A peer-reviewed version of this preprint was published in PeerJ on 29 November 2018.

View the peer-reviewed version (peerj.com/articles/5742), which is the preferred citable publication unless you specifically need to cite this preprint.

ShapeGTB: The role of local DNA shape in prioritization of functional variants in human promoters with machine learning

Maja Malkowska ¹, Julian Zubek ², Dariusz Plewczynski Corresp. ², ³, Lucjan S Wyrwicz Corresp. ¹

¹ Laboratory of Bioinformatics and Biostatistics, Maria Sklodowska-Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland
² Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland
³ Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Corresponding Authors: Dariusz Plewczynski, Lucjan S Wyrwicz
Email address: d.plewczynski@cent.uw.edu.pl, lucjan.wyrwicz@coi.pl

Motivation: The identification of functional sequence variations in regulatory DNA regions is one of the major challenges of modern genetics. Here, we report results of a combined multifactor analysis of properties characterizing functional sequence variants located in promoter regions of genes.

Results: We demonstrate that GC-content of the local sequence fragments and local DNA shape features play significant role in prioritization of functional variants and outscore features related to histone modifications, transcription factors binding sites, or evolutionary conservation descriptors. Those observations allowed us to build specialized machine learning classifier identifying functional SNPs within promoter regions – ShapeGTB. We compared our method with more general tools predicting pathogenicity of all non-coding variants. ShapeGTB outperformed them by a wide margin (AUC ROC 0.97 vs. 0.57-0.59). On the external validation set based on ClinVar database it displayed only slightly worse performance (AUC ROC 0.92 vs. 0.74-0.81). Such results suggest unique characteristics of mutations located within promoter regions and are a promising signal for the development of more accurate variant prioritization tools in the future.

Availability and implementation: The datasets and source code are publicly available at: https://github.com/zubekj/ShapeGTB.
ShapeGTB: an analysis of the local DNA shape importance in the exploration of predictive features for accurate classification of functional variants in human promoters.

Maja Malkowska¹#, Julian Zubek²#, Dariusz Plewczynski²,³, Lucjan Wyrwicz¹

¹Laboratory of Bioinformatics and Biostatistics, Maria Sklodowska-Curie Memorial Cancer Centre and Institute of Oncology, W.K. Roentgena 5, 02-781 Warsaw, Poland
²Laboratory of Functional and Structural Genomics, Center of New Technologies, University of Warsaw, Banacha 2C, 02-097 Warsaw, Poland
³Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

#contributed equally

ABSTRACT

Motivation: The identification of functional sequence variations in regulatory DNA regions is one of the major challenges of modern genetics. Here, we report results of a combined multifactor analysis of properties characterizing functional sequence variants located in promoter regions of genes.

Results: We demonstrate that GC-content of the local sequence fragments and local DNA shape features play significant role in prioritization of functional variants and outscore features related to histone modifications, transcription factors binding sites, or evolutionary conservation descriptors. Those observations allowed us to build specialized machine learning classifier identifying functional SNPs within promoter regions – ShapeGTB. We compared our method with more general tools predicting pathogenicity of all non-coding variants. ShapeGTB outperformed them by a wide margin (average precision 0.93 vs. 0.47-0.55). On the external validation set based on ClinVar database we observed that all methods decreased their performance (average precision 0.47 vs. 0.23-0.42). Such results suggest unique characteristics of mutations located within promoter regions and are a promising signal for the development of more accurate variant prioritization tools in the future.

Availability and implementation: The datasets and source code are publicly available at: https://github.com/zubekj/ShapeGTB.

Contacts: lwyrwicz@coi.pl or d.plewczynski@cent.uw.edu.pl

Supplementary information: Supplementary data are available at PeerJ online.

INTRODUCTION

The concept of personalized medicine has made the functional annotation of genomic variations one of the major goals of human genetics. The research inquiries are done both at individual level of low-throughput methods and large-scale population studies. The results of genome-wide association studies (GWAS) of complex human traits have exposed enrichment for variations in the regulatory elements, such as promoters, enhancers, insulators or intergenic regions. Although about 90% of single nucleotide polymorphisms (SNPs) are located in non-coding regions of human genome, the knowledge about their role in pathology of diseases is limited. In this article, we propose a method for functional prioritization of variants in human promoters, which represent around 1% of all SNPs identified by the 1000 Genomes Project (Ignatieva et al. 2014).

In recent years, several computational methods have been developed to address the challenging...
task of noncoding variants annotation. These methods differ in the adopted algorithms and utilized data. The main three approaches used by currently available tools are: functional annotations, sequence homology analysis and machine learning models integrating information from both sources. Especially the third integrating machine learning approach is worth investigating. The last decade has brought dramatic progress in application of machine learning algorithms in computational biology. Their versatile predictions have been utilized to link noncoding variations properties to their functional nature by i.e. genome-wide annotation of variants (GWAVA) (Ritchie, et al., 2014), combined annotation-dependent depletion (CADD) (Kircher, et al., 2014), deleterious annotation of genetic variants using neural networks (DANN) (Quang, et al., 2015), FATHMM-MKL (Shihab, et al., 2015), deltaSVM (Lee, et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015).

Promoters are one of the key regulatory elements of transcription initiation. Several resources indicate that promoter regions show distinct structural constrains when compared with non-promoters (Kanhere and Bansal, 2005; Goni, et al., 2007; Morey, et al., 2011; Gan, et al., 2012). The analysis by Freeman et al. shows that the sequence-dependent shape of DNA encodes histone affinity and dominates molecular recognition in the problem of nucleosome positioning (Freeman, et al., 2014). Since various DNA sequences can encode similar shapes (Gardiner, et al., 2004; Greenbaum, et al., 2007), correlation between DNA shape descriptors and biological functions becomes an interesting problem to investigate.

The development of DNAshape web server by Zhou et al. (Zhou, et al., 2013) allowed analyzing DNA structural features on a genomic scale. The method computes four DNA shape features: minor groove width (MGW), roll (Roll), propeller twist (ProT) and helix twist (HelT). Recent studies have showed that combining DNA sequence with DNA local shape improves the prediction accuracy of transcription binding sites in vitro (Rohs, et al., 2009; Dror, et al., 2014). Here, we address the question of the usefulness of such data in predicting functional effects of sequence variations in promoter regions of genes. We are convinced that the DNA shape features may largely contribute to solving a demanding problem of regulatory variants interpretation and assessment of their effects on disease pathology.

To test this hypothesis and demonstrate its applicability, we trained a machine learning classifier, which uses local shape to predict functional prioritization of promoter sites. In this paper, we compare structural predictor’s performance with sequence-based methods, and analyze in detail the statistical relevance of different types of features characterizing DNA molecule.

In the light of the unique promoter characteristics, inclusive GC distribution (Lenhard, et al., 2012; Andersson, et al., 2014), transcription factor binding site composition (Rada-Iglesias, et al., 2011; Shen, et al., 2012; Thurman, et al., 2012) and unique chromatin signatures (Heintzman, et al., 2007; Hon, et al., 2009), we focused our analysis on the regions located upstream of the transcription start site. To our best knowledge previously developed methods have not aimed the variant prioritization in promoter regions by local DNA shape features but rather focused on non-coding sequence variations without acknowledging genomic region.

**MATERIALS AND METHODS**

**Datasets**

To obtain the positive dataset we used single-nucleotide variants (SNVs) annotated as regulatory mutations in The Human Gene Mutation Database (HGMD®) professional version (release 2016.2) within 5 kilobases (kb) upstream from the annotated transcription start sites (TSS) and provided sequences (Stenson, et al., 2014). The total number of experimentally validated disease-
related variants in our dataset is equal to 1772. The control dataset contains SNVs from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) with a global minor allele frequency $\geq 1\%$. The overlapping elements of both sets were removed. Only variants lying within 5 kb upstream of TSS were selected for further analysis (Rosenbloom, et al., 2015). The sequences of neutral motifs (not associated with disease phenotype) were retrieved from Ensembl with BioMart (Kinsella, et al., 2011). The total number of negative examples in our dataset is equal to 3806. We ensured that positive and negative motif sets are matched in their basic properties (Kolmogorov–Smirnov two sample test results for GC-content distributions are as follows D-statistic=0.02, p-value=0.48, null hypothesis of identical distributions retained). Distributions of TSS distances in the two sets differed, but we made sure that it does not affect obtained results (see Supplementary Material 5).

Machine learning pipeline

We split the available data into training and test sets randomly keeping the ratio 8:2. Full training set contained 1417 positives and 3045 negatives, full test set contained 355 positives and 761 negatives. Training set was used to build feature ranking, train classifiers and optimize their parameters, while test set was left for final validation and for comparison with other prediction methods. To validate our methods internally on the training set we used a cross-validation strategy in which in each fold SNPs from a single chromosome formed test set and SNPs from other chromosomes formed training set. This eliminated possibility of overfitting during parameter tuning and feature selection procedures, and additionally demonstrated whether our method generalizes across different chromosomes.

We applied Monte Carlo feature selection (MCFS) algorithm (Draminski, et al., 2008) to perform feature importance ranking. It is a universal feature selection strategy combining random subspace methods with decision trees. A random subset of the original features is drawn in each iteration of the algorithm and an equivalent of random forest is induced using the selected variables. Feature importance ranking is constructed based on all induced trees. Additionally, meaningful interdependencies between features are discovered by calculating how often two features are used together to predict the class value. MCFS aims at finding all features relevant for the classification task, and it guarantees that with sufficient number of iterations all features can be tested. Following general guidelines by the authors of the algorithm, we set the number of iterations to 1000 and the subset of original features considered in each iteration to 0.25.

In the classification task gradient tree boosting was used (GTB) – a popular tree-based ensemble algorithm (Friedman and Meulman, 2003). It is known to perform very well in many domains, often outperforming methods such as random forest, support vector machines or neural networks (Sheridan, et al., 2016; Ladds, et al., 2016; Babajide Mustapha and Saeed, 2016). The key idea behind GTB is to build trees sequentially, training a tree at each step to explain the prediction error made by the combination of existing trees. Usually the trees are regularized to prevent overfitting. We used the state-of-art implementation provided by XGBoost library (Chen and Guestrin, 2016). Through cross-validation performed on the training set we selected optimal parameter values (number of trees – 300, maximal tree depth – 8, learning rate – 0.1).

Comparison with existing approaches

Presently, the field of prediction and prioritization of human noncoding regulatory variants still lacks a large, independent and publicly available gold-standard dataset for training, testing and validating existing in silico approaches. The comparison of our method to the current state-of-the-art methods is hampered even further by different aims and objectives. To our best knowledge all available tools were designed for genome-wide, regulatory variants prioritization and there are no
computational methods focused on promoter regions. Nonetheless, we compared performance of our algorithm with other tools on our own hold-out test set and on independent high-quality data from ClinVar database (Jan 5, 2017 release) after excluding variants present in our training data (Landrum, et al., 2016). Our hold-out test set contained 355 positives from HGMD and 761 negative examples from 1000 Genomes Project. External validation set contained 32 positive examples labeled as pathogenic in ClinVar database and 761 negative examples from 1000 Genomes Project (not present in our train set).

**Features groups**

We used the following feature groups to annotate each SNV in our pathogenic and control datasets (more detailed description can be found in Supplementary material 1 and 4):

1. **DNA sequence** (52 variables): 9-nt sequence motifs centered on the mutated nucleotide. The sequence was encoded using 4-bits binary coding. Additional 12 binary (4-nt by 3 mutations) variables indicated what type of mutation occurred (e.g. A → C, G → T, etc.).

2. **Local DNA shape features** (88 variables): helix twist, minor groove width, propeller twist, roll values in span of 9 nt. Differences (_diff) between reference and mutated scores were added as additional features.

3. **GC-content** (8 variables): GC-content in span of 7- and 9-nt for reference and mutated sequences separately. Differences between the reference and mutated scores were added as additional features.

4. **Histone modifications** (38 variables): ChIP-seq data for histone 3 lysine 9 acetylation (H3K9ac) and histone 3 lysine 4 trimethylation (H3K4me3) across 16 cell lines from ENCODE (Ram, et al., 2011). For H3K9ac, H3K4me3 or either modification mean values over all cell lines and binary variables indicating modifications in any cell line were added.

5. **Transcription Factor Binding Sites** (12 variables): TFBS ChIP-seq clusters (V3) from ENCODE data retrieving binding sites of top 10 TFs with the highest binding site coverage. Mean value over all TFs and 0-1 indicator of any TF occurrence were added in addition (ENCODE Project Consortium, 2012).

6. **Transcription factor binding disruption** (1 variable):
P-value of disrupting putative strongest transcription factor binding site due to mutation was calculated with Annotation of Regulatory Variants using Integrated Networks (ARVIN) algorithm (Gao, et al., 2018) using Cis-BP database (Weirauch et al., 2014).

7. **Maximum transcription factor binding log-odds ratio score** (1 variable):
Maximum TF binding log-odds ratio score for reference and mutated sequences among scores calculated with ARVIN algorithm (Gao, et al., 2018, Weirauch et al., 2014).

8. **DNase I hypersensitivity** (1 variable): ENCODE DNase clusters (V3) from 125 cell line types (John, et al., 2011; Thurman, et al., 2012; Rosenbloom, et al., 2013).

9. **Evolutionary conservation** (10 variables):
a) **GERP ++**: Genomic Evolutionary Rate Profiling scores (Davydov, et al., 2010).
b) **PhastCons**: PhastCons conservation score by vtools (San Lucas, et al., 2012).
c) **Z-score**: recalculated Z-score values defined in our previous work (Wyrwicz, et al., 2007) on whole genome human–mouse alignments (genome builds hg19 and mm9 (Chiaromonte, et al., 2002; Kent, et al., 2003; Schwartz, et al., 2003) from UCSC Genome Browser (Kent, et al., 2002) for the reference and mutated sequence and for window length 7 and 9. Differences of Z-scores for the reference and mutated sequence were added.

10. **Dinucleotide content** (16 variables):
Observed vs. expected frequencies of 16 possible pairs of nucleotides appearing in the short sequence motif.

RESULTS

Feature importance

From MCFS we obtained the ranking of all 227 features according to their relative importance in the classification problem. Each feature group contained multiple individual features with different ranks in the overall ranking. In the context of machine learning task, usefulness of a particular group should be determined by the best performing features from this group.

Figure 1 presents detailed feature ranking including all features from each group. Generally, features that contribute to the correct classification mostly belong to GC content group, shape group and sequence group. Other feature groups were of lesser importance (the full ranking is included as Supplementary material 2, feature names glossary as Supplementary material 4). The most important feature was the difference in GC-content between the reference and the mutated sequence fragment (rank 1). Features describing raw nucleotide sequence and dinucleotide content appeared in the middle of the ranking. Among the shape features those describing the closest neighborhood of the mutated nucleotide were the most important. This is not surprising because differences in shape are expected to have local effects on DNA properties. Among the shape features attributes concerning propeller twist were ranked as the most important, attributes concerning helix twist and roll followed, and attributes concerning minor groove width occurred lower in the ranking. What is notable, most of the features appearing among the top 20 concerned differences in shape properties between SNP and wild type. Features derived from transcription factors were less important than sequence-based features. Histone modifications, conservation scores and DNase I hypersensitivity score were not identified as particularly informative features.

To investigate the role of individual features we calculated Welch’s t-score capturing the relationship between particular feature and class value. Decrease of GC-content between the reference and the mutated sequence correlated negatively with functionality (t-score -8.2088 for decrease for motif length 7, t-score -11.3710 for decrease for motif length 9), while increase of propeller twist value correlated positively (t-score 9.7417 for increase immediately before the modified nucleotide, t-score 5.5047 for increase immediately after the modified nucleotide).

The role of each feature in a classification task lies not only in its correlation with class value, but also in how well it complements with other features. For example, Figure 2 presents joint distributions of the two most important features in the two classes (difference of GC-content between the reference and the mutated sequence, difference of propeller twist at 3rd position between mutated variant and wild type). For non-functional SNPs the features are uncorrelated, but there is a visible negative correlation for functional SNPs. MCFS allows studying that kind of dependencies through its interdependency discovery function. Full list of feature interdependencies and their relative strength is included as Supplementary material 3. Figure 3 presents graph of the strongest interdependencies among the top selected features (GCSCORE – GC composition, SEQ – sequence feature, ROLL – roll, HELT – helix twist, PROT – propeller twist). Difference in GC-content acts as a central hub and interacts strongly with all groups of shape features except minor groove width. The simplified intuition is that functional SNPs should increase GC content of the motif, and at the same time increase rotation of the DNA strand accordingly.

Classifier performance

Obtained feature ranking suggests that a large portion of information is contained in features derived from the DNA sequence, and features describing evolutionary conservation and functional
properties play less significant role. To verify this hypothesis, we performed a cross-validation experiment (with folds determined by chromosomes) on the train set by training gradient tree boosting (GTB) classifier on different combinations of feature groups. Calculated values of multiple performance measures are presented in Table 1.

Classifier based on all available features performed better than the classifier using only 25 best ranked features. Among individual feature groups GC content produced classifier with the largest AUC ROC (0.78). Combining GC content with shape features and sequence features allowed achieving AUC ROC 0.98. No other combinations of features performed better. These results show that shape features are more meaningful when combined with another feature. In further experiments classifier trained on sequence, shape and GC content was used. We named this classifier ShapeGTB.

We compared final ShapeGTB classifier with more general SNP prioritization methods, which did not focus specifically on promoter regions: CADD, FATHMM-MKL and DeepSEA. Figure 4 present precision-recall curves calculated on the hold-out test set constructed from our data (HGMD and 1000 Genomes Project) and for smaller experimental dataset (ClinVar and 1000 Genomes Project). Area under precision-recall curve can be interpreted as average precision (AP), and is an aggregated measure of classifier performance. It is preferred over AUC ROC when problem is characterized by large class imbalance. On the hold-out test set ShapeGTB outperformed general-purpose methods by a large margin (AP 0.93 vs. 0.47-0.55). On the external validation set ShapeGTB aggregated performance was comparable with FATHMM-MKL (AP 0.47 vs. AP 0.42). However, shapes of precision-recall curves for those methods were very different: FATHMM-MKL displayed high precision only for small subset of examples, while ShapeGTB precision was relatively stable even for large values of recall. Differences between results obtained for the two datasets suggest that ClinVar-derived positives have different characteristics and pose a greater challenge. We speculated that the gap between ShapeGTB and reference tools on the hold-out test is due to inclusion of shape features and their interactions with GC content. To verify this, we randomly permuted these features in our test set and evaluated performance of ShapeGTB again on permuted data sets. AP of ShapeGTB with GC-derived features permuted was 0.80, with shape-derived features permuted 0.44, and with both kinds of features permuted 0.35 (Figure 5). This once more corroborates the hypothesis that shape features together with GC content provide important information for distinguishing functional SNPs in our data set.

DISCUSSION AND CONCLUSIONS

Here, we report the influence of the combined multifactor analysis of DNA shape and other descriptors in prediction of functional effect of promoter variants. Previously, Parker et al. has demonstrated that the nucleotide alternations can significantly affect the DNA structure causing changes in protein binding affinity and phenotype (Parker, et al., 2009). From our analysis, it is clear that changes in the geometry of DNA molecule are important features for the task of prioritization of functional regulatory variants within promoter regions. General conclusions that can be drawn from our study are as follows: a) shape features work very locally, what is important is what happens in the closest neighborhood of the mutated nucleotide, b) DNA chain rotations are more important than minor groove width, c) differences of properties of the mutated variant and the reference motif are the most meaningful. This picture is inherently complicated with the presence of feature interdependencies – mostly between GC content and shape features. It is impossible to make predictions based on DNA shape alone, it is meaningful only with respect to the sequence content.
Interestingly, in our method the most informative indicator of variant functional impact is whether the introduced nucleotide changes the GC-content. The GC composition has been previously linked to DNA thermostability, bendability and potential for conformational transition between B- and Z-forms, that relate to chromatin accessibility (Vinogradov, 2003). The instances of GC-rich sequence motifs have been shown to play an important role in transcription regulation through their connection with nucleosome occupancy and TF binding (Peckham, et al., 2007; Wang, et al., 2012). In our opinion, high rank of GC-ratio derivatives is a result of promoter properties, which distinguish it from other regulatory elements (Lenhard, et al., 2012; Andersson, et al., 2014). GC-ratio may not be highly ranked if similar analysis would be performed on other regulatory elements, which are not associated with promoter regions (e.g. splicing elements or insulators).

There is a vast amount of literature on complex networks of relations between nucleotide types and various shape attributes (Yoon, et al., 1988; Florquin, et al., 2005; Rohs, et al., 2005; Samanta, et al., 2009). For instance, the distribution of water around the minor groove shows specificity to the DNA sequence as the availability of the hydrogen bond forming atoms changes. Variation in DNA sequence may affect DNA flexibility by influencing the magnitude of propeller twist. Specific base pairs combinations have different electrostatic potentials and prefer specific stacking geometry (Samanta, et al., 2009). The results of Tillo and Hughes have highlighted that GC-ratio influences nearly all aspects of DNA structure (Tillo and Hughes, 2009). The most pronounced dependency has been observed between GC-ratio and propeller twist (Ponomarenko, et al., 1999). Deb et al. previously reported the effect of an A/T base pair replacement by a G/C base pair on narrowing of minor grows through negative propeller twisting (Deb, et al., 1987). This pair has also been rated high in our feature interdependencies ranking. To sum up, it appears that only a specific configuration of local structural feature values can meet the requirements of a functional genomic element and that causative mutation substantially disrupt it consensus.

The data derived from ChIP-seq experiments and DNaseI hypersensitivity assays have relatively low resolution generally ranging from 200 to 8 kbp (Park, 2009; Pique-Regi, et al., 2011; ENCODE Project Consortium, 2012). Our analysis shows that histone modification and TFBS ChIP-seq peaks along with TF disruption p-value and DNaseI hypersensitivity data, being used in genome-wide setting, have no discriminative power for promoter region sequence variations. This is especially true for TSS-balanced version of our data sets (Supplementary material 5). It is important to stress that features based on histone modifications and TFBS have different meaning than those derived directly from DNA sequence and shape. The former may represent statistical relationships connected with high-level functioning of the organism, while the latter may correspond to low-level binding mechanisms and biophysical properties of the DNA. Our method is able to make successful predictions using only low-level features, which may inform the study of low-level mechanisms behind functional SNP mutations.

There is a strong need in the field for entirely independent, high-quality collection of regulatory elements variants categorized by type of non-coding sequence and functional status. Such collection would allow constructing reliable tests sets to validate and compare available methods. According to Li and Wang (2017) analysis, human genetic variants databases such as HGMD and ClinVar contain contradictory entries and incorrectly categorized variants due to the lack of primary review of evidence.

In our experiments, our method outperformed significantly the reference tools on our own dataset, and exhibited better recall on external dataset. However, caution is required in drawing final conclusions from the comparison. Our model targeted promoter regions specifically, while the other tools were trained on larger subsets of non-coding regions. It is also possible that our
validation set, at least partially, overlapped with training sets used by other algorithms. We believe that the main reason behind good performance of ShapeGTB is the inclusion of shape features. Without them the expected performance is on par with the other methods (AP 0.44 on hold-out test set).

In summary, we demonstrated that the local shape features of DNA surrounding single nucleotide coupled with the GC-content and sequence composition are sufficient for single nucleotide variant prioritization within promoter regions of human genes. Our results additionally confirmed the interdependencies between alternations in the GC-content and local DNA shape features. Given that the shape vectors implicitly reflect electrostatics, base stacking, hydration profiles (Przytycka and Levens, 2015), including DNA shape into model results in functional reduction of the number of features and therefore a great simplification of the method. We believe that local DNA shape features carry a vast amount of information and their applicability should be investigated further. In the future, we plan to extend our analysis on all types of regulatory elements in non-coding regions of human genome.

Acknowledgements
We gratefully acknowledge the authors of FATHMM-MKL method, especially Dr. Hashem Shihab, for sharing their control dataset with us.

REFERENCES


Ignatieva EV, Levitsky VG, Yudin NS, Moshkin MP, and Kolchanov NA. 2014. Genetic basis of
olfactory cognition: extremely high level of DNA sequence polymorphism in promoter regions of
the human olfactory receptor genes revealed using the 1000 Genomes Project dataset. Front
Psychol 5:247. 10.3389/fpsyg.2014.00247

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, and
Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor
binding patterns. Nat Genet 43:264-268. 10.1038/ng.759

Kanhere A, and Bansal M. 2005. Structural properties of promoters: similarities and differences
between prokaryotes and eukaryotes. Nucleic Acids Res 33:3165-3175. 10.1093/nar/gki627

duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U
SA 100:11484-11489. 10.1073/pnas.1932072100

human genome browser at UCSC. Genome Res 12:996-1006. 10.1101/gr.229102

Kinsella RJ, Kahari S, Haidar S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D,

framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310-
315. 10.1038/ng.2892

Ladds MA, Thompson AP, Slip DJ, Hocking DP, and Harcourt RG. 2016. Seeing It All:
Evaluating Supervised Machine Learning Methods for the Classification of Diverse Otariid
Behaviours. PLoS One 11:e0166898. 10.1371/journal.pone.0166898

relevant variants. Nucleic Acids Res 44:D862-868. 10.1093/nar/gkv1222

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P,
Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D,
Reddy TE, Rozowsky J, Shores N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov
10.1101/gr.136184.111

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, and Beer MA. 2015. A
method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47:955-961.
10.1038/ng.3331

Lenhard B, Sandelin A, and Carninci P. 2012. Metazoan promoters: emerging characteristics and
insights into transcriptional regulation. Nat Rev Genet 13:233-245. 10.1038/nrg3163

Li Q, and Wang K. 2017. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-

functional non-coding regions in the human genome through integrated analysis of annotation data.
Sci Rep 5:10576. 10.1038/srep10576

Morey C, Mookherjee S, Rajasekaran G, and Bansal M. 2011. DNA free energy-based promoter
prediction and comparative analysis of Arabidopsis and rice genomes. Plant Physiol 156:1300-


10.1016/j.cell.2014.08.009


Table 1 (on next page)

Cross-validation classification results for different feature groups on TSS-balanced data set.
<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>AUC_std</th>
<th>Accuracy</th>
<th>Accuracy_std</th>
<th>F1</th>
<th>F1_std</th>
<th>Precision</th>
<th>Precision_std</th>
<th>Recall</th>
<th>Recall_std</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.9764</td>
<td>0.0133</td>
<td>0.9258</td>
<td>0.0247</td>
<td>0.8803</td>
<td>0.0456</td>
<td>0.8840</td>
<td>0.0643</td>
<td>0.8792</td>
<td>0.0480</td>
<td>227.0</td>
</tr>
<tr>
<td>Best 25</td>
<td>0.9243</td>
<td>0.0345</td>
<td>0.8449</td>
<td>0.0418</td>
<td>0.7551</td>
<td>0.0785</td>
<td>0.7456</td>
<td>0.1079</td>
<td>0.7713</td>
<td>0.0710</td>
<td>25.0</td>
</tr>
<tr>
<td>Sequence</td>
<td>0.5555</td>
<td>0.0473</td>
<td>0.6162</td>
<td>0.0584</td>
<td>0.3170</td>
<td>0.0416</td>
<td>0.3766</td>
<td>0.0878</td>
<td>0.2834</td>
<td>0.0453</td>
<td>52.0</td>
</tr>
<tr>
<td>GC content</td>
<td>0.7765</td>
<td>0.0525</td>
<td>0.7051</td>
<td>0.0626</td>
<td>0.4934</td>
<td>0.0634</td>
<td>0.5560</td>
<td>0.1054</td>
<td>0.4546</td>
<td>0.0713</td>
<td>8.0</td>
</tr>
<tr>
<td>Shape</td>
<td>0.5571</td>
<td>0.0566</td>
<td>0.6251</td>
<td>0.0690</td>
<td>0.2546</td>
<td>0.0597</td>
<td>0.3574</td>
<td>0.0994</td>
<td>0.2039</td>
<td>0.0551</td>
<td>88.0</td>
</tr>
<tr>
<td>Conservation</td>
<td>0.5440</td>
<td>0.0416</td>
<td>0.6569</td>
<td>0.0522</td>
<td>0.2693</td>
<td>0.0764</td>
<td>0.4313</td>
<td>0.1547</td>
<td>0.2003</td>
<td>0.0545</td>
<td>10.0</td>
</tr>
<tr>
<td>TFBS ChIP-seq</td>
<td>0.5255</td>
<td>0.0482</td>
<td>0.6674</td>
<td>0.0755</td>
<td>0.2416</td>
<td>0.0707</td>
<td>0.4722</td>
<td>0.1589</td>
<td>0.1683</td>
<td>0.0550</td>
<td>12.0</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>0.5664</td>
<td>0.0641</td>
<td>0.6270</td>
<td>0.0690</td>
<td>0.3342</td>
<td>0.0702</td>
<td>0.3987</td>
<td>0.1069</td>
<td>0.2994</td>
<td>0.0844</td>
<td>38.0</td>
</tr>
<tr>
<td>DNase I</td>
<td>0.5846</td>
<td>0.0622</td>
<td>0.6662</td>
<td>0.0817</td>
<td>0.1474</td>
<td>0.0674</td>
<td>0.4088</td>
<td>0.1921</td>
<td>0.0914</td>
<td>0.0431</td>
<td>1.0</td>
</tr>
<tr>
<td>Dinucleotide content</td>
<td>0.5205</td>
<td>0.0615</td>
<td>0.6211</td>
<td>0.0614</td>
<td>0.2354</td>
<td>0.0798</td>
<td>0.3407</td>
<td>0.1323</td>
<td>0.1858</td>
<td>0.0647</td>
<td>16.0</td>
</tr>
<tr>
<td>Max TFBS log-odds ratio score + TF disruption pval</td>
<td>0.5141</td>
<td>0.0613</td>
<td>0.6773</td>
<td>0.0824</td>
<td>0.0364</td>
<td>0.0381</td>
<td>0.3812</td>
<td>0.3618</td>
<td>0.0193</td>
<td>0.0205</td>
<td>2.0</td>
</tr>
<tr>
<td>Sequence + GC content</td>
<td>0.7689</td>
<td>0.0404</td>
<td>0.6997</td>
<td>0.0465</td>
<td>0.5029</td>
<td>0.0578</td>
<td>0.5426</td>
<td>0.1159</td>
<td>0.4816</td>
<td>0.0477</td>
<td>60.0</td>
</tr>
<tr>
<td>Shape + GC content</td>
<td>0.9175</td>
<td>0.0313</td>
<td>0.8395</td>
<td>0.0333</td>
<td>0.7399</td>
<td>0.0627</td>
<td>0.7557</td>
<td>0.1052</td>
<td>0.7332</td>
<td>0.0583</td>
<td>96.0</td>
</tr>
<tr>
<td>Sequence + GC content + Shape</td>
<td>0.9787</td>
<td>0.0140</td>
<td>0.9446</td>
<td>0.0208</td>
<td>0.9124</td>
<td>0.0381</td>
<td>0.8894</td>
<td>0.0616</td>
<td>0.9400</td>
<td>0.0437</td>
<td>148.0</td>
</tr>
<tr>
<td>Sequence + GC content + Shape + TF disruption pval</td>
<td>0.9787</td>
<td>0.0132</td>
<td>0.9471</td>
<td>0.0231</td>
<td>0.9161</td>
<td>0.0400</td>
<td>0.8899</td>
<td>0.0624</td>
<td>0.9468</td>
<td>0.0401</td>
<td>149.0</td>
</tr>
<tr>
<td>Sequence + GC content + TF disruption pval + Max TFBS log-odds ratio score</td>
<td>0.9782</td>
<td>0.0139</td>
<td>0.9442</td>
<td>0.0189</td>
<td>0.9118</td>
<td>0.0318</td>
<td>0.8933</td>
<td>0.0595</td>
<td>0.9346</td>
<td>0.0374</td>
<td>150.0</td>
</tr>
<tr>
<td>Sequence + GC content + TFBS ChIP-seq</td>
<td>0.7902</td>
<td>0.0332</td>
<td>0.7206</td>
<td>0.0410</td>
<td>0.5252</td>
<td>0.0614</td>
<td>0.5698</td>
<td>0.0934</td>
<td>0.4933</td>
<td>0.0616</td>
<td>72.0</td>
</tr>
<tr>
<td>Sequence + GC content + Histone modifications</td>
<td>0.7981</td>
<td>0.0426</td>
<td>0.7249</td>
<td>0.0464</td>
<td>0.5359</td>
<td>0.0656</td>
<td>0.5882</td>
<td>0.1170</td>
<td>0.5054</td>
<td>0.0664</td>
<td>98.0</td>
</tr>
</tbody>
</table>
Figure 1 (on next page)

Mean importance of 5 best scoring features in each feature group.
Figure 2 (on next page)

Joint distributions of the two most important features in the two classes. WT-SNP difference corresponds to difference of scores between reference (wild type) and mutated (SNP) variants.
WT-SNP difference in GC content

WT-SNP difference in PROT score at 3rd position

class
Non-functional
Functional
Figure 3 (on next page)

The strongest feature interdependencies.
Figure 4 (on next page)

Precision-recall curves for different classifiers.
a) Hold-out test set (HGMD+1000G)
- CADD (AP 0.47)
- FATHMM-MKL (AP 0.55)
- DeepSEA (AP 0.54)
- ShapeGTB (AP 0.93)

b) External validation set (ClinVar+1000G)
- CADD (AP 0.23)
- FATHMM-MKL (AP 0.42)
- DeepSEA (AP 0.37)
- ShapeGTB (AP 0.47)
Figure 5 (on next page)

Precision-recall curves for variants of ShapeGTB in which feature vectors from specific feature groups were permuted (effectively reducing their usefulness).

-GC corresponds to classifier with GC-derived features permuted, -Shape corresponds to classifier.
a) Hold-out test set (HGMD+1000G)

ShapeGTB (AP 0.93)
ShapeGTB(-GC) (AP 0.80)
ShapeGTB(-Shape) (AP 0.44)
ShapeGTB(-GC,-Shape) (AP 0.35)

b) External validation set (ClinVar+1000G)

ShapeGTB (AP 0.47)
ShapeGTB(-GC) (AP 0.31)
ShapeGTB(-Shape) (AP 0.07)
ShapeGTB(-GC,-Shape) (AP 0.03)