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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

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 Malkowska[1#], Julian Zubek[2#], Dariusz Plewczynski[2,3], Lucjan Wyrwicz[1]

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

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 ≥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

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, Forrest ARR, Carninci P, Rehli M, and Sandelin A. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507:455-461. 10.1038/nature12787

Babajide Mustapha I, and Saeed F. 2016. Bioactive Molecule Prediction Using Extreme Gradient Boosting. Molecules 21. 10.3390/molecules21080983

Chiaromonte F, Yap VB, and Miller W. 2002. Scoring pairwise genomic sequence alignments. Pac Symp Biocomput:115-126.

Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, and Rohs R. 2016. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics 32:1211-1213. 10.1093/bioinformatics/btv735

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423. 10.1093/bioinformatics/btp163

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74. 10.1038/nature11247

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, and Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025. 10.1371/journal.pcbi.1001025

Deb S, Tsui S, Koff A, DeLucia AL, Parsons R, and Tegtmeyer P. 1987. The T-antigen-binding domain of the simian virus 40 core origin of replication. J Virol 61:2143-2149.

Draminski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, and Komorowski J. 2008. Monte Carlo feature selection for supervised classification. Bioinformatics 24:110-117. 10.1093/bioinformatics/btm486

Dror I, Zhou T, Mandel-Gutfreund Y, and Rohs R. 2014. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res 42:430-441. 10.1093/nar/gkt862

Florquin K, Saeys Y, Degroeve S, Rouze P, and Van de Peer Y. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Res 33:4255-4264. 10.1093/nar/gki737

Freeman GS, Lequieu JP, Hinckley DM, Whitmer JK, and de Pablo JJ. 2014. DNA shape dominates sequence affinity in nucleosome formation. Phys Rev Lett 113:168101. 10.1103/PhysRevLett.113.168101

Friedman JH, and Meulman JJ. 2003. Multiple additive regression trees with application in epidemiology. Stat Med 22:1365-1381. 10.1002/sim.1501

Gan Y, Guan J, and Zhou S. 2012. A comparison study on feature selection of DNA structural properties for promoter prediction. BMC Bioinformatics 13:4. 10.1186/1471-2105-13-4

Gao L, Uzun Y, Gao P, He B, Ma X, Wang J, Han S, and Tan K. 2018. Identifying noncoding risk variants using disease-relevant gene regulatory networks. Nat Commun 9:702. 10.1038/s41467-018-03133-y

Gardiner EJ, Hunter CA, Lu XJ, and Willett P. 2004. A structural similarity analysis of double-helical DNA. J Mol Biol 343:879-889. 10.1016/j.jmb.2004.08.092

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68-74. 10.1038/nature15393

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng JJ, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, and Snyder M. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489:91-100. 10.1038/nature11245

Goni JR, Perez A, Torrents D, and Orozco M. 2007. Determining promoter location based on DNA structure first-principles calculations. Genome Biol 8:R263. 10.1186/gb-2007-8-12-r263

Greenbaum JA, Pang B, and Tullius TD. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res 17:947-953. 10.1101/gr.6073107

Guenther MG, Levine SS, Boyer LA, Jaenisch R, and Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130:77-88. 10.1016/j.cell.2007.05.042

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, and Ren B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311-318. 10.1038/ng1966

Hon GC, Hawkins RD, and Ren B. 2009. Predictive chromatin signatures in the mammalian genome. Hum Mol Genet 18:R195-201. 10.1093/hmg/ddp409

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

Kent WJ, Baertsch R, Hinrichs A, Miller W, and Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484-11489. 10.1073/pnas.1932072100

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996-1006. 10.1101/gr.229102

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, and Flicek P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030. 10.1093/database/bar030

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J. 2014. A general 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

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, and Maglott DR. 2016. ClinVar: public archive of interpretations of clinically 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, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, and Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813-1831. 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-AMP Guidelines. Am J Hum Genet 100:267-280. 10.1016/j.ajhg.2017.01.004

Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, and Zhao H. 2015. A statistical framework to predict 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-

1315. 10.1104/pp.110.167809

Nishida H, Suzuki T, Kondo S, Miura H, Fujimura Y, and Hayashizaki Y. 2006. Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. Chromosome Res 14:203-211. 10.1007/s10577-006-1036-7

Nishizaki SS, and Boyle AP. 2017. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends Genet 33:34-45. 10.1016/j.tig.2016.10.008

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669-680. 10.1038/nrg2641

Parker SC, Hansen L, Abaan HO, Tullius TD, and Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. Science 324:389-392. 10.1126/science.1169050

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, and Weng Z. 2007. Nucleosome positioning signals in genomic DNA. Genome Res 17:1170-1177. 10.1101/gr.6101007

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, and Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21:447-455. 10.1101/gr.112623.110

Ponomarenko JV, Ponomarenko MP, Frolov AS, Vorobyev DG, Overton GC, and Kolchanov NA. 1999. Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics 15:654-668.

Przytycka TM, and Levens D. 2015. Shapely DNA attracts the right partner. Proc Natl Acad Sci U S A 112:4516-4517. 10.1073/pnas.1503951112

Quang D, Chen Y, and Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31:761-763. 10.1093/bioinformatics/btu703

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, and Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279-283. 10.1038/nature09692

Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein CB, Regev A, and Bernstein BE. 2011. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147:1628-1639. 10.1016/j.cell.2011.09.057

Ritchie GR, Dunham I, Zeggini E, and Flicek P. 2014. Functional annotation of noncoding sequence variants. Nat Methods 11:294-296. 10.1038/nmeth.2832

Rohs R, Sklenar H, and Shakked Z. 2005. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. Structure 13:1499-1509. 10.1016/j.str.2005.07.005

Rohs R, West SM, Sosinsky A, Liu P, Mann RS, and Honig B. 2009. The role of DNA shape in protein-DNA recognition. Nature 461:1248-1253. 10.1038/nature08473

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, and Kent WJ. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43:D670-681. 10.1093/nar/gku1177

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, and Kent WJ. 2013. ENCODE data in the

UCSC Genome Browser: year 5 update. Nucleic Acids Res 41:D56-63. 10.1093/nar/gks1172

Samanta S, Mukherjee S, Chakrabarti J, and Bhattacharyya D. 2009. Structural properties of polymeric DNA from molecular dynamics simulations. J Chem Phys 130:115103. 10.1063/1.3078797

San Lucas FA, Wang G, Scheet P, and Peng B. 2012. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics 28:421-422. 10.1093/bioinformatics/btr667

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W. 2003. Human-mouse alignments with BLASTZ. Genome Res 13:103-107. 10.1101/gr.809403

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, and Ren B. 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488:116-120. 10.1038/nature11243

Sheridan RP, Wang WM, Liaw A, Ma J, and Gifford EM. 2016. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J Chem Inf Model 56:2353-2360. 10.1021/acs.jcim.6b00591

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, and Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31:1536-1543. 10.1093/bioinformatics/btv009

Sivolob AV, and Khrapunov SN. 1995. Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. J Mol Biol 247:918-931. 10.1006/jmbi.1994.0190

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, and Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1-9. 10.1007/s00439-013-1358-4

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, and Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. Nature 489:75-82. 10.1038/nature11232

Tillo D, and Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10:442. 10.1186/1471-2105-10-442

Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. Nucleic Acids Res 31:1838-1844.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, and Weng Z. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22:1798-1812. 10.1101/gr.139105.112

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, and Hughes TR. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158:1431-1443. 10.1016/j.cell.2014.08.009

Wyrwicz LS, Gaj P, Hoffmann M, Rychlewski L, and Ostrowski J. 2007. A common cis-element in promoters of protein synthesis and cell cycle genes. Acta Biochim Pol 54:89-98.

Yoon C, Prive GG, Goodsell DS, and Dickerson RE. 1988. Structure of an alternating-B DNA helix and its relationship to A-tract DNA. Proc Natl Acad Sci U S A 85:6332-6336.

Zhou J, and Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12:931-934. 10.1038/nmeth.3547

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, and Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res 41:W56-62. 10.1093/nar/gkt437

Table 1(on next page)

Cross-validation classification results for different feature groups on TSS-balanced data set.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

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

Figure 1(on next page)

Mean importance of 5 best scoring features in each feature group.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

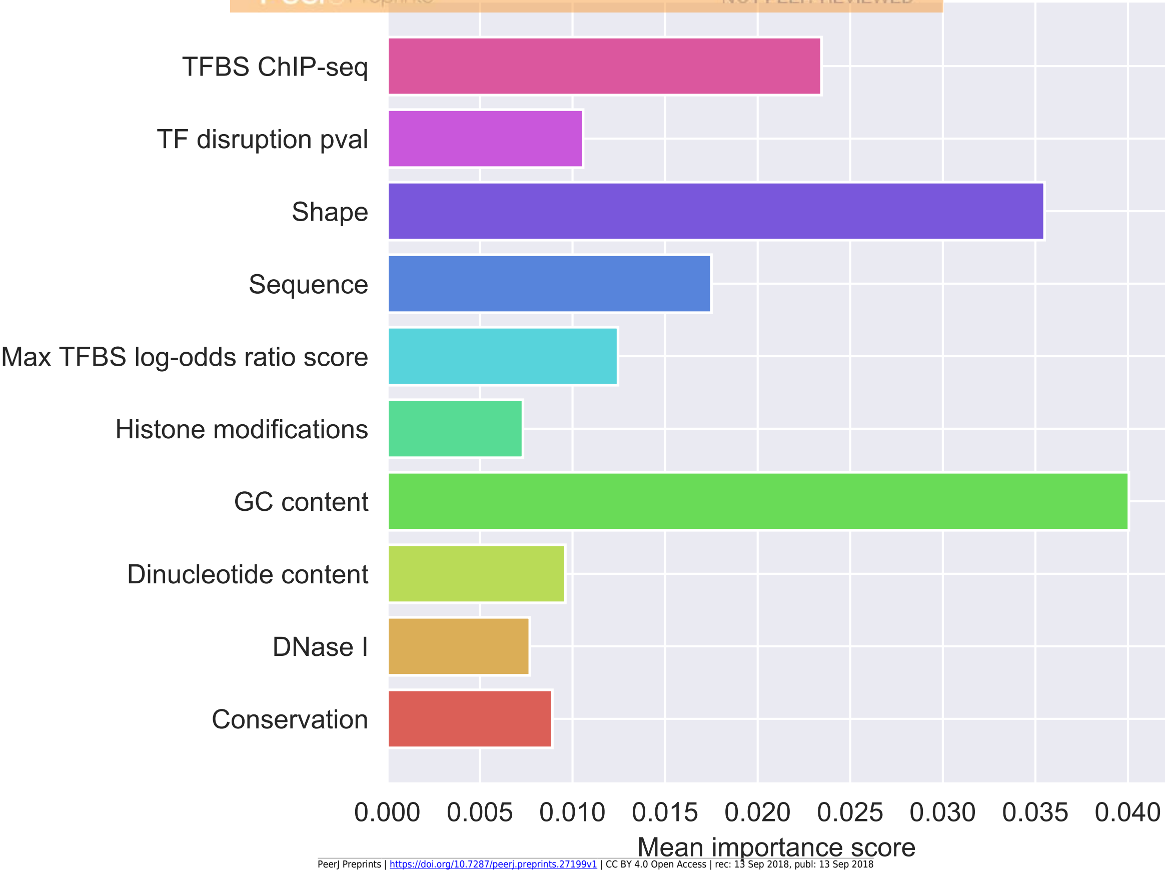
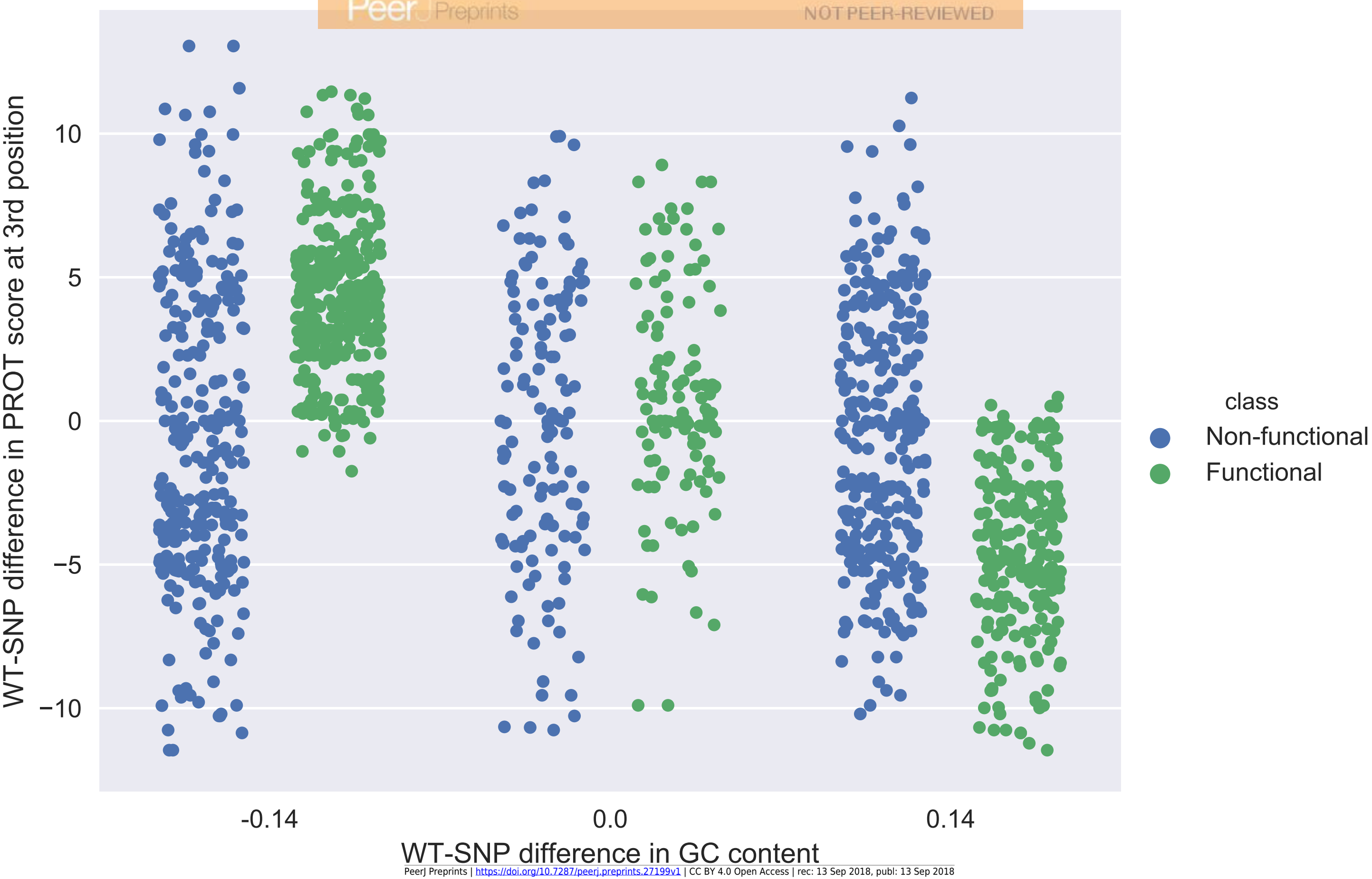

TFBS ChIP-seq
TF disruption pval
Shape
Sequence
Max TFBS log-odds ratio score
Histone modifications
GC content
Dinucleotide content
DNase I
Conservation
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Mean importance score

Figure 2(on next page)

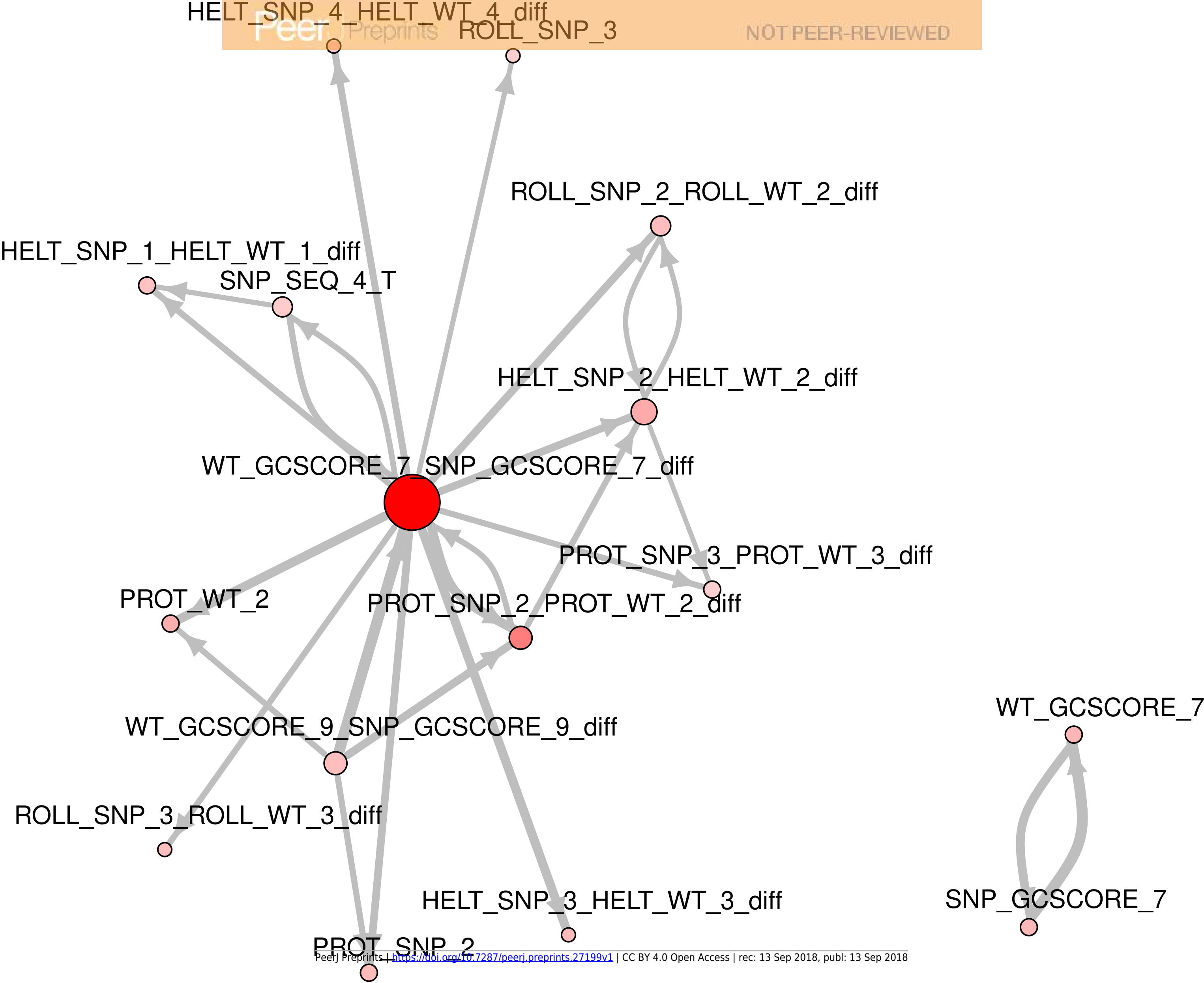
Joint distributions of the two most important features in the two classes. WT-SNP difference corresponds to difference of scores between reference (wild type) and mutated (SNP) variants.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

Figure 3(on next page)

The strongest feature interdependencies.


HELT_SNP_4_HELT_WT_4_diff
ROLL_SNP_3
ROLL_SNP_2_ROLL_WT_2_diff
HELT_SNP_1_HELT_WT_1_diff
SNP_SEQ_4_T
HELT_SNP_2_HELT_WT_2_diff
WT_GCSCORE_7_SNP_GCSCORE_7_diff
PROT_SNP_3_PROT_WT_3_diff
PROT_WT_2
PROT_SNP_2_PROT_WT_2_diff
WT_GCSCORE_7
WT_GCSCORE_9_SNP_GCSCORE_9_diff
ROLL_SNP_3_ROLL_WT_3_diff
HELT_SNP_3_HELT_WT_3_diff
SNP_GCSCORE_7
PROT_SNP_2

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018

Figure 4(on next page)

Precision-recall curves for different classifiers.

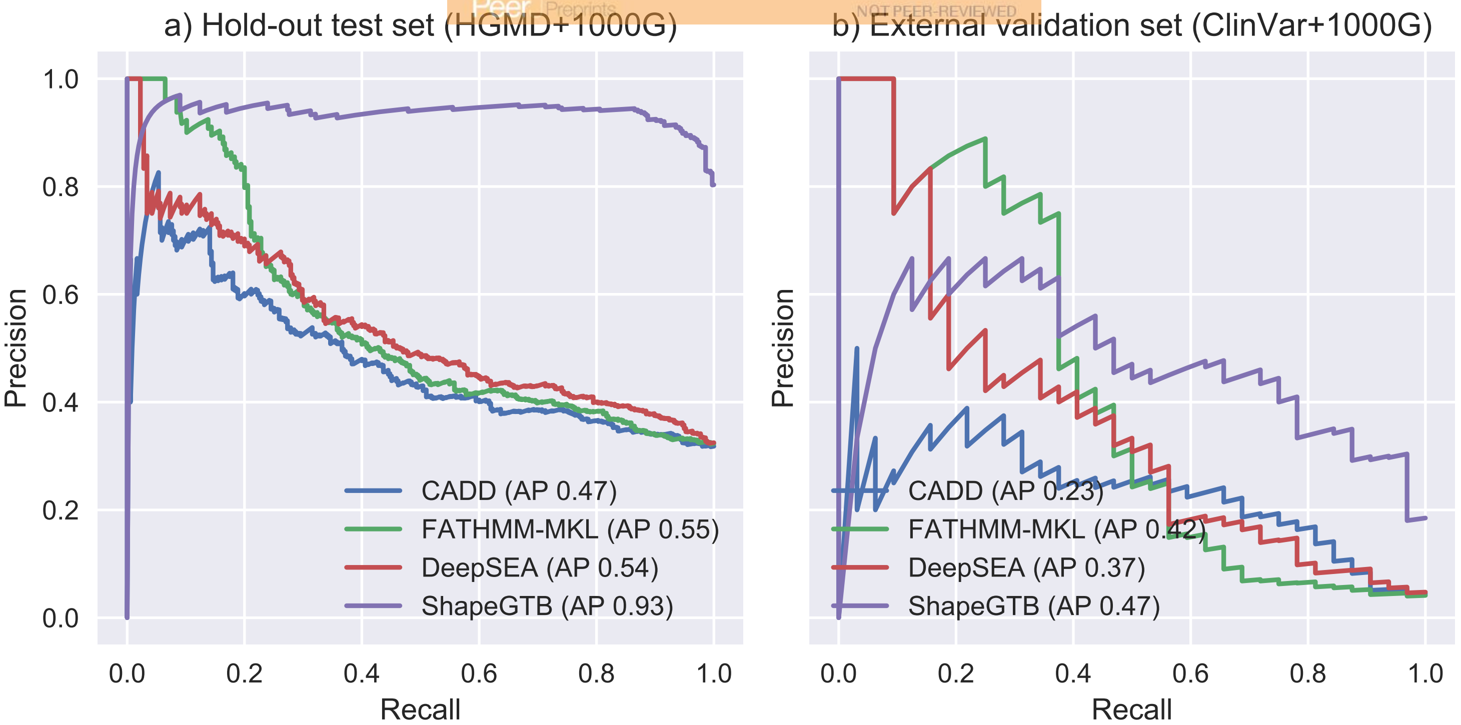
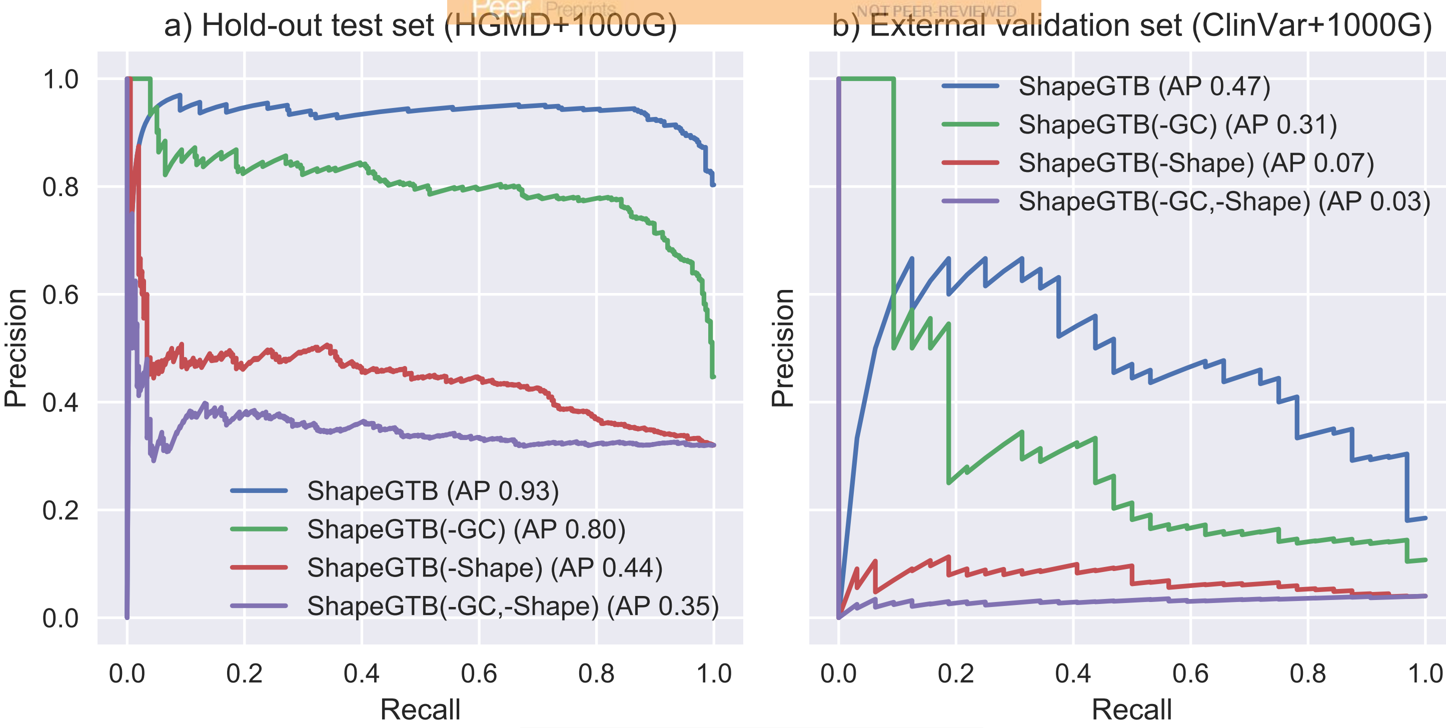

a) Hold-out test set (HGMD+1000G)
b) External validation set (ClinVar+1000G)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
CADD (AP 0.47)
FATHMM-MKL (AP 0.55)
DeepSEA (AP 0.54)
ShapeGTB (AP 0.93)
CADD (AP 0.23)
FATHMM-MKL (AP 0.42)
DeepSEA (AP 0.37)
ShapeGTB (AP 0.47)

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)

b) External validation set (ClinVar+1000G)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27199v1 | CC BY 4.0 Open Access | rec: 13 Sep 2018, publ: 13 Sep 2018