Received: 18 January 2019 | Revised: 17 April 2019 | Accepted: 15 May 2019

DOI: 10.1002/humu.23797

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

SPECIAL ARTICLE

Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay

Dustin Shigaki[1] | Orit Adato[2] | Aashish N. Adhikari[3] | Shengcheng Dong[4] | Alex Hawkins-Hooker[5] | Fumitaka Inoue[6] | Tamar Juven-Gershon[2] | Henry Kenlay[5] | Beth Martin[7] | Ayoti Patra[8] | Dmitry D. Penzar[9,10] | Max Schubach[11,12] | Chenling Xiong[6] | Zhongxia Yan[12] | Alan P. Boyle[4] | Anat Kreimer[6,13] | Ivan V. Kulakovskiy[9,10,14,15] | John Reid[5,16] | Ron Unger[2] | Nir Yosef[13] | Jay Shendure[7] | Nadav Ahituv[6] | Martin Kircher[7,11,12] | Michael A. Beer[1,8]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

[2]The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

[3]Department of Plant and Microbial Biology, University of California, Berkeley, California

[4]Department of Computational Medicine and Bioinformatics and Department of Human Genetics, University of Michigan, Ann Arbor, Michigan

[5]MRC Biostatistics Unit, University of Cambridge, UK

[6]Department of Bioengineering and Therapeutic Sciences and Institute for Human Genetics, University of California San Francisco, San Francisco, California

[7]Department of Genome Sciences, University of Washington, Seattle, Washington

[8]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland

[9]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

[10]School of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

[11]Berlin Institute of Health (BIH), Berlin, Germany

[12]Charité – Universitätsmedizin Berlin, Berlin, Germany

[13]Department of Electrical Engineering and Computer Science and Center for Computational Biology, University of California, Berkeley, California

[14]Institute of Mathematical Problems of Biology, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Russia

[15]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

[16]Alan Turing Institute, British Library, London, UK

Correspondence
Michael A. Beer, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD.
Email: mbeer@jhu.edu

Funding information
National Human Genome Research Institute, Grant/Award Numbers: NIH R13 HG006650, HG007348, NIH U41 HG007346, HG009380; Russian Foundation for Basic Research, Grant/Award Number: 18-34-20024

Abstract

The integrative analysis of high-throughput reporter assays, machine learning, and profiles of epigenomic chromatin state in a broad array of cells and tissues has the potential to significantly improve our understanding of noncoding regulatory element function and its contribution to human disease. Here, we report results from the CAGI 5 regulation saturation challenge where participants were asked to predict the impact of nucleotide substitution at every base pair within five disease-associated human enhancers and nine disease-associated promoters. A library of mutations covering all bases was generated by saturation mutagenesis and altered activity was assessed in a massively parallel reporter assay (MPRA) in relevant cell lines. Reporter expression was measured relative to plasmid DNA to determine the impact of variants. The challenge was to predict the functional effects of variants on reporter

expression. Comparative analysis of the full range of submitted prediction results identifies the most successful models of transcription factor binding sites, machine learning algorithms, and ways to choose among or incorporate diverse datatypes and cell-types for training computational models. These results have the potential to improve the design of future studies on more diverse sets of regulatory elements and aid the interpretation of disease-associated genetic variation.

KEYWORDS
enhancers, gene regulation, machine learning, MPRA, promoters, regulatory variation

1 | INTRODUCTION

Gene regulatory variants are known to play an important role in a number of common human diseases, including diabetes, neuropsychiatric disorders, autoimmune disorders, cardiovascular disease, and cancer. Although some disease-relevant variants have been identified and thoroughly characterized, this set provides insufficient data to test computational methods that aim to find such variants. Gene regulatory variants modulate the strength of interactions between enhancers and promoters and the transcription factors (TFs) that bind them, and alter the cell-specific transcriptional control of gene regulatory networks central to the proper development and functioning of human cells and tissues. Although we have a good basic understanding of the general molecular mechanisms of these interactions, quantitative and predictive models of cell-specific enhancer and promoter function are currently under active development.

Blind community assessments provide the most principled way to gauge the performance of the leading computational prediction models. The 2016 Critical Assessment of Genome Interpretation (CAGI 4) eQTL challenge (Beer, 2017; Kreimer et al., 2017; Tewhey et al., 2016; Zeng, Edwards, Guo, & Gifford, 2017) assessed the effect of common human variation on the enhancer activity in lymphoblast cell lines. It established that the top performing state-of-the-art models of the enhancer activity typically used machine learning methods e.g. gkm-SVM (Ghandi, Lee, Mohammad-Noori, & Beer, 2014; Lee et al., 2015), DeepBind (Alipanahi, Delong, Weirauch, & Frey, 2015), and/or DeepSEA (Zhou & Troyanskaya, 2015)) using features learned from epigenomic chromatin state data (DHS-seq, Histone modification ChIP-seq, TF ChIP-seq, or ATAC-seq) to build models of TF binding specificity (Beer, 2017; Kreimer et al., 2017; Zeng et al., 2017). Here, we significantly extend the earlier CAGI 4 study, and report the results of the 2018 CAGI 5 regulation saturation challenge. In this study, computational groups were asked to submit the predicted impact on expression for every possible base pair mutation within nine disease associated promoters (including TERT, LDLR, F9, HBG1) and five disease-associated enhancers (including IRF4, IRF6, MYC, SORT1) tested by massively parallel reporter assay (MPRA) in one of eight specified cell types (Kircher et al., 2018). This study expands on the CAGI 4 assessment in two key aspects. First, although the CAGI 4 assessments were all in the GM12878 lymphoblast cell line, the CAGI 5 assessment separately tests a wider range of elements (promoters and enhancers) in multiple disease-relevant cell types. Second, the CAGI 4 assessments tested common SNPs linked to GM12878 eQTLs, whereas the current CAGI 5 assessment mutates numerous bases in the 14 elements tested. This approach has the advantage that saturation mutagenesis can test mutations that are not common variants in the human population, are not subject to selection, and thus potentially have a larger impact on the enhancer or promoter activity, whether positive or negative.

Although the CAGI 4 regulation variation experiment established that machine learning models can predict MPRA experiments with moderate precision, the current CAGI 5 experimental design allows us to address some additional fundamental questions, which we hope will be used to improve future experiments to investigate human gene regulatory disease variants. The GM12878 cell line used in CAGI 4 is one of the most well-covered cell lines in the ENCODE epigenomic data sets, so the training data available was already well utilized by previously published models. The CAGI 5 assessment in multiple cell lines is thus potentially more challenging in terms of model training, because of the more diverse selection of cell types. One of our primary results is that multiple groups presented successful ways to incorporate multiple functional datasets of different types into the prediction models. In addition, we can ask to which degree promoters and enhancers have shared or distinct regulatory vocabularies (within the limitations of the sample size), and whether different training designs should be adopted for testing promoters versus distal enhancer regulatory variants.

2 | REGULATION SATURATION CHALLENGE

In the MPRA assay, the activity of the enhancer elements is characterized by a reporter assay linking a candidate enhancer sequence to a minimal promoter and a reporter gene whose 3′-untranslated region includes a unique sequence tag. The reporter vectors are introduced into cell lines as plasmids, and the reporter gene expression for each variant is examined relative to the amount of its plasmid DNA, which is variable because different elements tested have varying rates of synthesis and transfection. If the candidate sequence acts as an enhancer, it will

increase promoter activity and the reporter gene expression in the tissue/cell type of interest. Like enhancers, promoter candidate sequences are also cloned into a plasmid upstream of a tagged reporter (without an additional minimal promoter), and reporter expression is measured as RNA relative to the plasmid DNA to determine the impact of promoter variants.

The underlying MPRA libraries (50k-2 M) were derived from saturation mutagenesis of regulatory regions of up to 600 bp length. Changes to functional sequences from the template sequence with a rate of 1 per 100 bases were created by error-prone PCR, and the resulting PCR products were integrated into plasmid libraries containing random tag sequences. High-throughput sequencing was carried out to determine the tag association with the introduced enhancer/promoter sequence variants (Inoue & Ahituv, 2015; Patwardhan et al., 2012). Promoter and enhancer libraries were transfected into a cell line relevant to the disease phenotype (Table 1). Across three transfection replicates, RNA and DNA was collected and sequenced. The relative abundance of each transcribed RNA tag count in relation to DNA tag counts of the transfected plasmid library provides a digital readout of the transcriptional efficiency of the cis-linked mutant promote or enhancer. Specifically, a multiple linear regression model of log2(RNA) ~ log2(DNA) + N + offset (where RNA and DNA are counts observed for all tags, N is a binary matrix associating tags to sequence variants, and offset normalizes total DNA to RNA counts) was used to assign sequence effects. From this fit, coefficients (corresponding to the columns of matrix N) were assigned as the effects of each sequence variant. The coefficient/regression weight for a given nucleotide can be interpreted as the degree to which it contributes to the gene expression. A more detailed description of the MPRA experimental methods is given in (Kircher et al., 2018).

Participant groups were given the impact of the variants in selected subsets from each region to train their models, consisting of 25% of the sequence, and the remaining 75% of the sequence regions were used for evaluation. An example of the training regions selected and the expression impact for each base is shown in Figure 1 for SORT1, and for all regions tested in Figure S1. Although the reference sequence strand is shown, in all cases the experiment preserved the wild-type orientation of the sequence relative to the TSS, as indicated. The 25% training data is indicated in yellow. The participants were to submit a label of +1 to the variants in the testing set if there was a significant upregulating effect, −1 for a significant downregulating effect, or 0 for very little to no effect on expression. Because of the imprecise interpretation of what a "significant" effect could mean, we used as a primary metric of evaluation the Pearson correlation of the predicted labels (−1,0,1) with the continuous MPRA expression impact scores. We also calculated the AUROC treating this as a discretized classification task (1 vs −1, or 1 vs (0 and −1), or −1 vs (0 and 1), and so forth), but the relative ranking of prediction methods using correlation or AUROC were very similar. In retrospect, the discretization of predictions into three classes (−1,0,1) limited the sensitivity of our model comparisons. For our detailed comparisons among the top models, we asked participants to submit continuous prediction scores for their best-performing models. Each group was allowed to submit multiple separate prediction sets from different models.

3 | RESULTS

Seven groups submitted multiple prediction methods, as described in detail in Methods. Most groups used a combination of epigenomic

TABLE 1 Regions tested

Promoters				
Region	**hg19 coords**	**hg38 coords**	**Length**	**Cell line**
F9	chrX:138612624–138612923	chrX:139530465–139530764	300	HepG2
GP1BB	chr22:19710790–19711173	chr22:19723267–19723650	384	HEL 92.1.7
HBB	chr11:5248252–5248438	chr11:5227022–5227208	187	HEL 92.1.7
HBG1	chr11:5271035–5271308	chr11:5249805–5250078	274	HEL 92.1.7
HNF4A	chr20:42984160–42984444	chr20:44355520–44355804	285	HEK293T
LDLR	chr19:11199907–11200224	chr19:11089231–11089548	318	HepG2
MSMB	chr10:51548988–51549578	chr10:46046243–46046833	591	HEK293T
PKLR	chr1:155271187–155271655	chr1:155301396–155301864	469	K562
TERT	chr5:1295105–1295362	chr5:1294990–1295247	258	HEK293T, GBM
Enhancers				
Region	**hg19 coords**	**hg38 coords**	**Length**	**Cell line**
IRF4	chr6:396143–396593	chr6:396143–396593	451	SK-MEL-28
IRF6	chr1:209989135–209989734	chr1:209815790–209816389	600	HaCaT
MYC	chr8:128413074–128413673	chr8:127400829–127401428	600	HEK293T
SORT1	chr1:109817274–109817873	chr1:109274652–109275251	600	HepG2
ZFAND3	chr6:37775276–37775853	chr6:37807500–37808077	578	MIN6

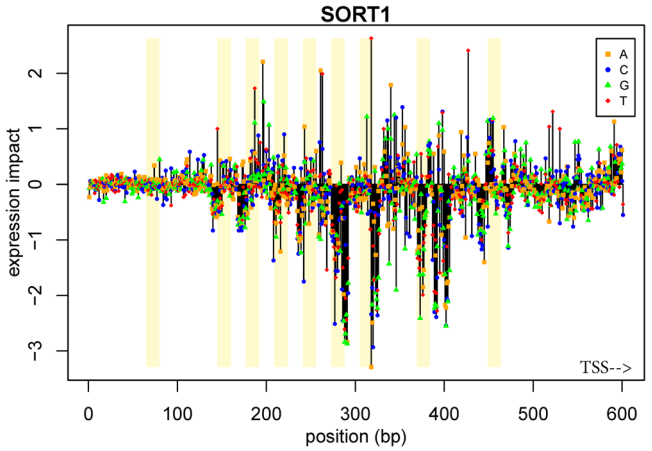

FIGURE 1 MPRA expression data for the SORT1 enhancer. The expression impact of each of the three mutations of each base in the 600 bp enhancer is shown. Clusters of negative impact regions occur near TFBS, which can be disrupted in many ways. Isolated positive impact regions indicate rare creation of TFBS. Training regions indicated in yellow. MPRA, massively parallel reporter assay

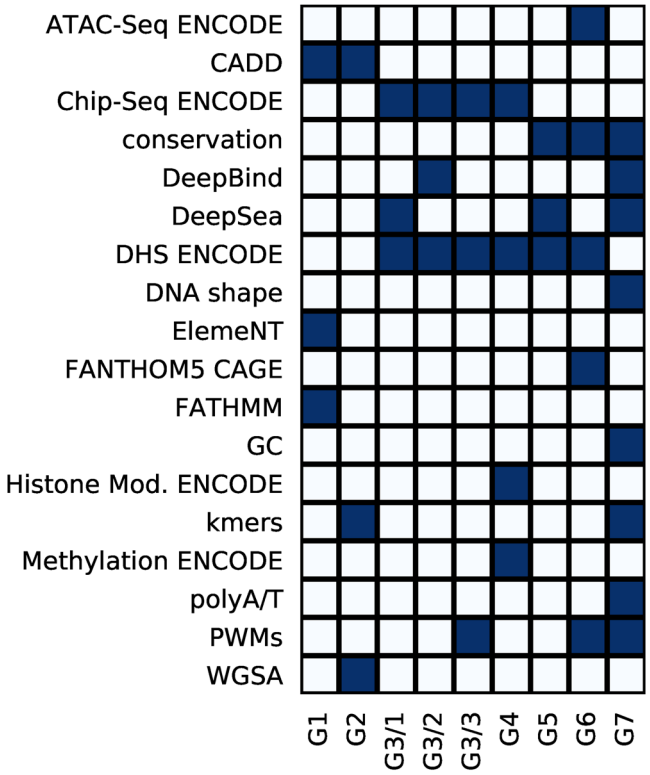

FIGURE 2 Summary of features used by prediction groups

state features (ENCODE DHS-seq, Histone modification ChIP-seq, TF ChIP-seq, and methylation, established marks of promoter and enhancer activity) and DNA sequence-based features (Position Weight Matrices, PWMs, for known TFs from databases, DeepSEA, and DeepBind scores from training machine learning methods to predict ENCODE accessibility and binding, evolutionary conservation, or constraint, kmers, DNA shape, and AT/GC content) to train their models. The conceptual challenge was how to best combine these diverse features to make predictions, and most groups used tree-based classifiers to learn the proper weighting of which of the features best predicted the impact of the mutations in the training data. The classes of prediction features used by each group are summarized in Figure 2. Different methods submitted by each group used slight variations of learning methods or combined feature subsets, so the broad set of methods submitted and wide range of performance allowed us to compare subsets of prediction methods to assess the informative value of different feature identification methods, or different feature subsets or feature selection methods.

The Pearson correlation between the discretized predictions (−1,0,1) and the MPRA expression impact for all regions are shown in Figure 3 and Table S1. A few submissions were late by a few minutes and are labeled 'L' but were fully included in the evaluation. Although there was some variability across regions, the top three methods G3/1, G5/5-7, and G7/3 L had average correlation C = (0.308, 0.255, and 0.318), respectively. Predictions for G5/5-7 were indistinguishable. The distinguishing feature of the top three performing methods is that they all used DNA sequence features derived from Deep Neural Networks (DNN) trained on ENCODE data (DeepSEA or similar network methods). Thus one of the main conclusions of this study is that machine learning-based DNA sequence features are the best predictors of mutation impact in enhancers and promoters, consistent with our previous findings (Beer, 2017; Inoue et al., 2017; Kreimer et al., 2017; Lee et al., 2015). The top three groups all did particularly well on F9 and TERT-GBM, which we will discuss below. None of the methods used gapped-kmer features, which have advantages (Ghandi, Mohammad-Noori, & Beer, 2014) relative to full-length kmers, which were used by groups G2 and G7. As discussed below, gkm-SVM also performed as well as the top submitted methods, which allowed us to use gkm-SVM (Ghandi, Lee et al., 2014) and deltaSVM (Lee et al., 2015) as a previously published benchmark method for comparison, and to evaluate which design choices and subsets of ENCODE training data are most informative without retraining the submissions from the various groups.

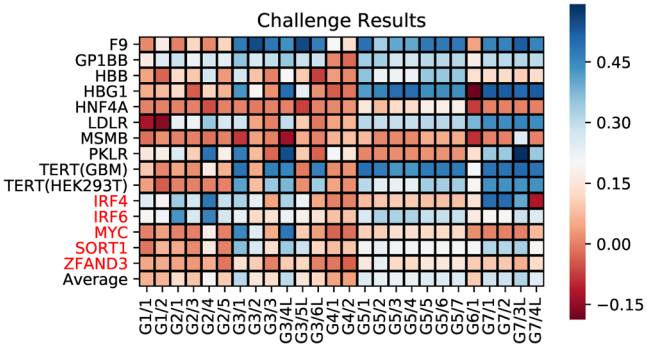

FIGURE 3 Overview of challenge results. The Pearson correlation between the discretized predictions (−1,0,1) and the MPRA expression impact for all regions (promoters labeled in black, enhancers labeled in red) and average correlation over all regions. The top three methods G3/1, G5/5-7, and G7/3 L had an average correlation C = (0.318, 0.255, and 0.318), respectively. MPRA, massively parallel reporter assay

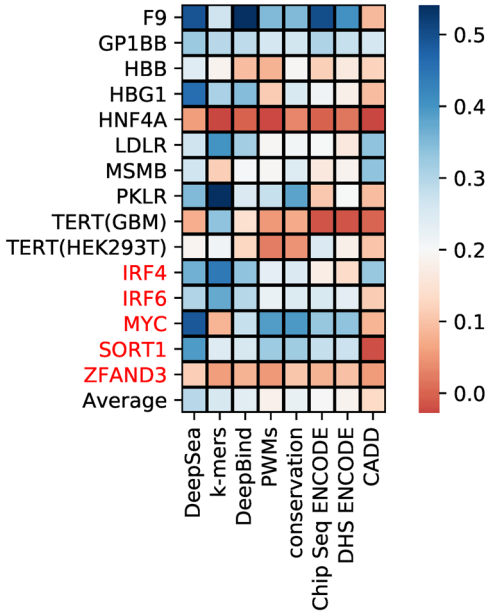

FIGURE 4 Average performance of various feature sets. We averaged the correlation of the best submission from each group using a given feature, as long as it was used by at least two groups. Methods which used DNN derived features (DeepSEA) tended to produce the best performance on average (C = 0.29). DNN, deep neural networks

To assess the predictive value of different types of features, we plot the average correlation of methods using a given feature in Figure 4 across all regions tested. We averaged the correlation of the best submission from each group using a given feature, as long as it was used by at least two groups. We included G3/1-3 because these methods used distinct features, which informed our assessment of relative performance. Methods which used DNN derived features (DeepSEA) tended to produce the best performance on average (C = 0.29). Groups 3 and 7 used the 919 delta *P* outputs from DeepSEA as features, and Group 5 used features from a neural network modeled similar to DeepSEA in training and network structure (Hawkins-Hooker, Kenlay, & Reid, 2018).

After the challenge submission closed, our preliminary analysis indicated that group-to-group differences in assigning cutoffs for the discretized prediction classes were in some cases limiting the robustness of our comparisons, so we asked the three groups with the best-performing methods to submit continuous scores, and reevaluated these continuous predictions, as shown in Figure 5 and Table S1. The Pearson correlation with continuous scores was significantly higher for all three methods, average correlation (C = 0.45, 0.28, 0.45) for method G3/1, G5/5-7, and G7/3 L, respectively. This presents a significant improvement relative to a previously published Method CADD (Kircher et al., 2014), which had C = 0.11.

We compared these three top-performing submissions to the previously published method, deltaSVM. Comparison of the leading three prediction groups indicated the important classes of features, and each leading group arrived at a similar method to successfully combine features from different cell types using the training data. The use of deltaSVM allows us to explore the effects of different training datasets and experimental design choices in more detail without having to retrain the submitted models. deltaSVM is usually trained on histone ChIP-seq, DHS-seq, ATAC-seq, or TF ChIP-seq data from a single cell type. After this approach, we chose the most closely matched cell line DHS-seq data set from ENCODE for all cell lines tested (HepG2, K562, HEK293, NHEK, Melano), except MIN6, for which we used ATAC-seq data from (Kycia et al., 2018), and trained gkm-SVM to determine sequence features. We then generated deltaSVM scores for each locus using the appropriate cell-specific trained gkm-SVM model. As shown in Figure 5, deltaSVM (dSVM) with discretized predictions trained on only DHS-seq from one cell type was slightly less accurate than the best performing models G3, G5, G7 (C = 0.30). To discretize the deltaSVM scores for a fair comparison, we used z-score > 1 for class + 1, z-score < −1 for class −1, and all others class 0. We then compared continuous scores from deltaSVM trained on DHS from one cell type, and performance improved, but it was still somewhat below the best performing submissions (C = 0.38). After the innovation introduced by the best prediction methods (groups 3, 5, and 7), we hypothesized that deltaSVM predictions could be improved further if we combined deltaSVM scores identified from models trained on more than one ENCODE data set. We trained gkm-SVM independently on all available ENCODE (DHS, histone, and TF datasets, 3,350 datasets

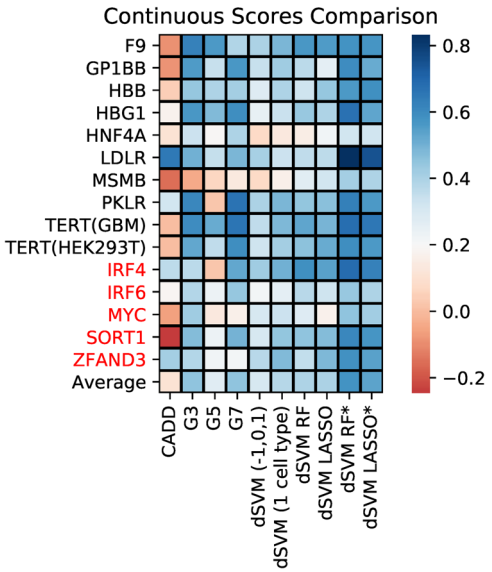

FIGURE 5 Continuous Scores Comparison. The correlation of the top three submissions was higher using continuous prediction scores (C = 0.45, 0.36, 0.45). We also compared with deltaSVM (dSVM) with discretized scores (C = 0.30) and continuous scores (C = 0.39) when trained on DHS from one cell type. Following the method used by the top three groups, combining deltaSVM scores from multiple epigenomic datasets with a RF or Lasso improved the performance to C = 0.40. Using a 50-50 training/test split increased the multiple datatype trained deltaSVM RF and Lasso performance dramatically to C = 0.578 and 0.562, respectively. RF, random forest

total, and trained separate models for the DHS and histone promoter and enhancer peaks), and then used both a Random Forest (RF) and Lasso classifier to learn which combinations of deltaSVM features best predicted the training mutation impact data. We evaluated the model on the held out test set. Combining deltaSVM scores from multiple epigenomic datasets with a RF or Lasso improved the performance slightly to C = 0.40. We further noticed that some of the regions which did poorly (MYC and HNF4a) had sparse training data, and to detect the proper feature importance, relevant binding sites should be disrupted in the training data. We then used a randomly sampled 50-50 training/test split to train the deltaSVM RF and Lasso models, and this increased performance dramatically to C = 0.578 and 0.562 (averaged over 10 randomly sampled splits, standard deviation = 0.046), as shown in Figure 5 and Figure S2. The most informative datasets for deltaSVM training as selected by the Lasso model are listed in Table 2. The most commonly selected ENCODE datasets were DHS and TF ChIP-seq. Training gkm-SVM on these datasets yield sequence features which in combination are most able to reproduce the training data. The weighted combination of these features in either the RF or Lasso model is also the most predictive model of test set mutation impact, as shown in Figure 5.

These comparisons motivated a simpler method to compare the importance of DNA sequence features trained in all ENCODE data types, and assess their comparative informative value for predicting the impact of mutations in each MPRA experiment. Although selection in the RF model is one measure of importance, more simply, we can learn deltaSVM scores trained on one ENCODE datatype in one cell-type or tissue, and calculate the correlation of these deltaSVM scores with mutation impact across each locus one at a time. The range of correlation for deltaSVM models trained on DHS (enhancer and promoter, e&p), H3K27ac (e&p), H3K4me1 (e&p), H3K4me3 (e&p), and TF ChIP-seq (all peaks) is shown in Figure 6 and Figure S3. For the TERT-GBM promoter, DHS promoter deltaSVM scores are most highly correlated with expression impact, followed by select TF ChIP-seq datasets. It is noteworthy that the range of correlation across all DHS promoter trained models is quite narrow, indicating that models trained on DHS promoters from any cell type are quite good at predicting mutation impact at the TERT-GBM promoter, and implying that promoter regulatory vocabulary might be less dependent on cell type. The same is not true of enhancers (SORT1), only a few cell-types yield high correlation, indicating that enhancer regulatory vocabulary is more cell-specific. To quantify this, in Figure 6c,d we show the mean correlation versus best correlation of DHS enhancer trained models and promoter trained models for all regions tested. The best performing promoter regions have significantly higher mean across all ENCODE promoter datasets, indicating that promoter performance is less dependent on training on a matched cell type, whereas enhancer performance is only high for a few matched cell types. Also interesting is that H3K27ac deltaSVM scores are systematically less predictive of mutation impact at both enhancers and promoters. In addition, although H3K4me3 is a promoter-specific mark, deltaSVM trained on H3K3me3 at promoters was only weakly correlated with promoter mutation impact. The informative value of H3K4me1 derived enhancer features was also weaker than DHS and TFs. A subset of deltaSVM scores derived from TF ChIP-seq data were among the most informative marks at both promoters and enhancers (see Table 2 and 3), but the range of ENCODE TF ChIP-seq datasets correlated with promoter MPRA impact was larger than at enhancers, implying that within the ENCODE TF ChIP-seq compendium are many TFs that bind at promoters. One of these is RNA Polymerase II, which is among the frequently ChIP-ed factors. The list of the ENCODE datasets whose deltaSVM scores were most highly correlated with mutation impact in each region tested is listed in Table 3. These simple correlation measures are largely consistent with the top Lasso selected deltaSVM features in Table 2. Among the most correlated datasets for deltaSVM training on promoters, five were enhancer DHS (17%), four were promoter DHS (13%), and 21 were TF ChIP-seq (70%). Among the most correlated datasets for deltaSVM training on enhancers, eight were enhancer DHS (53%) and seven were TF ChIP-seq (47%).

To investigate whether the contiguous bases in the training regions provided in the CAGI 5 experimental design could affect the detection of TFBS, we compared the performance of the deltaSVM RF and Lasso models trained on randomly selected bases versus training sets of equal size with no contiguous bases. In the latter case, we used a regular mask, eg. 000100010001..., to select bases for inclusion in the training set, in this case, 25% training, and repeated four times with each possible phasing. In Figure S4 we compare performance for training ratios 1:1, 1:2, 1:3, 1:4, 1:5, and 1:6 with random or regularly masked training set selection. Although performance dropped with reduced training set size, regularly spaced training sets performed slightly better than randomly selected training sets of equal size, presumably because there is more uniform coverage of TFBS disruption in the regular training sets, and more clustering in the random sets. However, this small difference in performance did not scale with the spacing between bases, as one might expect if neighboring bases within a TFBS were influencing performance.

4 | DISCUSSION

This MPRA computational challenge yielded several useful results. On the experimental side, saturation mutagenesis generated a broad range of mutation impact which allowed clear distinction among competing computational methods. Each region tested had negative impact scores, which reflected the disruption of clusters of multiple (~5-8) binding sites, and our impression from the success of these comparisons is that the longer regions tested in this experiment might more closely reflect the function of these regions in their native genomic context than experiments with shorter inserts. In terms of future challenge design, we recommend that continuous prediction scores be used for all assessments. The requested prediction confidence scores were difficult to incorporate into our analysis.

TABLE 2 Most informative ENCODE datasets by incorporation into deltaSVM Lasso model. Datasets that were given non-zero regression coefficients for each region when searching for at most ten non-zero coefficients using LASSO across multiple training and testing 50-50 data splits

F9 (HepG2)	LDLR (HepG2)	IRF4 (SK-MEL-28)
HepG2: ETV4 ChIP-seq	MCF-7: SREBF1 ChIP-seq	SK-MEL-5: DHS enhancers
HepG2: 3xFLAG-KAT8 ChIP-seq	HepG2: 3xFLAG-SP5 ChIP-seq	foreskin melanocyte: DHS enhancers (Roadmap)
K562: FOXK2 ChIP-seq	HEK293: eGFP-SP3 ChIP-seq	GM12878: RAD51 ChIP-seq
K562: ZNF592 ChIP-seq	HEK293: ZNF263 ChIP-seq	GM12878: ATF7 ChIP-seq
HepG2: ZHX2 ChIP-seq	HepG2: 3xFLAG-ZNF652 ChIP-seq	HT1080: DHS enhancers (Roadmap)
GP1BB (HEL 92.1.7)	**MSMB** (HEK293T)	**IRF6** (HaCaT)
K562: GABPB1 ChIP-seq	HEK293: eGFP-PRDM6 ChIP-seq	foreskin keratinocyte: DHS enhancers (Roadmap)
CMK: DHS enhancers	K562: ATF2 ChIP-seq	bronchial epithelial cell: DHS enhancers
K562: GATA2 ChIP-seq	adrenal gland female embryo: DHS enhancers (Roadmap)	keratinocyte: DHS enhancers
MCF-7: eGFP-KLF9 ChIP-seq	HeLa-S3: DHS enhancers	Peyer's patch: DHS enhancers
K562: GABPA ChIP-seq	mesenchymal stem cell: DHS enhancers	K562: ATF2 ChIP-seq
HBB (HEL 92.1.7)	**PKLR** (K562)	**MYC** (HEK293T)
HEK293: eGFP-SP2 ChIP-seq	K562: DHS enhancers	HeLa-S3: CTCF ChIP-seq
L1-S8R: DHS enhancers	MCF-7: eGFP-KLF9 ChIP-seq	HepG2: RAD21 ChIP-seq
K562: NFYB ChIP-seq	liver embryo: DHS enhancers	CWRU1: DHS enhancers
liver embryo: DHS enhancers	K562: DPF2 ChIP-seq	A549: SMC3 ChIP-seq
K562: DHS enhancers	K562: eGFP-ZNF148 ChIP-seq	K562: MAZ ChIP-seq
HBG1 (HEL 92.1.7)	**TERT** (GBM)	**SORT1** (HepG2)
GM12878: NFYB ChIP-seq	HepG2: GABPA ChIP-seq	HepG2: CEBPB ChIP-seq
HEK293: eGFP-KLF1 ChIP-seq	CMK: DHS enhancers	HepG2: FOXA1 ChIP-seq
K562: NFYB ChIP-seq	HepG2: 3xFLAG-SP5 ChIP-seq	K562: eGFP-ZNF148 ChIP-seq
HL-60: DHS enhancers	heart left ventricle: DHS enhancers	HepG2: KDM1A ChIP-seq
K562: IRF1 ChIP-seq	TH17: DHS promoters	HepG2: TCF7 ChIP-seq
HNF4A (HEK293T)	**TERT** (HEK293T)	**ZFAND3** (MIN6)
K562: NFYB ChIP-seq	K562: GABPB1 ChIP-seq	adrenal gland: DHS enhancers A549: USF1 ChIP-seq
K562: EGR1 ChIP-seq	HepG2: GABPA ChIP-seq	brain embryo: DHS enhancers (Roadmap)
K562: eGFP-ZFX ChIP-seq	HepG2: 3xFLAG-GABPA ChIP-seq	SK-N-SH: RFX5 ChIP-seq
HEK293: eGFP-KLF9 ChIP-seq	CMK: DHS promoters	cerebellar cortex: DHS enhancers
MCF-7: eGFP-KLF9 ChIP-seq	K562: eGFP-ETV1 ChIP-seq	Ammon's horn: DHS enhancers

Comparisons of the computational predictions revealed additional important insights for understanding enhancer/promoter function and how to build more accurate models of their role in human disease. All top performing models for mutation impact prediction used machine learning based DNN (or gkm-SVM) DNA sequence features trained on chromatin accessibility or chromatin state data. These models consistently outperformed models using sequence features derived from other sources: PWMs from existing databases, evolutionary conservation, kmers, or more generic sequence features (eg. GC content). The machine learning-based models also outperformed models using chromatin accessibility, chromatin state, or TF ChIP-seq data without using the epigenomic data to derive DNA sequence-based models. When the machine learning-based DNA sequence features are combined with proper importance weighting derived from another layer of machine learning on a subset of the mutation data used as training for each cell type, the overall prediction accuracy is high. Although we have shown that gapped-kmer features are equally or more informative than DNN-based features, we emphasize that deltaSVM was not evaluated in a blind prediction, but only after the challenge, as part of the model assessment. Although most of the 15 different experiments were well predicted, there was significant variation. MSMB, HNF4a, and MYC were the most difficult to predict, which may be because of the quality of the MPRA data for these experiments or biological characteristics of these sequences, such as the density of binding sites, or the specific genomic sequence interval tested.

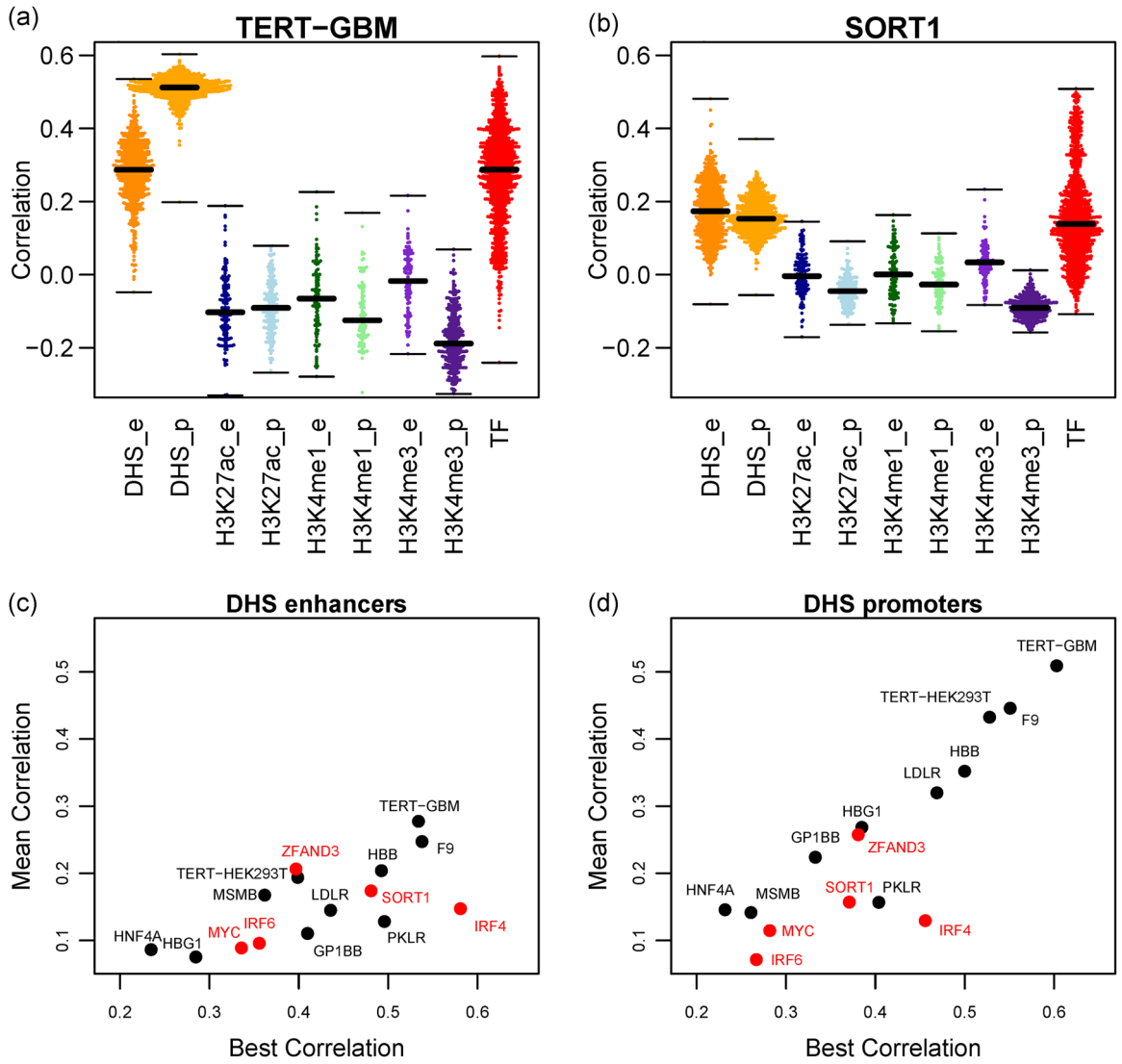

FIGURE 6 Correlation of deltaSVM scores trained on all ENCODE datasets with MPRA expression impact. In one promoter (a) TERT-GBM and one enhancer (b) SORT1, the full range of correlation of deltaSVM scores with expression impact is shown for different training datatypes (see Methods for a full description of datasets). (c,d) The correlation of deltaSVM scores with expression impact for the best single ENCODE DHS enhancer (c) and promoter (d) datasets, compared with the mean across all ENCODE datasets of that type. Although only a few enhancer datasets in relevant cell types yield high correlation, the mean performance for the best-predicted promoters is much higher than for enhancers, suggesting a more cell-type independent promoter vocabulary. MPRA, massively parallel reporter assay

Following the three most successful prediction models, we designed a method to combine gkm-SVM features derived from multiple ENCODE datatypes to predict the impact of mutations in cell type specific promoters and enhancers. This analysis demonstrates that combining DNA sequence features trained from multiple cell and datatypes improves the accuracy of mutation impact prediction, even if the cell types are not perfect matches to the cell type used in the MPRA. DHS alone does not produce optimal performance. This comparison with deltaSVM allowed us to assess the relative informative value of features derived from different cell types and different ENCODE assays. We found that DHS derived features were most informative, but that the integration of TF ChIP-seq derived features significantly improved performance. The question of which TFs to include in this computational exercise was addressed by using a training subset of the mutation data. For computational model assessment, this is an effective experimental design, but for future disease studies, where training data is unavailable, averaging mutation impact scores for DHS and known relevant and available TF ChIP-seq derived models in the cell type of interest is probably the best approach. Identifying which TFBS are learned in the DNN or gkm-SVM weights and using ChIP-seq data for these TFs is another possible approach.

We also found evidence of somewhat distinct TF vocabulary at enhancers and promoters: Promoter DHS trained feature models predicted promoters better, and enhancer DHS trained feature models predicted enhancers better. This suggests that separate training of peaks in enhancers and promoters is advantageous, and raises the potential concern that predictions on the basis of models

TABLE 3 Most informative ENCODE datasets by correlation of deltaSVM scores with mutation impact

Promoters				
Region	**Data set**	**Rank**	**Corr**	**Description**
F9	TF_E2_41	1	0.600	GAbisphenol A ChIP-seq on ethanol treated A549
F9	TF_E3_346	2	0.592	ETV4 ChIP-seq on human HepG2
F9	TF_E3_472	3	0.587	3xFLAG-KAT8 ChIP-seq on human HepG2
GP1BB	TF_E2_274	1	0.414	GABPA ChIP-seq on K562
GP1BB	DHS_E2_95e	2	0.410	CMK
GP1BB	DHS_E2_56e	3	0.398	K562
HBB	DHS_E3_182p	1	0.500	hematopoietic multipotent progenitor cell
HBB	DHS_E3_157p	2	0.494	L1-S8R
HBB	DHS_E2_1e	3	0.494	K562
HBG1	TF_E2_113	1	0.472	NFYB ChIP-seq on human GM12878
HBG1	TF_E2_290	2	0.457	NFYB ChIP-seq on human K562
HBG1	TF_E2_229	3	0.439	NFYA ChIP-seq on human K562
HNF4A	TF_E3_761	1	0.309	EGR1 ChIP-seq on human K562
HNF4A	TF_E2_233	2	0.305	EGR1 ChIP-seq on human K562
HNF4A	TF_E3_765	3	0.302	EGR1 ChIP-seq on human K562
LDLR	TF_E3_505	1	0.539	3xFLAG-SP5 ChIP-seq on human HepG2
LDLR	TF_E3_234	2	0.510	eGFP-SP3 ChIP-seq on human HEK293
LDLR	TF_E2_226	3	0.482	IRF1 ChIP-seq on IFN treated human K562
MSMB	DHS_E2_13e	1	0.362	fibroblast of villous mesenchyme
MSMB	TF_E3_244	2	0.361	eGFP-PRDM6 ChIP-seq on human HEK293
MSMB	TF_E3_267	3	0.332	eGFP-ZNF629 ChIP-seq on human HEK293
PKLR	DHS_E2_1e	1	0.496	K562
PKLR	DHS_E3_130e	2	0.495	liver embryo (59 days) and embryo (80 days)
PKLR	TF_E2_345	3	0.492	TBL1XR1 ChIP-seq on human K562
TERT-GBM	DHS_E2_95p	1	0.603	CMK
TERT-GBM	TF_E2_200	2	0.597	GABPA ChIP-seq on human HepG2
TERT-GBM	DHS_E3_11p	3	0.585	PC-3
TERT-HEK293T	TF_E3_666	1	0.554	GABPB1 ChIP-seq on human K562
TERT-HEK293T	TF_E2_200	2	0.550	GABPA ChIP-seq on human HepG2
TERT-HEK293T	TF_E3_407	3	0.547	3xFLAG-GABPA ChIP-seq on human HepG2
Enhancers				
Region	**Data set**	**Rank**	**Corr**	**Description**
IRF4	DHS_E3_110e	1	0.581	SK-MEL-5
IRF4	DHS_RM_214e	2	0.574	foreskin melanocyte male newborn
IRF4	DHS_RM_14e	3	0.572	foreskin melanocyte male newborn
IRF6	DHS_E2_126e	1	0.356	bronchial epithelial cell - retinoic acid
IRF6	DHS_E2_42e	2	0.351	keratinocyte female
IRF6	DHS_RM_17e	3	0.348	foreskin keratinocyte male newborn
MYC	TF_E2_152	1	0.395	CTCF ChIP-seq on human HeLa-S3
MYC	TF_E3_2	2	0.394	SMC3 ChIP-seq on human A549
MYC	TF_E2_188	3	0.39	RAD21 ChIP-seq on human HepG2
SORT1	TF_E2_175	1	0.508	CEBPB ChIP-seq on human HepG2
SORT1	TF_E3_401	2	0.498	SOX13 ChIP-seq on human HepG2
SORT1	TF_E3_473	3	0.492	3xFLAG-SOX13 ChIP-seq on human HepG2

(Continues)

TABLE 3 (Continued)

Enhancers				
Region	**Data set**	**Rank**	**Corr**	**Description**
ZFAND3	DHS_RM_146e	1	0.397	brain female embryo (85 days)
ZFAND3	TF_E2_394	2	0.392	RFX5 ChIP-seq on human SK-N-SH
ZFAND3	DHS_E3_166e	3	0.384	adrenal gland male adult (37 years)

trained on all sets of peaks without this distinction (like DeepSEA) may yield predictions that are less accurate at describing mutations in distal enhancer elements. Our results in Figure 5 show that the groups using features trained on all peaks (DeepSEA or DNN) are consistently better at predicting promoters, whereas deltaSVM separately trained on enhancers and promoters shows better performance than the submitted models on predicting mutation impact at the five distal enhancers tested.

Although the current MPRA assays yield extremely useful tests of base pair resolution cell-specific DNA regulatory element activity, these assays do not recapitulate the native 3D interactions between regulatory elements, for example, enhancer-promoter interactions, which also impact transcriptional output and the impact of regulatory mutations on human disease (reviewed, e.g., in (Bonev & Cavalli, 2016; Gorkin, Leung, & Ren, 2014)). These interactions do not appear to be completely specified by features within single elements (Xi & Beer, 2018) and modeling these interactions remains an active area for future investigation.

5 | METHODS

Group 1: Selected features were taken from the following databases: Combined Annotation Dependent Depletion (CADD; Kircher et al., 2014), Functional Analysis through Hidden Markov Models v2.3 (FATHMM; Shihab et al., 2013), and ElemeNT (Sloutskin et al., 2015). The prediction analyses were performed using WEKA 3.8 data mining software. More than 100 data features were created and downloaded from the abovementioned collections. In the training data, sequence variants that had a confidence level lower than 0.1 were considered as 0 ("No Effect") and variants that had a higher confidence level were marked as either 1 (increase in expression level) or −1 (decrease in expression level), on the basis of the sign of the change in the expression value. In the first submission method, the impact (−1,0,1) was predicted using a RF classifier on 27 features from CADD and FATHMM. In the second submission, each variant effect was predicted separately, that is one classifier tried to predict which variants would cause upregulation, and another classifier tried to predict which variants would cause downregulation. Both predictions were combined in the following way: For a particular variant in the testing set, if the labels of the classifiers were (0,0), 0 was assigned. If a variant's pair of predictions was (−1,0) or (0,1), −1 or 1 was assigned, respectively. In the case of conflict between pairs of labels (−1,1), the label with the highest prediction score was chosen. Fifteen features were selected to predict which variants cause downregulation and 11 features were used for upregulation prediction.

Group 2: The training and test set variants were annotated using features from WGSA v0.7 and CADD v1.3. In addition to capture sequence context, 3-mers and 5-mers centered on a given position, along with the mutated variant were included as additional categories. All categorical features were one-hot encoded. Imputation on missing values was performed using k-nearest neighbors (KNN, $n = 3$) from the "fancyimpute" python package. All models were implemented in Python using "scikit-learn" package. Model training was performed on the estimated variant effect as labels, provided by the CAGI5 challenge organizers (25% of the measured alleles). Submissions 1,2,3, and 5 utilized multi-class classification, and submission 4 used regression. For multi-class classification, three classes were defined on the basis of the experimental variant effect value: −1 if value < −0.1; 1 if value > 0.1; else 0. Learning was performed using "XGBClassifier" and "XGBRegressor" in the Python "xgboost" library. The hyperparameters for the xgboost algorithm were optimized using a Bayesian implemented in the "BayesSearchCV" class of the "scikit-optimize" package. For submission 1, the optimal set of xgboost parameters was obtained by "BayesSearchCV" using 100 iterations and three-fold stratified cross-validation.

In submission 1 (primary submission) and submissions 3–5, all the training data was used irrespective of the different loci. In submission 2, the training set was first split by locus and different multi-class "xgboost" classifiers were trained and predicted separately. In submission 3, the KNN imputation was performed on data from each locus separately before training. In submission 4, the problem was modeled as regression rather than multi-class classification. In submission 5, for all numeric features, the values were transformed by a rolling mean of a window of five consecutive bases.

Group 3: For each variant in training and test data, features from functional genomics data obtained by RegulomeDB (Boyle et al., 2012) were either binary or numerical values. The binary features indicated overlapping regions from ENCODE ChIP-Seq peaks, ENCODE DNase-Seq peaks, TF motif matching using PWM's, and DNase footprints. Numerical features from ChIP-Seq signals and information change of the matched PWM were also used (these were used for all submissions). In addition, submissions 1 and 4 used DeepSEA (Zhou & Troyanskaya, 2015) features, submissions 2 and 5 used DeepBind (Alipanahi et al., 2015) features, and submissions 3 and 6 used information change of all tested PWMs. A RF classifier (500 trees) was used to predict the direction of variant effects. The RF classifier also outputs a probability of prediction, which was used to calculate continuous scores. For continuous scores, if the predicted label was +1 or −1, the probability of +1 or −1, respectively, was used as the continuous prediction (the continuous prediction was

also given the same sign as the sign of the label). Otherwise, the difference between the probability of +1 and the probability of −1 was taken.

Group 4: This group used a DNN similar to DeepSEA trained on ENCODE data, but used DeepLIFT (Shrikumar, Greenside, & Kundaje, 2017) to extract features and score mutation impact. Subsequent neural network and SVM layers weighted these features using the training data for impact prediction.

Group 5: The features used were conservation, DNase hypersensitivity, and features generated from a neural network model similar to DeepSEA in network structure (Hawkins-Hooker et al., 2018). Conservation scores were retrieved from phyloP (Pollard, Hubisz, Rosenbloom, & Siepel, 2010), phastCons (Siepel et al., 2005), GerpN, and GerpRS (Davydov et al., 2010). DNase hypersensitivity for each regulatory element was determined by identifying the closest matching ENCODE cell type for which there is a DNase-hypersensitivity track. For DeepSEA scores, several different neural network architectures were trained on the basis of the genomic prediction benchmark detailed in the original DeepSEA paper. This group evaluated these networks on a region surrounding each variant twice, once with the reference allele and once with the alternate allele. Features were generated as the difference in activations between the two evaluations of internal and output layers of the networks. To predict the direction of change, this group used three different gradient boosting algorithms: "XGBoost", "CatBoost" and "LightGBM". The best features were determined by performing five-fold cross-validation on different subsets of features. Models were assessed by cross-validating one against many area-under-precision-recall-curve.

Group 6: Features were derived from DNase accessibility, ATAC-Seq data, conservation, FANTOM 5 CAGE, and motif analysis. DNase-Seq and ATAC-Seq profiles were retrieved from ENCODE. Conservation scores were generated from phyloP100way, phastCons100way, and MultiZ alignments. Dinucleotide PWMs from HOCOMOCO (Kulakovskiy et al., 2018) were used for the motif analysis. SPRY-SARUS and PERFECTOS-APE (Vorontsov, Kulakovskiy, Khimulya, Nikolaeva, & Makeev, 2019) were used to map motif occurrences within reporter regions and to assess the difference of motif P-values for alternative alleles. "XGboost" and "LightGBM" were used in multiclass prediction to determine the direction of expression change.

Group 7: The set of features included DeepSEA scores (all 919), DeepBind scores (all 515), ENCODE motifs, k-mers (length 5), number of unique motifs in a sequence, the density of unique motifs, poly A/T, GC content, and conservation. An ensemble of five RF classifiers and five "ExtraTreesClassifiers" with 1,000 trees. The square root of total features was used to predict direction of change. For continuous scores on promoters, again an ensemble of five RF regressors and five "ExtraTreesRegressors" (with the same parameters as above) was trained on the released training set. For continuous scores on enhancers, an ensemble of one RF regressor, one ExtraTreesRegressor (same parameters), and one gradient boosting regressor with 1,000 boosting estimators) was trained on the released training data.

5.1 | gkm-SVM

We called MACS2 peaks after combining replicates of ENCODE2 (ENCODE Consortium, 2012), ENCODE3, and Roadmap (Roadmap Epigenomics Consortium et al., 2015) human DHS, chromatin state, and TF data for hg38 downloaded from the DCC (www.encodeproject.org). We further separated enhancer (>2k from TSS) and promoter DHS and chromatin peaks, and removed datasets with fewer than 2,500 enhancer or promoter peaks, or fewer than 2,500 TF peaks (independent of position). This yielded the following number of datasets for each datatype: ENCODE2: (DHS.e, DHS.p, H3K27ac.e, H3K27ac.p, H3K4me1.e, H3K4me1.p, H3K4me3.e, TF) = (159, 163, 20, 24, 17, 17, 35, 91, 345); ENCODE3: (DHS.e, DHS.p, H3K27ac.e, H3K27ac.p, H3K4me1.e, H3K4me1.p, H3K4me3.e,TF) = (182, 196, 61, 67, 23, 23, 34, 68, 699); Roadmap: (DHS.e, DHS.p, H3K27ac.e, H3K27ac.p, H3K4me1.e, H3K4me1.p, H3K4me3.e, TF) = (313, 317, 65, 98, 66, 63, 31, 173, 0), for a total of 3,350 training datasets. We extended +/−150 bp from each MACS2 summit, trained on 300 bp regions, and ran gkm-SVM using parameters (-l 11 -k 7 -d 3 -t 2) using the gkm-SVM R-package (Ghandi et al., 2016) and ls-gkm (Lee, 2016) for large training sets. All test set AUROCs were high (median >0.9).

ACKNOWLEDGMENTS

M.B., D.S., and A.P. are supported by NIH R01 HG007348 and NIH U01 HG009380. I.V.K. is supported by RFBR 18-34-20024. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

ORCID

Aashish N. Adhikari http://orcid.org/0000-0003-4305-9494
Shengcheng Dong http://orcid.org/0000-0001-5728-8090
Ivan V. Kulakovskiy http://orcid.org/0000-0002-6554-8128
John Reid http://orcid.org/0000-0002-7762-6760
Ron Unger http://orcid.org/0000-0003-4153-3922
Nadav Ahituv http://orcid.org/0000-0002-7434-8144
Michael A. Beer http://orcid.org/0000-0001-9955-3809

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Beer, M. A. (2017). Predicting enhancer activity and variant impact using gkm-SVM. *Human Mutation*, *38*(9), 1251–1258. https://doi.org/10.1002/humu.23185

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678. https://doi.org/10.1038/nrg.2016.112

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

ENCODE Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Computational Biology*, *10*(7), e1003711. https://doi.org/10.1371/journal.pcbi.1003711

Ghandi, M., Mohammad-Noori, M., & Beer, M. A. (2014). Robust k-mer frequency estimation using gapped k-mers. *Journal of Mathematical Biology*, *69*, 469. https://doi.org/10.1007/s00285-013-0705-3

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., & Beer, M. A. (2016). gkmSVM: An R package for gapped-kmer SVM. *Bioinformatics*, *32*(14), 2205–2207. https://doi.org/10.1093/bioinformatics/btw203

Gorkin, D. U., Leung, D., & Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, *14*(6), 762–775. https://doi.org/10.1016/j.stem.2014.05.017

Hawkins-Hooker, A., Kenlay, H., & Reid, J. (2018). Projection layers improve deep learning models of regulatory DNA function. *BioRxiv*, 412734. https://doi.org/10.1101/412734

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, *106*(3), 159–164. https://doi.org/10.1016/j.ygeno.2015.06.005

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., & Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research*, *27*(1), 38–52. https://doi.org/10.1101/gr.212092.116

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J., & Ahituv, N. (2018). Saturation mutagenesis of disease-associated regulatory elements. *BioRxiv*, 505362. https://doi.org/10.1101/505362

Kreimer, A., Zeng, H., Edwards, M. D., Guo, Y., Tian, K., Shin, S., & Yosef, N. (2017). Predicting gene expression in massively parallel reporter assays: A comparative study. *Human Mutation*, *38*(9), 1240–1250. https://doi.org/10.1002/humu.23197

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., & Makeev, V. J. (2018). HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, *46*(D1), D252–D259. https://doi.org/10.1093/nar/gkx1106

Kycia, I., Wolford, B. N., Huyghe, J. R., Fuchsberger, C., Vadlamudi, S., Kursawe, R., & Stitzel, M. L. (2018). A common Type 2 diabetes risk variant potentiates activity of an evolutionarily conserved islet stretch enhancer and increases C2CD4A and C2CD4B expression. *The American Journal of Human Genetics*, *102*(4), 620–635. https://doi.org/10.1016/j.ajhg.2018.02.020

Lee, D. (2016). LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics*, *32*(14), 2196–2198. https://doi.org/10.1093/bioinformatics/btw142

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, *47*(8), 955–961. https://doi.org/10.1038/ng.3331

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., & Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, *30*(3), 265–270. https://doi.org/10.1038/nbt.2136

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., & Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *ArXiv:1704.02685 [Cs]*. Retrieved from http://arxiv.org/abs/1704.02685

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Sloutskin, A., Danino, Y. M., Orenstein, Y., Zehavi, Y., Doniger, T., Shamir, R., & Juven-Gershon, T. (2015). ElemeNT: A computational tool for detecting core promoter elements. *Transcription*, *6*(3), 41–50. https://doi.org/10.1080/21541264.2015.1067286

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., & Sabeti, P. C. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, *165*(6), 1519–1529. https://doi.org/10.1016/j.cell.2016.04.027

Vorontsov, I. E., Kulakovskiy, I. V., Khimulya, G., Nikolaeva, D. D., & Makeev, V. J. (2019,). *PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation*. 102–108. https://doi.org/10.5220/0005189301020108

Xi, W., & Beer, M. A. (2018). Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Computational Biology*, *14*(12), e1006625. https://doi.org/10.1371/journal.pcbi.1006625

Zeng, H., Edwards, M. D., Guo, Y., & Gifford, D. K. (2017). Accurate eQTL prioritization with an ensemble-based framework. *Human Mutation*, *38*(9), 1259–1265. https://doi.org/10.1002/humu.23198

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Shigaki D, Adato O, Adhikari AN, et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Human Mutation*. 2019;40:1280–1291. https://doi.org/10.1002/humu.23797