PLoS COMPUTATIONAL BIOLOGY

Provisional doi:10.1371/journal.pcbi.0030106.eor

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

Future Article URL: http://dx.doi.org/10.1371/journal.pcbi.0030106

A SURVEY OF GENOMIC PROPERTIES FOR THE DETECTION OF REGULATORY POLYMORPHISMS

Stephen B. Montgomery[1, 2], Obi L. Griffith[1], Johanna M. Schuetz[1], Angela Brooks-Wilson [1], Steven J.M. Jones[1]

[1] Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC Canada, V5Z 4E6
[2] Corresponding Author: sm8@sanger.ac.uk
Current Address: Wellcome Trust Sanger Institute,
Hinxton, Cambridge, United Kingdom
CB10 1SA

Abstract

Advances in the computational identification of functional non-coding 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 literature, a computationally-tractable set of regulatory polymorphisms within the ORegAnno database (www.oreganno.org). We have further used 104 regulatory SNPs (rSNPs) from this set and 951 polymorphisms of unknown function, from 2kb and 152bp non-coding 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 frequency, 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 distance to transcription start site.

Author summary (Non-technical synopsis)

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 utilized 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 existing literature. Furthermore, we use this collection to investigate and prioritize various properties that may be important for identifying novel regulatory polymorphisms.

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, helping to chart 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 catalogue genotypes has far outstripped our ability to implicate them in phenotypes. Currently, over 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 like 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 approaches are false positives. By extending this technique and using phylogenetic footprinting between mouse and human, it was demonstrated that from 10 SNPs which show significant allele-specific differences in Jaspar predictions, 7 also demonstrated electrophoretic mobility shift differences [23]. However, only 2 of the 7 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 EPD promoters where 0 of 10 experimentally-validated regulatory variants were in conserved binding sites [25].

A substantial challenge with developing strategies for identifying functional non-coding 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 [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 2kb). Our hypothesis is that using a combination of regulatory and population genetic 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 transcription start site, local repetitive density and content, sequence conservation, minor 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, where the properties of all rSNPs were compared to all the ufSNPs and the other where each property value of the positive SNPs and ufSNPs within an associated gene are compared to the average values for each property within that gene (termed here 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 gene that may discriminate functional SNPs from ufSNPs. In a 10-fold cross validated test, the SVM achieves a ROC value of 0.83 +/- 0.05 for the ALL analysis (sensitivity 0.82 +/- 0.08; specificity 0.71 +/- 0.13) and 0.78 +/- 0.04 for the GROUP analysis (sensitivity 0.72 +/- 0.19; specificity 0.68 +/- 0.07).

Methods

Data

Literature describing non-coding 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 which confirmed its direct role in altering gene expression. This selection criterion specifically excluded those polymorphisms where the experimental evidence could only confirm that the reported polymorphism was in linkage disequilibrium with a regulatory polymorphism. Each identified regulatory polymorphism was manually-curated in the ORegAnno database (www.oreganno.org). Subsequently, 104 polymorphisms were selected based on the criteria that they were SNPs (excluding 7 insertion-deletion polymorphisms), and within 2kb of the transcription start site of their associated gene (as annotated in version 37 of EnsEMBL [28]; Table 1). A 2kb region was chosen to maximize the number of rSNPs included while minimizing the size of sequence investigated; at 2kb, 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 10 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 2kb of the transcription start site were extracted from version 37 of EnsEMBL (dbSNP version 125) producing exactly 951 SNPs of unknown function (ufSNPs). The ufSNPs and rSNPs genomic locations have been mapped and are available as supplementary material.

Investigated Properties

Twenty-three 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 non-coding regions.

Test data design (ALL and GROUP)

Two types of analyses were conducted using the investigated properties. One, an all-versus-all approach, where the 104-rSNP and ufSNP set were compared *en masse*. The other, a group analysis, where the average value of each property within each upstream non-coding 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 was designed to identify global characteristics of rSNPs while the GROUP test data was 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; whereas, the GROUP test is design to identify if there are any features that distinguish rSNPs from other SNPs within the same upstream non-coding region.

Support Vector Machine

The ALL and GROUP test data were input to the Gist SVM implementation [29]. We excluded the logarithmic distance to transcription start site to prevent redundant classification with the raw distance to transcription start site. 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 gene 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 by controlling for the family-wise error rate (α=0.05 and B=10000) [31,32].

The performance of the Gist SVM classifier was measured using a receiver operating characteristic (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 ten-fold cross validation was performed to assess the overall performance of the SVM. The input data was randomly partitioned by transcript into 10 sets. Data from one set was 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 basepairs of the TSS which is much smaller when compared to our sampling distance of 2kb. Furthermore, it has been well-established in a previous study that distance to TSS is correlated to detection of regulatory polymorphisms (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 TFBS in these regions) [24]. For this reason, the discrimination potential of distance to 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 ten-fold cross validation, the unnormalized ROC values for the ALL test are 0.82 +/- 0.05 (unnormalized) and 0.83 +/- 0.05 (normalized), and for the GROUP test are 0.79 +/- 0.04 (unnormalized) compared to 0.78 +/- 0.07 (normalized).

Results

Property Ranking

104 regulatory SNPs (104-rSNPs) and 951 SNPs of unknown function (951-ufSNPs) in the upstream non-coding 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 transcription start site (property 13 and 14),
2) In 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 (property 11),
7) Regulatory Potential score (property 22),
8) In repeat (property 14),
9) ClustalW alignment distance (property 23).

However, a concern with the ALL analysis was that calculated property values for SNPs in individual upstream non-coding regions would not be comparable to those in other upstream non-coding 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 transcription start site (property 13 and 14),
2) Long repeat events (property 16),
3) In CpG island (property 19),
4) Minor allele frequency (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),
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 non-coding region. In both tests, distance to transcription start site was found to be the most significant discriminant. While 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 500bp assayed regions, 50% of rSNPs identified through transfection experiments were within 100bp of the transcription start site (n=40) [24].

Furthermore, several other properties are consistently identified as being significant after normalization against distance to transcription start site. 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 to the ufSNPs indicating that 1kb 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 non-significant 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 regulatory SNPs 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 104-rSNPs were surveyed in terms of repetitive elements is not known and future collections of discovered rSNPs should address this issue.

Both minor and derived allele frequency are also identified as significant discriminants. Unexpectedly, for genotyped SNPs, the minor allele frequency (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 minor allele frequency 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 either through population bottlenecks or positive selection. The former hypothesis is supported by the supplemental observation that when restricting populations to HapMap populations only, the Asian and European populations mirror this result while the African population has lower, on average, MAF. The latter hypothesis, however, supports a model of evolution of genetic susceptibility to common diseases explained by ancient alleles recently becoming disease predisposing 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 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 values lower than the ufSNP values (Figure 1) [35,36]. However, without normalization the significance of this value for the ALL and GROUP tests is similar (ALL, $p = 3.78e-05$; GROUP, $p = 1.96e-03$) 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. Additionally, 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 using oPOSSUM 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 that 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 non-generalizable 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 support vector machine 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 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 transcription start site achieved ROC values of 0.77 +/- 0.08 and 0.75 +/- 0.07, respectively. This result suggests that non-significant results may act to over-parameterize 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 base pairs immediately upstream of the TSS the average distance to the TSS for the ufSNPs was identical that of the rSNPs. This 152 base pair 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 500bp, 1kb and 1.5kb and noticed only a gradual reduction in performance of the tested properties for smaller window sizes (provided as supplemental data).

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 300bp of sequence from the TSS shows an expected increase around the 21-31 position where 7 rSNPs are located, twice as many as average. However, it is apparent that these types of binding sites are only overrepresented slightly when compared to the distribution of rSNPs at other positions (Figure 3).

Availability

All pipeline software has been programmed in Perl and is available under the LGPL at http://www.bcgsc.ca/chum under the name CHuM (Cis-acting Human mutation modules). All data is available from this site.

Discussion

This study introduces the largest publicly-available collection of regulatory polymorphisms--160 known regulatory polymorphisms from literature. Using this collection, we investigate 104 regulatory polymorphisms (104-rSNP) and 951 SNPs of unknown function (ufSNPs) in human 2kb upstream regions to identify properties that may discriminate functional from non-functional polymorphisms. We identify several properties that may be of utility for researchers attempting to determine the functional status of upstream non-coding SNPs. The most important properties detected suggest that rSNPs are close to the transcription start site, not within CpG islands, isolated from

repetitive elements, possess higher minor allele frequency and higher derived allele frequency, and are within comparatively more divergent regions. However, within a 152bp window, where an equal distribution of rSNPs and ufSNPs from the transcription start site is obtained, the significant results suggest that only repetitive element content and local divergence remain important (we have included as supplementary data information on how property significance changes with window size). We further combined each of the properties identified in the 2kb region to train a support vector machine to classify the functional status of the 104-rSNPs and 951 ufSNPs. We hypothesized that subtle differences in individual properties may be more important than any one property in detecting regulatory SNPs. 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. Of note, the strength of distance to transcription start site 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 demonstrated 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 2kb ALL and GROUP test were applied and distance to transcription start was excluded.

Through this work, several challenges are apparent with current predictive approaches to prioritize candidate regulatory polymorphisms. Of necessity to future analyses is a dataset of core-promoter polymorphisms that are non-functional 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 regulatory polymorphisms. We have also introduced the first gene regulation and population genetics-based approach 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 regulatory polymorphisms. 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 is poor criteria for SNP selection. However, we highlight 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 regulatory polymorphisms and an improved class

of bioinformatics tools for regulatory polymorphism analysis that address these challenges.

Acknowledgements

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

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 Spec No. 2: R207-214.
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, 3rd, 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.
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.
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: D520-522.
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: W242-248.
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: D39-45.
28. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34: D556-561.
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 Suppl 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 multtest Package and Applications to Genomics. Bioinformatics and Computational Biology Solutions Using R and Bioconductor: 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.
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. 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: W199-203.
42. 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.
43. 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.
44. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.
45. Goodsell DS, Dickerson RE (1994) Bending and curvature calculations in B-DNA. Nucleic Acids Res 22: 5497-5503.
46. Kozobay-Avraham L, Hosid S, Bolshoy A (2006) Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Res 34: 2316-2327.
47. 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.
48. 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.
49. Baldino F, Jr., Chesselet MF, Lewis ME (1989) High-resolution in situ hybridization histochemistry. Methods Enzymol 168: 761-777.
50. 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.

51. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, et al. (2006) Mice and men: their promoter properties. PLoS Genet 2: e54.
52. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253-259.
53. Fondon JW, 3rd, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101: 18058-18063.
54. 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.
55. 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.
56. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34: D590-598.
57. 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.
58. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. (2003) Distinguishing regulatory DNA from neutral sites. Genome Res 13: 64-72.
59. 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.

Figure Legends

Figure 1: CpG island positional bias. CpG island expectancy is plotted for each chromosome as a function of distance from the transcription start site. This type of data was used to normalize many of the features in this study for distance from TSS. In this figure, the expectancy of being in a CpG island at position -1 for any promoter region is ~0.5.

Figure 2: Receiver operating characteristic (ROC) curves for discriminating known regulatory SNPs from polymorphisms of unknown function. 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 utilized for training, followed by testing on the held-out 10%. The ALL SVM approach was used for training. Furthermore, each curve has 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 9 of 11 true positives and 56 of 69 true negatives. (Plots for each tested partition are available as supplementary material).

Figure 3: Histogram of positional bias of rSNPs for the first 300bp of sequence. The positions of rSNPs are plotted in a histogram for bin sizes of 10bp for the first 300bp of sequence from the transcription start site. 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.

Tables

Table 1: Investigated properties. 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. Sequence properties are properties of the genome location in which the SNP is located.

Table 2: Analysis of rSNP and ufSNP properties in the 2kb and 152bp upstream non-coding regions. Both ALL and GROUP test sets containing were analyzed using a Wilcoxon Rank Sum test in 2kb and 152bp region. In the 2kb region, 104-rSNPs and 951 ufSNPs were tested. In the 152bp region, 16 rSNPs and 21 ufSNPs were tested. The 152bp region was selected because it contained nearly equivalent mean distances from the transcription start site for both the rSNP and ufSNPs under study. In each cell, the value for the 2kb region test is on top of the value for the 152bp region test. Each value was corrected for multiple testing using the BioConductor MTP package by controlling for the family-wise error rate (α=0.05 and B=10000). The direction of difference between the two populations is also recorded and describes the relationship between the rSNPs and ufSNPs; here, '+' indicates that the rSNPs have higher mean values and '-' indicates the rSNPs have lower mean values.

	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 cut-off of 80%) for both alleles and calculating the absolute cumulative difference in predicted binding site scores. $$\Delta s = \sum_{factor=n;n\in p}^{p=i\cup j} \left| (score(r)_{factor} - score(v)_{factor} \right|$$ 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 >= 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 factors 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 1kb, evenly flanking, DNA sequence was retrieved from EnsEMBL (NCBI35). The EnsEMBL compara database was subsequently used to retrieve pre-calculated orthologous sequences from completed genomes (using BLASTZ_NET); specifically, sequences from chimpanzee, rhesus macaque, mouse, dog, rat and chicken were used. A Weeder [41] and MotifSampler [42] analysis was performed by separately inputting both canonical and variant human sequences (the 1kb 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 effects scoring. The maximum score was used to measure whether the polymorphism was in a high-scoring motif (irrespective of allele). To improve the probability of detecting the desired motif, Weeder was set to detect 500 motifs and MotifSampler was seeded with 25bp around the polymorphism. For MotifSampler, a background file was supplied containing 745 regulatory regions annotated in ORegAnno as of January 2006 (supplied as supplemental data). 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 [43]. A 1kb 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	
5	MotifSampler (Difference)	Allele-specific	Motif Discovery, Evolutionary conservation	
6	MotifSampler (Maximum)	Allele-specific	Motif Discovery, Evolutionary conservation	
7	DNA Bendability	Allele-specific	DNA Structure, Sequence	A DNA bendability and curvature analysis was performed on canonical and variant sequences (the 1kb sequences assembled for Weeder and

			composition	MotifSampler, above) using an implementation of the BEND algorithm called "banana" and packaged in the EMBOSS toolkit [44,45]. "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 [46,47].
8	DNA Curvature	Allele-specific	DNA Structure, Sequence composition	
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 [48,49]. The presence of functional transcription factor binding sites in GC-rich sequences has been previously demonstrated [50,51]. Similar to analyzing DNA bending and curvature, we utilized 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	
11	Minor Allele Frequency	Allele-specific	Population property	Minor allele frequencies 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. 79/104 rSNPs and 502/968 ufSNPs had genotype data.
12	Derived Allele Frequency	Allele-specific	Population property, Evolutionary conservation	

13	Local Repetitive Base Percentage	Sequence	Sequence characteristic	Local repetitive content of a 200-bp DNA segment centred 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 1kb; and 4) length less than 1kb that overlap 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 non-coding regions [52-55]. Each value was normalized to its expectancy at the calculated distance from the transcription start site in the associated chromosome (see "Distance Normalization").
14	In Repeat	Sequence	Sequence characteristic	
15	Short Repeat Events	Sequence	Sequence characteristic	
16	Long Repeat Events	Sequence	Sequence characteristic	
17	Distance to Transcription Start Site	Sequence	Regulatory sequence characteristic	The distance to the transcription start site (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-base pairs [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 Transcription Start Site (Log)	Sequence	Regulatory sequence characteristic	
19	In CpG Island	Sequence	Regulatory sequence characteristic	CpG islands were obtained from annotation in the UCSC genome browser [56]. 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 transcription start site in the associated chromosome (see

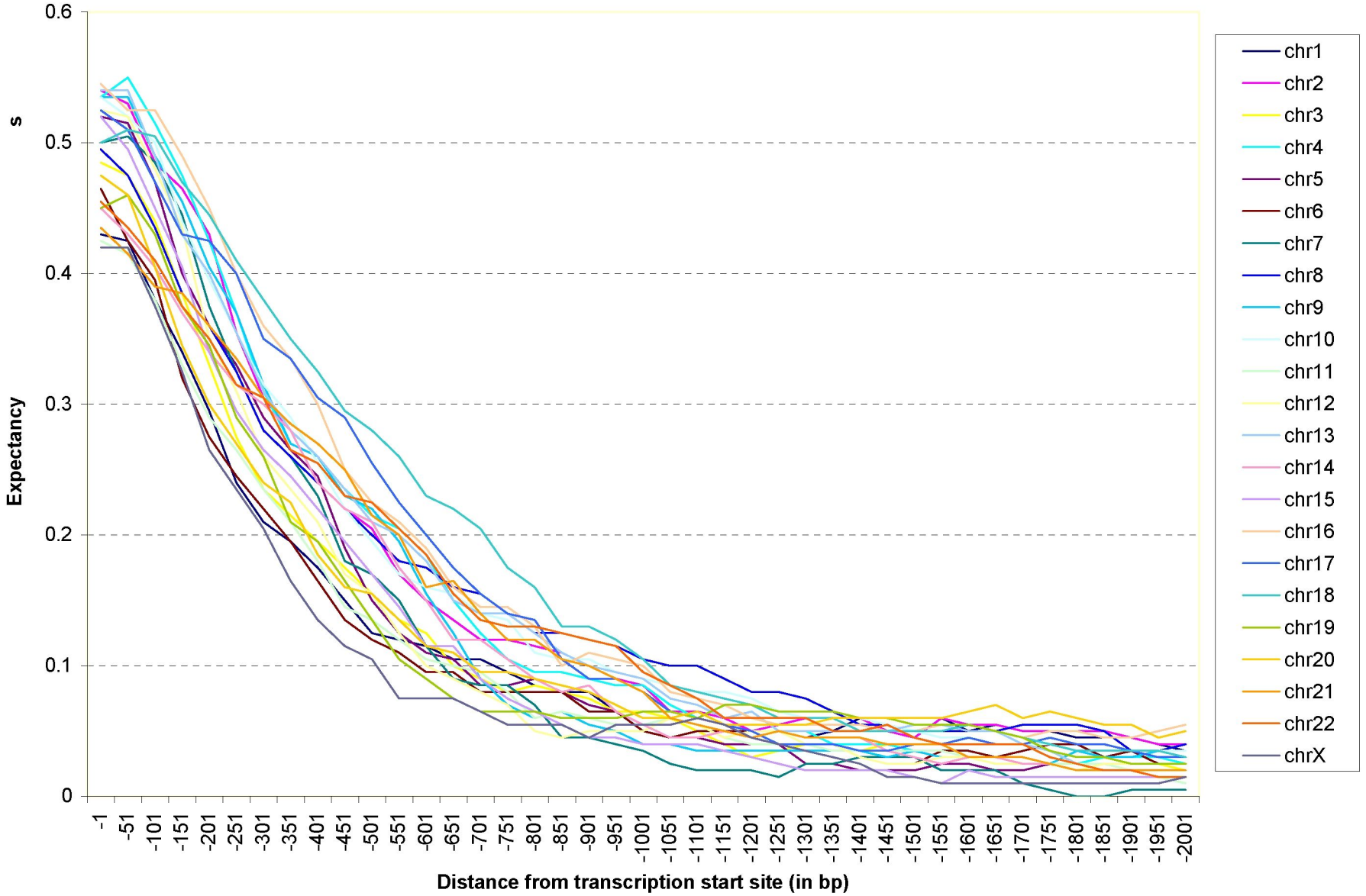
				"Distance Normalization").
20	DNaseI Hypersensitive site	Sequence	Regulatory sequence characteristic	DNaseI hypersensitive sites were obtained from predicted sites as per Noble *et al*. [30]. These sites were mapped from hg15 to hg17 coordinates using blast. Whether or not a polymorphism was in a DNaseI hypersensitive site was recorded. These values were normalized to its expectancy at the calculated distance from the transcription start site in the associated chromosome (see "Distance Normalization").
21	PhastCons	Sequence	Regulatory sequence characteristic, Evolutionary conservation	Conservation scores from both the PhastCons [57] and Regulatory Potential (RP) [58] 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 transcription start site in the associated chromosome (see "Distance Normalization").
22	RP (Regulatory Potential)	Sequence	Regulatory sequence characteristic, Evolutionary conservation	
23	ClustalW Alignment Distance	Sequence	Evolutionary conservation	Each orthologous sequence set for an individual polymorphism was aligned using ClustalW [59] and the total evolutionary distance was calculated from the associated phylogenetic tree. Since orthologues 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.

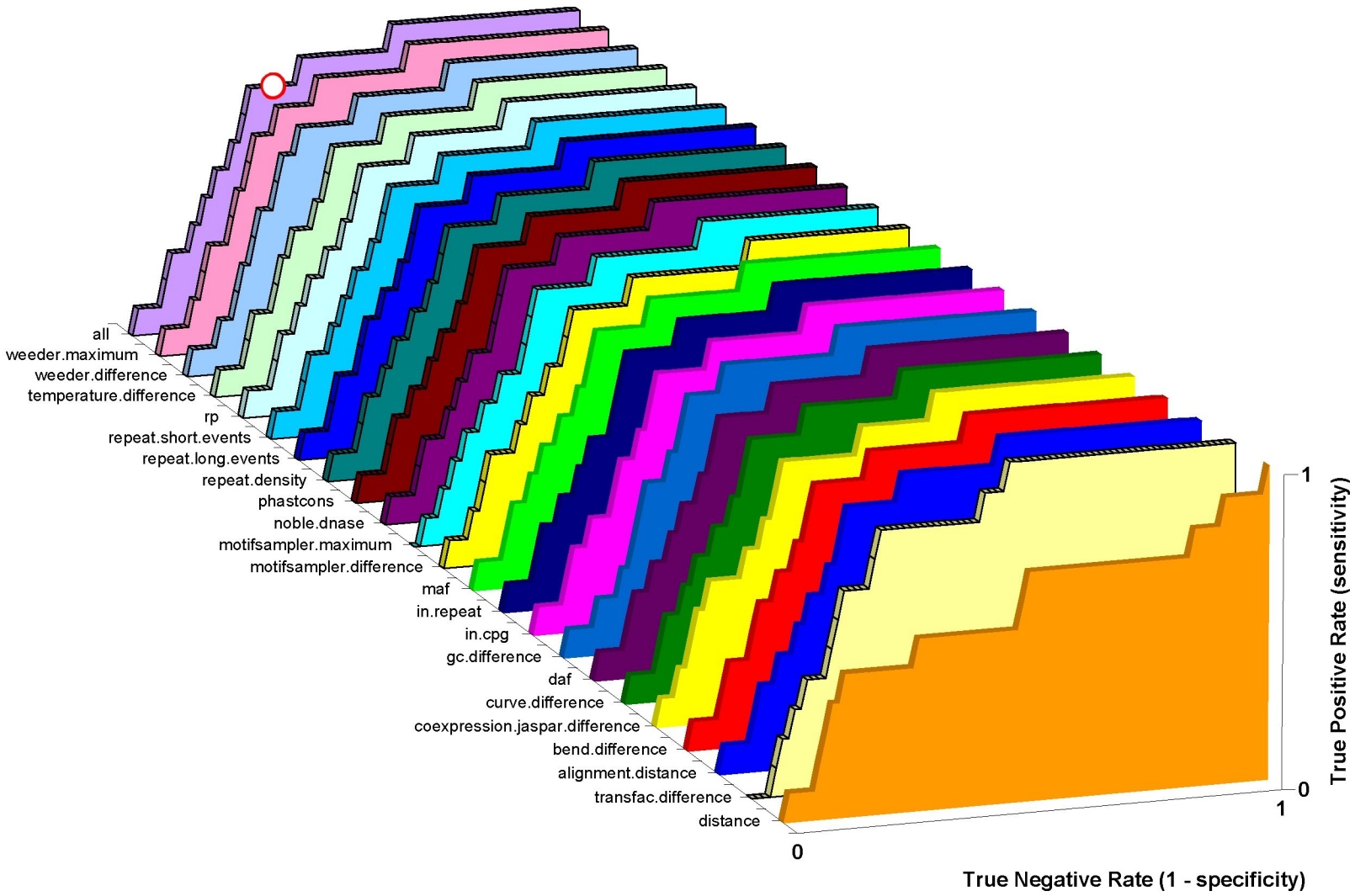
Table 1: Investigated properties. The properties are broken down into two types: allele-specific and sequence. Allele-specific properties are calculated as a difference in property values calculated by allele. Sequence properties are properties of the genome location in which the SNP is located.

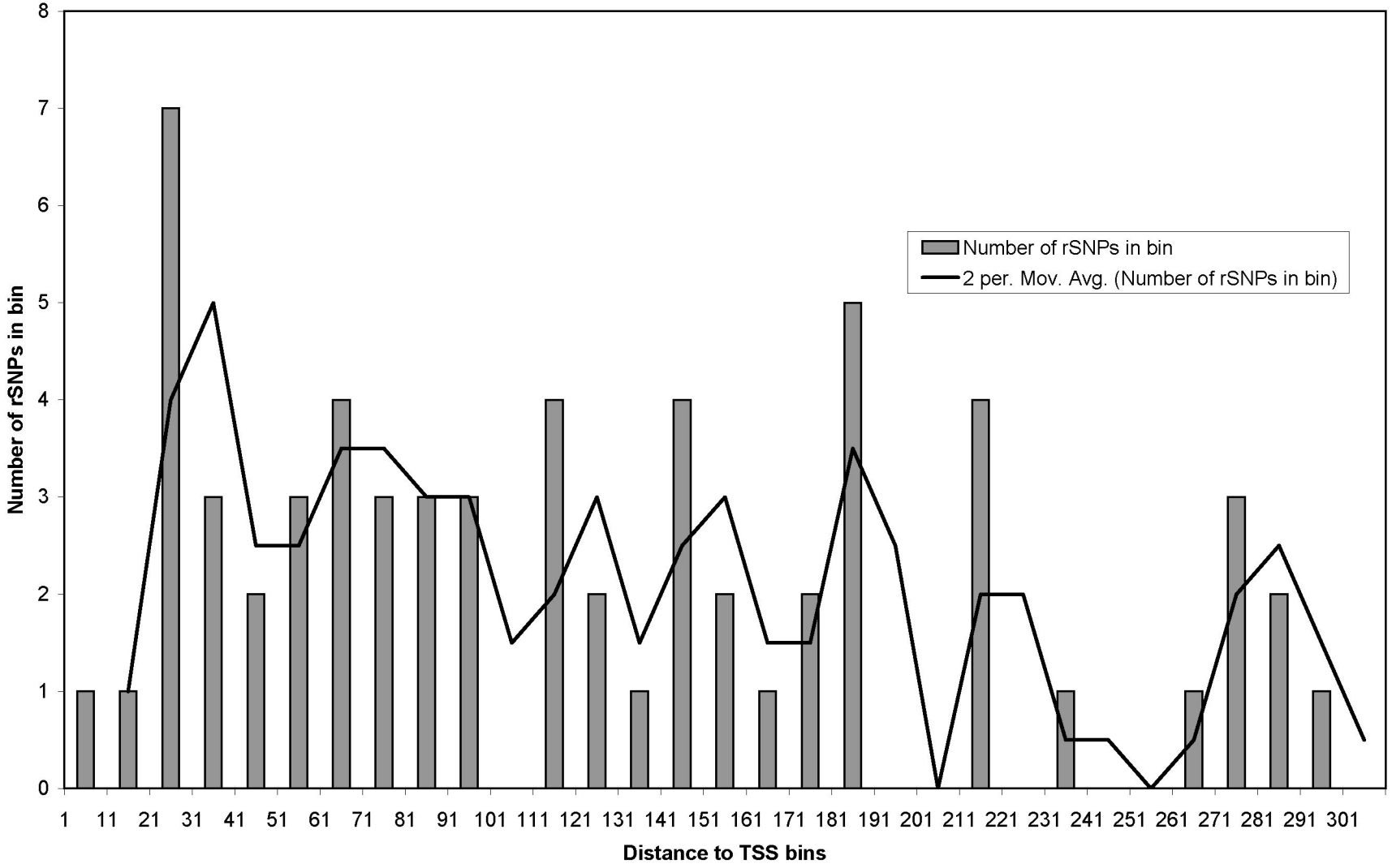
		ALL TEST			GROUP TEST		
	Investigated Property	**Wilcoxon Raw P-Value**	**Multiple Testing Corrected P-Value**	**Direction**	**Wilcoxon Raw P-Value**	**Multiple Testing Corrected P-Value**	**Direction**
1	TRANSFAC	0.958 0.206	0.712 0.603	- +	0.958 0.206	0.800 0.601	- +
2	oPOSSUM	0.161 0.576	0.493 0.816	- +	0.316 0.747	0.576 0.824	+ -
3	Weeder (Difference)	0.862 0.267	0.712 0.707	+ +	0.896 0.323	0.800 0.695	+ +
4	Weeder (Maximum)	0.296 0.267	0.496 0.707	- +	0.514 0.241	0.727 0.668	- +
5	MotifSampler (Difference)	0.275 0.308	0.496 0.714	+ +	8.27e-02 0.308	0.313 0.695	+ +
6	MotifSampler (Maximum)	0.733 0.047	0.712 0.211	- +	0.666 0.147	0.741 0.529	+ +
7	DNA Bendability	5.59e-02 0.975	0.261 0.816	- +	0.101 1	0.387 0.858	- -
8	DNA Curvature	0.201 0.668	0.496 0.816	- -	0.477 0.915	0.727 0.858	- -
9	GC Content	0.811 0.403	0.712 0.715	+ +	0.950 0.960	0.800 0.858	+ -
10	DNA Thermodynamics	0.201 0.713	0.496 0.816	+ +	0.138 0.892	0.496 0.858	+ -
11	Minor Allele Frequency	2.71e-05 0.719	**p < 1e-09** 0.816	+ +	1.09e-07 0.0853	**p < 1e-09** 0.283	+ +
12	Derived Allele Frequency	2.53e-05 1	**p < 1e-09** 0.828	+ +	9.52e-05 0.311	**p < 1e-09** 0.676	+ +
13	Local Repetitive Base Percentage	1.23e-07 0.290	**p < 1e-09** 0.587	- -	3.62e-06 0.016	**p < 1e-09** **p < 1e-09**	- -
14	In Repeat	1.73e-03 0.334	**p < 1e-09** 0.714	+ -	0.872 1	0.800 0.858	- -
15	Short Repeat Events	0.290 0.107	0.493 0.421	- -	1.51e-03 0.150	**8.70e-03** 0.528	- -
16	Long Repeat Events	p < 1e-09 0.522	**p < 1e-09** 0.810	+ -	p < 1e-09 0.282	**p < 1e-09** 0.695	+ -
17	Distance to Transcription Start Site	p < 1e-09 0.602	**p < 1e-09** 0.810	- -	p < 1e-09 0.187	**p < 1e-09** 0.586	- -
18	Distance to Transcription Start Site (Log)	p < 1e-09 0.602	**p < 1e-09** 0.816	- -	p < 1e-09 0.575	**p < 1e-09** 0.776	- -
19	In CpG Island	p < 1e-09 0.705	**p < 1e-09** 0.816	- +	p < 1e-09 0.250	**p < 1e-09** 0.6786	- -
20	DNaseI	1.91e-02	0.118	+	4.54e-03	**2.55e-02**	+

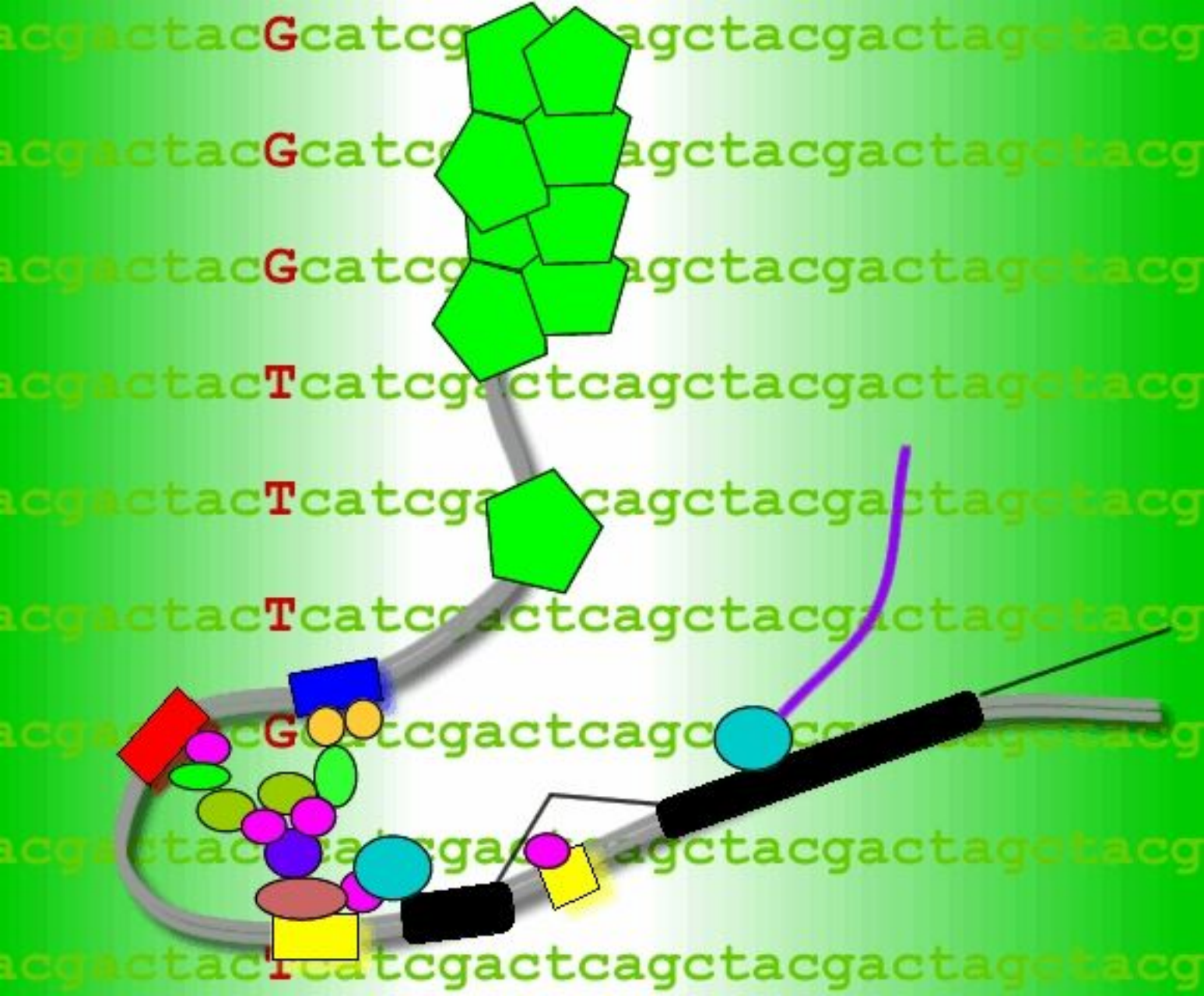
	Hypersensitive site	0.165	0.587	+	0.868	0.858	+
21	PhastCons	3.23e-02 0.061	0.188 0.268	+ +	0.192 0.150	0.576 0.529	+ +
22	RP (Regulatory Potential)	2.80e-05 0.100	**p < 1e-09** 0.443	+ +	0.114 0.554	0.507 0.777	- +
23	ClustalW Alignment Distance	3.68e-03 0.794	**1.34e-02** 0.292	+ =	9.64e-06 1.3e-04	**p < 1e-09** **p < 1e-09**	+ +

Table 2: Analysis of rSNP and ufSNP properties in the 2kb and 152bp upstream non-coding regions. Both ALL and GROUP test sets containing were analyzed using a Wilcoxon Rank Sum test in 2kb and 152bp region. In the 2kb region, 104-rSNPs and 951 ufSNPs were tested. In the 152bp region, 16 rSNPs and 21 ufSNPs were tested. The 152bp region was selected because it contained nearly equivalent mean distances from the transcription start site for both the rSNP and ufSNPs under study. In each cell, the value for the 2kb region test is on top of the value for the 152bp region test. Each value was corrected for multiple testing using the BioConductor MTP package by controlling for the family-wise error rate (α=0.05 and B=10000). The direction of difference between the two populations is also recorded and describes the relationship between the rSNPs and ufSNPs; here, '+' indicates that the rSNPs have higher mean values and '-' indicates the rSNPs have lower mean values.


0.6
0.5
0.4
0.3
0.2
0.1
0
s
Expectancy
-1
-51
-101
-151
-201
-251
-301
-351
-401
-451
-501
-551
-601
-651
-701
-751
-801
-851
-901
-951
-1001
-1051
-1101
-1151
-1201
-1251
-1301
-1351
-1401
-1451
-1501
-1551
-1601
-1651
-1701
-1751
-1801
-1851
-1901
-1951
-2001
Distance from transcription start site (in bp)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX


all
weeder.maximum
weeder.difference
temperature.difference
rp
repeat.short.events
repeat.long.events
repeat.density
phastcons
noble.dnase
motifsampler.maximum
motifsampler.difference
maf
in.repeat
in.cpg
gc.difference
daf
curve.difference
coexpression.jaspar.difference
bend.difference
alignment.distance
transfac.difference
distance
0
1
True Negative Rate (1 - specificity)
1
0
True Positive Rate (sensitivity)


Number of rSNPs in bin
2 per. Mov. Avg. (Number of rSNPs in bin)
Number of rSNPs in bin
Distance to TSS bins
0
1
2
3
4
5
6
7
8
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
261
271
281
291
301


Gcatcgagctacgactagctacg
Gcatcgagctacgactagctacg
Gcatcgagctacgactagctacg
Tcatcgactcagctacgactagctacg
Tcatcgagctacgactagctacg
Tcatcgactcagctacgactagctacg
Gcatcgactcagctacgactagctacg
Tcatcgactcagctacgactagctacg