regSNPs: a strategy for prioritizing regulatory single nucleotide substitutions

Mingxiang Teng1,2,†, Shoji Ichikawa4,†, Leah R. Padgett4, Yadong Wang1, Matthew Mort3, David N. Cooper5, Daniel L. Koller3, Tatiana Foroud3, Howard J. Edenberg3,6, Michael J. Econs3,4 and Yunlong Liu2,3,∗

1School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, 2Center for Computational Biology and Bioinformatics, 3Department of Medical and Molecular Genetics, 4Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA, 5Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK and 6Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

ABSTRACT

Motivation: One of the fundamental questions in genetics study is to identify functional DNA variants that are responsible to a disease or phenotype of interest. Results from large-scale genetics studies, such as genome-wide association studies (GWAS), and the availability of high-throughput sequencing technologies provide opportunities in identifying causal variants. Despite the technical advances, informatics methodologies need to be developed to prioritize thousands of variants for potential causative effects.

Results: We present regSNPs, an informatics strategy that integrates several established bioinformatics tools, for prioritizing regulatory SNPs, i.e. the SNPs in the promoter regions that potentially affect phenotype through changing transcription of downstream genes. Comparing to existing tools, regSNPs has two distinct features. It considers degenerative features of binding motifs by calculating the differences on the binding affinity caused by the candidate variants and integrates potential phenotypic effects of various transcription factors. When tested by using the disease-causing variants documented in the Human Gene Mutation Database, regSNPs showed mixed performance on various diseases. regSNPs predicted three SNPs that can potentially affect bone density in a region detected in an earlier linkage study. Potential effects of one of the variants were validated using luciferase reporter assay.

Contact: yunliu@iupui.edu

Supplementary information: Supplementary data are available at Bioinformatics online

Received and revised on March 27, 2012; accepted on April 30, 2012

1 INTRODUCTION

A key goal in human genetics is to identify the functional DNA variants that give rise to phenotypic differences among individuals. Recent studies of complex diseases and phenotypes have tended to focus on genome-wide association studies (GWAS) employing hundreds of thousands of single nucleotide polymorphisms (SNPs). GWAS target common DNA variants, which could either directly contribute to the clinical phenotype or provide an indirect proxy for functional variants, which are in linkage disequilibrium (LD) with the SNP being tested. Distinguishing between direct, mechanistic contributions emanating from the functional variants themselves and indirect associations resulting from LD is challenging and improved methods are needed. One feasible solution is to catalog all DNA variants in the LD region of the association, both common and rare, by utilizing next-generation sequencing (NGS) technology.

The large number of variants that will be identified generates an urgent need for bioinformatics and computational approaches capable of prioritizing the variants most likely to underlie the observed association, for further biological testing. Non-synonymous substitutions within coding regions directly affect protein structure and are likely to affect protein function; a variety of algorithms, including PolyPhen (Ramensky et al., 2002), SIFT (Ng and Henikoff, 2003), TopoSNP (Stitzel et al., 2004), PMUT (Ferrer-Costa et al., 2005), LS-SNP (Karchin et al., 2005), SNPeff ect v2.0 (Reumers et al., 2006), SNP3D (Yue et al., 2006) and PolyDoms (Jegga et al., 2007), were designed to identify functional non-synonymous substitutions. Synonymous coding variants may also exert phenotypic effect by influencing the conformation, splicing and stability of pre-mRNAs or by altering the expression level of a given protein (Capon et al., 2004; Hunt et al., 2009; Kimchi-Sarfaty et al., 2007). Coding sequences make up <2% of the human genome (Elgar and Vavouri, 2008). The regulatory component of the human genome is much less well defined but based upon conservation is 2 to 3 times larger than the coding region (Cooper et al., 2010). Regulatory variants can affect the transcription initiation rate (Mertens et al., 2009), microRNA binding and protein expression (Nicoloso et al., 2010), but are far more difficult to identify. It should therefore come as no surprise that many of the variants associated with common, complex disease by GWAS do not alter coding sequences but rather occur within the non-coding regions of genes or intergenic regions (Chen et al., 2010; Dickson et al., 2010; Glinskii et al., 2009; Hindorff et al., 2009; Johnson and O’Donnell, 2009).

Tools such as FASTSNP (Yuan et al., 2006), PupaSuite (Conde et al., 2006) and SNPlogic (Pico et al., 2009), have been designed to identify SNPs residing in known transcription factor binding sites
prioritizing transcriptional regulatory SNPs within the chromosomal results suggest that the proposed approach should be well suited for then further tested for influence on transcription with an estimated false discovery rate (FDR) < contribute to the variation of femoral neck and lumbar spine bone linkage and association analysis to harbor genetic factors that a region on chromosome 1q that had previously been found through operating characteristics (ROC) curve ranging from 53.7% in diseases, with the area under the curve (AUC) of the receiver collection of gene mutations underlying or associated with human of analysis, we selected 13 disease states with at least two related genes respectively; both contain experimental validated data. For the purposes et al was randomly selected to represent putatively non-binding sequences for them do not have detailed sequences for TFBSs. These yielded 301 TFBSs. information on experimentally validated human TFBSs, was used as the features of protein–DNA binding sites, and requires gene expression to integrates expression quantitative trait loci (eQTLs) and GWAS empirical methodology that accounts for local LD structure and requires gene expression to integrate existing bioinformatics tools. We tested our model by assessing its ability to identify mutations logged in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009), a comprehensive collection of gene mutations underlying or associated with human inherited disease. The accuracy of our approach varies among diseases, with the area under the curve (AUC) of the receiver operating characteristics (ROC) curve ranging from 53.7% in schizophrenia to 77.9% in breast cancer. We applied our method to a region on chromosome 1q that had previously been found through linkage and association analysis to harbor genetic factors that contribute to the variation of femoral neck and lumbar spine bone mineral density (BMD) in premenopausal white women (Ichikawa et al., 2008). Three SNPs were selected as likely to be functional, with an estimated false discovery rate (FDR) <25%. They were then further tested for influence on transcription in vitro. Our results suggest that the proposed approach should be well suited for prioritizing transcriptional regulatory SNPs within the chromosomal regions found to be associated with a disease or phenotype.

2 METHODS

2.1 Data source and materials

SNPs in HapMap release 27 derived from Utah residents with ancestry from Northern and Western Europe (CEU) were used. Promoter regions were defined as between upstream 3000 bp and downstream 1000 bp from transcriptional start sites of genes based on RefSeq annotation. A total of 96,069 promoter SNPs were identified; 10,000 were randomly selected as the background when generating the empirical P-values for the significance of effects on binding affinity of each transcription factor. While applying regSNPs to a BMD-related region, 51 promoter SNPs were analyzed. The TRANSFAC 9.2 database (Wingender et al., 1996), which contains information on experimentally validated human TFBSs, was used as the source of binding information for known transcription factors. We selected 323 of the 459 human transcription factors for further study, because the rest of them do not have detailed sequences for TFBSs. These yielded 301 TFBSs. A negative control set of 10,000 promoter sequences in the human genome was randomly selected to represent putatively non-binding sequences for each TFBS.

The OMIM database (McKusick, 2007) and the HGMD 2009.1 database (Stenson et al., 2009) provide disease-related genes and mutations, respectively, both contain experimental validated data. For the purposes of analysis, we selected 13 disease states with at least two related genes in OMIM and at least nine promoter mutations documented in HGMD for model evaluation.

2.2 Evaluating the effect of a SNP on the binding affinity of a transcription factor

For a given SNP V, we estimated its effect on the binding of a particular transcription factor t, denoted by the matching score \( S_{V}(V) \), by evaluating the binding affinity differences between reference and alternative alleles; the binding affinity was calculated using the position weight matrices (PWMs) documented in the TRANSFAC database using previously published methods (Wang et al., 2008).

\[
S_{V}(V) = \max_{i} \left( \sum_{j=1}^{T} \log \left( \frac{\frac{1}{2} \sum_{a=1}^{2} \sum_{a'=1}^{2} \sum_{b=1}^{B} \sum_{i=1}^{I} \frac{N_{a,a',b}}{N_{a',a,b}}}{\frac{1}{2} \sum_{a=1}^{2} \sum_{a'=1}^{2} \sum_{b=1}^{B} \sum_{i=1}^{I} \frac{N_{a,a',b}}{N_{a',a,b}}} \right) \right) \]

where \( w \) is the width (base pair) of the binding site, \( k \) represents the index of the 2w potential binding sites that contain the candidate variants on both the positive and negative strands, \( N_{i} \) is the total number of experimentally validated binding sequences for each TFBS in the TRANSFAC database, \( |N_{i}| \) is the number of counts of the i-th nucleotide at the j-th position in the PWM; \( d_{i} \) represents the percentage of the i-th nucleotide (A, C, G or T) in the human genome.

For each TFBS, we calculated the distributions of matching scores for both binding (\( b_{+} \)) and non-binding (\( b_{-} \)) events, based on the matching scores with the experimentally determined binding sites (documented in the TRANSFAC database) and randomly selected genomic sequences. The potential for a specific variant to change binding affinity of a TF was estimated as:

\[
\Delta S_{V}(V) = \log \left( \frac{\int_{SV}^{fb} SV_{i} \times SV_{j} \times SV_{k} \times SV_{l} \times SV_{m}}{\int_{SVA}^{fn} SVA_{i} \times SVA_{j} \times SVA_{k} \times SVA_{l} \times SVA_{m}} \right) \]

where \( S_{SV} \) and \( S_{SVA} \) denotes the matching scores [defined in Equation (1)] of the specific transcription factor (\( t \)) binding sites on the reference and alternative alleles, respectively. Equation (2) therefore represents the logarithmic odds ratio that two alleles fell within different distributions (binding or non-binding). A positive and negative \( \Delta S_{V} \) implies that the alternative allele will result in a gain or loss of binding affinity, respectively. For each TF binding site, a P-value \( P_{SV}(V) \) was further calculated for each variant, by calculating the percentage of 10,000 randomly selected HapMap (Gibbs et al., 2003) promoter SNPs that have \( \Delta S_{V} \) scores greater than the variant \( \Delta S_{V} \) being evaluated.

2.3 Final ranking estimation for each variant

We combined the rankings from the previous two steps of regSNPs to derive a final score, \( PS(V) \), for each SNP; this score weights the likelihood that the observed SNP affects the binding of a transcription factor as well as the likelihood that the transcription factors are important in the corresponding phenotype.

\[
PS(V) = \min \left( 1 - (1 - P_{SV}(V)) \times (1 - P_{tf}(V)) \right) \]

where \( \psi \) represents all the transcription factors in the TRANSFAC database, \( P_{SV}(V) \) indicates the significance of effects on binding affinity of transcription factor \( t \), while \( P_{tf}(V) \) denotes the ranking scores, as the output of Endeavour, a knowledge-learning tool for gene prioritization (Aerts et al., 2006), to represent the significance of prioritized gene (transcription factor) \( i \). A lower PS score implies a stronger relationship between the candidate SNP and the disease/phenotype being studied.

2.4 ROC curve of each disease

One thousand iterations, using a different negative set of randomly sampled regulatory SNPs were generated for each of the 13 disease states (e.g., diabetes) under study. For each iteration, we first ranked all candidate variants (both experimentally validated and randomly selected) by their final PS
We propose a novel, integrative approach that employs a SLC39A1 The 2-kb region upstream of were considered statistically significant. Luciferase Assay kit (New England Biolabs). Data from three Gaussia in Centro LB 960 Luminometer (Berthold Technologies, Oak Ridge, TN, Hochberg, 1995). was further calculated using Benjamini–Hochberg methods (Benjamini and post-transfection (20 μl) were assayed in triplicate, using Lipofectamine 2000 (Invitrogen). Media collected at 24-h 3 RESULTS

3.1 Principles of SNP prioritization by regSNPs

We propose a novel, integrative approach that employs a combination of existing bioinformatics tools to prioritize known SNPs found in or around the promoters (defined as −3000 to +1000 bp from the transcriptional start site, see Section 2) of high priority candidate genes related to a specific phenotype or disease. Our strategy contains three major steps (Fig. 1):

3.1.1 Test whether the SNP is within a target sequence of a known transcription factor and prioritize its effect on binding affinity. If a SNP falls within a TFBS, we computed the binding affinities of both the reference and alternative alleles to the transcription factor, as measured by the matching scores (Syt and Syl) calculated as Equation (4) (Section 2), based on the sum of the logarithmic transformations of the frequencies of each nucleotide in a known binding site that was documented as a PWM in the TRANSFAC database (Wingender et al., 1996). This is a standard estimation of protein–DNA binding affinity. For each TFBS, we calculated the distributions of matching scores for both binding and non-binding events (Fig. 2A and Section 2). Intuitively, if one specific SNP were predicted to cause the matching score to shift from within the non-binding distribution to within the binding distribution, or vice versa, then this SNP could contribute to the gain or loss of binding affinity, respectively. Based on this principle, we calculated the effect of a SNP on the binding affinity of a transcription factor (ΔS(V)) calculated as Equation (4) in Section 2), and subsequently compared the effect to 10000 randomly selected negative SNPs, which are not known to influence transcription factor binding, to calculate a P-value (see Section 2) for quantifying the likelihood that the observed SNP would change the protein–DNA binding affinity of a particular transcription factor (Fig. 2B).

3.1.2 Prioritize transcription factors that are related to a specific disease/phenotype. In this step, we used Endeavour (Aerts et al., 2006), a knowledge-learning tool for gene prioritization, to rank

Fig. 1. Procedures of prioritization by regSNPs. In the first step (A), the candidate SNPs are prioritized by their effects on the binding affinity of each transcription factor. Second (B), a set of training genes, collected from OMIM or Ingenuity, is used to prioritize the genome-wide transcription factors with respect to the studied phenotype/disease, using Endeavour. In the third step (C), overall rankings, characterizing the transcription regulatory roles of candidate SNPs in phenotype/disease, are calculated based on the integration of ranks of SNPs in each transcription factor (step 1) and ranks of transcription factors to phenotype/disease (Step 2). In principle, if one SNP highly alters the binding of one transcription factor which is highly related to the phenotype/disease being studied, the SNP should be prioritized as one regulatory SNP with strong confidence. TSS: transcription start site.
transcription factors by their relevance to the phenotype being investigated. *Endeavour* prioritizes candidate genes on the basis of how similar a candidate gene was to a profile derived from genes already known to be involved in the specific biological process (training set). Here, human transcription factors selected from the TRANSFAC database (Supplementary Data 1) were treated as candidate genes. Genes known to be involved in certain biological disease/phenotype features were collected from OMIM (Online Mendelian Inheritance in Man) (McKusick, 2007) and/or Ingenuity Pathway Analysis (IPA) software (http://www.ingenuity.com/), and were then used as the training set. The *Endeavour* program assigned a rank to each transcription factor based upon how likely it was to be involved in the disease/phenotype being studied.

### 3.1.3 Identify a subset of SNPs that are likely to affect the binding of transcription factors important to the phenotype/disease

In this step, we defined a score for each SNP that weighted the likelihood that the observed SNP affected the binding of a transcription factor as well as the likelihood that the transcription factors were important to the phenotype/disease of interest [PS calculated as Equation (4) in Section 2]. The lower the score is, the higher the likelihood that the SNP in question would affect transcriptional regulation related to the disease/phenotype. To evaluate the statistical significance of candidate SNPs, we performed permutation analysis by randomly selecting HapMap SNPs (Gibbs et al., 2003) within promoter regions of human protein coding genes, and by randomizing the ranking of the transcription factors with respect to phenotypes (Section 2). This allowed calculation of the FDR for each candidate SNP.

### 3.2 Evaluation of regSNPs using promoter mutations in the HGMD

The HGMD (Stenson et al., 2009) documents known mutations causing human inherited disease, as well as disease-associated and/or functional polymorphisms reported in the literature. It does not include mutations that lack phenotypic consequences. To evaluate the performance of our strategy for prioritizing regulatory SNPs, we tested the ability of regSNPs to select disease-causing regulatory SNPs/mutations (positive set) from an equal number of randomly selected regulatory HapMap SNPs (negative set).

We selected for evaluation 13 different disease states (Table 1), based on having ≥9 disease-causing mutations from promoter regions recorded in HGMD and with at least two disease-associated genes in OMIM. For each disease state, we extracted the disease-causing promoter regulatory SNPs/mutations from HGMD and constructed a negative dataset of an equal number of randomly selected regulatory SNPs from the HapMap database. For example, in the case of diabetes, 40 SNPs/mutations in promoter regions that have been reported (in HGMD) to be associated with diabetes were the positive dataset, while 40 randomly selected regulatory SNPs from HapMap were used as the negative set. It should be appreciated that it is conceivable, albeit rather unlikely, that some of the selected SNPs in the negative set may have exert a functional effect in the development of diabetes. Then, the balanced evaluation set (positive set + negative set) for each respective disease state was assessed using regSNPs. A PS score [Equation (4) in Section 2] was calculated for each variant, where lower values corresponded to a more significant relationship with the disease. Then, a ROC curve was constructed based upon the true positives, true negatives, false positives and false negatives, calculated at different score thresholds. For each disease state, we performed 1000 iterations with a different negative set randomly sampled from the HapMap promoter SNPs (Section 2). The average AUC of ROCs from the 1000 iterations was plotted as shown in Figure 3A. The precision of the proposed methodology varied among the 13 diseases. Breast cancer showed the highest precision (AUC = 77.9%), whereas schizophrenia had the lowest (AUC = 53.7%).

A major advantage of *regSNPs* is that it not only selects the DNA variants that are associated with a specific phenotype, but also identifies the transcription factors whose DNA binding affinities are putatively affected by the variant in question. Supplementary Data 2 lists the 69 different transcription factors whose DNA binding affinity was predicted to be affected by the 251 selected HGMD

![Fig. 2. The effect of one SNP on the binding affinity of a specific transcription factor (A). Binding patterns of TFBS M00803 for E2F1. The red distribution represents matching scores from binding events and the blue distribution represents matching scores from non-binding events. For a candidate SNP, $S_{V}^{A}$ and $S_{V}^{B}$ characterize the binding affinities of the reference and alternative alleles to the transcription factor, respectively. Then, four intersects from these binding patterns corresponds to the candidate SNP, including $f_{A}(S_{V}^{A})$, $f_{A}(S_{V}^{B})$, $f_{B}(S_{V}^{A})$ and $f_{B}(S_{V}^{B})$. And the SNP’s effect on the binding affinity of the transcription factor ($\Delta S_{V}$) can be calculated based on the two binding patterns and four intersects, as characterized in Equation (2). (B) Distribution of $\Delta S_{V}$ score from 10,000 random SNPs which may or not take influence on M00761 binding. A two-tail P-value calculation method is used to calculate candidate SNPs' significant P-value.](image-url)
Transcription factors associated with more than two diseases are inside the circle area of disease rectangles, those with fewer are outside solid small circles. Factors, 13 diseases and 147 associations are included in this network, where factors and diseases are represented by circles and rectangles, respectively. An alphabetical rank of transcription factors was assigned in vertical (vertical axis) of the test promoter mutations in diseases. Thirteen diseases as disease states, including breast cancer, diabetes, hemophilia, prostate cancer, Alzheimer disease, obesity, thalassaemia, hypertension and myocardial infarction. In addition, a total of 10 transcription factors were found to be associated with more than 3 of the 13 disease states, including E2F1 (7), p300 (7), SOX-9 (6), p53 (6), PPARα (peroxisome proliferator-activated receptor α), VDR (vitamin D receptor), C/EBPβ (4), FoxO3A (Forkhead box O3, 4), RXRα (Retinoic acid receptor α, 4) and YY1 (4).

3.3 Prioritizing SNPs in the genetic regions associated with BMD

We applied regSNPs to prioritize regulatory SNPs within the genetic regions known to be associated with BMD from a list of candidate SNPs in the HapMap database. We previously reported that a linkage disequilibrium block on chromosome 1q is associated with BMD (Ichikawa et al., 2008). This region was initially identified through linkage analysis using spine BMD as the phenotype of interest (Econs et al., 2004). Later, SNPs were genotyped across the linkage region and evidence of association with both spine and femoral neck BMD was identified in a 230 kb LD block (Ichikawa et al., 2008). Due to the extensive LD within the region, we could not determine which of the 11 candidate genes within this region was contributing to the variation in BMD. According to the HapMap CEU population records, 51 SNPs were in gene promoter regions within this block. We then used regSNPs to ascertain which of these SNPs would be most likely to affect transcription factor binding affinities, our tacit assumption being that SNPs which affect transcription factor binding would also have a high likelihood of altering gene expression/regulation.

First, all 51 promoter SNPs within the 230 kb LD block in chromosome 1q were evaluated for their potential to affect the binding affinities of known human transcription factors. In this study, we focused on 323 human transcription factors, whose PWMs, a commonly used representation of the DNA binding motifs, were available in the TRANSFAC database version 9.2 (Wingender et al., 1996). The empirical P-value (Section 2) was calculated for each SNP-transcription factor pair.

Second, existing bioinformatics tools, including the IPA software (http://www.ingenuity.com/) and Endeavour (Aerts et al., 2006) were used to evaluate the potential impact of individual transcription factors based upon their relevance to bone biology. We used IPA software to compile a list of genes whose functions are related to bone biology. A total of 869 genes were retrieved using IPA software to evaluate the potential impact of individual transcription factors based upon their relevance to bone biology. We used IPA software to rank the 323 transcription factors according to how relevant a transcription factor was to the 869 genes that are related to biology, based upon prior information including functional annotations, protein–protein interactions, expression data and other information in literature (Aerts et al., 2006). Endeavour provided a score that evaluated the relationship between different transcription factors and bone biology. Top ranking transcription factors include p53, STAT1, p65 and E2F1 etc.; the full ranking of these transcription factors is listed in Supplementary Data 4.

Third, the regSNPs scores were calculated by jointly considering the capabilities of the 51 SNPs to affect transcription factor binding and the relevance of the affected transcription factors to bone biology (Section 2). Three SNPs received lowest scores with FDR ≤25%. These three SNPs, rs34612917, rs11265251 and rs6661009, reside in the regulatory regions of the following genes: SLC39A1 (solute carrier family 39, zinc transporter, member 1),
A, 41.2% C) exhibit much higher minor allele frequencies, which may give rise to gain of binding of p65 and E2F1, respectively. RU: relative units. * alternative alleles that are predicted to disrupt binding of transcription factors affect the binding affinities of p53 (tumor protein p53), p65 (v-rel reticuloendotheliosis viral oncogene homolog A) and E2F1 (E2F transcription factor 1), respectively. The sequence logos, a form of graphical representation of nucleic acid multiple sequence alignment (Schneider and Stephens, 1990), for the binding sites of these three transcription factors, rsID of SNPs and allele frequencies of SNPs are shown in Figure 4A. Clearly, all three SNP sites are located at positions that allow less degeneracy (i.e. larger letters), and therefore have the potential to dramatically alter the binding affinities of their cognate transcription factors. The allele frequencies of these three sites vary quite markedly. Thus, with rs34612917, only 0.5% of the Caucasian population contains the T allele, a variant that may cause the loss of binding of p53. The other two SNPs, rs11265251 (60.6% A, 39.4% G) and rs6661009 (58.8% A, 41.2% C) exhibit much higher minor allele frequencies, which may give rise to gain of binding of p65 and E2F1, respectively. Therefore, we have the potential to dramatically alter the binding affinities of their cognate transcription factors. We tested the promoter activity. Consistent with this interpretation, the expression level of the TPM3 gene was found to be low in kidney cells and in all fetal tissues, as documented in the microarray gene expression database at the UCSC Genome Browser website (Fujita et al., 2011). In contrast, luciferase assays demonstrated that SNP rs6661009 significantly altered the activity of the SLC39A1 promoter (Fig. 4B). It has been reported that E2F1 can serve as both positive and negative regulator on their target genes (Crowe et al., 2001; Croxton et al., 2002; Stanelle et al., 2002).

4 DISCUSSION

In this study, we proposed a three-step bioinformatics approach to identify functional SNPs within the regulatory regions following GWAS. Promoter regions were used as an example of regulatory regions to illustrate the procedures involved. First, we tested whether the identified variant is within the target sequence of a known transcription factor and hence could affect its binding affinity. Second, we prioritized those transcription factors that are related to a specific phenotype. Third, we identified a subset of sequence variants that are likely to affect the binding of a transcription factors important to the disease/phenotype under study. We tested the performance of regSNPs in identifying the promoter mutations in 13 diseases, using experimentally validated SNPs from HMGD. The results of this analysis showed that the prioritization was sufficiently high and could be useful, at least in some disease states (e.g. breast cancer: 77.9%). Our analysis also shows that some transcription factors (e.g. HNF4α) may play a key role in multiple disease states. Finally, our model identified three putative functional promoter SNPs, with a FDR estimated to be 25%, in a region that influences BMD. We demonstrated that one of these SNPs, rs6661009, significantly altered luciferase activity between the A and C alleles, suggesting that regSNPs has identified a causal SNP associated with BMD. The direction of effect, in which the variant predicted to reduce binding of a transcription factor increased promoter activity, could be due to assembly of an alternative promoter complex in the cells used. It is also reported that the E2F1 binding can be associated with both positive and negative regulation.

Table 2. Top SNPs in a bone-related LD block in chromosome 1q

| SNP   | PS(%) | Gene               | Location | TFBS | TF  | FDR  |
|-------|-------|--------------------|----------|------|-----|------|
| rs34612917 | 0.00062 | SLC39A1            | −1692    | M00761 | p53 | 0.25 |
| rs11265251 | 0.00098 | TPM3              | −1365    | M00854 | p65 | 0.25 |
| rs6661009  | 0.00110 | SLC39A1/CREB3L4    | −551/342 | M00803 | E2F1 | 0.25 |

Three SNPs are significant at FDR <25%. Score is the final PS score of prioritization; location is relative to transcription start site (TSS); while negative location indicates upstream of TSS and positive location indicates downstream of TSS; TFBS is the identity of the TFBS in TRANSFAC database; TF is the name of the transcription factor; FDR is FDR based on permutation analysis.

Fig. 4. Top prioritized SNPs in BMD and luciferase reporter assay validation. (A) Detailed binding alterations by three top prioritized SNPs in BMD. Three transcription factors, with detailed disrupted positions in their binding sites, are shown in the form of sequence logos. The names of transcription factors, rsID of SNPs and allele frequencies of SNPs are marked. For each SNP, the left allele and right allele with an arrow separated, are the reference and alternative alleles, respectively, based on Caucasian population in HapMap. (B) Relative promoter strengths determined by luciferase reporter assay. RU: relative units. *P < 0.01 compared with alternative alleles that are predicted to disrupt binding of transcription factors.
on the target genes (Crowe et al., 2001; Croxton et al., 2002; Stanelle et al., 2002).

regSNPs offers several major improvements over other methods used for SNP prioritization. First, it is a phenotype-specific approach to prioritize functional regulatory variants, while most existing methods (Conde et al., 2006; Lee and Shatkay, 2009; Pico et al., 2009; Yuan et al., 2006) only generate putative candidates based on the sequence composition of the TFBSs, irrespective of whether the specific transcription factor is functionally important in the disease or phenotypic status. By integrating different bioinformatics tools and databases (Endeavour and IPA/Omim), regSNPs has the advantage of narrowing the search to those transcription factors that are likely to influence the phenotype of interest. Second, unlike the previous methods, which only yield the putative SNPs, regSNPs identifies SNPs as well as the transcription factors which are most likely to influence the phenotype. The identified transcription factors associated with specific phenotypes provide candidate proteins for further research. Third, statistical evaluation is included in the regSNPs output that allows quantification of the false positive rate that can be placed in the prioritizations. Thus, at each step, a P-value indicates the likelihood that a given SNP will alter transcription factor-binding affinity, the likelihood of an association between a transcription factor and a specific phenotype, and the likelihood of a particular SNP exerting a regulatory role in the context of a specific phenotype. This statistical information will enable investigators to prioritize particular genes/SNPs for more in depth molecular study.

When we used regSNPs to analyze promoter mutations from HOMD to test the program, we found variable utility across disease states, ranging from 53.7% (schizophrenia) to 77.9% (breast cancer). When we used regSNPs to analyze promoter mutations from HOMD to test the program, we found variable utility across disease states, ranging from 53.7% (schizophrenia) to 77.9% (breast cancer).

regSNPs prioritizes candidate SNPs identified by the GWAS, or novel or rare transcription factors. Despite these, (e.g. disease/phenotype-related genes, and binding models for likely to be more sensitive to the influence of polymorphic variation are more heavily reliant upon genetic regulation, and are therefore likely to be more sensitive to the influence of polymorphic variation within promoter regions. Breast cancer is a widely studied disease, reported to be highly associated with gene regulation (Dunning et al., 1999). Thus, when ranking validated disease-associated promoter mutations (positives) with randomly selected SNPs (negatives), reasonably high precision can be achieved. In contrast, in the case of complex traits, such as hypertension, their pathophysiological mechanisms involve multiple genetic factors together with a potential environmental influence (Kosugi et al., 2009; Rahmouni et al., 2005; Risser et al., 2009; Sesso et al., 2008); in such cases, the ability of regulatory variant identification may be limited.

Limitations of regSNPs include its reliance on existing knowledge (e.g. disease/phenotype-related genes, and binding models for transcription factors). Despite these, regSNPs provides an important strategy for prioritizing causal regulatory DNA variants with respect to specific disease/phenotype of interest. It is especially valuable for prioritizing candidate SNPs identified by the GWAS, or novel or rare variants discovered from the direct human genome sequencing.

ACKNOWLEDGEMENTS

The authors thank the valuable discussion from Dr Lang Li in the Center for Computational Biology and Bioinformatics at Indiana University School of Medicine.

Funding: National Institutes of Health grants AA017941 and AG041517.

Conflict of Interest: none declared.

REFERENCES

Artsi, S. et al. (2006) Gene prioritization through genomic data fusion. Nat. Biotechnol., 24, 837–844.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B, 57, 289–300.

Capuzzi, P. et al. (2010) Large-scale associations of non-synonymous SNPs with complex traits: none declared.

Benjamin, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B, 57, 289–300.

Capuzzi, P. et al. (2010) Large-scale associations of non-synonymous SNPs with complex traits: none declared.

Benjamin, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B, 57, 289–300.
M. Teng et al.

Pico, A. R. et al. (2009) SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. *Nucleic Acids Res.*, 37, D803–D809.

Rahmouni, K. et al. (2005) Obesity-associated hypertension: new insights into mechanisms. *Hypertension*, 45, 9–14.

Ramsisky, V. et al. (2002) Human non-synonymous SNPs server and survey. *Nucleic Acids Res.*, 30, 3994–3999.

Reamer, J. et al. (2006) SPN effect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, 22, 2183–2185.

Riemsma, U. et al. (2009) Dietary fats and prevention of type 2 diabetes. *Prog. Lipid Res.*, 48, 44–51.

Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18, 6097–6100.

Sesso, H. D. et al. (2008) Alcohol consumption and the risk of hypertension in women and men. *Hypertension*, 51, 1080–1087.

Stanic, J. et al. (2002) Gene expression changes in response to E2F1 activation. *Nucleic Acids Res.*, 30, 1859–1867.

Stimson, P. D. et al. (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.*, 1, 13.

Stitzel, L. O. et al. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, 32, D520–D522.

Wang, G. et al. (2008) Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics*, 9 (Suppl. 2), S22.

Wingender, E. et al. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, 24, 238–241.

Yuan, H. Y. 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, F. et al. (2006) SNP-3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7, 166.