Ranking single nucleotide polymorphisms by potential deleterious effects

Phil Hyoun Lee, MSc, Hagit Shatkay, PhD
Computational Biology and Machine Learning Lab,
School of Computing, Queen's University, Kingston, ON, Canada

Abstract

Identifying single nucleotide polymorphisms (SNPs) that are responsible for common and complex diseases such as cancer is of major interest in current molecular epidemiology. However, due to the tremendous number of SNPs on the human genome, to expedite genotyping and analysis, there is a clear need to prioritize SNPs according to their potentially deleterious effects to human health. As of yet, there have been few efforts to quantitatively assess the possible deleterious effects of SNPs for effective association studies. Here we propose a new integrative scoring system for prioritizing SNPs based on their possible deleterious effects in a probabilistic framework. We also provide the evaluation result of our system on the OMIM (Online Mendelian Inheritance in Man) database, which is one of the most widely-used databases of human genes and genetic disorders.

Introduction

Much effort in current epidemiology, medicine, and pharmarcogenomics is focused on the identification of genetic variations that are involved in common and complex diseases. Specifically, single nucleotide polymorphisms (SNPs), which are substitutions of a single nucleotide at a specific position on the genome occurring in more than 1% of the human population, are in the core of such studies, as they form the majority of genetic variations in the human. Reliable identification of disease-causing SNPs is expected to enable early diagnosis, personalized treatments, and targeted drug design[1].

Typically, the first step toward identifying causal SNPs for common and complex human diseases, involves case-control association studies[2]. However, due to the sheer number of SNPs on the human genome, estimated at over ten milion[3], it is often required, when conducting association studies, to prioritize SNPs based on their potential deleterious functional effects[1]. 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 associated with disease, a substantial amount of genotyping and analysis overhead can be reduced.

However, for the vast majority of SNPs, no experimental evidence is currently available to substantiate their deleterious effects. As such, web-services and public databases that provide computationally predicted putative deleterious effects of SNPs have been developed and widely used[4]. These tools examine whether a SNP resides in functional genomic regions 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. The utility of these computational tools has been empirically demonstrated in several genetic variation studies[5-6].

Such tools and systems, which prioritize functionally significant SNPs, suffer from two main limitations: First, they provide only partial information about the functional significance of SNPs. That is, they each examine the putative deleterious effects of SNPs with respect to a *single* biological function, for example, either protein coding or transcriptional regulation. Thus, to comprehensively analyze the functional significance of SNPs, researchers must spend much time and effort to separately apply multiple tools, and interpret/integrate their (often conflicting) predictions.

Second, while current systems classify SNPs into distinct groups (e.g., 'deleterious' or 'neutral'), they do not numerically score or rank SNPs according to their functional significance. Budget considerations often force researchers to select a limited number of SNPs on the target genomic region for conducting association studies. When the number of putatively deleterious SNPs presented by current tools is larger than this pre-specified limit, with no additional ranking information, selecting only some of them is not straightforward. As a result, researchers must rely on other resources to finalize their decision.

To address these limitations, we propose a new integrative scoring system for ranking SNPs based on their putative deleterious effects. To do this, we assess SNPs with respect to four major bio-molecular functional categories of genomic regions: splicing, transcription, translation, and post-translation modification. We attempt to overcome the

incompleteness and possible false findings of an individual bioinformatics tool by combining the assessment results from multiple independent prediction tools. Most significantly, we assign a specific numerical score to each SNP, representing its putative deleterious effects to human health. Using this score, a limited subset of the most functionally significant SNPs can be ranked and selected.

We applied our system to 123,697 SNPs located on 607 disease-susceptibility genes obtained from the OMIM database[7]. Splice sites and coding regions are most enriched with potentially deleterious SNPs, which is consistent with established findings. We further demonstrate the utility of our scoring system by showing that the functional significance score for known disease-associated SNPs from OMIM is significantly higher than the score assigned to randomly selected SNPs on the same gene. Finally, we discuss the impact of our work and possible directions for future research.

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 assessing the FS score, we first introduce some basic notations. Suppose that we are given p SNPs on the target genomic region. 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}. The true (and unknown) functional category of SNP X_j is then represented by another discrete random variable, Y_j, whose value is 1 when X_j is deleterious and 0 otherwise. We note that we do not know the true functional category Y_j of SNP X_j. We thus estimate it using m bioinformatics tools that predict, for each SNP X_j, the functional label (i.e., `deleterious' or 'neutral') along four major bio-molecular functions: *protein coding*, *splicing regulation*, *transcriptional regulation* or *post-translation modification*.

For each of the m tools, and each of the SNPs, we define two discrete random variables, U_{ij} and S_{ij}. (i=1,...,m; j=1,…,p). The variable U_{ij} denotes the label assigned to the j^{th} SNP by the i^{th} tool, that is, U_{ij}=1 when the i^{th} tool predicts SNP X_j to be deleterious, and 0 otherwise. The variable S_{ij} represents a confidence score on the assigned label. The higher the value of S_{ij}, the more strongly the tool supports its own prediction, U_{ij}. As different tools use different confidence scales, we define a normalized confidence score, $\overline{S_{ij}}$, whose value is between 0 and 1, as follows:

$$\overline{S_{ij}} = \frac{S_{ij} - \min_k S_{ik}}{\max_j S_{ij} - \min_j S_{ij}}.$$

For each SNP and tool, we also define a random variable C_{ij} to indicate whether the SNP may affect any other bio-molecular function due to its genomic location (obtained from dbSNP). The value of C_{ij} is 1 if and only if either: a) SNP X_j is located on a protein coding region, and the i^{th} tool examines the deleterious effects of SNPs on either protein coding, exonic splicing regulation, or post-translation modification; b) SNP X_j is located on a splice site, and the i^{th} tool examines the deleterious effects of SNPs on intronic splicing regulation; c) SNP X_j is located on either an intronic region, 5'/3' un-translated regions of a gene (UTR), or directly upstream or downstream from a gene, and the i^{th} tool examines the effects of SNPs on transcriptional regulation; or d) SNP X_j is located on any intergenic regions whose function is currently unspecified. (As we do not know the function of the region, we need to examine the putative effects of SNPs with respect to all four bio-molecular functions.)

Last, for each tool, we define a continuous random variable TR_i, corresponding to the Tool Reliability (TR) score for the i^{th} tool. This score represents how likely the tool is to correctly predict deleterious SNPs. The computation of the TR score is explained in the next section.

Based on the variables U_{ij}, $\overline{S_{ij}}$, C_{ij}, and TR_i, the FS score of SNP X_j is defined and calculated as follows:

Definition 1. Functional Significance (FS) score: *Given a SNP X_j, a set of m functional prediction results and normalized confidence scores for the SNP X_j, {(U_{1j}, $\overline{S_{1j}}$),…,(U_{mj}, $\overline{S_{mj}}$)} from m distinct prediction tools, and a set of m Tool Reliability Scores for the tools, {TR_1,…, TR_m} the Functional Significance score of SNP X_j, denoted by FS_j, is defined as:*

$$FS_j \equiv \frac{\sum_{i=1}^{m} C_{ij} \cdot TR_i \cdot (U_{ij} \cdot \overline{S_{ij}})}{\sum_{i=1}^{m} C_{ij} \cdot TR_i}.$$

That is, the FS score of a SNP is the weighted average of the normalized confidence scores obtained from the different prediction tools – regarding the SNP being deleterious – where the weight is the reliability score of each tool. We note that by multiplying U_{ij}, the confidence score of each tool is

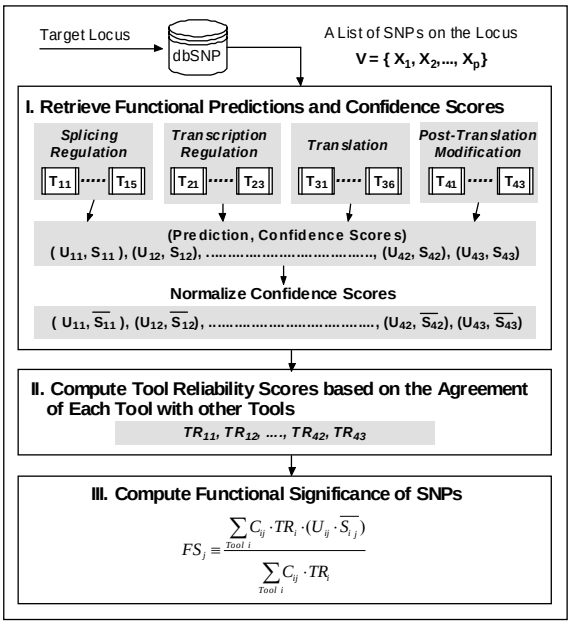

Figure 1. Outline of our assessment process.

counted only when the tool indeed predicts that the SNP X_j is deleterious.

Methods

Our system executes three main steps to assess the functional significance of SNPs. Figure 1 outlines the process. In step I, functional categories of SNPs, (i.e., either *deleterious* or *neutral*) and supporting confidence scores are obtained from *m* external prediction tools. In step II, the reliability of each tool is computed based on the tool's agreement with the prediction of other tools. In step III, the functional significance scores of SNPs are computed as an average of the normalized supporting confidence scores, weighted by the reliability of each tool. We further describe each step in the following subsections.

STEP I. Retrieving Predicted Functional Information

Given a set of *p* SNPs, $\{X_1, \ldots, X_p\}$, we first retrieve from the F-SNP database[8], their predicted functional labels (i.e., *deleterious* or *neutral*) along with respective confidence scores, obtained from 16 publicly available web-based services and databases. These 16 tools are grouped into four functional categories as follows:

- *Splicing Regulation*: SNPs in splicing regulatory sites may interfere with splicing regulation, resulting in unintentional exon skipping or intron retention;
- *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;
- *Translation*: SNPs in protein coding regions may cause a deleterious amino acid substitution (i.e., nonsynonymous SNPs) or interfere with protein translation (i.e., nonsense SNPs);
- *Post-Translation Modification*: SNPs in protein coding regions may alter post-translation modification sites (e.g., phosphorylation, o-glycosylation, or tyrosine sulfation sites), interfering with proper post-translation modification.

STEP II. Computing Tool Reliability

The Tool Reliability score, TR_i denotes how likely the i^{th} tool is to correctly predict deleterious SNPs. We measure the Tool Reliability score using the conditional probability as defined below:

$$TR_i \equiv \Pr(Y_j = 1 \mid U_{ij} = 1).$$

If the true labels of the SNPs, Y_j ($j=1,\ldots,p$), are known, this score can be statistically estimated. For example, using a maximum likelihood approach, TR_i can be estimated as the ratio between the number of correctly predicted deleterious SNPs and the total number of deleterious SNPs predicted by the tool. However, in most cases we do not know the true functional categories of SNPs. We thus estimate the probability $Pr(Y_j=1|U_{ij}=1)$ using the method proposed by Long et al in their theoretical work[9] on classification. When class labels are unknown, they propose to evaluate the prediction accuracy of a classifier based on the extent that the classifier tends to agree with other classifiers, and prove that the conditional probability $Pr(U_{ij}=1|Y_j=1)$ can be calculated in this context as follows:

$$\Pr(U_{ij}=1 \mid Y_j=1) = \Pr(U_{ij}=1) + \sqrt{\frac{1-\Pr(Y_j=1)}{\Pr(Y_j=1)} \cdot \frac{(q_{ik} - q_i \cdot q_k)\cdot(q_{il} - q_i \cdot q_l)}{(q_{kl} - q_k \cdot q_l)}} \quad (1)$$

where k and l represent the indices for any two distinct tools ($i \neq k \neq l$), and q_k and q_{kl} denote $Pr(U_{kj}=1)$ and $Pr(U_{kj}=1 \wedge U_{lj}=1)$, respectively. The maximum likelihood approach is used to estimate these probabilities. For a detailed proof of Equation 1), see the work by Long et al[9].

Using Bayes' rule and Equation 1), we compute the Tool Reliability score for the i^{th} tool, TR_i, as follows:

$$TR_i \equiv \Pr(Y_j = 1 \mid U_{ij} = 1)$$
$$= \Pr(U_{ij} = 1 \mid Y_j = 1) \cdot \frac{\Pr(Y_j = 1)}{\Pr(U_{ij} = 1)} \quad \text{(by Bayes' rule)}$$
$$= \Pr(Y_j = 1) + \sqrt{\frac{\Pr(Y_j = 1)}{(1 - \Pr(Y_j = 1))^{-1}} \cdot \frac{(q_{ik} - q_i \cdot q_k) \cdot (q_{il} - q_i \cdot q_l)}{(q_{kl} - q_k \cdot q_l)(q_i)^2}}.$$
(by Eq. 1)

We use uninformative priors for $Pr(Y_j=1)$ and $Pr(U_{ij}=1)$ over all SNPs, and as such, the Tool Reliability score is invariant to each SNP. To compute an estimate of the prior probability of the SNP being deleterious, for $Pr(Y_j=1)$, we take a conservative maximum likelihood approach. That is, for each tool examining the effects of a SNP on a specific bio-molecular function, the fraction of SNPs that are *unanimously* predicted to be deleterious by *all* the tools examining the same function is used as an estimate for $Pr(Y_j=1)$.

STEP III. Computing Functional Significance

Given the prediction results and confidence scores obtained in step I and the Tool Reliability score (TR_i) computed in step II, the Functional Significance (FS) score of SNP X_j can be computed as shown in Definition 1.

Results

We applied our method to 123,697 SNPs located on 607 disease-susceptible genes, for which the OMIM[7] database provides references to biomedical literature reporting the SNPs to be disease-causing or showing a positive statistical correlation with common disorders (downloaded Feb. 18, 2008). The list of SNPs linked to each of the 607 genes and their primary information such as genomic locations were downloaded from the dbSNP[2] database (build 126).

For each genomic location, Figure 2 shows the percentage of low FS scoring vs. high FS scoring SNPs that reside in the location among the examined 123,697 SNPs. The X-axis denotes 6 distinct types of genomic regions, while the Y-axis shows the percentage of SNPs whose FS scores are lower than 0.5 (gray bars) vs. the percentage of SNPs whose scores are at least 0.5 (black bars) on each region type. For clarity, the percentage is displayed up to 12%. The majority of the examined SNPs are located within intronic regions (81.7%), but the FS score for most SNPs in intronic regions is lower than 0.5

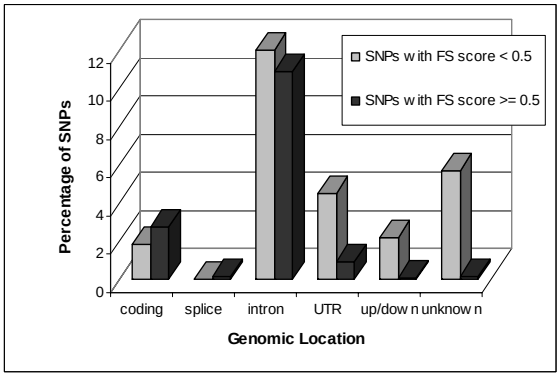

Figure 2. The percentage of low FS scoring vs. high FS scoring SNPs according to six genomic locations.

(70.84%). A similar tendency is noted in 5'/3' untranslated regions (UTR), upstream/downstream of a gene, 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 putative deleterious SNPs. That is, an FS score of at least 0.5 is assigned to 99% of the SNPs in splice sites and to 55% of SNPs in coding regions. This scoring pattern is consistent with the broadly accepted assumption that mutations in splice sites and coding regions would have direct effects on gene function[4].

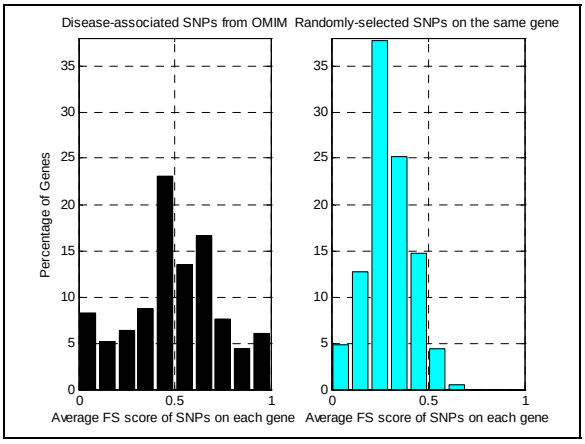

Figure 3. The distribution of the average FS scores of disease-associated SNPs on each gene, compared to that of randomly selected SNPs on the same gene.

Next, we examined whether the average FS score of SNPs known to be disease-causing or disease-associated in each of 607 genes (obtained from OMIM) is different from that of SNPs selected uniformly at random on the same gene. Figure 3 shows the distribution of the average FS scores. The

X-axis represents the average FS score for each group of SNPs on the same gene, binned into 10 equal intervals, while the Y-axis represents the percentage of genes whose average FS score corresponds to each bin.

As is clearly seen in Figure 3, the distribution of the average FS scores of known disease-causing or associated SNPs is significantly different from that of randomly selected SNPs on the same gene (p-value 1.0303e-055 using the paired t-test with 5% significance level). We note that for 48.3% of the SNPs that are disease associated, the average FS score is at least 0.5, while only 4.9% of randomly selected SNPs are assigned such a high score.

We note that the FS score assigned to about half of the known disease-causing or associated SNPs is still below 0.5. There are two possible explanations for this seemingly inappropriate FS score. First, even though some SNPs, obtained from the OMIM database[7], show a positive statistical correlation with common disorders in some association studies, they may not all be real disease-causing mutations. Some of these SNPs may simply be linked to the actual disease-causing mutations, or may represent false positive findings.

Second, while the disease-associated SNPs may indeed be disease-causing mutations, our current scoring scheme may not capture them properly. For example, besides the bio-molecular functions that we currently examine, there could be other genetic mechanisms that have a profound impact on human pathogenesis. As such, disease-associated SNPs with low FS scores should not be ruled out until biological experiments confirm their role.

Conclusion

We have presented a new scoring system for assessing the putative deleterious effects of SNPs. Our integrative scoring method combines assessment results from multiple independent computational tools in a probabilistic framework, which takes into account the certainty of each prediction as well as the reliability of the different tools. An empirical study over 607 disease-associated genes taken from the OMIM[7] database shows that our system provides distinct scoring patterns that are consistent with well-established findings about functional SNPs. We expect our scoring system to be a valuable resource for facilitating effective association studies for common and complex genetic disorders.

In the near future we plan to conduct more rigorous evaluation studies, and report comparisons to other function-assessment systems for SNPs (some of those were already carried out but not shown here for lack of space). We also plan to continue the computational assessment of potential deleterious effects of SNPs identified on the human genome, and provide the scoring results via the F-SNP database[8].

Acknowledgement

This work was supported by HS's NSERC Discovery grant 298292-04, CFI New Opportunities Award 10437, and PHL's Queen's University Duncan and Urlla Carmichael Fellowship.

References

1. Bhatti P, Church DM, Rutter JL, et al. Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. Am J Epidemiol. 2006;164(8):794–804.
2. Pharoah PD, Dunning AM, Ponder BA, et al. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer. 2004;4:850-60.
3. Sherry S, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
4. Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet. 2004;5:589-97.
5. Yeo G, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. P Natl Acad Sci USA. 2004;101(44):15700–5.
6. Brunham LR, Singaraja RR, Pape TD, et al. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLOS Genet. 2005;1(6):739–47.
7. McKusick-Nathans Institute of Genetic Medicine, John's Hopkins Uni.and NCBI, NLM. Online Mendelian Inheritance in Man, OMIM ™. http://www.ncbi.nlm.nih.gov/omim/.
8. Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res. 2008;36(Database issue):D820-4.
9. Long PM, Varadan V, Gilman S, et al. Unsupervised evidence integration. Proc. 2005 ICML, 19:521-8.