# Meta-analysis of massive parallel reporter assay enables functional regulatory elements prediction

*Zhongxia Yan*
*Nir Yosef, Ed.*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-79
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-79.html

May 18, 2018

# Meta-analysis of massive parallel reporter assay enables functional regulatory elements prediction

by Zhongxia Yan (equal contribution with Anat Kreimer)

## Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**. Note that this is a revision of an unpublished paper which can be found at http://biorxiv.org/content/early/2017/11/03/202002.

Approval for the Report and Comprehensive Examination:

**Committee:**

Professor Nir Yosef

Research Advisor

(Date)

* * * * * * *

Professor Joseph Gonzalez

Second Reader

(Date)

## ABSTRACT

Deciphering the potential of non-coding loci to influence the regulation of nearby genes has been the subject of intense research, with important implications in understanding the genetic underpinnings of human diseases. Massively parallel reporter assays (MPRAs) can measure the activity of thousands of regulatory DNA sequences and their variants in a single experiment. With the increase in the number of publicly available MPRA datasets, one can now develop functional-based models which, given a DNA sequence, predict its regulatory activity. Here we performed a comprehensive meta-analysis of several MPRA datasets in a variety of cellular contexts. We apply an ensemble of methods to accurately predict the MPRA output in each context and observe that the most predictive features are consistent across datasets. We then demonstrate that predictive models trained in one cellular context can be used to predict MPRA output in another. Finally, we identify the factors that are predictive across all or some of the datasets.

## INTRODUCTION

Massive Parallel Reporter Assays (MPRA) (Weingarten-Gabbay and Segal 2014), which allow for cost effective, high-throughput activity screening of thousands of sequences and their variants for regulatory activity (Mogno et al. 2013; Patwardhan et al. 2012; Kheradpour et al. 2013; Patwardhan 2012; Melnikov et al. 2014; Sharon et al. 2012; Smith et al. 2013) have become a major tool for the functional characterization of gene regulatory elements. In these assays, a library of putative regulatory elements is cloned alongside DNA barcodes or the sequence itself can be used as the barcode (Arnold et al. 2013). Libraries can either be transfected or infected into cells and the activity associated with a given regulatory element is assessed by sequencing the transcribed barcodes. Since MPRA is still a nascent technology, the development of computational tools that take advantage of its existing datasets could help improve future MPRA candidate sequence selection, enhance our ability to predict functional regulatory sequences and increase our understanding of the regulatory code and how its alteration can lead to a phenotypic consequence.

Previous works have used single MPRA datasets to better predict functional sequences or regulatory grammar (Lee et al. 2015; Grossman et al. 2017; Sharon et al. 2012). For example, the Critical Assessment of Genome Interpretation (CAGI) consortium, which launched the expression quantitative trait loci (eQTL) causal SNP challenge (Kreimer et al. 2017; Beer 2017; Zeng et al. 2017). The main lessons learned from this community effort highlighted that the use of ensemble of methods, specifically non-linear methods, generally yielded better performance and features that are related to transcription factor (TF) binding and chromatin accessibility as top predictors for MPRA activity. Interestingly, methods that use predicted, rather than observed features (e.g.

epigenetic properties predicted from DNA sequence (Alipanahi et al. 2015; Zhou and Troyanskaya 2015; Zeng et al. 2016)) were shown to be the most accurate, even more than using experimentally derived epigenetic properties as features.

While these lessons provided an important first step, their focus has been on a single MPRA dataset in a specific cellular context. Critical questions therefore remain as to how generalizable the insights from MPRA experiments are – either across datasets (possibly from different cellular contexts) or by exploring the function of endogenous DNA loci. Here, we present a first comprehensive analysis of several MPRA datasets of endogenous loci collected by different labs in different cell types. We derive a large set of properties to characterize each putative regulatory region and compare the performance of different methods and features for predicting MPRA output. We investigate the capacity of our models to be transferable across datasets, which allowed us to distinguish between determinants of MPRA activity that are dependent on the cellular context (e.g., protein milieu in the cell) vs. ones that are intrinsic to the DNA sequence.

## METHODS

### MPRA Datasets

We perform all experiments on five publicly available MPRA datasets and one unpublished dataset collected by different labs. In all cases, the MPRA constructs were designed to test endogenous human DNA sequences, and not synthetic elements (Smith et al. 2013). Unless otherwise stated, the MPRA experiment was performed in an episomal context. Thus, each element tested in each dataset is associated with a source genomic region. The length of elements varies between 121 to 171 base pairs. The datasets are defined below:

1) *K562* – putative regulatory regions (Kwasnieski et al. 2014) selected from ENCODE-based annotated regions in K562 cells (ENCODE Project Consortium 2012; Hoffman et al. 2012; Ernst and Kellis 2010). This set includes 600 regions annotated as enhancers, 600 as weak enhancers, 300 as repressed, and 284 scrambled negative controls. All these sequences were tested in K562 cells.
2) *LCL*-eQTL – 3,044 regions (Tewhey et al. 2016) that contain an eQTL in Lymphoblastoid Cell Lines (LCLs). Notably, this dataset was used as the primary source for the CAGI eQTL causal challenge (Kreimer et al. 2017).
3) *HepG2-eQTL* – the same set of elements (Tewhey et al. 2016) as above, tested in HepG2 cell line instead of LCL.
4) *HepG2-chr* – 2,236 candidate liver enhancers (Inoue et al. 2017), tested in chromosomal context.
5) *HepG2-epi* – the same set of elements (Inoue et al. 2017) as above, tested in episomal context.
6) *hESC* – 2,268 putative enhancer regions (Inoue et al. unpublished), tested in chromosomal context in neural embryonic stem cells.

## Quantifying Activity of Regions

For each dataset, we obtain the raw counts of barcodes observed by the MPRA experiment for RNA and DNA; multiple barcodes are associated with a single genom ic region. To obtain a single quantitative measure of transcriptional activity for each genomic region from the MPRA RNA and DNA counts, we input the counts into an unpublished method called MPRAnalyze (Fischer et al. unpublished). This measure of transcriptional activity is coined the name “alpha” by MPRAnalyze and somewhat resembles the ratio between the counts of transcribed RNA and the counts of the initial DNA while adjusting for experimental settings. For all analyses utilizing the quantitative measure of expression, we preprocess by taking base-2 logarithm of alpha.

We also define a binary active / inactive label for transcriptional activity from alpha. If a dataset has control regions (*K562* and *hESC*), we first calculate a robust version of the standard score from the alpha values by subtracting the median over the control regions and dividing by the median absolute deviation (MAD) of the control regions. If no control region exists for the dataset, we use the median and MAD over all regions instead of just the control regions in the previous step. We then compute the survival function for each standard score and apply the Benjamini-Hochberg (BH) correction. The active regions are then defined as regions with a false discovery rate (FDR) of less than 0.05.

Featurization

We featurize each element in each dataset with various methods utilizing the element's sequence and genomic locus:

1) *Experimental* – 1095 binary features representing whether the genomic region overlaps with experimentally measured tracks of transcription factor (TF) binding sites (TFBS), histone binding sites, and DNase-hypersensitivity sites downloaded from ENCODE (ENCODE Project Consortium 2012). Some epigenetic factors covered include DNase, Ctcf, Ezh2, and H3k4me3; each of these factors is measured in multiple cells, so each factor is associated with multiple features.
2) *DeepBind* – 515 predicted TF binding quantifications generated by a neural network

model trained on protein-binding microarrays (Alipanahi et al. 2015).

***3)*** *DeepSea* – 919 TF binding, DNA accessibility, and histone modification quantifications generated by a convolutional neural network model trained on chromatin profiling data (Zhou and Troyanskaya 2015).

***4)*** *Motifs* **–** 2065 predicted motif hits (ENCODE Project Consortium 2012) from simple DNA-binding motif scoring (Grant et al. 2011).

***5)*** *5-mers* – 1024 binary features, each associated with a 5-mer (permutation of 5 base pairs), each indicating whether the corresponding 5-mer exists in the sequence.

***6)*** *Summary* – a small collection of features either directly derived from the element's properties or summarizing one of the features above.

  ***a.*** *#GC; #polyA, #polyT* – number of G/C in the sequence; length of longest polyA/T subsequence.

  ***b.*** *#5-mers* – number of distinct 5mers in the sequence.

  ***c.*** *MGW, Roll, ProT, HelT* – DNA shape features (Zhou et al. 2013) quantifying minor groove width, roll, propeller twist, and helix twist.

  ***d.*** *Conservation* – evolutionary conservation score of region as predicted by phastCons (siepel et al. 2005)

  ***e.*** *Closest Gene Expression* – expression (TPM) of the closest gene from RNA-seq data in the corresponding cell type

  ***f.*** *Promoter, Exon, Intron, Distal* – binary features indicating whether the element intersects a promoter, exon, and intron. *Distal* is defined to be 1 if the element does not intersect with promoter, exon, and intron.

***g.*** *#motifs, Motif Density* – number of significant DNA-binding ENCODE motifs in the sequence, maximum number of motifs within a 20 bp window in the sequence

***h.*** *#deepsea-top, #deepbind-top* – number of TFs quantifications above 90th percentile across all the regions in *DeepSea* / *DeepBind*.

***i.*** *#tf-high, #tf-med, #tf-low* – number of TFs that are bound above 90th percentile in *DeepBind* and rank in the top, middle, or bottom 100 (out of 515) for RNA-seq TPM in the relevant cell type.

***j.*** *<factor> [Cell] Mean, TFBS Shuffled Mean* – mean across subsets of *Experimental* features. <factor> can be *TFBS*, *DNase*, *Ctcf*, *Ezh2*, *H2az*, *H3k4me1*, *H3k4me2*, *H3k4me3*, *H3k9ac*, *H3k9me1*, *H3k9me3*, *H3k27ac3*, *H3k27me3*, *H3k36me3*, *H3k79me2*, *H4k20me1*, *P300*. For these factors we take the mean of the binary overlaps over all corresponding [, cell-type specific to the dataset's cell-type,] *Experimental* features. *TFBS Shuffled Mean* is the mean across n **not** cell-type specific, randomly chosen *TFBS* features, where n is the number of features in *TFBS Cell Mean*. As a note, we only use these features for analysis when evaluating the correlation of individual features with MPRA activity; we do not use these features with the full classification and regression models, as we already use the *Experimental* features for those models.

Statistical tests

We examine the predictivity of features and accuracy of prediction models using several statistical tests. For regression task – e.g. predicting quantitative activity – we applied several correlation measures (Pearson, Spearman, Kendall) considering either the entire test data or regions at the top 25% of quantitative activity; we also applied Spearman correlation by first binning quantitative activity by quintiles. We refer to these seven tests as the *regression tests*. For classification task – e.g. predicting active or not active – we record the AUROC (area under receiver operating characteristic curve) and AUPRC (area under precision recall curve); we refer to these two tests as the *classification tests*. The significance of each *regression task* was evaluated by the respective statistical test *Q-values*, which are obtained from *P-values* via the Benjamini–Hochberg correction.

## Model Description

We predict the quantitative activity from element features with four regression models and their ensemble. The four models are a linear regressor with ElasticNet regularization (Zou and Hastie 2005) with 0.5 as the L1 and L2 regularization coefficients and a RandomForest regressor (Breiman 2001), an ExtraTrees regressor (Geurts et al. 2006), and a GradientBoosting regressor (Hastie et al. 2009), each with 1000 estimators. The ensemble method is implemented by taking the average prediction of all four regression models.

For the classification task, we use a RandomForest classifier (Breiman 2001) and an ExtraTrees classifier (Geurts et al. 2006), each with 1000 estimators, as well as their ensemble. The ensemble method averages the predicted probability from each classifier.

For both regression and classification, we define a shuffle model with the same composition as an ensemble model but shuffles the labels of the training set before training. This allows us to quantify the probability of producing our ensemble results by chance.

## RESULTS

### Predictive features for MPRA activity are consistent across datasets

We examine the 56 *Summary* features (**Methods**) individually in two ways: **1)** we test how well each feature correlates with the quantitative MPRA output using the seven *regression tests* and **2)** we test how well each feature discriminates between active and inactive regions using the two *classification tests*. We rank each feature for each of the nine tests and then take the median of these ranks to obtain a dataset-specific feature ranking. We take the median across all dataset-specific ranking to obtain a global ranking for the features and sort the features according to increasing global rank (**Figure 1**). The dataset-specific feature rankings, Spearman values, and AUROC values all agree well with the global rank, so the global feature ranking is robust across datasets, with *TFBS Mean* and *DNase Mean* as the most predictive features.

Furthermore, we found that limiting the set of epigenetic features in a manner specific to the cell type under investigation (e.g. for the *K562* dataset, *TFBS Cell Mean* only takes the mean over K562 cell-type TFBS features) leads to reduced accuracy, compared with the more simple approach of taking all available data regardless of cell type of origin; this observation is consistent with previous work on enhancer annotation (Erwin et al. 2014). In addition, we compute *TFBS Shuffled Mean* with 100 different sets of randomly selected features and evaluate each trial with the *regression tests* and *classification tests*. We find that when compared to *TFBS Shuffled Mean*, *TFBS Cell Mean* is better than *TFBS Shuffled Mean* 276 out of 600 times, which shows that the mean over cell-type specific features does not perform significantly better than the mean over the mean over the same number of non-cell-type specific features.

To further explore cell-type specificity in the context of TF binding, we defined cumulative features of TF binding that are cell-type specific; we stratified the TFs into three groups according to their expression level in the cell type of interest (low / intermediate / high) and sum over the number of binding sites in each group. While these three features *#tf-high, #tf-med, #tf-low* had a strong correlation (especially *#tf-high*) with MPRA activity (**Figure 1**), they are still less predictive than *TFBS Mean* (the simple mean across all TFBS-related features). Consistently, we find several cell-type agnostic features such as *GC* content and *#motif* that are predictive of MPRA activity as well.

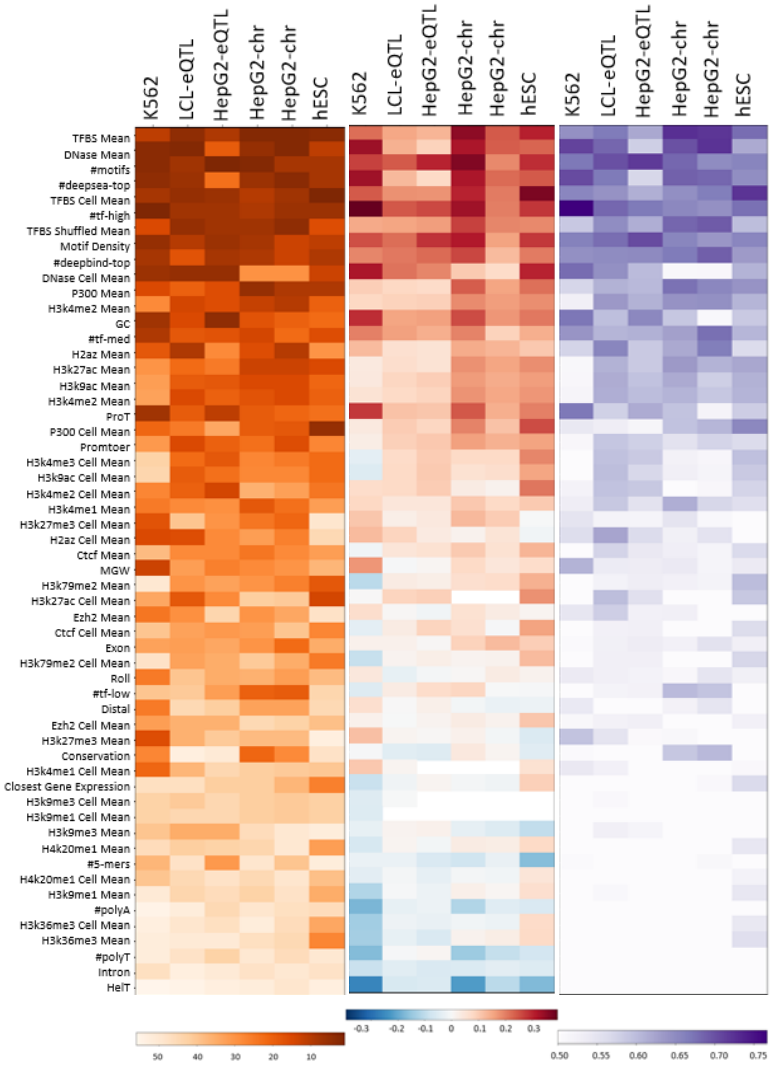


Figure 1: Individual feature correlation with MPRA output. The features are sorted from lowest to highest median ranking across all dataset. The three heatmaps are colored according to **1)** the median rank within each dataset **2)** the Spearman correlation value for the regression task **3)** the AUROC value for the classification task.

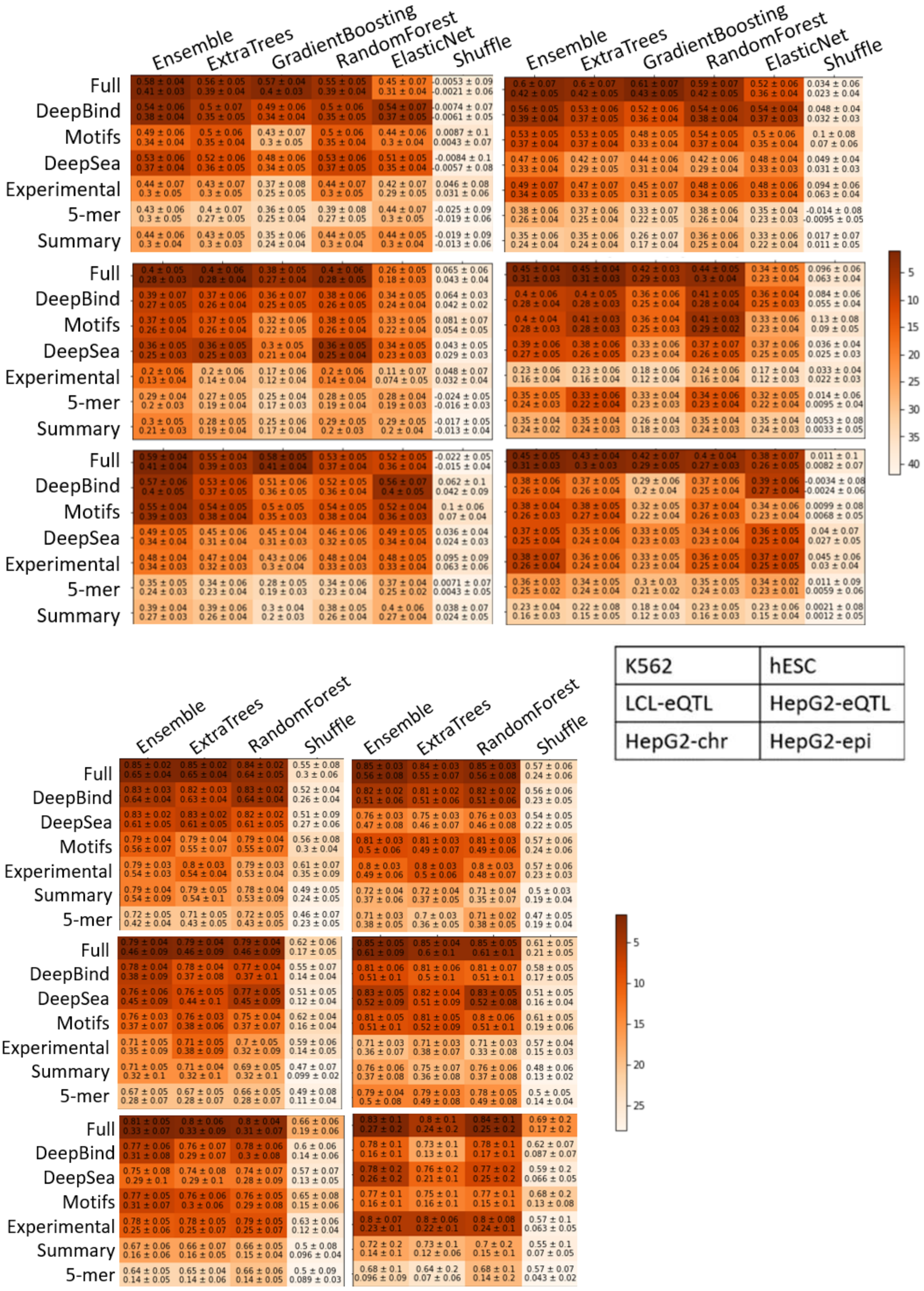
## Predictive models of MPRA activity are similar across datasets

We next turned to the construction of supervised predictive models that combine multiple features to increase accuracy. Importantly, we do not use the evaluation of individual *Summary* features (from **Figure 1**) during model construction (e.g., for feature selection), thus avoiding circularity. We train and evaluate each regressor (see **Method**: *ElasticNet*, *RandomForest*, *ExtraTrees*, *GradientBoosting*, and *ensemble*) on subsets of our featurization (see **Method**: *Experimental, DeepBind, DeepSea, Motifs, 5-mer, Summary*) as well as the full featurization (*Full*), and similar for each classifier (see **Method**: *RandomForest, ExtraTrees, ensemble*) in the classification task.

Consistent with our results above, we observe similar trends in the performance of the different feature sets and methods across datasets, with ensemble methods usually at the top (**Figure 2**). Among the feature subsets, the predicted TF binding properties according to *DeepBind* are top performers, and the concatenation of all feature classes results yields the best performance. As above, we noticed that limiting the epigenetic features to be cell type specific does not increase accuracy.

We note that the *Shuffle* models performs significantly worse than our *Ensemble* models, which shows that the *Ensemble* model results are not trivially obtained.

Figure 2: Performance of different models with different feature combinations across datasets. Both heatmaps are colored according to the median of the within-dataset rank across all tests. (top) Accuracy of five regression models and their ensemble on various feature subsets. The first printed statistic is the mean and standard deviation of the Spearman correlation while the second is the mean and std of Kendall (bottom) Accuracy of two classification models and their ensemble on various feature subsets. The first and second printed statistics are the mean and std of AUROC and AUPRC.



| | |
|---|---|
| K562 | hESC |
| LCL-eQTL | HepG2-eQTL |
| HepG2-chr | HepG2-epi |

### Transferring knowledge between cell types

To evaluate how well our models can be applied to a new cellular context where MPRA data does not exist, we tested models trained in each dataset on the remaining datasets. Based on the results in Figure 2, we used take the set of features to be *Full* and use the *ensemble* model for both the regression and classification tasks. We observe that MPRA prediction is robust across datasets (**Figure 3**) with slightly reduced prediction power compared to the supervised settings.

We examine the testing performance of training on *HepG2-chr* vs training on *HepG2-epi*, in both a supervised and transfer learning context. We found that training on *HepG2-chr* always showed better supervised learning results when training the *ensemble* model on *Full* features; *HepG2-chr* also showed better transfer learning results than *HepG2-epi* 37 out of 40 times (comparing results across the Pearson, Spearman, Kendall, AUROC, and AUPRC tests) when testing on the other four datasets and same regions are used for training in both datasets. These results suggest that MPRA done in chromosomal context reflects better the endogenous environment of the sequence, stressing the importance of implementing this approach (Inoue et al. 2017).

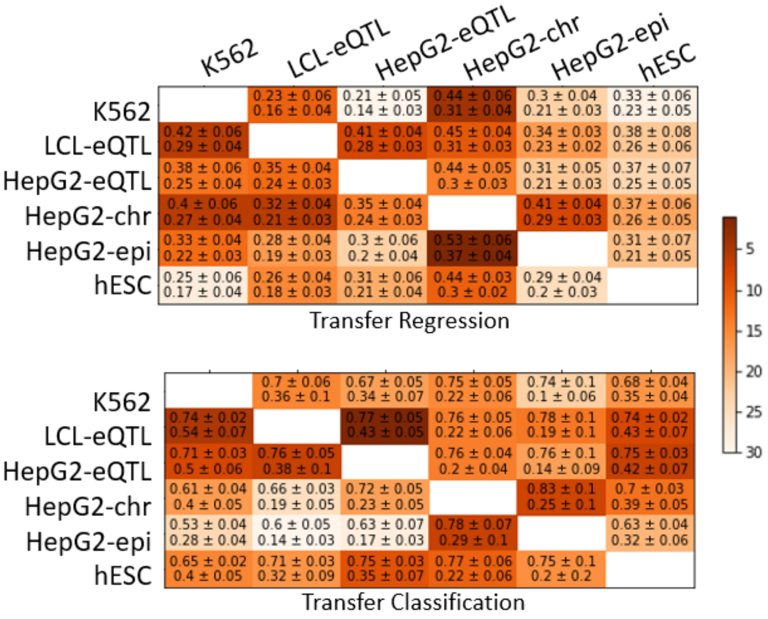


Figure 3: Transfer learning between cell types. An ensemble model with all features and is used for training on one cell type and testing on a different cell type. Again each cell is colored according to the median rank over the relevant correlators. The first and second statistics are mean and std for Spearman and Kendall for regression, and AUROC and AUPRC for classification.

## Contributions of Individual TFs

To further investigate the prospects of transfer learning, we examine the contribution of individual TFs. For each TF and each MPRA region we define a binary score for whether the *DeepBind* TF binding prediction is significantly correlated with quantitative activity (Alipanahi et al. 2015). This score reflects the individual TF binding's predictivity of MPRA activity. We then ranked the TFs based on their predictive ability across datasets, thus revealing several TFs whose binding is generally informative of regulatory activity of MPRA constructs in all cellular contexts in this study (**Figure 4**). For instance, two TF families with a dataset-wide high predictive capacity, that is supported by both motif-predicted and experimentally-evaluated binding sites are JUN and FOS. Proteins of the FOS family dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1, which has been implicated in a wide range of cellular processes, including cell growth, differentiation, and apoptosis across different cell types (Ameyar et al. 2003). The predictive TFs that are common across data sets are also highly expressed across all the three cell types, as indicated by RNA-seq data (**Figure 4**). More generally, the gene expression of TFs is overall consistent with their predictivity, whereby more predictive factors have overall higher expression as measured by RNA-seq (ENCODE Project Consortium 2012) (**Figure 4** – right four columns) across all cell types (Wilcoxon rank sum test of top vs. bottom 50 factors: p-value of 3.9e-06, 1.36e-5, 8.7e-4, and 8.0e-4 for K562, LCL, HepG2, and H1hESC respectively).

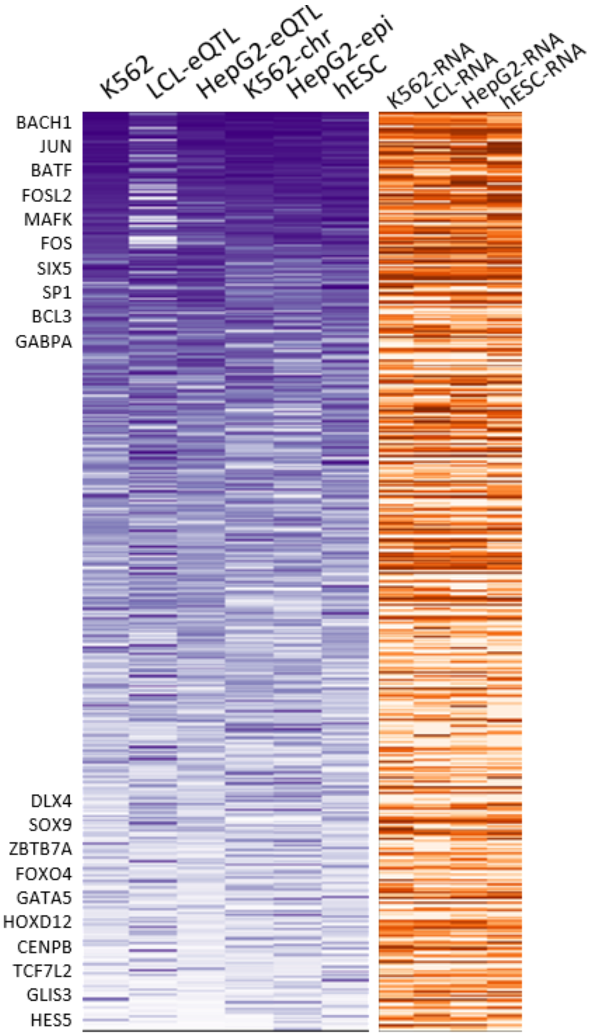


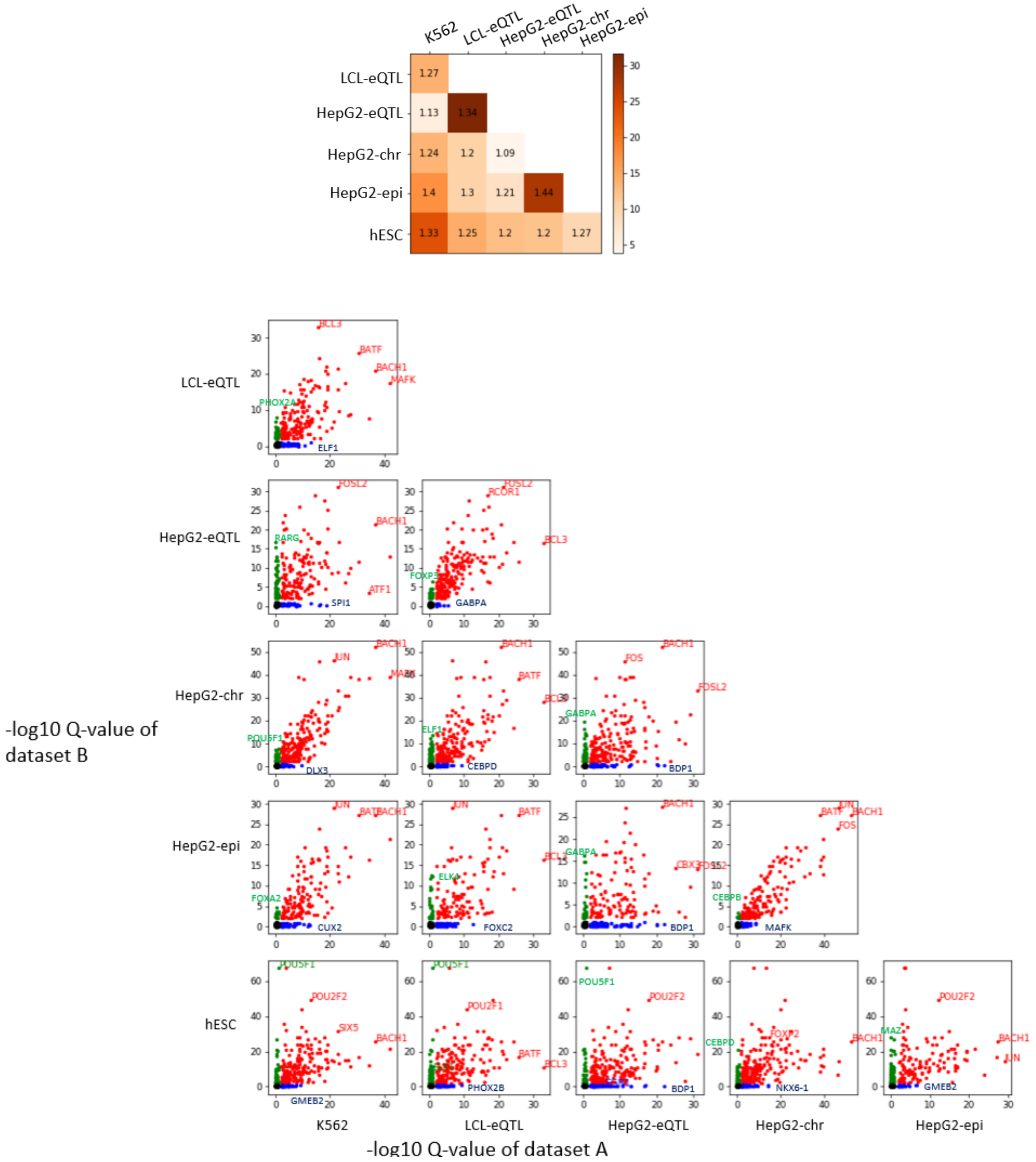
Figure 4: Contribution of individual TF binding for predicting regulatory activity of MPRA constructs. TFs are sorted according to the median rank (across correlators) across datasets. (left panel) Heatmap of the within-dataset rankings. (right panel) the TF's ranking of gene expression measured by RNA-seq in each one of the four cell lines. Names of the common top/bottom 10 factors is indicated on the left.

### Exploring common and distinct TF binding between datasets

We first explore dataset similarity in terms of the composition of predictive features. For each dataset, we find the set of *DeepBind* features that are significantly correlated with quantitative expression (Spearman Q-value less than 0.05) and call this set of features *predictive*. We then compare across pairs of dataset if to determine if there is significant overlap in *predictive* features; for each pair of dataset, we calculate the enrichment ratio and the hypergeometric probability that the amount of overlap in *predictive* features occurs by chance (**Figure 5A**). Overall, we see that there is significant overlap in *predictive* features across every pair of dataset. Unsurprisingly, the *HepG2-chr* and *HepG2-epi* pair and *LCL-eQTL* and *HepG2-eQTL* pair had the strongest *predictive* feature overlap, which suggests that the same genomic region in different cell-types have similar sets of *predictive* features.

We also examine the features that differ in predictivity between pairs of datasets (**Figure 5B**), and provide a list of top factors predictive for at least one of the datasets. In some cases, we find proteins whose function is related to the cell type under investigation. For instance, when comparing *K562* vs *LCL-eQTL* for factor predictivity, we observed that the ETS family TF ETV6, a proto-oncogenes implicated with chromosomal rearrangements associated with leukemia (Safran et al. 2003) is predictive in only K562. When comparing *K562* to *HepG2-eQTL*, we find that RARG – a retinoic acid receptor which belongs to the nuclear hormone receptor family and is associated with liver risk phenotype (Roberts et al. 2010) – is predictive in *HepG2-eQTL* but not *K562*.

Figure 5: We assign a binary feature predictivity indicating whether the Q-value of the Spearman correlation between each TF in *DeepBind* and the quantitative activity is below a threshold. **(A)** Similarity between the datasets. The Q-value threshold is 0.05 here. Each heatmap cell is colored according to the negative-log10 of the adjusted hypergeometric P-value, which measures significant of the overlap in predictive features. **(B)** TF showing differential predictivity. We call the TFs with Q-values below 0.01 as predictive and the TFs with Q-values above 0.1 as not predictive, discarding the rest of the TFs. We plot the TFs that are predictive in both (red), predictive in one (blue or green), and not predictive in both according to their -log10 Q-value in each dataset.



### Studying the effects of small genetic variants on transcription of nearby genes

MPRA can be used to study the transcriptional effects of small variants that commonly occur in regulatory regions, namely SNPs and small indels (Tewhey et al. 2016). We wanted to know if we can predict these effects - starting from the synthetic setting of MPRA. An important feature of the LCL-eQTL and HepG2-eQTL datasets (Tewhey et al. 2016) is that each of the sequences (which come from the reference human genome) is matched with an alternative allele (single nucleotide variants (SNVs) or short indels) (Lappalainen et al. 2013) that was tested by MPRA as well. Here, we test the ability of our models to determine the amount of shift in MPRA transcriptional activity, comparing each reference allele to its alternative. We focus on the LCL-eQTL dataset, which was featured in the CAGI challenge, and for which the results of competing methods are available (Kreimer et al. 2017). Our method first applied the ensemble regression model above to predict transcriptional activity of the reference and alternative alleles, separately. Next, we train a logistic regression using the absolute difference between those predicted expression values as a feature to predict whether there is a significant allelic variation. This strategy lead to favorable results (0.67 AUROC, 0.45 AUPRC), compared to other participants in the CAGI challenge (0.65 AUROC, 0.45 AUPRC). We achieved good performance initially without needing to balance the 0 and 1 classes in the training set allelic variation. However, after discovering that the test set has different composition of 0 and 1 classes than the training set, we add an attempt to balance the effect of the 0 and 1 classes in the training set by weighing the effect of each region by the inverse frequency of its label in the training set (while keeping all other parameters constant). We achieve the best classifier with this method (0.69 AUROC, 0.47 AUPRC) but acknowledge that we experimented with this method after finding out that the composition of 0 / 1 classes in the test set is different than the training set.

## DISCUSSION

MPRA holds a great promise to be a key functional tool that will increase our understanding of gene regulatory elements and the consequences of nucleotide changes on their activity. While previous studies already used MPRA to construct predictive models of transcriptional regulation, its generalizability across cellular contexts and its applicability for studying the endogenous genome have not yet been systematically evaluated. Here, we study MPRA data from several cellular systems to determine which features are reflective of the cellular context (e.g., protein milieu in the cell), and which are intrinsic to DNA sequence. We explore the extent by which knowledge on regulatory activity in one cellular context can be used to make predictions in a held out cellular context.

Our work highlights genome accessibility and TF binding as the strongest predictors of regulatory activity, with no observed advantage to cell type specific features. When applying prediction models, we observe that performance is improved when using an ensemble of all features, with no significant prediction improvement when using cell type specific features. These results imply that part of the signal observed in MPRA studies is not cell type specific. Interestingly, models trained with chromosomal MPRA data yield better predictions across datasets than those trained on episomal MPRA data, stressing the importance of this experimental approach that conveys a more reliable representation of the endogenous settings.

When training on one cell type and predicting on another cell type, we observe overall slightly lower, but robust results, with regions enriched in cell type specific signal being harder to predict.

Notably, we detect a communal component across datasets with a group of TFs being top predictors, as well as some cell-specific factors that seem to be involved in phenotypes associated with the corresponding cell type (e.g. immune functions for LCL factors).

## REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–838.

Ameyar M, Wisniewska M, Weitzman JB. 2003. A role for AP-1 in apoptosis: the case for and against. *Biochimie* **85**: 747–752.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077.

Beer MA. 2017. Predicting enhancer activity and variant impact using gkm-SVM. *Hum Mutat* **38**: 1251–1258.

Breiman L. 2001. *Mach Learn* **45**: 5–32.

Chuvpilo S. 1997. Three NF-ATc isoforms with individual properties are expressed in lymphoid cells. *Immunol Lett* **56**: 349.

Craig MP, Sumanas S. 2016. ETS transcription factors in embryonic vascular development. *Angiogenesis* **19**: 275–285.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817–825.

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. 2014. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* **10**: e1003677.

Fischer D, Ashuac T, Yosef N. Unpublished. MPRAnalyze. https://github.com/YosefLab/MPRAnalyze

Géraud C, Schledzewski K, Demory A, Klein D, Kaus M, Peyre F, Sticht C, Evdokimov K, Lu S, Schmieder A, et al. 2010. Liver sinusoidal endothelium: a microenvironment-dependent differentiation program in rat including the novel junctional protein liver endothelial differentiation-associated protein-1. *Hepatology* **52**: 313–326.

Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Mach Learn* **63**: 3–42.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE, et al. 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci U S A* **114**: E1291–E1300.

Hastie T, Tibshirani R, Friedman JH. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473–476.

Inoue F, Kreimer A, Ahituv N, Yosef N. Unpublished. Massively parallel characterization of regulatory dynamics during neural induction.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38–52.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**: 800–811.

Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, Welch R, Wainberg M, Mohan R, Sinnott-Armstrong NA, et al. 2017. Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum Mutat*. http://dx.doi.org/10.1002/humu.23197.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**: 1595–1602.

Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta

M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.

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

Leslie R, O'Donnell CJ, Johnson AD. 2014. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**: i185–94.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Melnikov A, Zhang X, Rogov P, Wang L, Mikkelsen TS. 2014. Massively Parallel Reporter Assays in Cultured Mammalian Cells. *J Vis Exp*. http://dx.doi.org/10.3791/51719.

Mogno I, Kwasnieski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23**: 1908–1915.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**: D77–82.

Patwardhan RP. 2012. *Massively Parallel Functional Dissection of Regulatory Elements*.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265–270.

Roberts KE, Kawut SM, Krowka MJ, Brown RS Jr, Trotter JF, Shah V, Peter I, Tighiouart H, Mitra N, Handorf E, et al. 2010. Genetic risk factors for hepatopulmonary syndrome in patients with advanced liver disease. *Gastroenterology* **139**: 130–9.e24.

Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, et al. 2003. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **31**: 142–146.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521–530.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**: 1021–1028.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**: 15545–15550.

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**: 1519–1529.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Weingarten-Gabbay S, Segal E. 2014. The grammar of transcriptional regulation. *Hum Genet* **133**: 701–711.

Zeng H, Edwards MD, Guo Y, Gifford DK. 2017. Accurate eQTL prioritization with an ensemble-based framework. *Hum Mutat*. http://dx.doi.org/10.1002/humu.23198.

Zeng H, Hashimoto T, Kang DD, Gifford DK. 2016. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32**: 490–496.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931–934.

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

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J R Stat Soc*

*Series B Stat Methodol* **67**: 301–320.