



Predicting functional variants in enhancer and promoter elements using RegulomeDB

Journal:	<i>Human Mutation</i>
Manuscript ID	humu-2018-0650_CAGI5.R1
Wiley - Manuscript type:	Special Article
Date Submitted by the Author:	26-Apr-2019
Complete List of Authors:	Dong, Shengcheng; University of Michigan, Department of Computational Medicine and Bioinformatics Boyle, Alan; University of Michigan, Department of Computational Medicine and Bioinformatics; University of Michigan, Department of Human Genetics
Key Words:	variation, functional genomics, gene regulation, MPRA, machine learning

SCHOLARONE™
Manuscripts

Predicting functional variants in enhancer and promoter elements using RegulomeDB

Shengcheng Dong¹ and Alan P. Boyle^{1,2*}

1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

2. Department of Human Genetics, University of Michigan, Ann Arbor, MI

*e-mail: apboyle@umich.edu

Grant numbers

NIH U41 HG009293, U41 HG007346, and R13 HG006650.

Abstract

Here we present a computational model, SURF (Score of Unified Regulatory Features), that predicts functional variants in enhancer and promoter elements. SURF is trained on data from massively parallel reporter assays and predicts the effect of variants on reporter expression levels. It achieved the top performance in the Fifth Critical Assessment of Genome Interpretation “Regulation Saturation” challenge. We also show that features queried through RegulomeDB, which are direct annotations from functional genomics data, help improve prediction accuracy beyond transfer learning features from DNA sequence-based deep learning models. Some of the most important features include DNase footprints, especially when coupled with complementary ChIP-seq data. Furthermore, we found our model achieved good performance on predicting allele specific transcription factor binding events. As an extension to the current scoring system in RegulomeDB, we expect our computational model to prioritize variants in regulatory regions, thus help the understanding of functional variants in noncoding regions that lead to disease.

Keywords

variation, functional genomics, gene regulation, MPRA, machine learning

Introduction

Evidence from Genome Wide Association Studies (GWAS) has provided us with insights into human phenotypes by identifying variation statistically associated with diseases (Welter et al., 2014). However,

1 GWAS is confounded by linkage disequilibrium when identifying the causal variants. Thus, it is desirable to
2 extend these studies beyond association to an understanding of biological impact. Unfortunately,
3
4 determining the function of these variants remains a major challenge, especially for single-nucleotide
5
6 polymorphisms (SNPs) in non-coding regions of the genome, where most of these GWAS variants fall
7
8 (Hindorff et al., 2009; Hnisz et al., 2013).
9

10
11
12 The advent of functional genomics assays has assisted us in mapping disease causative SNPs from
13
14 GWAS. By intersecting the position of variants with regulatory elements identified from these assays,
15
16 computational tools have been developed to prioritize SNPs in non-coding regions (Nishizaki & Boyle,
17
18 2017). Tools such as RegulomeDB (Boyle et al., 2012), GWAS3D (Li, Wang, Xia, Sham, & Wang, 2013),
19
20 and HaploReg (Ward & Kellis, 2012) have reduced time-consuming experiments for validation. Machine
21
22 learning methods have been widely applied to integrate the annotations from functional genomics assays in
23
24 a more sophisticated way, and thus produce more robust and accurate predictions (Kircher et al., 2014;
25
26 Lee et al., 2015). More recently, the rapid development of deep learning techniques has enabled mining in
27
28 high-dimensional sequences data. Some examples include DeepSEA (Zhou & Troyanskaya, 2015),
29
30 DeepBind (Alipanahi, DeLong, Weirauch, & Frey, 2015), DanQ (Quang & Xie, 2016), Define (Wang, Tai, E,
31
32 & Wei, 2018), and Basenji (Kelley et al., 2018). However, since data sets used for training in those
33
34 algorithms vary, comparisons across different models can become a problem considering there is currently
35
36 no gold-standard for evaluation (Nishizaki & Boyle, 2017).
37
38
39
40
41

42 One independent method for evaluating the performance of these tools is through the use of massively
43
44 parallel reporter assays (MPRA) wherein libraries that are derived from PCR-based saturation mutagenesis
45
46 have been applied to test the effect of variants in a putative regulatory region. These assays can measure
47
48 the functional effect of variants on the expression level of a reporter construct in a high-throughput manner
49
50 allowing for rapid testing of large numbers of variants. Kircher and collaborators performed MPRA for
51
52 17,500 single nucleotide variants (SNVs) in 9 promoters and 5 enhancers with clinical relevance (Inoue &
53
54 Ahituv, 2015; Patwardhan et al., 2009; Tewhey et al., 2016). This dataset allows for an unbiased
55
56 comparison of computational tools used for variant prioritization and was used in this manner for the Fifth
57
58
59
60

1 Critical Assessment of Genome Interpretation (CAGI5) “Regulation Saturation” challenge. Participants
2 were asked to predict the functional effects of variants in these regulatory regions as measured by the
3 reporter expression.
4
5
6
7

8 We present a machine learning-based computational framework, SURF (Score of Unified Regulatory
9 Features), which combines features from RegulomeDB and DeepSEA, to predict the effect of variants on
10 expression in promoters and enhancers. Our model achieved the top performance in the CAGI5
11 “Regulation Saturation” challenge. We also demonstrate that direct features from functional genomics data
12 improve the prediction accuracy in addition to features from DNA sequence-based deep learning models.
13
14
15
16
17
18
19
20

21 **Background**

22 **Datasets in CAGI5 Regulation Saturation Challenge**

23 The regulation saturation challenge assessed 17,500 SNVs in 5 human disease associated enhancers
24 (IRF4, IRF6, MYC, SORT1, ZFAND3) and 9 promoters (F9, GP1BB, HBB, HBG, HNF4A, LDLR, MSMB,
25 PKLR, TERT) in a massively parallel reporter assay (**Fig. 1A**). The MPRA libraries were derived from
26 saturation mutagenesis of regulatory regions up to 600bp length, with a random change rate of 1 per 100
27 bases.
28
29
30
31
32
33
34
35
36
37

38 Approximately 25% of all measured SNVs were used for training (4,650 SNVs in total), and the remaining
39 75% of the data were held from competitors and used for testing by an independent assessor. The count of
40 transcribed RNA and DNA of the transfected plasmid library was modeled by applying multiple linear
41 regression (**Fig. 1B**). The coefficients (“effect size”) and re-scaled p-values (“confidence score”) from
42 regression were provided in the training set. The SNV with a confidence scores greater or equal to 0.1 (i.e.
43 p-value of 10^{-5}) was defined as “has an expression effect”.
44
45
46
47
48
49
50
51
52

53 **Tasks in CAGI5 Regulation Saturation Challenge**

54 For each variant in testing set, the participants were asked to submit prediction of effect of the variant in
55 one of the three cases: repressive, activating, or no effect (“Direction”), and the probability of a correct
56
57
58
59
60

1 assignment of the prediction (“P_Direction”). The participants also needed to submit a prediction of the
2 confidence score for each variant, as well as the standard error of the prediction (“SD”).
3
4
5

6 **Methods**

7
8 For each variant in training and test data, we created features from functional genomics data obtained from
9 RegulomeDB (Boyle et al., 2012). We also used sequence-based features from DeepSEA (Zhou &
10 Troyanskaya, 2015). We further trained a random forest model to predict direction of variant effects and
11 confidence score (**Fig. 1**).
12
13
14
15
16
17
18

19 **Features**

20
21 The first six features were created by querying each variant through RegulomeDB (Boyle et al., 2012). All
22 ENCODE data represented in RegulomeDB is from the 2012 freeze and subsequent publication. We
23 assigned binary values to represent if the position of the queried variant overlaps the following functional
24 genomics regions:
25
26
27
28

29 **1. Transcription factor (TF) binding site**

30
31 TF CHIP-seq peaks were from ENCODE data.
32
33

34 **2. Open chromatin site**

35
36 DNase-peaks were from ENCODE data.
37
38

39 **3. TF motifs**

40
41 TF motif matches were called using positional weight matrices (PWM) from RegulomeDB (Boyle et al.,
42 2012). Positional weight matrices were from TRANSFAC (Matys et al., 2006), JASPAR CORE (Bryne et
43 al., 2008), UniPROBE (Newburger & Bulyk, 2009) and Jolma et al (Jolma et al., 2013).
44
45
46

47 **4. Matched TF motif**

48
49 TF motif matches were obtained as described in feature 3, but further requiring the PWM motif matching
50 with a TF binding peak of the same TF from CHIP-seq in the same position.
51
52

53 **5. DNase footprint**

54
55 DNase footprints were called by combining PWMs and DNase-seq data sets. We used footprint calls
56
57
58
59
60

1 from Boyle et al (Boyle et al., 2011), Pique-Regi et al (Pique-Regi et al., 2011) and Piper et al (Piper et
2 al., 2013).

3 4 **6. Matched DNase footprint**

5
6 DNase footprints were obtained as described in feature 5, but further requiring the PWM motif matching
7 with a TF binding peak from ChIP-seq in the same position.
8
9

10
11
12 We also included additional numeric features:

13 14 **7. ChIP-seq signal**

15
16 We calculated the maximum TF ChIP-seq signal from feature 1 for each position in the regulatory
17 regions.
18
19

20 21 **8. Maximum information content change of TF motif**

22
23 For each variant, we calculated the information content change of PWMs called in feature 3 and took
24 the one with maximum absolute value.
25
26

27 28 **9. Maximum information content change of matched TF motif**

29
30 For each variant, we calculated the information content change of matched PWMs called in feature 4
31 and took the one with maximum absolute value.
32
33

34 35 **10. DeepSEA scores**

36 We passed a vcf file of all variants through DeepSEA model (from <http://deepsea.princeton.edu/>) to
37 predict chromatin effects of each mutation on 919 functional genomics features, including chromatin
38 accessibility, TF binding and histone modification. We used the difference between reference and
39 alternative alleles of those 919 functional genomics features in our model. We also included the
40 functional significance score for each variant, which considers chromatin effects as well as evolutionary
41 conservation.
42
43
44
45
46
47
48
49
50

51 **Random forest training**

52
53 A random forest model was trained to make predictions for both direction of effects and confidence scores.
54
55 Specifically, we used the R package *randomForest* version 4.6-12 with ntree=500 (Liaw & Wiener, 2002).
56
57
58
59
60

For direction prediction, we first classified training data from all studied regulatory regions into three groups using the following criteria:

1. Repressive (-1): confidence greater than or equal to 0.1 and effect size smaller than 0 (736 in total).
2. Activating (+1): confidence greater than or equal to 0.1 and effect size greater than 0 (374 in total).
3. No effect (0): confidence smaller than 0.1 (3,540 in total).

We then trained three binary classifiers for each label with a random forest model and predicted the label with the highest probability. We assigned “P_Direction” column with the prediction probability from the model. In order to generate a confidence prediction, we trained a random forest regression model on confidence scores and calculated the standard deviation of predictions from 500 trees in “SD” column.

Performance evaluation

Group performance was evaluated on correlation coefficients and the area under the receiver operating characteristic (AUROC). Pearson and Spearman correlation coefficients were calculated for predicted direction and effect size from MPRA on variants in test set in the same way as the assessors. Three categories of AUROC were assessed: variants with positive effects versus negative effects, variants with positive effects versus all variants, and variants with negative effects versus all variants. Predicted directions were treated as labels and effect sizes were used as probability scores. To increase the sensitivity of model comparisons, we also provided continuous value predictions as requested by the assessors, which are a transformation from “P_Direction”:

$$Direction' = \begin{cases} P_Direction & \text{if } Direction = 1 \\ -P_Direction & \text{if } Direction = -1 \\ 1 - P_Direction & \text{if } Direction = 0 \text{ and } D_{-1} < D_{+1} \\ P_Direction - 1 & \text{if } Direction = 0 \text{ and } D_{-1} > D_{+1} \end{cases},$$

where D_i is the probability of class i ($i = -1, 0, +1$) from random forest model.

Pearson correlation with continuous predictions were reevaluated among top three methods by the assessors (Supp. Table S1).

Allele specific transcription factor (TF) binding analysis

Allele specific TF binding sites were defined as variants that result in stronger binding of a TF to one allele at heterozygous sites in an individual. We applied AlleleDB pipeline to call allele specific TF binding sites using ChIP-seq data downloaded from ENCODE project (Chen et al., 2016). 1,814 allele specific binding sites were called in GM12878 cell line from 76 TFs at an FDR of 5%. To test the performance of our binary classifier trained on CAGI5 data, we also built a control set including 10,783 variants having equal ChIP-seq read counts on two alleles at heterozygous sites. For all 48,630 heterozygous sites, we calculated the allelic ratio defined by the ratio between number of ChIP-seq reads from the allele with stronger binding affinity and total number of reads from two alleles. For cases where multiple TFs shared a heterozygous variant, we took the maximum ratio.

Results

SURF outperforms other groups in CAGI5 regulation saturation challenge

SURF combines features from RegulomeDB, which directly intersects variants with functional genomics annotations, and DeepSEA, which generates transfer learning features from genomics assays. For assessment, both Pearson and Spearman correlation coefficients were calculated for predicted direction and effect size from MPRA on test data. To examine how false positive rate changes with true positive rate, the area under the receiver operating characteristic (AUROC) was also calculated (**Table 1**). Overall, we were close to group 7 on correlation coefficients, and we outperformed all groups in terms of all three categories of AUROC, especially in the case when distinguishing between variants with positive and those with negative effects on expression level. In addition, we note that it is generally easier to predict negative effects compared with positive effects, which might be because there were more examples with negative effects in training set.

Model performance in different enhancers and promoters

We assessed our performance in each of the 5 enhancers and 9 promoters (**Fig. 2**). Continuous value predictions were used for calculating Pearson correlation with effect sizes. We observe no evident difference in performance between enhancers and promoters, but predictions on enhancers are more consistent in terms of AUROC performance. Also, our model performance has no strong association with

1 cell types. The four regions in HEK293T (HNF4A, MSMB, TERT and MYC) have a wide range of
2 performance. Overall, we predicted most accurately in regions of: MYC (HEK293T), PKLR (K562) and HBB
3 (HEL_92.1.7). Interestingly, the cell line HEL_92.1.7 has no corresponding functional genomics data from
4 the ENCODE project. In addition, ZFAND3 data is from mouse pancreatic beta cell lines (MIN6). These
5
6 imply our model is able to predict these effects from the available data in other cell types.
7
8
9

10 **Features from RegulomeDB provide complementary information to DeepSEA scores**

11 We next analyzed the predictive importance of RegulomeDB features. We calculated Pearson correlation
12 of features and absolute value of effect sizes in test data (**Fig. 3A**). All features have positive correlation,
13 which is consistent with the fact that the variants in functional regulatory elements have a higher chance of
14 affecting the expression level downstream. Among all binary features from RegulomeDB, features such as
15 matched TF motif and matched DNase footprint have the highest correlation coefficients, which indicates
16 that integrating sequence information with evidence from functional genomics data directly into one feature
17 assists prediction accuracy. We further examined two of the most predictive features in the region of MYC
18 enhancer, where we achieved the best AUROC compared with other enhancers and promoters. As shown
19 in **Fig. 3B**, these two features from RegulomeDB, DNase footprint and matched DNase footprint, are
20 largely in agreement with the position of variants leading to significant change of gene expression beyond
21 DeepSEA scores.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 **Predicting allele specific TF binding events**

41 To test the generality of our model, we next evaluated how SURF performs on predicting allele specific TF
42 binding events identified from ChIP-seq data. We collected 1,848 variants associated with allele specific
43 binding in GM12878 cell line, and then generated prediction scores using the binary classifier we trained
44 from variants with no effects versus the rest of the variants in CAGI5 training set. Overall, our model is able
45 to predict allele specific binding events with a fairly good performance (AUROC=0.6218; AUPRC=0.2298).
46 We further relaxed our thresholds to examine the performance on a wider spectrum of allelic ratio, which is
47 defined by the ratio between number of ChIP-seq reads from the allele with stronger binding affinity and
48 total number of reads from two alleles. We found a significant difference in prediction scores for
49
50
51
52
53
54
55
56
57
58
59
60

heterozygous sites showing balanced (allelic ratio smaller than 0.6) and imbalanced (allelic ratio equal or larger than 0.9) TF binding affinity (**Fig. 4**, p-value = 9.735e-311 from a t-test).

Discussion

Understanding the function of variants in noncoding regions remains a major challenge to interpret results from GWAS studies. The CAGI5 Regulation Saturation challenge has provided a valuable dataset for developing prediction models on regulatory variants leading to significant effects on expression level. Here we described our model, SURF, based on our existing resource RegulomeDB, that achieves the top performance in this challenge (Table 1). However, one limitation of the evaluation with AUROC is that the imbalance rate was different across groups, which makes it hard to compare. A more accurate comparison is the correlation between continuous prediction scores and effect sizes from MPRA, which is shown in Supp. Table S1 but only available from three groups.

We found that the direct annotations from functional genomics data queried through RegulomeDB enables the improvement of prediction beyond the transfer learning features from the DeepSEA model. One possible reason to explain the improvement is that the chromatin features from underrepresented cell types in deep learning model are compensated by direct annotations from RegulomeDB. Thus, continued working on RegulomeDB resource, including updates and expansion of available data from ENCODE project, will enable us to develop prediction models with better accuracy. For example, 3D chromatin interaction data illustrating loops between enhancers and promoters can be used to assign target genes of variants in regulatory elements. In addition, ATAC-seq as an alternative method for studying chromatin accessibility will potentially give us complementary information to DNase-seq.

Furthermore, instead of obtaining general features through all available cell types in RegulomeDB as we did in this challenge, it is possible to query features in a cell type-specific way to improve performance. Although a previous study suggests that limiting features to be cell type specific does not increase prediction accuracy for MPRA data (Kreimer, A., et al. 2017), it is worth exploring further whether this is due to the limitation of MPRA to capture cell type-specific activity. Another strategy is to integrate cell type-

1 specific features with a generic model trained with all available cell types, thus taking advantage of a
2 sufficient set of training data as well as a retention of cell type-specific information.
3
4
5

6 The initial premise behind the development and scoring in the RegulomeDB tool was that functional
7 genomics data is key to understanding and prioritizing variants that may be disrupting transcription factor
8 binding and thus having a direct effect on gene expression. We have shown that these data have aided our
9 model to perform well on MPRA training data and improve the ability to predict allele specific TF binding
10 events. Multiple studies have successfully applied RegulomeDB to infer regulatory variants in cancer
11 genomes (Melton, Reuter, Spacek, & Snyder, 2015; Sharma, Jiang, & De, 2018), and continued work is
12 needed with the increasing availability of cancer whole genome data. Encouraged by these results, we are
13 currently developing a newer version of RegulomeDB, which will provide all the features we used in this
14 challenge, including the allelic scores such as information content change of TF motifs. We will also make
15 our prediction scores available to general users, thus to help research on prioritizing non-coding variants in
16 various contexts.
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 **Acknowledgements**

32 SD and APB were supported by NIH U41 HG009293. The CAGI experiment coordination is supported by
33 NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. We thank the organizers of the
34 Fifth Critical Assessment of Genome Interpretation for hosting the challenge. We also thank Adam Diehl,
35 Sierra Nishizaki, Ningxin Ouyang and Samuel Zhao for constructive feedback.
36
37
38
39
40
41
42
43

44 **References**

- 45 Alipanahi, B., DeLong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of
46 DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8), 831-838. doi:
47 10.1038/nbt.3300
48
49 Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., . . . Snyder, M. (2012).
50 Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, 22(9),
51 1790-1797. doi: 10.1101/gr.137323.112
52
53
54
55
56
57
58
59
60

- 1 Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., . . . Furey, T. S. (2011). High-resolution
2 genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*,
3
4 21(3), 456-464. doi: 10.1101/gr.112656.110
5
- 6 Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., . . . Sandelin, A. (2008).
7
8 JASPAR, the open access database of transcription factor-binding profiles: new content and tools in
9
10 the 2008 update. *Nucleic Acids Res*, 36(Database issue), D102-106. doi: 10.1093/nar/gkm955
11
12
- 13 Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., . . . Gerstein, M. (2016). A
14
15 uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.
16
17 *Nat Commun*, 7, 11101. doi: 10.1038/ncomms11101
18
- 19 Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A.
20
21 (2009). Potential etiologic and functional implications of genome-wide association loci for human
22
23 diseases and traits. *Proc Natl Acad Sci U S A*, 106(23), 9362-9367. doi: 10.1073/pnas.0903103106
24
- 25 Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., . . . Young, R. A. (2013).
26
27 Super-enhancers in the control of cell identity and disease. *Cell*, 155(4), 934-947. doi:
28
29 10.1016/j.cell.2013.09.053
30
- 31 Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*,
32
33 106(3), 159-164. doi: 10.1016/j.ygeno.2015.06.005
34
35
- 36 Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., . . . Taipale, J. (2013). DNA-binding
37
38 specificities of human transcription factors. *Cell*, 152(1-2), 327-339. doi: 10.1016/j.cell.2012.12.009
39
- 40 Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential
41
42 regulatory activity prediction across chromosomes with convolutional neural networks. *Genome*
43
44 *Res*, 28(5), 739-750. doi: 10.1101/gr.227819.117
45
- 46 Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general
47
48 framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3),
49
50 310-315. doi: 10.1038/ng.2892
51
- 52 Kreimer, A., Yan, Z., Ahituv, N., & Yosef, N. (2017). Meta-analysis of massive parallel reporter assay
53
54 enables functional regulatory elements prediction. *BioRxiv*, doi: 10.1101/202002
55
56
57
58
59
60

- 1 Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A
2 method to predict the impact of regulatory variants from DNA sequence. *Nat Genet*, *47*(8), 955-961.
3
4 doi: 10.1038/ng.3331
5
- 6 Li, M. J., Wang, L. Y., Xia, Z., Sham, P. C., & Wang, J. (2013). GWAS3D: Detecting human regulatory
7 variants by integrative analysis of genome-wide associations, chromosome interactions and histone
8 modifications. *Nucleic Acids Res*, *41*(Web Server issue), W150-158. doi: 10.1093/nar/gkt456
9
10
11
12
- 13 Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News* *2*(3), 18-22.
14
- 15 Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., . . . Wingender, E. (2006).
16 TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic*
17 *Acids Res*, *34*(Database issue), D108-110. doi: 10.1093/nar/gkj143
18
19
20
- 21 Melton, C., Reuter, J. A., Spacek, D. V., & Snyder, M. (2015). Recurrent somatic mutations in regulatory
22 regions of human cancer genomes. *Nat Genet*, *47*(7), 710-716. doi: 10.1038/ng.3332
23
24
- 25 Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data
26 on protein-DNA interactions. *Nucleic Acids Res*, *37*(Database issue), D77-82. doi:
27
28
29
30
31
- 32 Nishizaki, S. S., & Boyle, A. P. (2017). Mining the Unknown: Assigning Function to Noncoding Single
33 Nucleotide Polymorphisms. *Trends Genet*, *33*(1), 34-45. doi: 10.1016/j.tig.2016.10.008
34
35
- 36 Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution
37 analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol*, *27*(12),
38
39
40
41
42
- 43 Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., & Ott, S. (2013). Wellington: a novel method
44 for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res*,
45
46
47
48
- 49 Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate
50 inference of transcription factor binding from DNA sequence and chromatin accessibility data.
51
52
53
54
55
56
57
58
59
60

- 1 Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for
2 quantifying the function of DNA sequences. *Nucleic Acids Res*, *44*(11), e107. doi:
3
4 10.1093/nar/gkw226
5
- 6 Sharma, A., Jiang, C., & De, S. (2018). Dissecting the sources of gene expression variation in a pan-
7 cancer analysis identifies novel regulatory mutations. *Nucleic Acids Res*, *46*(9), 4370-4381. doi:
8
9 10.1093/nar/gky271
10
- 11 Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., . . . Sabeti, P. C. (2016). Direct
12
13 Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay.
14
15 *Cell*, *165*(6), 1519-1529. doi: 10.1016/j.cell.2016.04.027
16
17
- 18 Wang, M., Tai, C., E, W., & Wei, L. (2018). DeFine: deep convolutional neural networks accurately quantify
19
20 intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding
21
22 variants. *Nucleic Acids Res*, *46*(11), e69. doi: 10.1093/nar/gky215
23
24
- 25 Ward, L. D., & Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and
26
27 regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*,
28
29 *40*(Database issue), D930-934. doi: 10.1093/nar/gkr917
30
31
- 32 Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The
33
34 NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*,
35
36 *42*(Database issue), D1001-1006. doi: 10.1093/nar/gkt1229
37
- 38 Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based
39
40 sequence model. *Nat Methods*, *12*(10), 931-934. doi: 10.1038/nmeth.3547
41
42
43

44 Figure legends

45
46 **FIGURE 1** The workflow of our method. A) The effect of variants in promoters and enhancers was tested
47 through massively parallel reporter assays (MPRA). B) Effect size modeled from regression for each
48 variant was provided with 25% of data (white area) used for training and 75% of data (grey area) hidden
49 from participants. C) A multiclass random forest model is trained by combining features from RegulomeDB
50 and DeepSEA on training data. D) Prediction of variants with significant effects (circled points) is made
51 from random forest models.
52
53
54
55
56
57
58
59
60

1
2 **FIGURE 2** Performance across regions. Cell type names are appended at the end of promoter and
3
4 enhancer regions. The average performance across all regions is also shown.
5
6
7

8 **FIGURE 3** Features from RegulomeDB facilitate prediction. A) Pearson correlation of features from
9
10 RegulomeDB and absolute value of effect sizes from MPRA in test data. B) A region of the MYC enhancer
11
12 in HEK293T cell line showing measured MPRA data with SNVs having significant effect circled. Two binary
13
14 features from RegulomeDB (DNase footprint and DNase footprint with matched TF ChIP-seq peak) show
15
16 agreement with the position of these variants. DeepSEA scores also identify some of the functional variants
17
18 in this enhancer.
19
20
21
22

23 **FIGURE 4** Boxplot of prediction scores for heterozygous sites showing balanced and imbalanced TF
24
25 binding affinity from two alleles. Allelic ratio is calculated by the number of ChIP-seq reads from the allele
26
27 with stronger binding affinity divided by total number of reads from two alleles.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

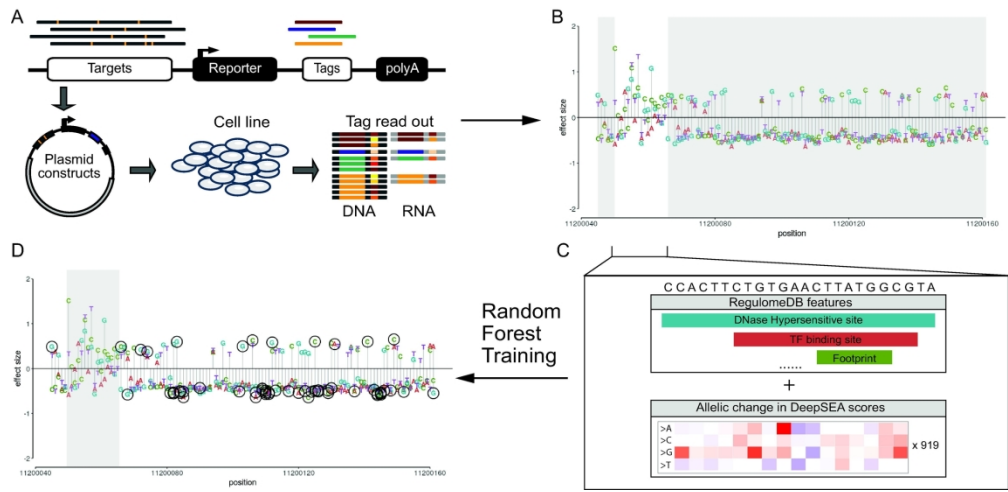


FIGURE 1 The workflow of our method. A) The effect of variants in promoters and enhancers was tested through massively parallel reporter assays (MPRA). B) Effect size modeled from regression for each variant was provided with 25% of data (white area) used for training and 75% of data (grey area) hidden from participants. C) A multiclass random forest model is trained by combining features from RegulomeDB and DeepSEA on training data. D) Prediction of variants with significant effects (circled points) is made from random forest models.

206x99mm (300 x 300 DPI)

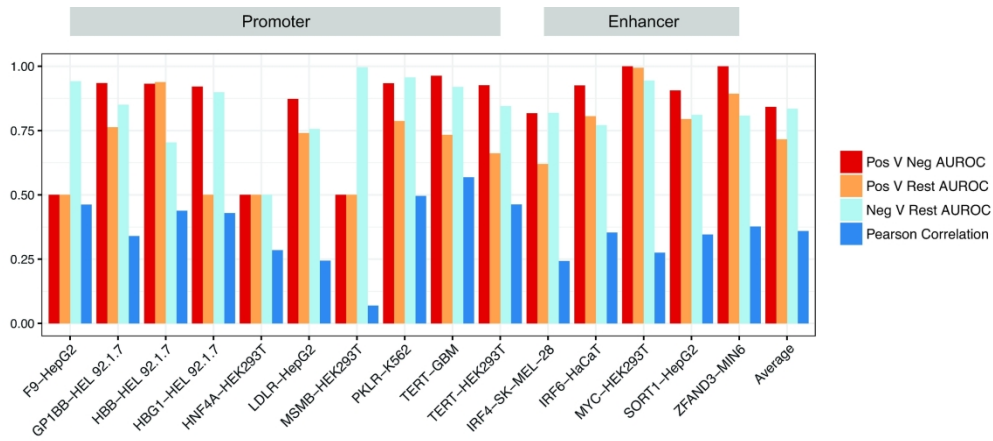


FIGURE 2 Performance across regions. Cell type names are appended at the end of promoter and enhancer regions. The average performance across all regions is also shown.

254x111mm (300 x 300 DPI)

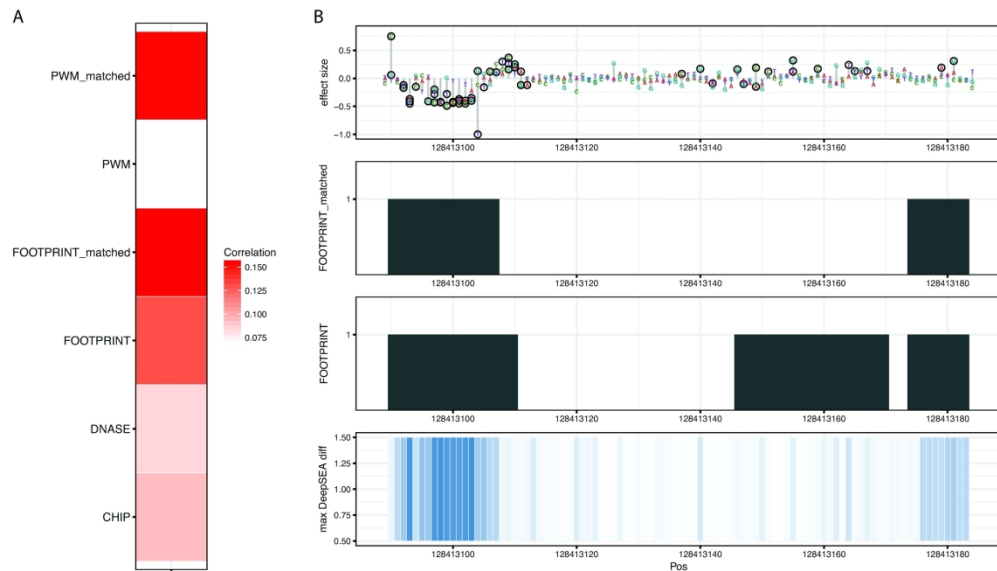


FIGURE 3 Features from RegulomeDB facilitate prediction. A) Pearson correlation of features from RegulomeDB and absolute value of effect sizes from MPRA in test data. B) A region of the MYC enhancer in HEK293T cell line showing measured MPRA data with SNVs having significant effect circled. Two binary features from RegulomeDB (DNase footprint and DNase footprint with matched TF ChIP-seq peak) show agreement with the position of these variants. DeepSEA scores also identify some of the functional variants in this enhancer.

223x127mm (300 x 300 DPI)

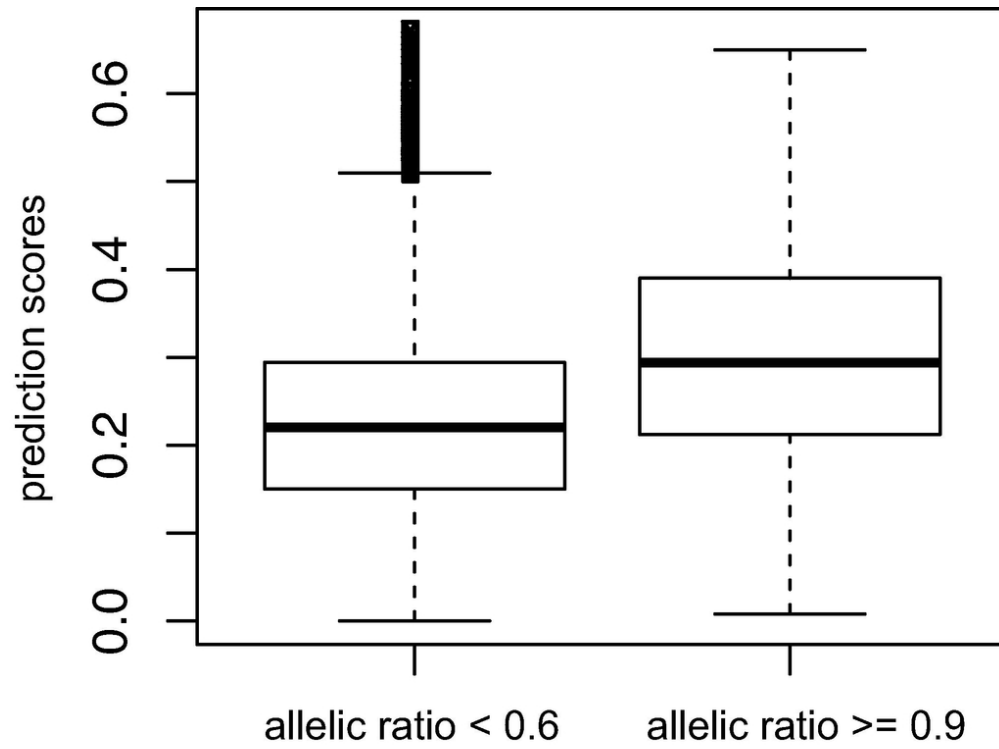


FIGURE 4 Boxplot of prediction scores for heterozygous sites showing balanced and imbalanced TF binding affinity from two alleles. Allelic ratio is calculated by the number of ChIP-seq reads from the allele with stronger binding affinity divided by total number of reads from two alleles.

86x67mm (300 x 300 DPI)

Table 1. Correlation and AUROC for predicting direction of variant effects across all participated groups. The best submission of each group was selected and the best performance of each category is bolded. AUPRC and correlation with continuous prediction scores were calculated in Supp. Table S1.

Participant (lab-submission)	Pearson correlation	Spearman correlation	Pos V Neg AUROC	Pos V Rest AUROC	Neg V Rest AUROC
3-4 (our group)	0.301	0.239	0.842	0.716	0.835
7-3	0.318	0.249	0.762	0.706	0.776
5-6	0.255	0.235	0.714	0.608	0.691
1-2	0.069	0.046	0.544	0.553	0.636
6-1	0.103	0.094	0.537	0.544	0.584
4-2	0.041	0.033	0.556	0.528	0.571

Supp. Table S1. AUPRC for predicting direction of variant effects across all participated groups and Pearson correlation with continuous scores (only available with first three groups).

Participant (lab-submission)	Pos V Neg AUPRC	Pos V Rest AUPRC	Neg V Rest AUPRC	Pearson correlation with continuous scores
3-4 (our group)	0.637	0.097	0.308	0.452
7-3	0.611	0.165	0.312	0.451
5-6	0.639	0.261	0.434	0.277
1-2	0.446	0.051	0.147	NA
6-1	0.007	0.004	0.680	NA
4-2	0.576	0.063	0.079	NA