, still suffer from two main limitations. First, most systems examine the putative deleterious effects of SNPs with respect to only a *single* biological function, such as protein-coding or splicing regulation (but not both). Thus, to comprehensively analyze the functional significance (FS) of SNPs, researchers must spend much time and effort to separately apply multiple tools, and interpret/integrate their (often conflicting) predictions.

Second, while many systems classify SNPs into qualitatively distinct groups (e.g. 'deleterious' versus 'neutral'), few systems numerically score or rank SNPs according to their FS. Budget considerations often force researchers to select a limited number of SNPs on the target genomic region for conducting association studies. When the number of putatively deleterious SNPs presented by current tools is larger than this prespecified limit, without additional ranking information, selecting only some of them is not straightforward. As a result, researchers must rely on other resources, such as the published literature, to finalize their decision.

*To whom correspondence should be addressed.

To address these limitations, we propose a new integrative scoring system for ranking SNPs based on their putative deleterious effects. We aim to provide more comprehensive information about the FS of SNPs. We thus assess the deleterious effects of SNPs with respect to four major bio-molecular functional categories: splicing, transcription, translation and post-translational modification. We attempt to overcome the incompleteness and possible false findings of any individual bioinformatics tool by combining the assessment results from multiple independent prediction tools within a probabilistic framework. Most significantly, we assign a specific numerical score to each SNP, representing its putative deleterious effects. Using this score, a limited subset of the most functionally significant SNPs can be ranked and selected.

We applied our system to 112 949 SNPs located on 580 disease-susceptibility genes obtained from the OMIM (Online Mendelian Inheritance in Man) database. Consistent with previous findings (Xu *et al.*, 2005), our results show that splice sites and exonic regions are most enriched for potentially deleterious SNPs. We further demonstrate the utility of our scoring system by showing that the FS score of known disease-related SNPs from OMIM is significantly higher than the score assigned to randomly selected neutral SNPs. We also show the improved performance of our system through a comparative study based on two evaluation measures. Finally, we discuss the impact of our work, and outline directions for future research.

2 PROBLEM DEFINITION

We aim to quantitatively measure the potential deleterious effects of SNPs on the bio-molecular function of their genomic region. For simplicity, we refer to the assessed score as the functional significance (FS) score of each SNP.

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

For each of the p SNPs and q tools, we define two random variables, δ_{ij} and $S_{ij}\,(i=1,\ldots,p;\,j=1,\ldots,q)$. The variable δ_{ij} denotes the label assigned to the i-th SNP by the j-th tool, that is, $\delta_{ij}=1$ when the j-th tool predicts SNP X_i to be deleterious, and 0 otherwise. The variable S_{ij} represents the tool's own confidence score with respect to the assigned label. The higher the value of S_{ij}, the more strongly the tool supports its own prediction, δ_{ij}. As different tools use different confidence scales, we define another random variable, $\bar{S}_{ij}$, representing a normalized confidence score. The normalization procedure is explained in Section 3.

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

Last, for each tool, we define a continuous random variable $TR_j\,(j=1,\ldots,q)$, corresponding to the *tool reliability* (TR) score for the j-th tool. This score represents how likely the tool is to correctly categorize SNPs as deleterious. The computation procedure of the TR score is explained in Section 3.

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

DEFINITION 2.1. FS score of SNP X_i

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

That is, for each bio-molecular functional category k, we compute the *weighted average* of the confidence of each prediction tool with respect to the deleterious[1] effect of the SNP, where the weight is the reliability score of each tool. Note that although summation is done over all the tools (i.e. $j=1$ to q, where q is the total number of tools, regardless of the bio-molecular functional categories that they examine), F_{jk} allows only the ones that examine the specific bio-molecular functional category k to be considered. The maximum score, over all examined bio-molecular functions, is then assigned as the FS score for the SNP.

3 METHODS

Our system conducts three main steps to calculate the FS score of SNPs, as outlined in Figure 1. In Step I, the functional labels for the SNPs (i.e. 'deleterious' or 'neutral'), predicted by q bioinformatics tools, are retrieved. Confidence scores associated with these predictions are also retrieved, when available. In step II-1, the reliability score of each tool is computed based on its tendency to agree with other tools' predictions. In Step II-2, the confidence scores, obtained in Step I, are normalized to a value between 0 and 1, as explained below. In Step III, the FS score of SNPs is computed as shown in Definition 2.1. We further describe each step below.

Step I. Retrieving Predicted Labels and Confidence Scores: Given a set of p SNPs, $\{X_1,\ldots,X_p\}$, we first retrieve their predicted functional labels (i.e. 'deleterious' or 'neutral') and corresponding confidence scores from 16 publicly available web services and databases, as illustrated in Figure 2. We obtain the genomic location of each SNP (e.g. exon, intron, splice site, 5′/3′-untranslated regions (UTRs) of a gene or directly upstream or downstream from a gene) from the dbSNP database (build 126) (Sherry *et al.*, 2001). According to the genomic location, each SNP is examined for its possible deleterious effects with respect to the corresponding bio-molecular functional category as follows:

- *Protein Coding*: SNPs in exonic regions may alter protein structure and/or function by creating a new start or stop codon (i.e. nonsense SNPs) or a deleterious amino acid substitution (i.e. missense SNPs).
- *Splicing Regulation*: SNPs in (canonical) splice sites may disrupt splicing regulation, resulting in exon skipping or intron retention. SNPs

[1]We note that by multiplying by δ_{ij}, the confidence score of each tool is counted only when the tool predicts the SNP to be deleterious.

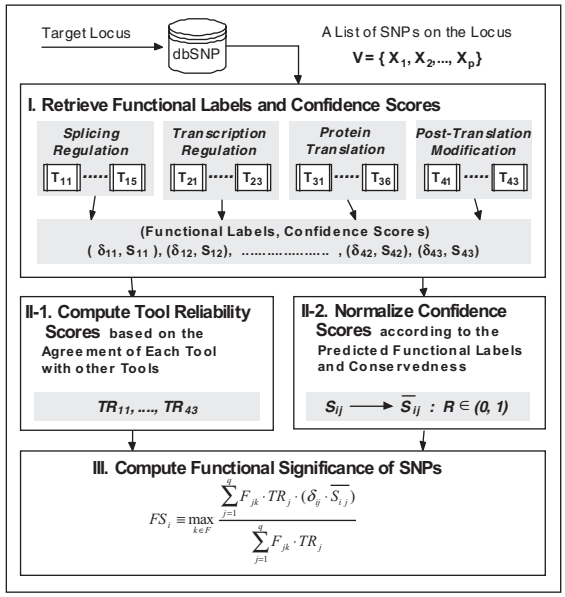

Fig. 1. Outline of our assessment process. In Step I, we retrieve the predicted functional labels of SNPs from integrated tools, along with their confidence scores. In Step II, we compute the tool reliability, and normalize the confidence scores. In Step III, we compute the FS score of SNPs as stated in Definition 2.1.

in exonic splice sites may interfere with alternative splicing regulation by changing exonic splicing enhancers or silencers.

- *Transcriptional Regulation*: SNPs in transcription regulatory regions (e.g. transcription factor binding sites, CpG islands, microRNAs, etc.) can alter binding sites, and thus disrupt proper gene regulation.
- *Post-Translational Modification*: SNPs in protein-coding regions may alter post-translational modification sites, interfering with proper post-translational modification.

As shown in Figure 2, the confidence scores for the SNPs that create a new start or stop codon (i.e. nonsense SNPs) or the SNPs that occur in the first two or in the last two bases of intronic splice sites (i.e. canonical splice sites) are set to one. This is because their deleterious effects to either *Protein Coding* or *Splicing Regulation* is unequivocal. Nonsense SNPs lead to a premature termination of amino acid peptides, often resulting in loss of protein function (Yamaguchi-Kabata *et al.*, 2008). The change to the canonical splice sites is also known to be detrimental as suggested by the high selection pressure on the splice sites among mammalian genomes (Burset *et al.*, 2000). We note that other SNP prioritization studies (Bhatti *et al.*, 2006; Xu *et al.*, 2005; Yuan *et al.*, 2006) assign the highest rank or score of functional impact to these two kinds of SNPs, as well. For the remaining cases, the confidence scores are obtained from the tools that are used to assess the corresponding bio-molecular functions.

Step II-1. Computing Tool Reliability: The tool reliability score, TR_j denotes how likely the j-th tool is to correctly predict deleterious SNPs ($j=1,\ldots,q$). We express the tool reliability score using the conditional probability as defined below:

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

That is, for each tool j, we calculate the conditional probability of any SNP X_i to actually be deleterious ($Y_i=1$) when the tool predicts so ($\delta_{ij}=1$). If the true labels of the SNPs, $Y_1,\ldots,Y_p$, are known, this score can be statistically estimated. For example, using a maximum likelihood approach, TR_j can be estimated as the ratio between the number of correctly predicted deleterious SNPs and the total number of SNPs predicted to be deleterious by the tool. However, in most cases we do not know the true functional categories of SNPs. We thus estimate the probability $\Pr(Y_i=1\,|\,\delta_{ij}=1)$ using the theoretical work proposed by Long and his colleagues (2005) on classification. When class labels are unknown, they propose to estimate the prediction accuracy of a classifier based on the extent to which the classifier tends to agree with other classifiers. They prove that the conditional probability $\Pr(\delta_{ij}=1\,|\,Y_i=1)$ can be calculated in this context as follows:

$$\Pr(\delta_{ij}=1\,|\,Y_i=1)=$$

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

where m and n represent the indices of any two distinct tools ($m\neq n\neq j$), $u_{jm}\stackrel{\text{def}}{=}\Pr(\delta_{ij}=1,\delta_{im}=1)$, and $u_j\stackrel{\text{def}}{=}\Pr(\delta_{ij}=1)$. For the detailed proof of Equation (1), we refer to the work by Long *et al.* (2005). Using Bayes' rule and Equation (1), we compute the tool reliability score of the j-th tool, TR_j, as follows:

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

Note that we use the *same* uninformative priors, $\Pr(Y_i=1)$ and $\Pr(\delta_{ij}=1)$ for all SNPs X_i, and as such, the tool reliability score is independent of the SNP X_i. To estimate $\Pr(Y_i=1)$, which is the prior probability of any SNP X_i to be deleterious, we take a conservative maximum likelihood approach. That is, for each tool assessing the effect of a SNP on a specific bio-molecular function, the fraction of SNPs that are *unanimously* predicted to be deleterious by all the tools assessing the same function is used as an estimate for $\Pr(Y_i=1)$, $(1\leq i\leq p)$.

Step II-2. Normalizing Confidence Scores: To account for the fact that different tools use different scales to report their confidence scores, we normalize the obtained confidence scores S_{ij} to be between 0 and 1 as follows:

$$\bar{S}_{ij}=\frac{1}{2}\cdot\left(\delta_{ij}+(1-C_{ij})\cdot\frac{(S_{ij}-\min_i S_{ij})}{(\max_i S_{ij}-\min_i S_{ij})}\right)$$

where

$$\begin{cases} C_{ij}=1, & \text{if } X_i \text{ resides on a non-conserved regulatory site;} \\ C_{ij}=0, & \text{otherwise,} \end{cases}$$

and $(1\leq i\leq p;\ 1\leq j\leq q)$. In principle, when SNP X_i is predicted to be deleterious (i.e. $\delta_{ij}=1$), the confidence score S_{ij} is converted to a value between 0.5 and 1; otherwise (i.e. $\delta_{ij}=0$), S_{ij} is converted to a value between 0 and 0.5. We note that, for the SNPs occurring in regulatory regions, we examine whether the SNP's region is *conserved* across multiple species (i.e. chimpanzee, dog, mouse, rat, chicken, zebrafish and fugu)— information that is obtained from GoldenPath (Kuhn *et al.*, 2007)—to reduce the effects of possible false positive predictions. As is widely known, there is a high rate of false positive findings of regulatory sites by *in silico* prediction tools due to their relatively short length of DNA sequences (typically 6- to 10mer) (Zhang *et al.*, 2003). Thus, when the predicted regulatory sites are not within a conserved region, the confidence score for the SNPs in the region is set to 0.5, reflecting our uncertainty regarding the functionality of the region, and the consequent, lack of confidence about potential deleterious effects of SNPs on the function.

We also note that some prediction tools, such as SNPeffect (Reumers *et al.*, 2005) or LS-SNP (Karchin *et al.*, 2005), do not provide confidence scores.

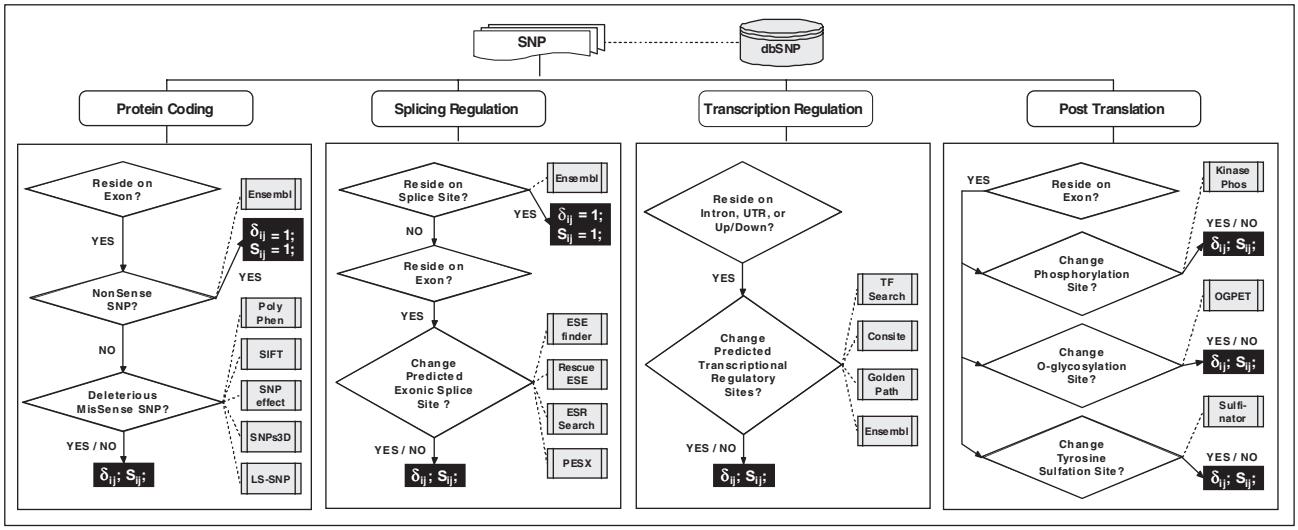

Fig. 2. The prediction flow-chart for four major bio-molecular functional categories. For the Protein Coding category, Ensembl (Hubbard *et al.*, 2007) is used to identify nonsense SNPs, and the web services, PolyPhen (Ramensky and Sunyaev, 2002), SIFT (Ng and Henikoff, 2001), SNPeffect (Reumers *et al.*, 2005), SNPs3D (Yue *et al.*, 2006), LS-SNP (Karchin *et al.*, 2005) are used to predict deleterious missense SNPs. For Splicing Regulation, Ensembl (Hubbard *et al.*, 2007) is used to identify SNPs in canonical splice sites, and ESEfinder (Cartegni *et al.*, 2003), RescueESE (Yeo and Burge, 2004), ESRSearch (Fairbrother *et al.*, 2002), and PESX (Zhang *et al.*, 2005) are used to examine SNPs in exonic splice sites. For Transcriptional Regulation, TFSearch (Akiyama, 1998), ConSite (Sandelin *et al.*, 2004), GoldenPath (Kuhn *et al.*, 2007), Ensembl (Hubbard *et al.*, 2007) are used to identify SNPs changing transcriptional regulatory sites. For Post-Translational Modification, KinasePhos (Huang *et al.*, 2005), OGPET (Gerken *et al.*, 2004), and Sulfinator (Monigatti *et al.*, 2002) are used.

For these systems, we impute the confidence scores using the confidence scores for the same SNP obtained from other tools. Suppose that the j-th tool, which examines the possible effects of SNP X_i on the bio-molecular functional category k, does not provide a confidence score on its prediction. Among the other tools that provide the confidence scores for the same function k, let us denote the index of the tool whose tool reliability score is highest as t. The imputed value is calculated as:

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

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

Step III. Computing Functional Significance: Given the prediction results obtained in Step I and the tool reliability and normalized confidence scores computed in Step II, the FS score of SNP X_i is computed as stated earlier in Definition 2.1.

4 EXPERIMENTS AND RESULTS

We applied our method to 112 949 SNPs located on 580 disease-susceptible genes for which the OMIM database references the biomedical literature that report the existence of SNPs on these genes that are either disease causing or associated with common disorders. The list of SNPs linked to the 580 genes, along with their primary information (e.g. genomic location), were downloaded from the dbSNP database (build 126) (Sherry *et al.*, 2001). The number of known *disease-causing* or *disease-associated* SNPs on these 580 genes is 1399 (the list was obtained from ftp://ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/Omim-VarLocusIdSNP.bcp.gz). The remaining 111 550 SNPs are not yet identified to be related to disease. For simplicity, we refer to the former set (of 1399 SNPs) as *disease-related* SNPs, and the latter set (of 111 550 SNPs) as *neutral* SNPs. We note, however, that currently known disease-related SNPs can explain only a fraction of the genetic basis of human disease, and as such, the latter set of SNPs that are temporarily classified as neutral may include functionally significant SNPs with deleterious effects that are not yet identified.

In Section 4.1, we summarize the scoring results by our system for all 112 949 SNPs, and show the distinguishing features of disease-related SNPs compared with neutral SNPs. In Section 4.2, we further validate that our integrative scoring system improves upon the state-of-the-art when applied to the same set of SNPs.

4.1 Review of the scoring results

First, we examine the scores that our system assigns to disease-related SNPs compared with neutral SNPs. Figure 3 shows the distribution of the FS score for disease-related SNPs (shown on the left) along with that of neutral SNPs (shown on the right). The figure clearly shows that the distribution of the FS scores for disease-related SNPs is significantly[2] different from that of neutral SNPs on the same genes. In particular, the median FS score for neutral SNPs is 0.1764, whereas, for disease-related SNPs, the median rises to 0.5. Moreover, 48.39% of disease-related SNPs are assigned an FS score >0.5, whereas only 2.2% of neutral SNPs are assigned such a high score.

Next, we examine the FS score distribution for SNPs based on their functional genomic regions. Figure 4a shows, for each genomic

[2]The difference is also statistically significant, with a P-value of practically 0, according to the Kolmogorov–Smirnov two-side test with $\alpha = 0.05$.

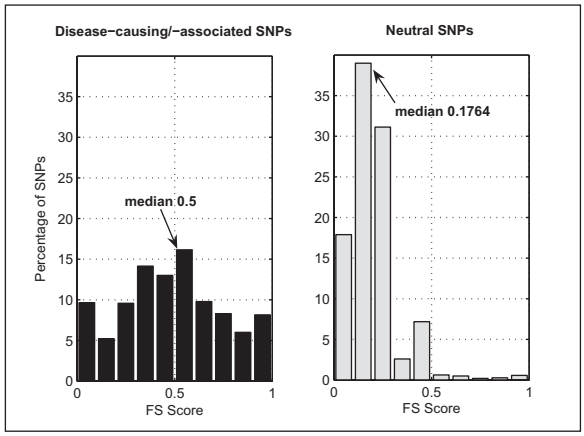

Fig. 3. The distribution of FS scores for disease-related SNPs and for neutral SNPs, assigned by our system. The *X*-axis represents the FS score for each group of SNPs, binned into 10 equal intervals, while the *Y*-axis represents the percentage of SNPs whose FS score corresponds to each bin.

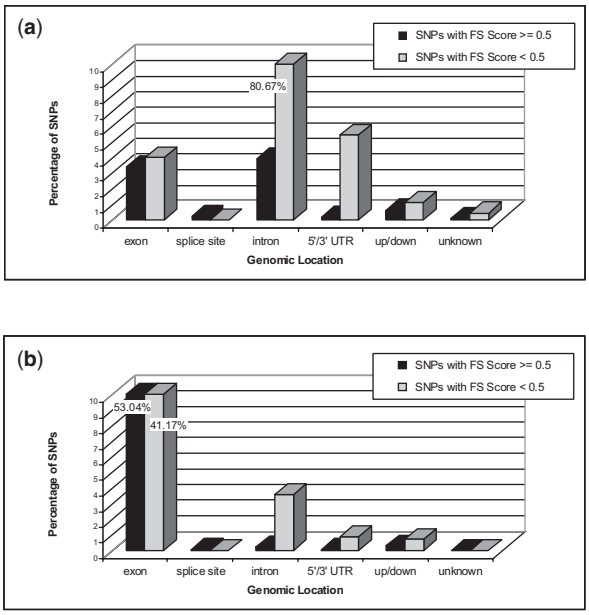

Fig. 4. The distribution of low FS scoring versus high FS scoring SNPs based on functional genomic locations (**a**) neutral SNPs (111 550 SNPs) and (**b**) known disease-related SNPs (1399 SNPs). The *X*-axis denotes six types of genomic regions that are used in the decision procedure (shown in Fig. 2), while the *Y*-axis shows the percentage of SNPs whose FS scores are at least 0.5 (black bars) versus the percentage of SNPs whose scores are lower than 0.5 (gray bars) on each region type.

region, the percentage of high FS scoring versus low FS scoring neutral SNPs on the region. Figure 4b shows this score distribution for disease-related SNPs. For clarity, the percentage is displayed only up to 10%.

As shown in Figure 4a, the majority of neutral SNPs are located within intronic regions, and the FS score for most intronic SNPs is <0.5 (80.67%). A similar tendency is noted in 5′/3′-UTRs, upstream

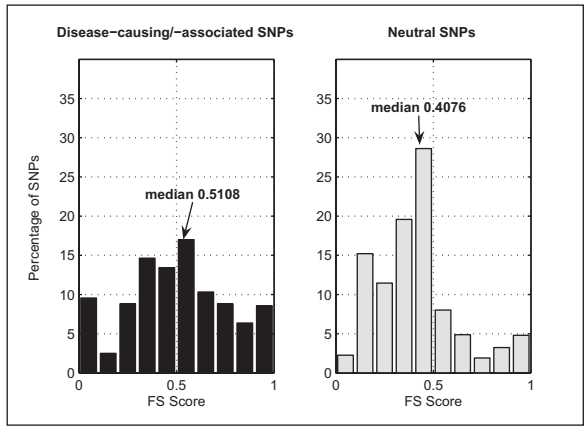

Fig. 5. The distribution of the assessed FS scores for *exonic* SNPs. The *X*-axis represents the FS score for each group of SNPs, binned into 10 equal intervals, while the *Y*-axis represents the percentage of SNPs whose FS score corresponds to each bin.

or downstream from genes, and in currently unspecified regions. In contrast, despite the relatively smaller number of SNPs on splice sites and on coding regions, these regions are enriched for high-scoring putatively deleterious SNPs. That is, an FS score of at least 0.5 is assigned to all SNPs in canonical splice sites and to 46.07% of the SNPs in coding regions. This scoring pattern is consistent with previous findings that mutations in splice sites and coding regions are likely to have direct impact on gene function (Cartegni *et al.*, 2003; Reumers *et al.*, 2005; Yue *et al.*, 2006).

Meanwhile, Figure 4b shows the FS score distribution for disease-related SNPs as a function of their genomic regions. In contrast to the case for neutral SNPs, shown above, most disease-related SNPs are located within exons (94.21%). This is indeed expected, as most association studies that validated these SNPs to be disease-related, have focused on protein-coding SNPs, whose functional effects are relatively easy to pinpoint due to their direct impact on protein products. Aside for the outstanding proportion of exonic SNPs, disease-related SNPs show a similar scoring pattern to that of neutral SNPs. Most SNPs on intronic, 5′/3′-UTR and up/downstream regions are assigned an FS score <0.5, but more than half of SNPs in exonic regions (53.04%) and all SNPs in canonical splice sites are assigned an FS score of at least 0.5.

As is clear from the data shown above, most disease-related SNPs are located on exons, while most (currently assigned) neutral SNPs are located within introns. We thus need to examine whether the difference in FS-score distributions between the two sets of SNPs, shown in Figure 3, is an artifact of the difference in their genomic region. Figure 5 shows the distribution of assigned FS scores, this time only for 1318 *exonic* SNPs that are already known to be disease-related (shown on the left) and for 8228 *exonic* SNPs assumed to be neutral (shown on the right). As expected, the median score for exonic SNPs is higher than that of SNPs in all regions, both for disease-related SNPs and for neutral SNPs. Nevertheless, still only 22.86% of neutral exonic SNPs are assigned an FS score >0.5, while the ratio rises to 56.30% for disease-related exonic SNPs. The Kolmogorov–Smirnov test with 5% significance level confirms that the two groups of exonic SNPs are unlikely to share a common score distribution (*P*-value 1.30e-079).

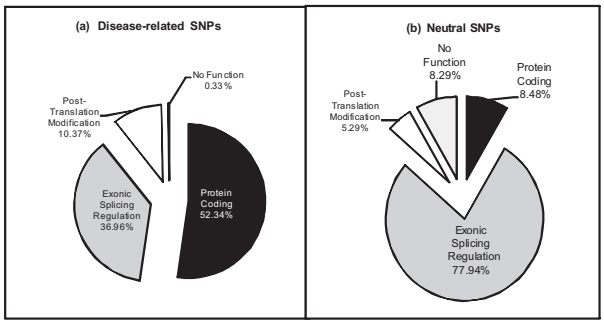

Fig. 6. The distribution of bio-molecular functions affected by disease-related exonic SNPs (shown on the left) and by neutral exonic SNPs (shown on the right).

Last, we examine what kinds of bio-molecular functions the two groups of exonic SNPs mainly affect. Recall that SNPs in exonic regions may affect either Protein Coding, (exonic) Splicing Regulation or Post-Translational Modification (as summarized in Figure 2), and we assign the maximum score over the three functional categories to each exonic SNP as its final FS score (as stated in Definition 2.1). We thus examine the proportion of the three bio-molecular functions that are used to assign the final FS scores for disease-related exonic SNPs and for neutral exonic SNPs. Figure 6 summarizes the results. In the case of disease-related SNPs, more than half of the exonic SNPs affect Protein Coding, and about 37% of the SNPs affect exonic Splicing Regulation. Conversely, only 8.48% of neutral exonic SNPs affect Protein Coding, while more than two-thirds of them affect (exonic) Splicing Regulation. In either case, Post-Translational Modification seems to be a minor cause for potential deleterious effects of SNPs.

4.2 A comparative study

To validate that our scoring system improves upon the state-of-the-art, we compare our system with two public web services that numerically score deleterious effects of SNPs: SNPselector (Xu *et al.*, 2005) and FastSNP (Yuan *et al.*, 2006). SNPselector provides a numeric score for each SNP, called *function score*, which designates the possible effects of SNPs on gene transcript structure or on protein product. The score is a real number between 0.6 and 1.0; the higher the score is, the more deleterious the effects of the SNPs are expected to be. FastSNP is another web service for SNP function analysis and prioritization. It assigns to each SNP an integer score between 0 and 5, called *risk rank*, which quantifies how likely the SNP is to have functional effects leading to disease phenotypes. Last, as a baseline performance, we compute the FS score of SNPs using a simple majority vote. For example, when one-third of the tools that examine the deleterious effects of SNPs on protein coding predict the SNP to be deleterious, a value of 1/3 is assigned as its FS score. Our scoring scheme is distinguishable from this simple majority vote as it takes into account the certainty of each prediction (through normalized confidence scores) as well as the reliability of each tool (through tool reliability scores).

To compare the three scoring schemes with ours, we generated test datasets using the following sampling procedure. For each disease-related SNP X_i, one neutral SNP is selected uniformly at random in the same functional region on the same gene as X_i. This selection is done for all disease-related SNPs. As a result, a dataset of 1399 SNP pairs, one disease-related and one randomly selected neutral, is generated. We repeat this procedure M times, generating M test datasets (here, $M = 100$). We note that, by limiting the random selection to the same functional region on the same gene, we reduce the bias that may arise due to the differences in the functional or chromosomal regions.

Table 1. The results of a comparative study based on two evaluation measures, Higher score and Paired t-test.

	Evaluation Measure	
System	Higher score (%)	Paired t-test (avg. P-value)
Our System	63.82	1.00 (0.00)
FastSNP	61.15	1.00 (3.61e-127)
SNPselector	55.39	1.00 (6.91e-125)
Simple majority vote	45.42	0.93 (0.01)

Using the test datasets, we examine how well each system distinguishes disease-related SNPs from neutral SNPs. Intuitively, a better scoring system would assign a higher functional score to disease-related SNPs than to neutral SNPs. First, we measure this tendency by directly computing the percentage of disease-related SNPs that are assigned a higher FS score than their paired, randomly selected neutral SNPs, averaged over M test datasets. We refer to this measure as *Higher score* (%). Second, using the paired t-test, we examine whether the disease-related SNPs and the neutral SNPs in each dataset share the FS score distributions with the same mean. We separately conduct the paired t-test on each of the M datasets, and compute the proportion of the rejected tests along with their average P-value. The rejection implies that the FS score distribution of disease-related SNPs and that of likely neutral SNPs are distinctive. Therefore, scoring schemes with a high proportion of rejected t-tests are preferred. We refer to this second measure as *Paired t-test*.

Table 1 summarizes the results of our comparative study. Overall, our system improves upon all the compared systems with respect to both evaluation measures. In the case of the Higher score measure, our system assigns higher FS scores to about 64% of known disease-related SNPs than to neutral SNPs. FastSNP comes second, and SNPselector and Simple Majority Vote follow. The score difference between our system and the compared systems is also statistically significant (P-values are 6.96e-038, 4.82e-105, and 5.26e-174 for FastSNP, SNPselector and Simple Majority Vote, respectively, using the paired t-test, $\alpha = 0.05$). It is notable that our system greatly outperforms Simple Majority Vote, which demonstrates the utility of the confidence and the tool reliability scores, integrated into our scoring scheme.

In the case of the Paired t-test measure, the first three systems, namely, our system, FastSNP and SNPselector perform the same; all of the paired t-tests were rejected with a significance level 0.05. However, the average P-value of the rejected tests is smallest (i.e. asymptotically zero) for our system among the three, which means that the score distribution of disease-related SNPs and that of neutral SNPs are most disparate when their FS scores are assigned by our system. In the case of Simple Majority Vote, only 93% of the paired

t-tests were rejected. The average *P*-value for the rejected tests is also the largest among all the compared systems.

5 DISCUSSION

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

Two main features distinguish our system from others. First, we integrate multiple tools to overcome the incompleteness or erroneous predictions of individual prediction tools. While a single tool may fail to capture the deleterious effects of many SNPs, a combination of multiple independent tools, which are based on different resources and algorithms, are less likely to all make the same error. Thus the tools are likely to complement each other, and as we have demonstrated in our results, typically compensate for each other's errors. As a result, the effect of possible false negative or false positive predictions in any single tool is reduced when computing the combined FS score.

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

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

Second, while the disease-related SNPs may indeed be disease-causing mutations, our current scoring scheme may not capture them properly. For example, in addition to the bio-molecular functions that we currently examine, there could be other genetic mechanisms that have a profound impact on human pathogenesis. We thus plan to update our system through combining other epidemiological resources, such as literature information, as well as integrating more prediction tools for each bio-molecular function.

We note that our integrative scoring system is primarily intended for candidate gene-based association studies, where there is already a region in which some SNPs are likely to have deleterious functional effects. In the case of genome-wide association studies, our system can be used to prioritize a subset of SNPs that needs further investigation after indication for disease association has been detected for a genomic region.

We currently provide the calculated FS score of 112 949 SNPs through our public web-based database service, F-SNP (available at http://compbio.cs.queensu.ca/F-SNP). We will continue assigning FS scores to other SNPs, and update the scoring results. In addition, we plan to integrate our scoring system with our earlier tagging SNP prioritization approach (Lee and Shatkay, 2006) for association studies. By combining the two most representative selection approaches for SNPs, we expect to provide a comprehensive SNP prioritization system for facilitating effective association studies on common and complex genetic disorders.

Funding: HS's NSERC Discovery grant 298292-04; CFI New Opportunities Award 10437.

Conflict of Interest: none declared.

REFERENCES

Akiyama,Y. (1998) TFSEARCH: Searching Transcription Factor Binding Sites. Available at Web Service: http://www.cbrc.jp/research/db/TFSEARCH.html (last accessed date March 3, 2009).

Bhatti,P. *et al.* (2006) Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. *Am. J. Epidemiol.*, **164**, 794–804.

Burset,M. *et al.* (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

Cartegni,L. *et al.* (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.

Fairbrother,W.G. *et al.* (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

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

Huang,H. *et al.* (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.

Hubbard,T.J.P. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Karchin, R. *et al.* (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.

Kuhn,R. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.

Lee,P. and Shatkay,H. (2006) BNTagger: improved tagging SNP selection using Bayesian networks. *Bioinformatics* (Special issue on *Proceedings of the 14th Annual International Conference on Intelligent Systems for Molecular Biology)*, **22**, e211–e219.

Long,P.M. *et al.* (2005) Unsupervised evidence integration. In *Proceedings of the 22nd international conference on Machine learning*, Vol.119, ACM, New York, NY, USA, Bonn, Germany, pp. 521–528.

Monigatti,F. *et al.* (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, **18**, 769–770.

Ng,P. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Ramensky,V. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acid Res.*, **30**, 3894–3900.

Rebbeck,T. R. *et al.* (2004) Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.*, **5**, 589–597.

Reumers,J. *et al.* (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res..*, **33**, D527–D532.

Sandelin,A. *et al.* (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, **32**, W249–W252.

Sherry,S. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Xu,H. *et al.* (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics*, **21**, 4181–4186.

Yamaguchi-Kabata,Y. *et al.* (2008) Distribution and effects of nonsense polymorphisms in human genes. *PLOS One*, **3**, e3393.

Yeo,G. and Burge,C. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl Acad. Sci.*, **101**, 15700–15705.

Yuan,H. *et al.* (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.*, **34**, W635–W641.

Yue,P. *et al.* (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.

Zhang,L.-H. *et al.* (2003). Finding regulatory sequences. *Int. J. Biochem.*, **35**, 95–103.

Zhang,X.H.-F. *et al.* (2005) Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell Biol.*, **25**, 7323–7332.