A Survey of Genomic Properties for the Detection of Regulatory Polymorphisms

Stephen B. Montgomery[¤*], Obi L. Griffith, Johanna M. Schuetz, Angela Brooks-Wilson, Steven J. M. Jones

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada

Advances in the computational identification of functional noncoding polymorphisms will aid in cataloging novel determinants of health and identifying genetic variants that explain human evolution. To date, however, the development and evaluation of such techniques has been limited by the availability of known regulatory polymorphisms. We have attempted to address this by assembling, from the literature, a computationally tractable set of regulatory polymorphisms within the ORegAnno database (http://www.oreganno.org). We have further used 104 regulatory single-nucleotide polymorphisms from this set and 951 polymorphisms of unknown function, from 2-kb and 152-bp noncoding upstream regions of genes, to investigate the discriminatory potential of 23 properties related to gene regulation and population genetics. Among the most important properties detected in this region are distance to transcription start site, local repetitive content, sequence conservation, minor and derived allele frequencies, and presence of a CpG island. We further used the entire set of properties to evaluate their collective performance in detecting regulatory polymorphisms. Using a 10-fold cross-validation approach, we were able to achieve a sensitivity and specificity of 0.82 and 0.71, respectively, and we show that this performance is strongly influenced by the distance to the transcription start site.

Citation: Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJM (2007) A survey of genomic properties for the detection of regulatory polymorphisms. PLoS Comput Biol 3(6): e106. doi:10.1371/journal.pcbi.0030106

Introduction

Our ability to identify the molecular mechanisms responsible for specific genetic traits within our population will be enhanced by our imminent ability to decipher each individual's genome. This is evident from recent advances in sequencing and genotyping technologies, which allow an increasing number of variants to be sampled for association and linkage (reviewed in [1–3]) and contribute a growing number of sources of variation and their frequencies to public databases each year. As new variants are identified, each becomes a molecular window into our past, present, and future—each aids in tracing our genetic heritage and in charting the footsteps of our common evolution, and possesses the potential to predict disease or drug susceptibilities, ideally acting as an early-warning system in preventative medical practice (reviewed in [4,5]). However, our ability to catalog genotypes has far outstripped our ability to implicate them in phenotypes. Currently, more than 6 million unique single-nucleotide polymorphisms (SNPs) are included in version 126 of dbSNP [6]; of these SNPs, only a very small fraction have been associated with a phenotype using genetic association or linkage analysis. This is because association studies are costly, time-consuming, and dependent on the frequency of the genotype in the sampled population. Furthermore, many SNPs are not necessarily expected to have a function. To select candidates for functional validation, computational methods have been developed to identify SNPs that alter the protein-coding structure of genes [7–16]. These types of computational methods tend to prioritize putative functional SNPs by identifying those SNPs that alter a protein's amino acid sequence, are located within well-conserved regions or functional protein domains, and alter the biochemical structure of the protein. However, very few methods identify regulatory SNPs (rSNPs) that alter the expression of genes. Such rSNPs have been implicated in the etiology of several human diseases, including cancer [17,18], depression [19], systemic lupus erythematosus [20], perinatal HIV-1 transmission [21], and response to type 1 interferons [22]. This work aims to extend computer-based techniques to identify this particular class of functional variants within the core promoter regions of human genes.

Conventional computational approaches to rSNP classification have predominantly relied on allele-specific differences in the scoring of transcription factor weight matrices as supplied from databases such as TRANSFAC and Jaspar [15,16,23]. SNPs located within matrix positions possessing high information content are assumed more likely to be functional. Support for this hypothesis to date, however, has been restricted to single-case examples. Furthermore, a recent study has failed to detect significant weight matrix signals in 65% of regulatory polymorphisms ($n = 40$) [24]. However, the prevailing hypothesis in computational regulatory element prediction has been that the majority of predictions using unrestricted application of matrix-based

Editor: Yitzhak Pilpel, Weizmann Institute of Science, Israel

Received October 16, 2006; **Accepted** April 25, 2007; **Published** June 8, 2007

A previous version of this article appeared as an Early Online Release on April 25, 2007 (doi:10.1371/journal.pcbi.0030106.eor).

Copyright: © 2007 Montgomery et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: MAF, minor allele frequency; ROC, receiver operating characteristic; rSNP, regulatory single-nucleotide polymorphism; SVM, support vector machine; TSS, transcription start site; ufSNP, SNP of unknown function

* To whom correspondence should be addressed. E-mail: sm8@sanger.ac.uk

¤ Current address: Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

Author Summary

Computational techniques are used in biology to prioritize DNA sequence variants (or polymorphisms) that may be responsible for population diversity and the manifestation of species-specific traits. Predominantly, they have been used to predict the class of polymorphisms that alter protein function through allele-specific changes to amino acid composition. However, polymorphisms that alter gene expression have been increasingly implicated in manifestation of similar traits. Prioritization of these polymorphisms is challenged, though, by the lack of knowledge regarding the mechanisms of gene regulation and the paucity of characterized regulatory polymorphisms. Our work attempts to address this issue by assembling a collection of regulatory polymorphisms from the existing literature. Furthermore, we use this collection to investigate and prioritize various properties that may be important for identifying novel regulatory polymorphisms.

approaches are false positives. By extending this technique and using phylogenetic footprinting between mouse and human, it was demonstrated that from ten SNPs that show significant allele-specific differences in Jaspar predictions, seven also demonstrated electrophoretic mobility shift differences [23]. However, only two of the seven had a marked effect in reporter gene assays. Conservation alone has also been demonstrated as a poor discriminant of function in a study of regulatory polymorphisms in Eukaryotic Promoter Database promoters, where zero of ten experimentally validated regulatory variants were in conserved binding sites [25].

A substantial challenge with developing strategies for identifying functional noncoding variants has been the shortage of characterized regulatory variants. Few studies have successfully identified the causative variant(s) after a susceptibility haplotype is identified. To address this problem, we have assembled the largest openly available collection of functional regulatory polymorphisms within the ORegAnno database (http://www.oreganno.org) [26]. From this dataset, we have examined several features of these SNPs as they relate to polymorphisms of unknown function (ufSNPs) within the promoter regions of associated genes (up to 2 kb). Our hypothesis is that using a combination of regulatory and population genetics properties, the discriminative efficacy of individual properties can be evaluated, and significant predictors of rSNP function can be chosen. Within our assayed set, we have found that the best discriminants are the distance to the transcription start site (TSS), local repetitive density and content, sequence conservation, minor allele frequency (MAF) and derived allele frequency, and CpG island presence. Notably, the unrestricted application of a matrix-based approach is demonstrated to be one of the least effective classifiers.

We have used this dataset of rSNPs and their properties to train a support vector machine (SVM) classifier. Two approaches were used to train the classifier: one in which the properties of all rSNPs were compared with that of all the ufSNPs, and one in which each property value of the positive SNPs and ufSNPs within an associated gene were compared with the average values for each property within that gene (referred to here as the "All" and "Group" approaches, respectively). The All approach is designed to determine if there are any properties that are important across the test set, while the Group approach is designed to determine if there are important directional shifts in values within a promoter that may discriminate functional SNPs from ufSNPs. In a 10-fold cross-validated test, the SVM achieves a receiver operating characteristic (ROC) value of 0.83 $\pm$ 0.05 for the All analysis (sensitivity, 0.82 $\pm$ 0.08; specificity, 0.71 $\pm$ 0.13) and 0.78 $\pm$ 0.04 for the Group analysis (sensitivity, 0.72 $\pm$ 0.19; specificity, 0.68 $\pm$ 0.07).

Methods

Data

Literature describing noncoding polymorphisms responsible for allele-specific differences in gene expression was surveyed from PubMed [27]. From this literature, 160 regulatory polymorphisms were identified in 103 publications; each was selected based on experimental evidence that confirmed its direct role in altering gene expression. This selection criterion specifically excluded those polymorphisms in which the experimental evidence could only confirm that the reported polymorphism was in linkage disequilibrium with an rSNP. Each identified rSNP was manually curated in the ORegAnno database. Subsequently, 104 polymorphisms were selected based on the criteria that they were SNPs (excluding seven insertion–deletion polymorphisms), and were within 2 kb of the TSS of their associated gene (as annotated in version 37 of EnsEMBL [28]; Table 1). A 2-kb region was chosen to maximize the number of rSNPs included while minimizing the size of sequence investigated; at 2 kb, the addition of a single further rSNP would increase the surveyed region by 43%, whereas the previous addition resulted in an increase of 9%. At this window size, 39 rSNPs were excluded from analysis. An additional ten polymorphisms were excluded because of deprecated annotation of the gene or discordant genomic location with the associated gene. In total, the remaining 104-rSNP set contained polymorphisms involved in altering the expression of 78 different transcripts.

Using each of the 78 transcripts, SNPs within 2 kb of the TSS were extracted from version 37 of EnsEMBL (dbSNP version 125), producing exactly 951 ufSNPs. The ufSNP and rSNP genomic locations have been mapped (see Table S1).

Investigated Properties

A total of 23 different properties of relevance to assessing regulatory function were calculated for each SNP in both the 104-rSNP and ufSNP sets (Table 1). These properties were selected to represent a cross-section of well-documented methodologies for assessing the functional significance of both allele-specific changes and DNA sequences within noncoding regions.

Test Data Design (All and Group)

Two types of analyses were conducted using the investigated properties. One was an all-versus-all approach, where the 104-rSNP and ufSNP sets were compared en masse. The other was a group analysis, where the average value of each property within each upstream noncoding region was first calculated, and then the individual SNP properties within that region were recalculated as the difference from this average. The All test data were designed to identify global characteristics of rSNPs, while the Group test data were

Table 1. Investigated Properties

Property Number	Investigated Property	Type	Methodology	Description
1	TRANSFAC	Allele-specific	Database, matrix similarity	An allele-specific TRANSFAC (version 7.2) analysis was performed by individually running TRANSFAC (for all transcription factor–binding matrices with a prediction level cutoff of 80%) for both alleles and calculating the absolute cumulative difference in predicted binding site scores. $\Delta s = \Sigma_{factor=n;n\in p}^{p=i,j} |(score(r)_{factor} - score(v)_{factor}|$. Here, Δs is the absolute cumulative difference in the predicted binding site scores between the set of predicted factors, *i*, from the reference allele and the set of predicted factors from the variant allele, *j*. For example, in situations where a binding site is predicted for both alleles, the calculated score is the magnitude of the difference between the allele-specific scores; it is the absolute difference between *score*(*r*) and *score*(*v*). If a binding site, however, is predicted for only one allele, the magnitude is the value of the prediction score (either *score*(*r*) or *score*(*v*)). This calculation generalizes many similar, previously reported methods based on allele-specific weight matrix calculations [16,23,38].
2	oPOSSUM	Allele-specific	Database, matrix similarity, coexpression	Coexpression data was extracted from the Tmm coexpression set published by Pavlidis et al. [39]. This set was chosen because it comprises a large cross-section of microarray experiments from various human cell lines. For each target gene, coexpressed genes were broadly selected based on at least one study reporting coexpression (i.e., Tmm score $\geq$ 1). oPOSSUM was run to short-list a set of transcription factor–binding matrices for allele-specific analysis (as in the TRANSFAC test above) [40]. This property was designed to assess whether a subset of transcription factor–binding sites selected based on biological relevance would improve assessing the functional significance, if any, of the allele-specific changes.
3	Weeder (difference)	Allele-specific	Motif discovery, evolutionary conservation	For each SNP, a 1-kb, evenly flanking, DNA sequence was retrieved from EnsEMBL (NCBI35). The EnsEMBL compara database was subsequently used to retrieve precalculated orthologous sequences from completed genomes (using BLASTZ_NET [41]); specifically, sequences from chimpanzee, rhesus macaque, mouse, dog, rat, and chicken were used. A Weeder [42] and MotifSampler [43] analysis was performed by separately inputting both canonical and variant human sequences (the 1-kb sequences with the respective alleles in situ) with the set of associated orthologues and separately recording the difference in predicted scores (difference) and the maximum score (maximum) for predicted motifs overlapping the polymorphism. The difference score was used to measure how an allele-specific change affects scoring. The maximum score was used to measure whether the polymorphism was in a high-scoring motif (regardless of allele). To improve the probability of detecting the desired motif, Weeder was set to detect 500 motifs, and MotifSampler was seeded with 25 bp around the polymorphism. For MotifSampler, a background file was supplied containing 745 regulatory regions annotated in ORegAnno as of January 2006 (supplied as Text S1). Weeder and MotifSampler were both selected because of their different approaches to motif discovery (Weeder is enumerative and MotifSampler is based on optimizing an objective function) and because they have been previously demonstrated to have moderately complementary performance characteristics [44]. A 1-kb region was selected to allow duplicated motifs to contribute to the scoring function and to permit relaxed positional constraint on contributing motif location.
4	Weeder (maximum)	Allele-specific	Motif discovery, evolutionary conservation	(same as above)
5	MotifSampler (difference)	Allele-specific	Motif discovery, evolutionary conservation	(same as above)
6	MotifSampler (maximum)	Allele-specific	Motif discovery, evolutionary conservation	(same as above)
7	DNA bendability	Allele-specific	DNA structure, sequence composition	A DNA bendability and curvature analysis was performed on canonical and variant sequences (the 1-kb sequences assembled for Weeder and MotifSampler, above) using an implementation of the BEND algorithm called "banana" and packaged in the EMBOSS toolkit [45,46]. "Banana" predicts bending and curvature of a normal B-DNA double helix. The magnitude of the allele-specific difference between each was reported. The effects of DNA structure on gene regulation in mammalian systems remains largely unascertained; however, previous characterization in bacterial systems has demonstrated its role in creating conditions suitable for transcription factor binding [47,48].
8	DNA curvature	Allele-specific	DNA structure, sequence composition	(same as above)

Table 1. Continued.

Property Number	Investigated Property	Type	Methodology	Description
9	GC content	Allele-specific	DNA structure, sequence composition	The effects on local GC content and thermodynamic stability (melting temperature) of the DNA sequence were assessed using the "dan" application packaged in the EMBOSS toolkit [44]. For thermodynamic stability calculations, "dan" uses free energy values calculated from nearest-neighbor thermodynamics [49,50]. The presence of functional transcription factor–binding sites in GC-rich sequences has been previously demonstrated [51,52]. Similar to analyzing DNA bending and curvature, we used thermodynamic stability calculations to measure whether allele-specific changes to the kinetics of the DNA sequence would be functionally constrained.
10	DNA thermodynamics	Allele-specific	DNA structure, sequence composition	(same as above)
11	Minor allele frequency	Allele-specific	Population property	MAFs were obtained from dbSNP (version 125) directly using the "eutils" service. Each allele frequency was calculated by averaging frequencies across all available populations. Derived alleles were obtained by aligning a 1-kb human region centered on the polymorphism with orthologous chimpanzee sequence in ClustalW. They were then matched with previously calculated allele frequencies. A total of 79 of 104 rSNPs and 502 of 968 ufSNPs had genotype data.
12	Derived allele frequency	Allele-specific	Population property, evolutionary conservation	(same as above)
13	Local repetitive base percentage	Sequence	Sequence characteristic	Local repetitive content of a 200-bp DNA segment centered on the assayed polymorphism was calculated using repetitive annotation curated in EnsEMBL. Four different metrics were assessed in this region: (1) the percentage of repetitive bases; (2) whether the polymorphism was in a repeat or not; (3) the number of repeats of length greater than 1 kb; and (4) length less than 1 kb that overlaps this region (we made this distinction as an estimate of the disruptive potential of smaller versus larger repeats). Repetitive content was investigated in this study because of its known role in altering gene regulation and mirroring selective constraint in noncoding regions [53–56]. Each value was normalized to its expectancy at the calculated distance from the TSS in the associated chromosome (see "Distance Normalization").
14	In repeat	Sequence	Sequence characteristic	(same as above)
15	Short repeat events	Sequence	Sequence characteristic	(same as above)
16	Long repeat events	Sequence	Sequence characteristic	(same as above)
17	Distance to TSS	Sequence	Regulatory sequence characteristic	The distance to the TSS, as annotated by EnsEMBL, was recorded. Distance to TSS has been previously identified as a significant discriminant of regulatory polymorphisms; a study of 674 haplotypes in 247 gene promoters reported that sequence variants altering expression by 1.5-fold or more are preferentially located within the first 100 bp [24]. Both the raw distance and the logarithm of the distance were used. We hypothesized that the logarithm of the distance to the TSS might more naturally reflect this properties importance within the promoter region. The logarithm of the distance was not included in SVM training.
18	Distance to TSS (log)	Sequence	Regulatory sequence characteristic	(same as above)
19	In CpG island	Sequence	Regulatory sequence characteristic	CpG islands were obtained from annotation in the UCSC Genome Browser [57]. Whether or not a polymorphism was in a CpG island was recorded. This value was normalized to its expectancy at the calculated distance from the TSS in the associated chromosome (see "Distance Normalization").
20	DNaseI hypersensitive site	Sequence	Regulatory sequence characteristic	DNaseI hypersensitive sites were obtained from predicted sites as per Noble et al. [30]. These sites were mapped from hg15 to hg17 coordinates using blast. Whether or not a polymorphism was in a DNaseI hypersensitive site was recorded. These values were normalized to its expectancy at the calculated distance from the TSS in the associated chromosome (see "Distance Normalization").
21	PhastCons	Sequence	Regulatory sequence characteristic, evolutionary conservation	Conservation scores from both the PhastCons [58] and Regulatory Potential (RP) [59] methods were obtained from the UCSC Genome Browser. The local conservation of the polymorphism, as calculated by these scores, was recorded. PhastCons and RP scores were selected to mirror what a typical UCSC Genome Browser user would use to assess genome conservation when prioritizing potential rSNPs. These values were normalized to their expectancy at the calculated distance from the TSS in the associated chromosome (see "Distance Normalization").
22	RP (Regulatory Potential)	Sequence	Regulatory sequence characteristic, evolutionary conservation	(same as above)

Table 1. Continued.

Property Number	Investigated Property	Type	Methodology	Description
23	ClustalW alignment distance	Sequence	Evolutionary conservation	Each orthologous sequence set for an individual polymorphism was aligned using ClustalW [60], and the total evolutionary distance was calculated from the associated phylogenetic tree. Since orthologs were retrieved in a standardized way from the EnsEMBL compara database, the total evolutionary distance is comparable as a measure of sequence mutability. For example, conserved sequences should have a low evolutionary distance as computed from their ClustalW alignment, whereas variable regions should have a high evolutionary distance.

The properties are broken down into two types: allele-specific and sequence. Allele-specific properties are calculated as a difference in property values calculated by allele, and sequence properties are properties of the genome location in which the SNP is located.
doi:10.1371/journal.pcbi.0030106.t001

designed to look for directional trends within the sampled region that might be indicative of SNP importance. For example, the All test is able to ask whether rSNPs have generic features that would distinguish them from any other promoter SNP; the Group test is designed to identify whether there are any features that distinguish rSNPs from other SNPs within the same upstream noncoding region.

SVM

The All and Group test data were input to the Gist SVM implementation [29]. We excluded the logarithmic distance to the TSS to prevent redundant classification with the raw distance to the TSS. Gist was run using the default parameters as described previously [30]. Of note, the Gist SVM requires that every value in the test and training parameter space is filled. To reflect the null hypothesis, that there are no differences between the ufSNPs and rSNPs, the All SVM was filled with promoter-specific average values wherever data could not be calculated. Likewise, the Group SVM was filled with zero values wherever data could not be calculated, indicating no divergence from average within the GROUP test set.

Performance Measurement

The individual importance of each property in discriminating regulatory polymorphisms was assessed in the All and Group test sets using a Wilcoxon rank sum test. Each value was corrected for multiple testing using the BioConductor MTP package (http://www.bioconductor.org) by controlling for the family-wise error rate ($\alpha = 0.05$ and $B = 10{,}000$) [31,32].

The performance of the Gist SVM classifier was measured using a ROC curve. ROC scores of 1 indicate perfect discrimination, while those at 0.5 indicate random classification of the input SNPs. ROC performance measurements have been previously described in detail elsewhere [30].

A 10-fold cross-validation was performed to assess the overall performance of the SVM. The input data was randomly partitioned by transcript into ten sets. Data from one set were excluded, and the remaining nine sets were trained on for each fold validation. This analysis was performed for each set to cover the entire training site and to calculate an average ROC value for the SVM.

Distance Normalization

We were concerned that several properties may be indirect measurements of distance from the TSS, and that any discrimination strategy would be limited to characterizing this property alone. This concern is a particular challenge since distance ascertainment bias exists; most SNPs surveyed were within a few hundred base pairs of the TSS, which is much smaller when compared with our sampling distance of 2 kb. Furthermore, it has been well established in a previous study that distance to the TSS is correlated to detection of rSNPs (it is unknown if this is because they are more likely to affect essential transcription factor–binding sites, or because there is a higher density of transcription factor–binding sites in these regions) [24]. For this reason, the discrimination potential of distance to the TSS could not be ignored. To adjust for bias, however, we calculated the expectancy of observing a feature at a particular distance from the TSS for each individual chromosome (Figure 1; CpG islands are shown as an example of this trend). This expectancy value was used to normalize the observation values for several of the properties in this study (identified in Table 1). This was performed by subtracting the expectancy value from the observed value. The impact of this normalization is negligible when comparing normalized ROC values against unnormalized ROC values; using a 10-fold cross-validation, the unnormalized ROC values for the ALL test are 0.82 ± 0.05 (unnormalized) and 0.83 ± 0.05 (normalized), and values for the GROUP test are 0.79 ± 0.04 (unnormalized) compared with 0.78 ± 0.07 (normalized).

Results

Property Ranking

A total of 104 rSNPs and 951 ufSNPs in the upstream noncoding regions of 78 genes were compiled to test properties that discriminate polymorphisms with effects on gene expression. A multiple testing–corrected Wilcoxon rank sum test was used to analyze the All test data (Table 2). Analyzing the All test data identified several properties of significance in discriminating between rSNPs and ufSNPs ($p < 0.05$). The properties of significance in the All test data, in order of importance, were: 1) distance to the TSS (properties 13 and 14); 2) in a CpG island (property 19); 3) long repeat events (property 16); 4) local repetitive base percentage (property 13); 5) derived allele frequency (property 12); 6) minor allele frequency (MAF; property 11); 7) Regulatory Potential score (property 22); 8) in a repeat (property 14); and 9) ClustalW alignment distance (property 23).

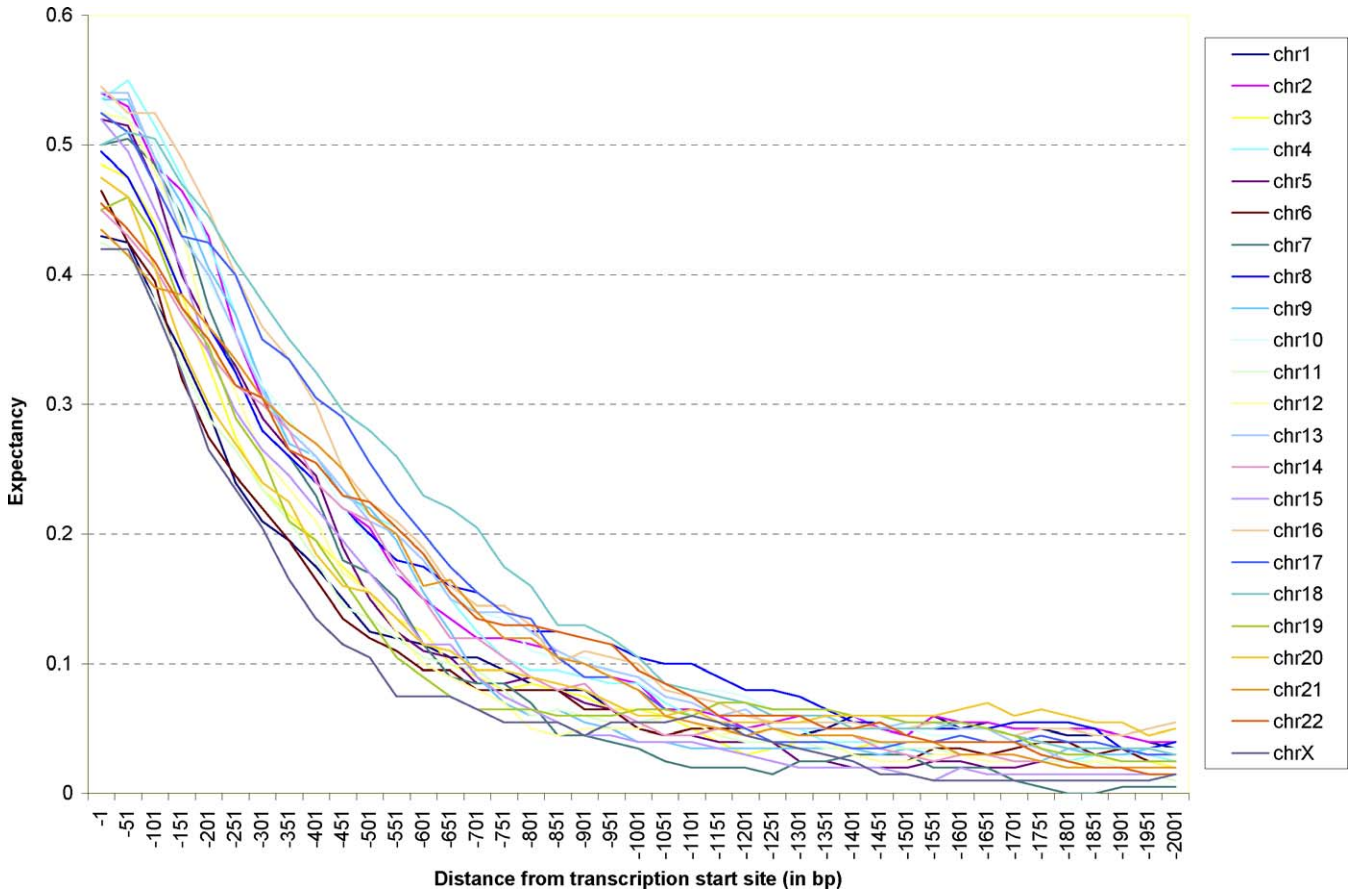

Figure 1. CpG Island Positional Bias

CpG island expectancy is plotted for each chromosome as a function of the distance from the TSS. This type of data was used to normalize many of the features in this study for distance from the TSS. In this figure, the expectancy of being in a CpG island at position −1 for any promoter region is ~0.5.
doi:10.1371/journal.pcbi.0030106.g001

However, a concern with the All analysis was that calculated property values for SNPs in individual upstream noncoding regions would not be comparable with those in other upstream noncoding regions due to differences in background property values. To address this, a multiple testing–corrected Wilcoxon rank sum test was also used to analyze the Group test data (Table 2). The properties of significance ($p < 0.05$) in the Group test data, in order of importance, were: 1) distance to the TSS (properties 13 and 14); 2) long repeat events (property 16); 3) in a CpG island (property 19); 4) MAF (property 11); 5) local repetitive base percentage (property 13); 6) ClustalW alignment distance (property 23); 7) derived allele frequency (property 12); 8) short repeat events (property 15); and 9) DNaseI hypersensitive site (property 20).

Both lists are highly concordant and demonstrate several properties that may be of utility when prioritizing SNPs for functional analysis either across the genome or within an individual upstream noncoding region. In both tests, distance to the TSS was found to be the most significant discriminant. While it is possible that ascertainment bias in the 104-rSNP set contributes to the strength of this discriminant in our study, this property has also been independently identified as an important discriminant in a previous study where, in 500-bp assayed regions, 50% of rSNPs identified through transfection experiments were within 100 bp of the TSS ($n = 40$) [24].

Furthermore, several other properties are consistently identified as being significant after normalization against distance to TSS. One property, ClustalW alignment distance, was identified in both the All and Group tests as being significant. The mean value of ClustalW alignment distance was slightly higher for the tested rSNPs compared with the ufSNPs, indicating that 1-kb multiple alignments centered on the tested rSNPs were more divergent than those centered on ufSNPs. This result is concordant with previous analyses of conservation around rSNPs ($n = 10$) [25]. However, trends in the other conservation scores used in this study, while nonsignificant in discriminating between the tested rSNP and ufSNPs, conversely suggest that the tested rSNPs are more conserved than ufSNPs. Since these metrics use tighter window sizes than those used for calculating the ClustalW alignment distance, this result suggests that increased mutation around an rSNP may be more informative than the conservation status of the rSNP itself.

Another property of significance was repetitive element content. Our results indicate that the tested rSNPs were less likely to be in or around repetitive elements. This suggests that regions that are likely to contain a transcription factor–binding site are less likely to accrue repetitive elements and be subject to dysregulation. We note, however, that ascertainment bias by which the 104-rSNPs set was surveyed in terms of repetitive elements is not known, and future collections of discovered rSNPs should address this issue.

Both MAF and derived allele frequency are also identified as significant discriminants. Unexpectedly, for genotyped

Table 2. Analysis of rSNP and ufSNP Properties in the 2-kb and 152-bp Upstream Noncoding Regions

Property Number	Investigated Property	Region	All Test			Group Test		
			Wilcoxon Raw *p*-Value	Multiple Testing–Corrected *p*-Value	Direction	Wilcoxon Raw *p*-Value	Multiple Testing–Corrected *p*-Value	Direction
1	TRANSFAC	2 kb	0.958	0.712	−	0.958	0.800	−
		152 bp	0.206	0.603	+	0.206	0.601	+
2	oPOSSUM	2 kb	0.161	0.493	−	0.316	0.576	+
		152 bp	0.576	0.816	+	0.747	0.824	−
3	Weeder (difference)	2 kb	0.862	0.712	+	0.896	0.800	+
		152 bp	0.267	0.707	+	0.323	0.695	+
4	Weeder (maximum)	2 kb	0.296	0.496	−	0.514	0.727	−
		152 bp	0.267	0.707	+	0.241	0.668	+
5	MotifSampler (difference)	2 kb	0.275	0.496	+	8.27×10^{-2}	0.313	+
		152 bp	0.308	0.714	+	0.308	0.695	+
6	MotifSampler (maximum)	2 kb	0.733	0.712	−	0.666	0.741	+
		152 bp	0.047	0.211	+	0.147	0.529	+
7	DNA bendability	2 kb	5.59×10^{-2}	0.261	−	0.101	0.387	−
		152 bp	0.975	0.816	+	1	0.858	−
8	DNA curvature	2 kb	0.201	0.496	−	0.477	0.727	−
		152 bp	0.668	0.816	−	0.915	0.858	−
9	GC content	2 kb	0.811	0.712	+	0.950	0.800	+
		152 bp	0.403	0.715	+	0.960	0.858	−
10	DNA thermodynamics	2 kb	0.201	0.496	+	0.138	0.496	+
		152 bp	0.713	0.816	+	0.892	0.858	−
11	MAF	2 kb	2.71×10^{-5}	$<1 \times 10^{-9}$	+	1.09×10^{-7}	$<1 \times 10^{-9}$	+
		152 bp	0.719	0.816	+	0.0853	0.283	+
12	Derived allele frequency	2 kb	2.53×10^{-5}	$<1 \times 10^{-9}$	+	9.52×10^{-5}	$<1 \times 10^{-9}$	+
		152 bp	1	0.828	+	0.311	0.676	+
13	Local repetitive base percentage	2 kb	1.23×10^{-7}	$<1 \times 10^{-9}$	−	3.62×10^{-6}	$<1 \times 10^{-9}$	−
		152 bp	0.290	0.587	−	0.016	$<1 \times 10^{-9}$	−
14	In repeat	2 kb	1.73×10^{-3}	$<1 \times 10^{-9}$	+	0.872	0.800	−
		152 bp	0.334	0.714	−	1	0.858	−
15	Short repeat events	2 kb	0.290	0.493	−	1.51×10^{-3}	8.70×10^{-3}	−
		152 bp	0.107	0.421	−	0.150	0.528	−
16	Long repeat events	2 kb	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	+	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	+
		152 bp	0.522	0.810	−	0.282	0.695	−
17	Distance to TSS	2 kb	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−
		152 bp	0.602	0.810	−	0.187	0.586	−
18	Distance to TSS (log)	2 kb	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−
		152 bp	0.602	0.816	−	0.575	0.776	−
19	In CpG island	2 kb	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−	$<1 \times 10^{-9}$	$<1 \times 10^{-9}$	−
		152 bp	0.705	0.816	+	0.250	0.6786	−
20	DNaseI hypersensitive site	2 kb	1.91×10^{-2}	0.118	+	4.54×10^{-3}	2.55×10^{-2}	+
		152 bp	0.165	0.587	+	0.868	0.858	+
21	PhastCons	2 kb	3.23×10^{-2}	0.188	+	0.192	0.576	+
		152 bp	0.061	0.268	+	0.150	0.529	+
22	RP (Regulatory Potential)	2 kb	2.80×10^{-5}	$<1 \times 10^{-9}$	+	0.114	0.507	−
		152 bp	0.100	0.443	+	0.554	0.777	+
23	ClustalW alignment distance	2 kb	3.68×10^{-3}	1.34×10^{-2}	+	9.64×10^{-6}	$<1 \times 10^{-9}$	+
		152 bp	0.794	0.292	+	1.3×10^{-4}	$<1 \times 10^{-9}$	+

Both All and Group test sets were analyzed using a Wilcoxon rank sum test in 2-kb and 152-bp regions. In the 2-kb region, 104 rSNPs and 951 ufSNPs were tested. In the 152-bp region, 16 rSNPs and 21 ufSNPs were tested. The 152-bp region was selected because it contained nearly equivalent mean distances from the TSS for both the rSNPs and ufSNPs under study. Each value was corrected for multiple testing using the BioConductor MTP package by controlling for the family-wise error rate ($\alpha = 0.05$ and $B = 10{,}000$).

The direction of difference between the two populations is also recorded and describes the relationship between the rSNPs and ufSNPs; + indicates that the rSNPs have higher mean values; − indicates the rSNPs have lower mean values.

doi:10.1371/journal.pcbi.0030106.t002

SNPs, the MAF was higher in the 104-rSNP set than in the ufSNP set. Previous analyses of MAF have suggested that most functional SNPs are positioned around 6% [33] or possess no allele frequency bias [24]. In this study, the average MAF was approximately 22%. Since a subset of the 104-rSNP set has been derived from association studies, it is possible that ascertainment bias may explain part of this result as researchers may preferentially be choosing higher MAF SNPs because of their greater statistical power. Of further interest, the derived allele frequency was higher in the 104-rSNP set than in the ufSNP set. This could suggest that many of the derived alleles have been driven to higher frequencies due to new variants increasing in frequency in our population, through either population bottlenecks or positive selection. The former hypothesis is supported by the supplemental observation that when restricting populations to HapMap (http://www.hapmap.org) phase I populations only, the Asian and European populations mirror this result, while the

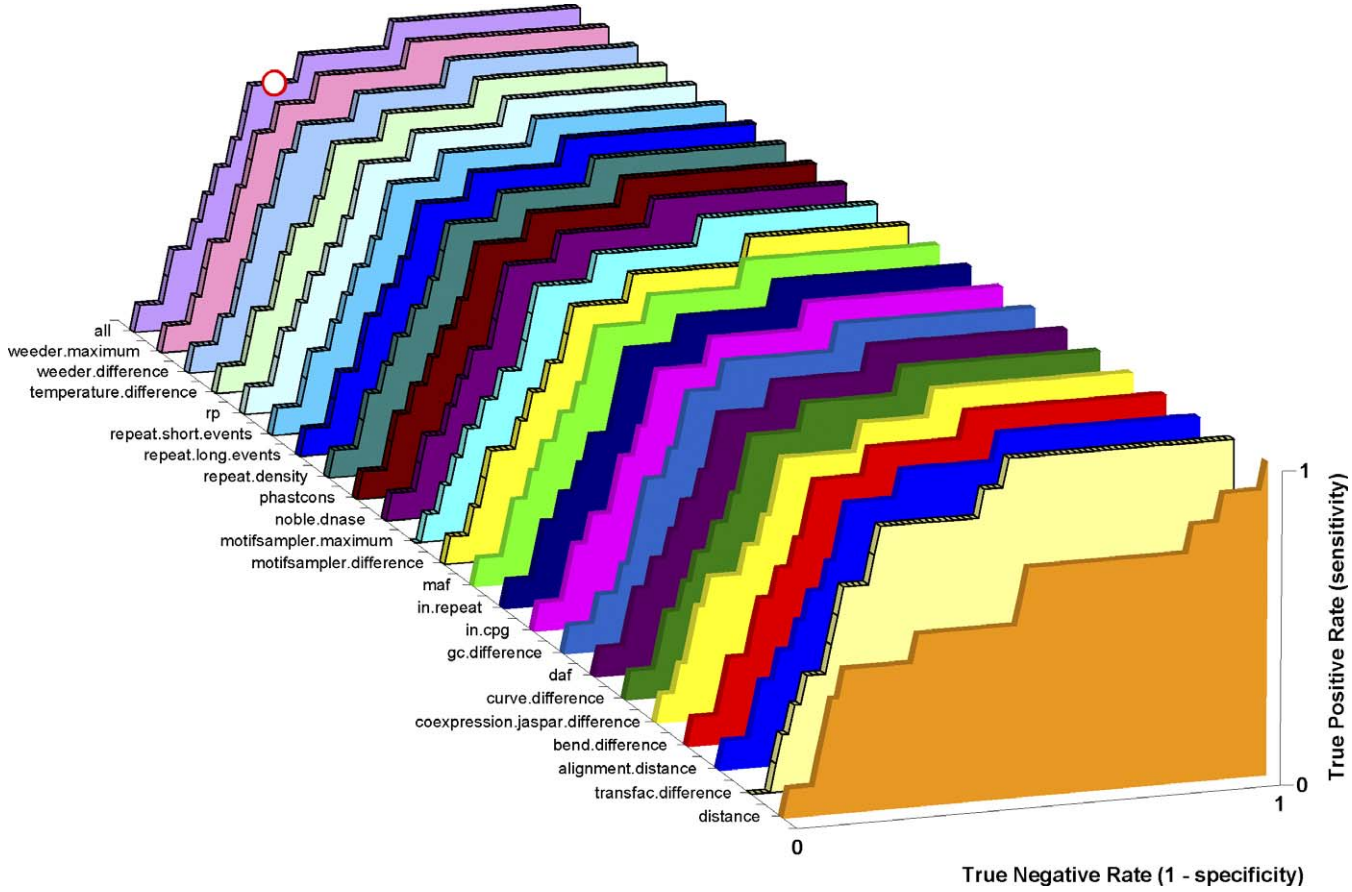

Figure 2. ROC Curves for Discriminating Known rSNPs from ufSNPs

Representative ROC curves were calculated by training an SVM on a 90% subset of the 104-rSNP and ufSNP datasets. Here, 93 rSNPs and 882 ufSNPs were used for training, followed by testing on the held-out 10%. The ALL SVM approach was used for training. Furthermore, each curve had one tested property held out to demonstrate the impact of various properties on training. Notably, many curves are the same except for a marked reduction in performance when the "Distance to TSS" property is held out. The area under the "all" curve is 0.830. The dot on the "all" curve marks the location of the decision boundary selected by the SVM. At this boundary, the SVM identifies nine of 11 true positives and 56 of 69 true negatives. (Plots for each tested partition are available at http://www.bcgsc.ca/chum/gistplots.html).

doi:10.1371/journal.pcbi.0030106.g002

African population has lower MAFs on average. The latter hypothesis, however, supports a model of evolution of genetic susceptibility to common diseases explained by ancient alleles recently becoming predisposed to disease due to changes in human lifestyle and life expectancy [34].

Another interesting result was that SNPs in the 104-rSNP set were less likely to be in CpG islands than were ufSNPs. Since CpG expectancy was normalized from average values at specific distances from the TSS of associated genes across individual chromosomes, an admixture of CpG and CpG-less promoters would drive the 104-rSNP set values lower than the ufSNP set values (Figure 1) [35,36]. However, without normalization, the significance of this value for the All and Group tests is similar (All, $p = 3.78 \times 10^{-5}$; Group, $p = 1.96 \times 10^{-3}$), suggesting that the rSNPs are in fact less likely to be in CpG islands.

Many tested properties fell below our significance threshold in these tests. Of interest, both weight matrix–based approaches did not discriminate well. In addition, our definition of coexpression was significantly broad as to allow multiple coexpressed partners for any given gene; this may have reduced the overall effectiveness of reducing transcription factor–binding profiles using this information. However, the performance of the coexpression-filtered approach was moderately better than the TRANSFAC approach alone. This suggests that targeted analysis of specific, biologically relevant transcription factors may further increase the discriminating ability of this approach. This should also act as a warning to those who have in the past applied the TRANSFAC approach to this problem indiscriminately. Furthermore, none of the DNA structural or stability analyses used were successfully discriminatory. This analysis could indicate that not only do these features have nongeneralizable effects using the data in this study, but since these analyses also measure local sequence composition, no particularly important effect is caused by specific base changes.

SVM Cross-Validation

To evaluate whether the combination of the tested properties would enhance discrimination of rSNPs from ufSNPs, we trained a SVM for the ALL and GROUP test data. We tested the classification performance of SVMs by 10-fold cross-validation. For each SVM, the mean area under the ROC curve was 0.83 ± 0.05 and 0.78 ± 0.04, respectively. Both suggest good performance. It is significant, however, that when removing distance from the classification, the

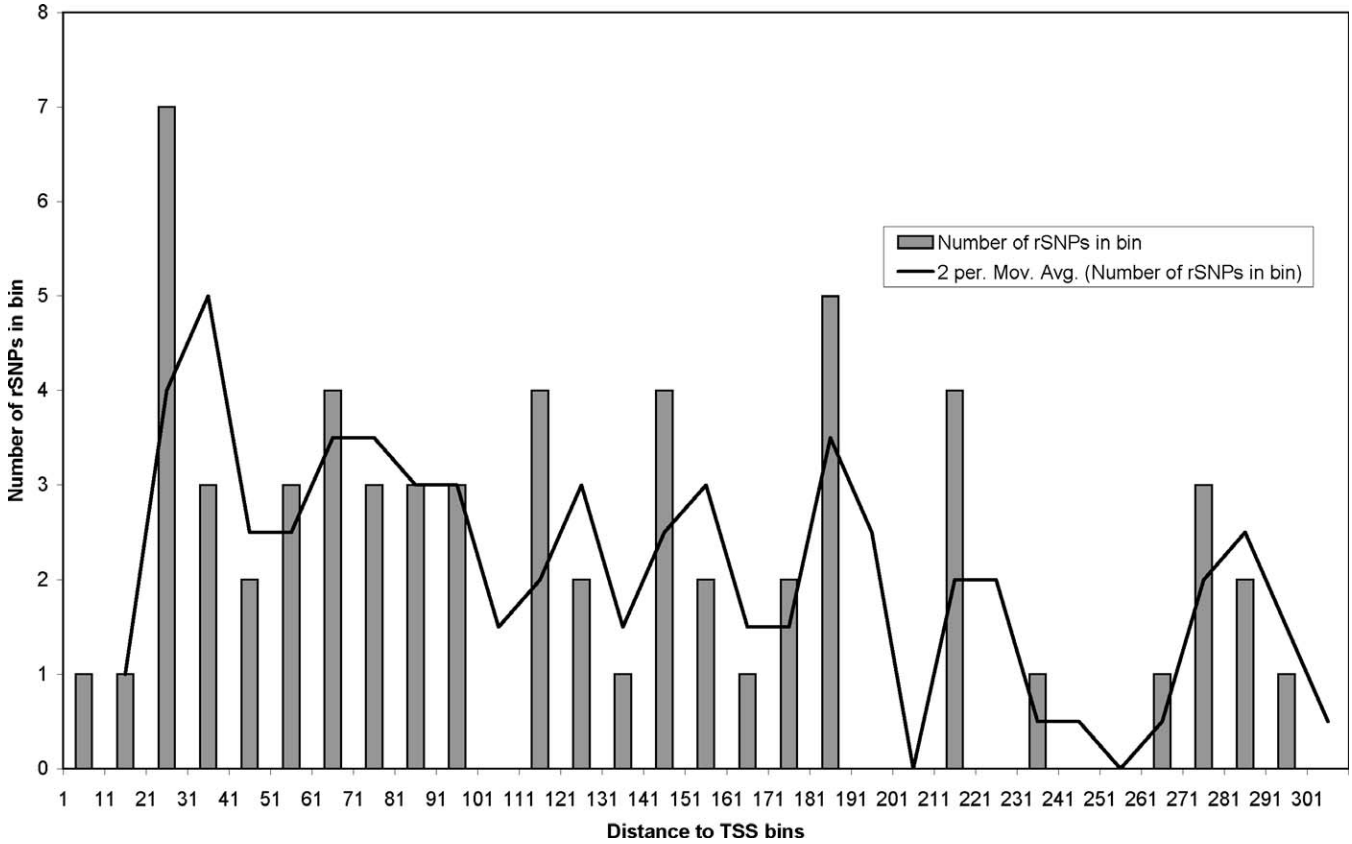

Figure 3. Histogram of Positional Bias of rSNPs for the First 300 bp of Sequence

The positions of rSNPs are plotted in a histogram for bin sizes of 10 bp for the first 300 bp of sequence from the TSS. A blip is seen at position 21–31, where it is likely that TATA and CCAAT box–binding sites are located. These types of rSNPs, however, are only slightly overrepresented in this study and from this graph are not expected to significantly bias the outcome.

doi:10.1371/journal.pcbi.0030106.g003

performance of each test drops to 0.52 ± 0.09 and 0.48 ± 0.07, respectively (Figure 2). This reduction in performance should not be taken to indicate that other properties identified in the multiple testing–corrected Wilcoxon rank sum test are not actually discriminatory since 10-fold cross-validation of All and Group test SVMs built with only the properties identified as significant using the multiple testing–corrected Wilcoxon rank sum test ($p < 0.05$) and excluding distance to the TSS achieved ROC values of 0.77 ± 0.08 and 0.75 ± 0.07, respectively. This result suggests that non-significant results may act to overparameterize the SVM model and mask subtle, true discriminatory signals.

Distance Analysis

To address the issue of distance bias further, we fortuitously identified that, across our dataset, in the 152 bp immediately upstream of the TSS, the average distance to the TSS for the ufSNPs was identical to that of the rSNPs. This 152-bp window therefore represented a region with no observable distance biases, albeit a greatly reduced subset in size; at this window size, only 16 rSNPs and 21 ufSNPs were available for analysis. When analyzed using a multiple testing–corrected Wilcoxon rank sum test for both All and Group test sets, only two properties were significant ($p < 0.05$): repetitive element density (property 13) and ClustalW alignment distance (property 23) (Table 2). We further tested window sizes of 500 bp, 1 kb, and 1.5 kb and noticed only a gradual reduction in performance of the tested properties for smaller window sizes (see Table S2).

We also examined the position of identified rSNPs to characterize possible bias. Our expectation was that well-established transcription factor–binding sites such as the TATA and CCAAT boxes may be overrepresented and contribute to lower distance values. A histogram of rSNPs for the first 300 bp of sequence from the TSS shows an expected increase around the 21–31 position where seven rSNPs are located, twice as many as average. However, it is apparent that these types of binding sites are only over-represented slightly when compared with the distribution of rSNPs at other positions (Figure 3).

Availability

All pipeline software has been programmed in Perl and is available under the Lesser GNU Public Licence at http://www.bcgsc.ca/chum under the name CHuM (*cis*-acting human mutation modules). All data are available from this site.

Discussion

This study introduces the largest publicly available collection of rSNPs—160 known rSNPs from literature. Using this collection, we investigate 104 rSNPs and 951 ufSNPs in human 2-kb upstream regions to identify properties that may discriminate functional from nonfunctional polymorphisms.

We identify several properties that may be useful to researchers attempting to determine the functional status of upstream noncoding SNPs. The most important properties detected suggest that rSNPs are close to the TSS, are not within CpG islands, are isolated from repetitive elements, possess higher MAF and higher derived allele frequency, and are within comparatively more divergent regions. However, within a 152-bp window, where an equal distribution of rSNPs and ufSNPs from the TSS is obtained, the significant results suggest that only repetitive element content and local divergence remain important (we have included in Table S2 information on how property significance changes with window size). We further combined each of the properties identified in the 2-kb region to train an SVM to classify the functional status of the 104-rSNP set and 951-ufSNP set. We hypothesized that subtle differences in individual properties may be more important than any one property in detecting rSNPs. It is of note, despite mentioned ascertainment biases, that our sensitivity and specificity for the All test was 0.82 ± 0.08 and 0.71 ± 0.13, respectively, and for the Group test was 0.72 ± 0.19 and 0.68 ± 0.07, respectively. Also of note, the strength of the distance to the TSS as a discriminatory property was demonstrated in both tests when removal of the property significantly reduced the effectiveness of the classifier to near random performance. However, we observed that this reduction in performance was recovered in part when only the properties identified as significant through the multiple testing–corrected Wilcoxon rank sum test ($p < 0.05$) in the 2-kb All and Group tests were applied, and the distance to the TSS was excluded.

Through this work, several challenges are apparent with current predictive approaches to prioritize candidate rSNPs. Necessary to future analyses is a dataset of core promoter polymorphisms that are nonfunctional across a broad range of cell types; since our negative control set was a neutral set, it is assured that more accurate performance metrics can come from addition of a reliable negative control set. Furthermore, recent analysis of allelic expression difference has demonstrated that the effects of rSNPs may be highly context-specific such that function in one cell line may not imply function in others; to address this complication, future analysis may require expanded collections of cell line–specific positive and negative rSNPs [37]. Future studies of promoter polymorphisms will also need to take advantage of known transcription factor–binding sites. Such information will be invaluable in dissecting the causal nature of many of the properties.

In summary, this study introduces a new dataset for the investigation of rSNPs. We have also introduced one of the first gene regulation and population genetics–based approaches to classifying rSNPs in the core promoter regions of human genes. We identify the utility of different gene regulation and population genetics properties in discriminating literature-curated rSNPs. Such results are increasingly essential to researchers seeking criteria for prioritizing SNPs to test in association, binding, or expression assays. Furthermore, we provided evidence that popular methodological practices based on identification of allele-specific differences in position weight matrices through unrestricted application of the TRANSFAC database are poor criteria for SNP selection. However, we highlight the fact that because of the lack of extensive unbiased collections of rSNPs, it still remains challenging to dissect the existing effects of investigator or methodological biases in evaluating the importance of these properties. We hope that this work will stimulate active discussion and both the development of expanded collections of rSNPs and an improved class of bioinformatics tools for rSNP analysis that address these challenges.

Supporting Information

Figure S1. Mapped rSNPs and ufSNPs

The locations of the tested rSNPs and ufSNPs are plotted upstream of their respective genes.

Found at doi:10.1371/journal.pcbi.0030106.sg001 (6.4 MB PNG).

Table S1. Tested rSNPs

The tested rSNP data is listed with information describing experimental evidence, associated gene, and dbSNP number, if available.

Found at doi:10.1371/journal.pcbi.0030106.st001 (59 KB PDF).

Table S2. Performance of Genomic Properties at 500-bp, 1,000-bp, and 1,500-bp Window Sizes

Different upstream window sizes were selected for All and Group analyses. The results of the Wilcoxon rank sum test for these windows are summarized and displayed as figures.

Found at doi:10.1371/journal.pcbi.0030106.st002 (44 KB XLS).

Text S1. Background File for MotifSampler

Promoters annotated in ORegAnno were assembled into this background file for MotifSampler analysis.

Found at doi:10.1371/journal.pcbi.0030106.sd001 (5 KB RTF).

Acknowledgments

We would like to acknowledge Manolis Dermitzakis and Wyeth W. Wasserman for support and feedback during the development of this work.

Author contributions. SBM and SJMJ conceived and designed the experiments, performed the experiments, and wrote the paper. SBM, JMS, ABW, and SJMJ analyzed the data. SBM, OLG, ABW, and SJMJ contributed reagents/materials/analysis tools.

Funding. We gratefully acknowledge funding from Genome Canada, Genome British Columbia, and the BC Cancer Foundation. SBM was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Michael Smith Foundation for Health Research (MSFHR), and the European Molecular Biology Organization. OLG was supported by the Canadian Institutes of Health Research (CIHR), NSERC, and MSFHR. JS was supported by the Chan Sisters Foundation, NSERC, and MSFHR. SJMJ was supported by MSFHR.

Competing interests. The authors have declared that no competing interests exist.

References

1. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: Theoretical and practical concerns. Nat Rev Genet 6: 109–118.
2. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–108.
3. Belmont JW, Gibbs RA (2004) Genome-wide linkage disequilibrium and haplotype maps. Am J Pharmacogenomics 4: 253–262.
4. Miller RD, Kwok PY (2001) The birth and death of human single-nucleotide polymorphisms: New experimental evidence and implications for human history and medicine. Hum Mol Genet 10: 2195–2198.
5. Sadee W, Dai Z (2005) Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet 14: R207–R214.
6. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: The NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.
7. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812–3814.
8. Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591–597.

9. Yue P, Melamud E, Moult J (2006) SNPs3D: Candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
10. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2006) SNPeffect v2.0: A new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics 22: 2183–2185.
11. Chang H, Fujita T (2001) PicSNP: A browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. Biochem Biophys Res Commun 287: 288–291.
12. Li JL, Li MX, Guo YF, Deng HY, Deng HW (2006) JADE: A distributed Java application for deleterious genomic mutation (DGM) estimation. Bioinformatics 22: 1926–1927.
13. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: A topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520–D522.
14. Mooney SD, Altman RB (2003) MutDB: Annotating human variation with functionally relevant data. Bioinformatics 19: 1858–1860.
15. Freimuth RR, Stormo GD, McLeod HL (2005) PolyMAPr: Programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat 25: 110–117.
16. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, et al. (2004) PupaSNP Finder: A web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 32: W242–W248.
17. Miao X, Yu C, Tan W, Xiong P, Liang G, et al. (2003) A functional polymorphism in the matrix metalloproteinase-2 gene promoter (–1306C/T) is associated with risk of development but not metastasis of gastric cardia adenocarcinoma. Cancer Res 63: 3987–3990.
18. Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, et al. (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119: 591–602.
19. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the *5-HTT* gene. Science 301: 386–389.
20. Prokunina L, Castillejo-Lopez C, Oberg F, Gunnarsson I, Berg L, et al. (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. Nat Genet 32: 666–669.
21. Kostrikis LG, Neumann AU, Thomson B, Korber BT, McHardy P, et al. (1999) A polymorphism in the regulatory region of the CC-chemokine receptor 5 gene influences perinatal transmission of human immunodeficiency virus type 1 to African-American infants. J Virol 73: 10264–10271.
22. Saito H, Tada S, Ebinuma H, Wakabayashi K, Takagi T, et al. (2001) Interferon regulatory factor 1 promoter polymorphism and response to type 1 interferon. J Cell Biochem 81: 191–200.
23. Mottagui-Tabar S, Faghihi MA, Mizuno Y, Engstrom PG, Lenhard B, et al. (2005) Identification of functional SNPs in the 5-prime flanking sequences of human genes. BMC Genomics 6: 18.
24. Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, et al. (2005) Strong bias in the location of functional promoter polymorphisms. Hum Mutat 26: 214–223.
25. Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, et al. (2003) Functional analysis of human promoter polymorphisms. Hum Mol Genet 12: 2249–2254.
26. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, et al. (2006) ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites, and regulatory variation. Bioinformatics 22: 637–640.
27. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39–D45.
28. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34: D556–D561.
29. Pavlidis P, Wapinski I, Noble WS (2004) Support vector machine classification on the web. Bioinformatics 20: 586–587.
30. Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J (2005) Predicting the in vivo signature of human gene regulatory sequences. Bioinformatics 21 (Supplement 1): i338–343.
31. Reimers M, Carey VJ (2006) Bioconductor: An open source framework for bioinformatics and computational biology. Methods Enzymol 411: 119–134.
32. Pollard KS, Dudoit S, van der Laan MJ (2005) Multiple testing procedures: R multitest package and applications to genomics. In: Gentleman R, et al., editors. Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer Science + Business Media. p. 251–272.
33. Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, et al. (2003) A population threshold for functional polymorphisms. Genome Res 13: 1873–1879.
34. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: The ancestral-susceptibility model. Trends Genet 21: 596–601.
35. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412–1417.
36. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
37. Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, et al. (2007) Extreme context-specificity in differential allelic expression. Hum Mol Genet 16: 537–546.
38. Tomso DJ, Inga A, Menendez D, Pittman GS, Campbell MR, et al. (2005) Functionally distinct polymorphic sequences in the human genome that are targets for p53 transactivation. Proc Natl Acad Sci U S A 102: 6431–6436.
39. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085–1094.
40. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, et al. (2005) oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res 33: 3154–3164.
41. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. (2003) Human–mouse alignments with BLASTZ. Genome Res 13: 103–107.
42. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32: W199–W203.
43. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17: 1113–1122.
44. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
45. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
46. Goodsell DS, Dickerson RE (1994) Bending and curvature calculations in B-DNA. Nucleic Acids Res 22: 5497–5503.
47. Kozobay-Avraham L, Hosid S, Bolshoy A (2006) Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Res 34: 2316–2327.
48. Olivares-Zavaleta N, Jauregui R, Merino E (2006) Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. Genomics 87: 329–337.
49. Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A 83: 3746–3750.
50. Baldino F Jr, Chesselet MF, Lewis ME (1989) High-resolution in situ hybridization histochemistry. Methods Enzymol 168: 761–777.
51. Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B (2004) Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5: 34.
52. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, et al. (2006) Mice and men: Their promoter properties. PLoS Genet 2: e54.
53. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253–259.
54. Fondon JW III, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101: 18058–18063.
55. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901–913.
56. Chiaromonte F, Yang S, Elnitski L, Yap VB, Miller W, et al. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. Proc Natl Acad Sci U S A 98: 14503–14508.
57. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: Update 2006. Nucleic Acids Res 34: D590–D598.
58. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
59. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. (2003) Distinguishing regulatory DNA from neutral sites. Genome Res 13: 64–72.
60. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31: 3497–3500.