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

Anat Kreimer[1,2,5], Zhongxia Yan[1], Nadav Ahituv[2] & Nir Yosef[1,3,4,5]
[1] Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, California USA; [2] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA; [3] Ragon Institute of MGH MIT and Harvard, Cambridge MA, USA; [4] Chan Zuckerberg Biohub Investigator; [5] Corresponding author.
Email: anat.kreimer@berkeley.edu, niryosef@berkeley.edu

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 MPRA datasets that are publically available, 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 first applied an ensemble of methods to accurately predict the MPRA output in each context, and observed that the most predictive features are consistent across data sets. We then demonstrate that predictive models trained in one cellular context can be used to accurately predict MPRA output in another, with mild loss of accuracy attributed to features that depend on the cell type. Finally, we study the extent to which MPRA can assist with the identification of genetic modifications that are associated with transcriptional changes to nearby genes. Our analysis provides insight into how MPRA data can be leveraged to highlight functional regulatory regions throughout the genome and can guide efficient design of future MPRA experiments.

INTRODUCTION

Massive Parallel Reporter Assays (MPRA)[1], which allow for cost effective, high-throughput activity screening of thousands of sequences and their variants for regulatory activity [2–8] 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 [9]. 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. For example, the Critical Assessment of Genome Interpretation (CAGI) consortium, which launched the expression quantitative trait loci (eQTL) causal SNP challenge [10]. 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 [11–13]) 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 are the insights from MPRA experiments - 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 collected by different labs in different cell types, and examining endogenous loci. 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 show that MPRA activity is predictable and that prediction methods tend to perform consistently well when tested on different datasets, with better performance for non-linear methods and favorable results when using an ensemble approach. Consistently, the predictive capacity of individual features is comparable across datasets with TF binding and epigenetic properties being the top predictors. Interestingly, we found that models trained with chromosomal- integrated MPRA data tends to yield higher accuracy than those trained on episomal data, suggesting that a chromosomal setting has a better reflection of the endogenous environment of the sequence.

We next turned to 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. Here, we demonstrate that predictive models trained in one cellular context can be used to accurately predict the MPRA output in another and that their accuracy is diminished for predicting sequences that are enriched in cell type specific properties such as cell type specific TF binding.

Finally, we wanted to evaluate the applicability of these approaches to study the function of regulatory regions in the endogenous genome. We first tested the ability of our framework to detect the effects of small variants (single nucleotide variants (SNV) or short insertions or deletions (indels)) on MPRA activity, and achieved better accuracy than the state of the art methods [14]. We then examined how MPRA data can be leveraged for prediction of expression quantitative trait loci (eQTL) across the genome. Consistent with the recent literature, we find that the predictive power of MPRA in this case is limited. To gain better understanding of this negative result, we explicitly looked for characteristics of eQTL regions that are difficult to distinguish from variants that are not associated with changes to gene expression. We find that features related to the sequence physical environment (e.g. closest gene expression) is characteristic of hard to predict regions and that models trained with chromosomal rather than episomal data slightly improve performance. These results suggest that the nature of MPRA experiment, where the location of the integration is not controlled for, can provide a partial explanation on why leveraging MPRA for eQTL prediction has limited performance.

RESULTS

We used five publicly available MPRA datasets collected by different labs using a range of experimental methodologies and cell types. In all cases, the MPRA constructs were designed to test endogenous human DNA sequences, and not synthetic elements [8]. The length of the tested elements varies between 121 to 171 bp. We treat the "output" of each MPRA construct (in each data set) in two ways: (1) a quantitative measure of transcriptional output, representing the amount of transcribed RNA per construct. This quantity is computed as the ratio between the estimated abundances of transcribed RNA and the construct's DNA. (2) a binary active/ inactive label, defined separately for each dataset.

The first dataset [15], which we refer to as **K562,** consists of putative regulatory regions selected from ENCODE- based annotated regions in K562 cells [16–18]. 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, and if a region has a larger expression than the 90th percentile of the scrambled sequences, it is labeled as active. The second and third

datasets, which we refer to as **LCL-eQTL** and **HepG2-eQTL** [19] consist of 3,044 sequences that contain an eQTL in Lymphoblastoid Cell Lines (LCLs). The same sequences were tested in LCL and HepG2 cells, thus forming the two datasets. Active regions were determined by the significance (FDR<1%) of differential abundance of transcripts versus plasmid input [20]. Notably, the LCL-eQTL dataset was used as the primary source for the CAGI eQTL causal challenge [10]. The fourth and fifth datasets [21] include 2,236 candidate liver enhancers, tested in either episomal or chromosomal context. We refer to these datasets as **HepG2-epi** (for MPRA plasmids) and **HepG2-chr** (for MPRA integrated in the genome). Active sequence are defined by comparing to negative controls, as in dataset 1.

Predictive features for MPRA activity are consistent across data sets

We first defined a set of features that characterize each MPRA sequence and inspected each feature individually (**Methods** and **Supplementary Table 1**). Overall, we examined 56 features that can be divided into four categories (similarly to ref [10]): (1) Experimentally measured epigenetic properties (e.g., chromatin accessibility) (2) Computationally-derived epigenetic properties (e.g., predicted TF binding sites [11,12]). (3) DNA k-mer frequencies (k=5). (4) Other loci specific features (e.g., genomic annotation). We examine the predictive capacity of each feature in each dataset in two ways. In the *regression* task we test how well the feature correlates with the quantitative MPRA output, using seven different statistical tests. In the *classification* task we test how well it discriminates between active and inactive regions, using three complementary measures (**Figure 1**). We rank the features per statistical test and then take the median of these ranks to obtain a dataset-specific feature score. Overall we found the feature ranking to be robust across datasets, with chromatin accessibility and the number of TF binding sites as the most predictive. Furthermore, we found that limiting the set of epigenetic features (classes 1 and 2) in a manner specific to the cell type under investigation (e.g., using only K562 ChIP-seq for the K562 data set) leads to reduced accuracy, compared with the more simple approach of taking all available data, regardless of cell type of origin, which is consistent with previous work on enhancer annotation [22]. To further explore this 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 summed over the number of binding sites in each group. While these three features (especially the high group) had a strong correlation with MPRA activity (**Figure 1, Supplementary Table 1**), they are still less accurate than a simple sum of all TF binding sites. Consistently, we find several cell-type agnostic features such as GC content and proximity to TSS that are predictive of MPRA activity as well (**Supplementary Figure 1,2**).

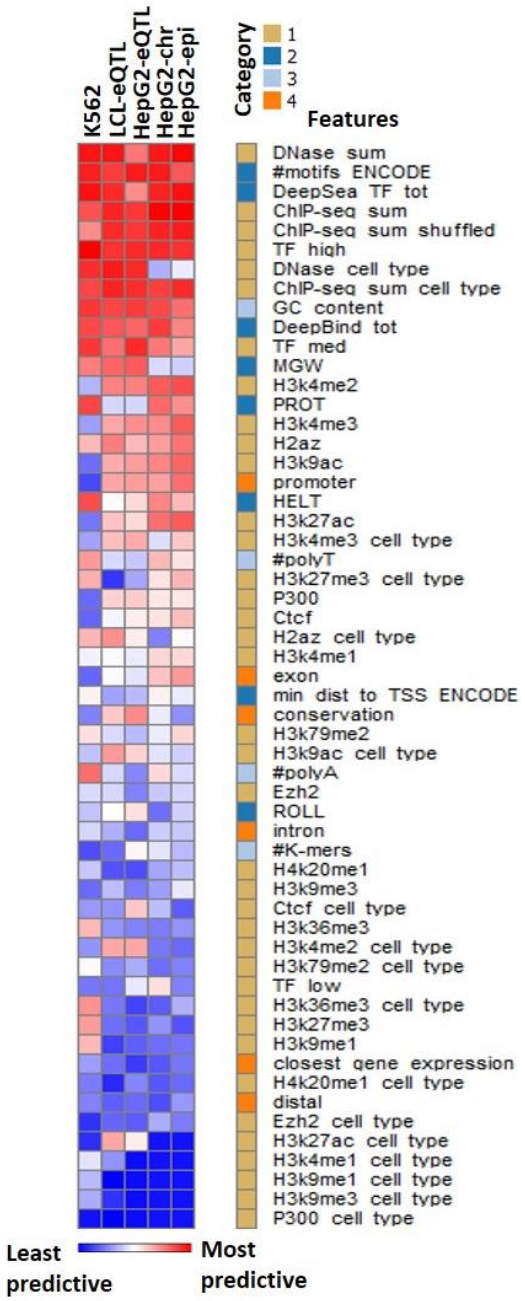


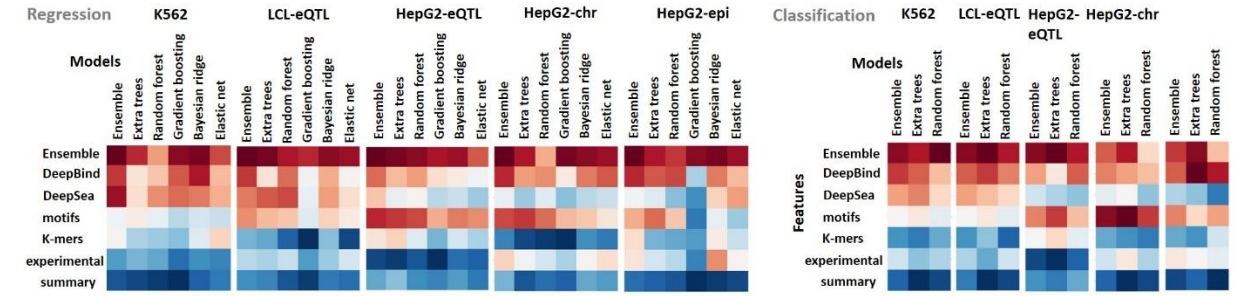
Figure 1: Single feature predictability of MPRA output. (A) Ranking of features importance across datasets. Features are divided to 4 categories (Methods). (1) Yellow squares- experimentally measured epigenetic properties. (2) Blue squares - predicted epigenetic properties. (3) Light blue square - DNA k-mer frequencies. (4) Orange squares - other locus-specific properties.

Predictive models of MPRA activity are similar across data sets

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 features (from Figure 1) during model construction (e.g., for feature selection), thus avoiding circularity. For the regression task, we used five models (**Methods**), two of which are linear (elastic net [23], Bayesian ridge regression) and three nonlinear (random forest [24], extra trees [25] and gradient boosting [26]). For classification, we used two nonlinear methods: random forest [24] and extra trees [25].

For evaluation, we applied these methods either individually or as an ensemble, across multiple datasets, and using different subsets of our 56 features. The features subsets are defined in accordance with the four categories above, with the exception that category 2 (predicted epigenetic properties) was divided into three subsets: (1) TF binding predicted by the DeepBind algorithm [11]; (2) TF binding, DNA accessibility, and histone modifications predicted by the DeepSea algorithm [12]; (3) TF binding predicted by motif hits [16] and category 4 to include both locus specific and summary features. Reassuringly, the accuracy of our top models (for regression and classification) applied on the LCL-eQTL dataset, matched that of the top ranking group in the CAGI challenge [14].

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**, and **Supplementary Table 2**). Among the feature subsets, the predicted epigenetic properties according to DeepBind and DeepSea are top performers, and the union 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 (**Supplementary Figure 3**).



Statistical test	ρ ensemble	ρ ensemble	ρ ensemble	ρ ensemble	ρ ensemble
Pearson all (top 25%)	0.60 (0.55)	0.59 (0.53)	0.59 (0.51)	0.62 (0.25)	0.51 (0.20)
Spearman all (top 25%)	0.57 (0.52)	0.46 (0.37)	0.45 (0.54)	0.65 (0.32)	0.53 (0.19)
Binned	0.55	0.43	0.42	0.61	0.52
Kendall all (top 25%)	0.40 (0.38)	0.32 (0.25)	0.31 (0.37)	0.46 (0.22)	0.37 (0.12)

Statistical test	ensemble	ensemble	ensemble	ensemble	ensemble
AUCROC	0.84	0.79	0.92	0.81	0.71
AUCPR	0.67	0.47	0.60	0.71	0.50

Figure 2: Performance of different models with different feature combinations across datasets. (A) Accuracy of five regression models and their ensemble using mean prediction. (B) Accuracy of two classification models and their ensemble using mean prediction of the probability. Features: DeepSea predictions, DeepBind predictions, ENCODE motifs predictions, 5-mer representation, epigenetic features from ENCODE, summary features combined: polyA/T, GC content, DNAshape, conservation, distance to TSS, sum of motif hits, TF anonymization and expression of the closest gene. For each statistical test we rank its correlation coefficient/AUC and report the median of these ranks for each cell of the matrix. The tables below each heatmap report the correlation coefficient for regression and the AUC for classification for the ensemble approach.

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 ones. Based on the results in Figure 2, we used the preferred strategy of taking all feature classes (without exclusion according to cell type) and using the ensemble of all prediction methods. We observe that MPRA prediction is robust across datasets (**Figure 3**) with slightly reduced prediction power compared to the within- dataset cross validation setting. We note that even in this context of transfer learning, cell type specific data (e.g., aggregating TFs according to expression or using epigenetic measurements only from the respective cell type) does not improve prediction (**Supplementary Figure 3**). Interestingly, models constructed using the chromosomal MPRA data tends to yield higher accuracy either in the within dataset (cross validation) setting and in the between dataset (transfer learning) setting. These results suggest that MPRA done in chromosomal context reflects better the endogenous environment of the sequence, stressing the importance of implementing this approach [21].

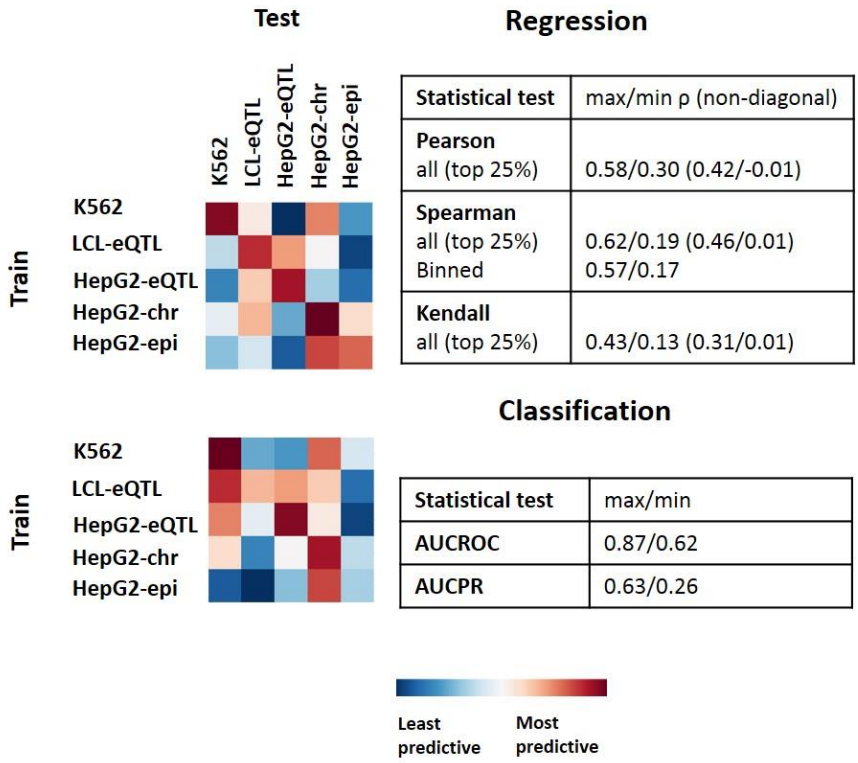


Figure 3: Transfer learning between cell types. An ensemble model with all features and is used for training on one cell type and testing on a different cell type. Diagonal predictions: for K562, HepG2-chr, HepG2-epi datasets we trained the models on randomly selected 80% of the data and tested it on the remaining 20%. For LCL-eQTL and HepG2-eQTL datasets we used separate training and test sets as defined in CAGI eQTL causal SNP challenge [10]. Off diagonal predictions: we trained on one dataset and tested in a different dataset. Same as in Figure 2, for each statistical test we rank its output and report the median of these ranks for each cell of the matrix. The corresponding regression and classification tables show the max/min correlation coefficient/AUC for the different statistical tests.

To further investigate the prospects of transfer learning, we went beyond the general feature classes described above and investigate the contribution of individual TFs. For each TF and each MPRA region we define a binary score (for bound/ un- bound), using either predicted sites [11] or ones detected by ChIP-seq [16]. 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**, **Supplementary Figures 4-6**, **Supplementary Table 4,5**). 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 [27]. 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 [16] (**Figure 4** - 3 right columns) across all cell types (Wilcoxon rank sum test of top vs. bottom 100 factors: p-value of 2.6e-07, 4.8e-08, and 9.6e-06 for K562, LCL and HepG2 respectively).

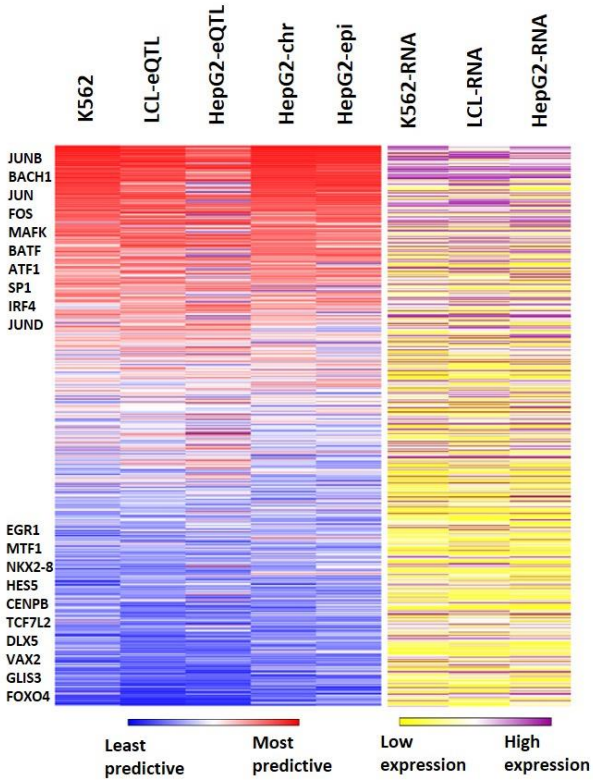


Figure 4: Contribution of individual TF binding for predicting regulatory activity of MPRA constructs. Ranking of the contribution of individual TF binding (515 factors predicted by DeepBind) across datasets (left panel) and their corresponding ranking of gene expression measured by RNA-seq in each one of the 3 cell lines (right panel). Names of the common top/bottom 10 factors is indicated on the left.

Exploring cell- type specific properties of MPRA activity

To examine TFs whose predictivity is restricted to a certain cell type, we conducted a comparative analysis, focusing on the K562, LCL-eQTL and HepG2-eQTL datasets (**Figure 5**). For each of 3

possible pairwise comparisons, we define four categories of TFs: cell type specific factors, common factors that are significant in both cell types, and factors that are not predictive in either cell type. Examining the difference in expression patterns between these categories reassures that common factors tend to have higher expression in all cell types than non-significant factors (p-values 1e-05, 1e-06, and 0.05 respectively) (**Figure 5 lower panel**). Furthermore, factors that are predictive only in one cell type do not tend to be more highly expressed in this cell type compared to the others, implying, unsurprisingly, that a simple expression- based criteria is not sufficient to determine cell type specificity. However, among the factors that are cell type specific, we find proteins whose function is related to the cell type under investigation. For instance, in K562 (**Figure 5A,B**) we observed the ETS family TFs: ETV6 and ETS2 (a paralog of GABPA), both of which are proto-oncogenes implicated with chromosomal rearrangements associated with leukemia [28]. In LCL (**Figure 5B,C**) we observed SP3 and ETV2, which play a role in vascular development and angiogenesis [29]. For HepG2 (**Figure 5B,C**), we find the TFs TFEC - a liver-sinusoidal-endothelium specific transcriptional regulator [30], and RARG - a retinoic acid receptor which belongs to the nuclear hormone receptor family, and is associated with liver risk phenotype [31].

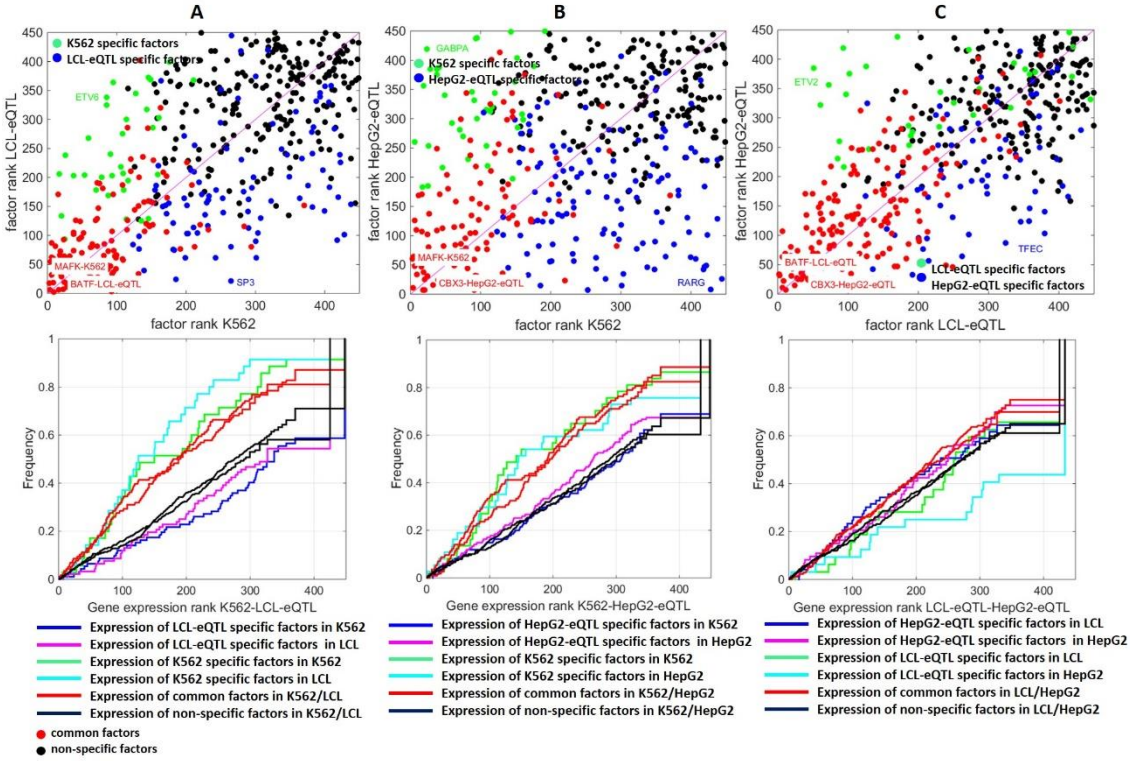


Figure 5: Comparison of contributing factors and their expression between three cell types using the following datasets: K562, LCL-eQTL, HepG2-eQTL. (A) Comparing K562 with LCL-eQTL. (B) Comparing K562 with HepG2-eQTL. (C) Comparing LCL-eQTL with HepG2-eQTL. X-axis shows the factor rank in one cell type, Y-axis shows the factor rank in another cell type. Cell type specific, common and non-significant factors are represented by green/blue, red, black dots respectively. The green/blue factor name is one of

the top 10 factors in one cell type with the highest degree difference from the second cell type. The red factor name is the top factor in each cell type (upper panel). Cumulative distribution of the corresponding expression rank of each class of factors in each of the two compared cell types (lower panel).

To further explore the extent of cell type specificity of MPRA, we took advantage of the LCL-eQTL and HepG2-eQTL, which include the same set of genomic regions. We first found that simply using the MPRA score of a region, as measured in LCL provides a good prediction to its activity in HepG2 (**Supplementary Figure 7A,B**), although this prediction is inferior to a classifier trained on other regions in HepG2 (**Figure 3**).

To understand the contribution of cell type specific factors we first examine the distribution of 3 region categories in these two datasets (**Figure 6A**): common regions (i.e. active regions in both datasets), cell type specific regions (i.e. regions active in one of the datasets), non-cell type specific regions (i.e. regions not active in both datasets). Focusing only on regions that are cell type specific, we record the experimental transcription factor binding (from ChIP-seq data [32]) in those regions. We compare the binding profiles of TFs between the cell type specific regions between LCL and HepG2. We find 35 significant factors (chi-square test p-value < 0.05). The top 10 factors that have higher binding rate in LCL specific regions (green dots) are: POU2F2, EBF1, BATF, NFKB, RUNX3, PML (immuno functional associated factors [28]), NFIC, PAX5C20, p300, FOXM1, and 2 factors that have higher binding rate in HepG2 specific regions (blue dots): RXRA, Rad21 (factors expressed in liver [28]). The 33 factors that have higher binding rate in LCL regions are enriched with lymphocyte activation [33]. Notably, the factors with the highest fold expression of 50 when comparing binding in LCL vs. HepG2 is NFATC1 which can regulate the activation, proliferation, differentiation and programmed death of lymphoid cells [34]. Factors specific to LCL have expression rank that is for some factors higher in LCL than in HepG2, but similar for others (**Figure 6B**). These results are consistent with the observation that dataset specific factors are not expressed significantly different between cell types (**Figure 5** lower panel).

To further explore where and why does transfer learning work, we examined prediction performance for active regions in LCL-eQTL. To this end, we define the "hardness" of the region based on the difference between the predicted score (in [0, 1]) and the class label (1 for active and 0 for not-active region) (**Methods**). We compared this score between regions predicted using the supervised cross- validation approach (i.e. training on the same cell type) or a transfer learning approach (i.e. training on a different cell type) (**Figure 6C**). Reassuringly, we observe that regions that are hard to predict in transfer learning. but easy to predict by training on the same data set (i.e., supervised, corss- validation) are enriched with cell type specific signals, such as the accesibility of the LCL chromatin, measured with DNAse-seq (**Figure 6D**).

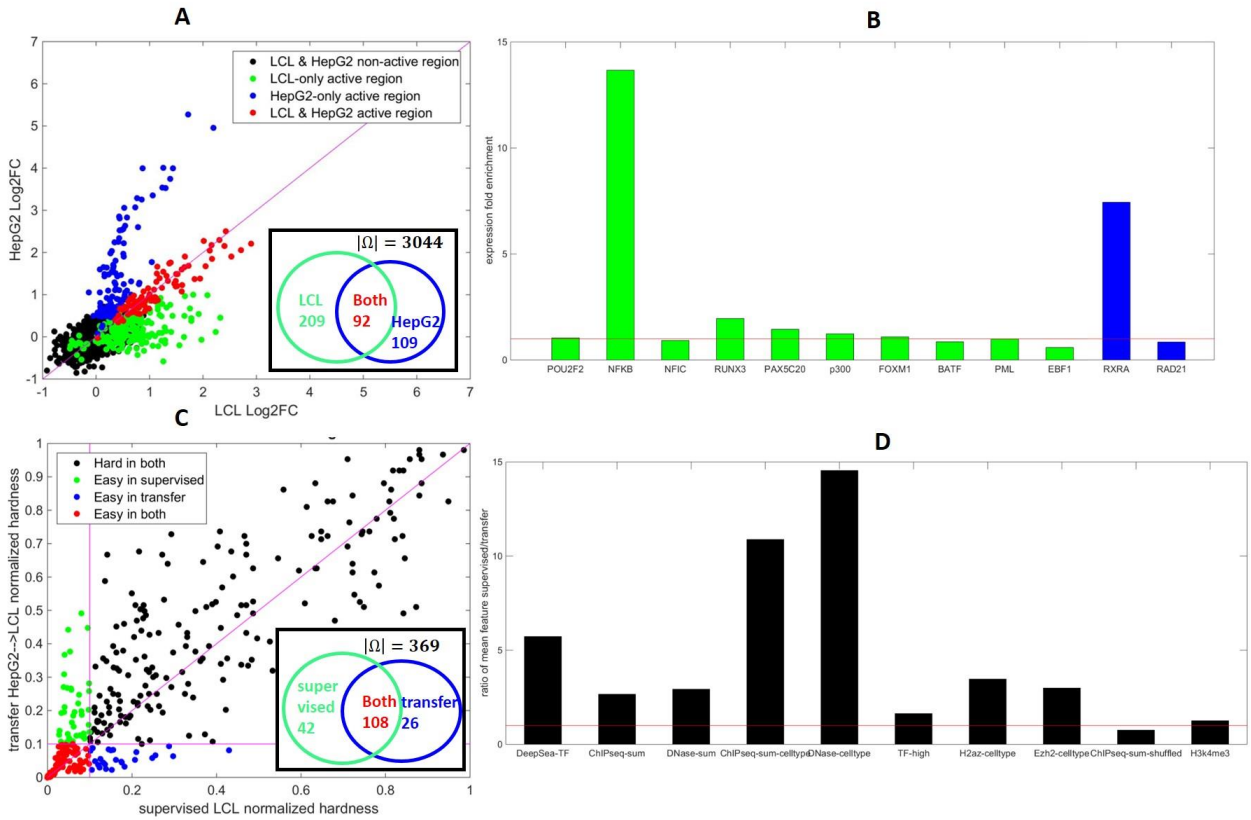


Figure 6: MPRA specificity. (A) MPRA activity of the same regions in HepG2 cells vs. LCLs. Black dots are non-active regions in both cell lines. Green dots are LCL only active regions. Blue dots are HepG2 only active regions. Red dots active regions in both datasets, lower right corner: Venn diagram representing the number of regions in each category. (B) Expression fold enrichment in LCL/HepG2 of top 10 and 2 factors that have significantly higher binding rate in LCL (green bars) and HepG2 (blue bars) respectively. For each cell type factors are sorted from the most (left) to the least (right) significant. (C) Comparing hardness to predict using a supervised approach in a cross validation within the same cell type vs. transfer learning between cell types - 0.1 is the cutoff to call a "hard" region. (D) The ratio (supervise/transfer) of the features mean, showing the top 10 features (out of 18 with FDR<0.05) that vary significantly in regions that are easy in the supervised vs. the transfer approach. Sorted from the most to the least significant.

Overall, these results support the notion that in many cases sequence- intrinsic properties rather than cell- type specific ones may be the dominant determining factors of MPRA activity. We also find that the cell type specific component can be captured through chromatin accessibility signal and histone marks that are cell type specific and by the activity of TFs whose function (but not necessarily expression profile) is associated with the cell type under investigation.

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 [19]. We wanted to know if we can predict these effects - starting from the synthetic setting of MPRA.

An important feature of the the LCL-eQTL and HepG2-eQTL datasets [19] 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) [35] 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 [10]. 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, compared to other participants in the CAGI challenge (**Supplementary Table 3**). Unsurprisingly, however, the absolute performance is substantially lower than that achieved in the task of predicting the transcription of individual sequences, which can be expected as this task relates to a much more nuanced signal.

We next tested how well transfer learning works in predicting allelic variation. Using the LCL-eQTL reference and alternative alleles for training and then testing on the HepG2-eQTL, achieved an accuracy level only moderately lower than the within- cell type scenario (AUROC of 0.61 and AUPRC of 0.41; **Supplementary Table 3**).

Detecting eQTL

After examining predictive ability in the synthetic context of MPRA, we aimed to move to the endogenous genome and distinguish between eQTLs and matched control SNPs. We first noted that the eQTL effect size is not correlated with neither MPRA levels (**Figure 7A**) or the difference between reference and alternative alleles (**Supplementary Figure 8**), at least as observed for the subset of LCL eQTLs, which makes this classification a challenging task. To this end, we explored the best combination of models and features for predicting eQTLs from the GRASP (Genome-Wide Repository of Associations between SNPs and Phenotypes) dataset [36] and controls from 1000 genomes project [37] curated by Zhou and Troyanskaya [12] (**Methods**). Our eQTL predictors are trained in two stages; First, predict the activity of the reference and alternative alleles using models trained on any of our five datasets. Second, decide whether the difference between the alleles is significant using a logistic regression model, trained using the LCL-eQTL reference/alternative dataset (see **Figure 7B** and **Methods** for details). We find that the Extra Trees regressor trained on K562 MPRA data with DeepSea and class 4 features yielded comparable results to an existing method [14] but the overall performance is low for all the variations tested (**Supplementary Figure 9A**).

Considering this low performance we were interested to examine which features characterize

eQTL regions that are hard to detect using MPRA data, but are otherwise easy to predict in a supervised setting (i.e., cross validation using the eQTL labels), which is known to lead to a better accuracy. We define the "hardness" of the region based on the difference between the predicted score (in [0, 1]) and the class label (1 for eQTL and 0 for control region). We then record the hardness of a region, either in the MPRA unsupervised setting (**Supplementary Figure 10**), or in a supervised, cross- validation setting (**Methods**). Consistent with our results above, we find that the MPRA- based hardness measure is consistent when using the different MPRA datasets as a training source (**Figure 7C**, left).

Focusing on eQTL regions that are easy to detect in a supervised setting, we looked for features that correlate with their hardness in the unsupervised, MPRA- based setting (**Figure 67**, right). Importantly, this subset of regions included a range of MPRA regions' "hardness" (**Supplementary Figure 10A**), and can be thus used to demonstrate what are the determinants of low performance when using an unsupervised approach that leverages MPRA predictions as features. Among the top features we find the region complexity (number of unique K-mers) and the number of TF binding sites, indicating that information rich eQTL regions tend to be easier to identify by MPRA. Moreover, when exploring the extreme examples of the easiest/hardest to predict eQTL/control SNPs (**Supplementary Figure 9B**), we notice that falsely classified variants are observed in unannotated regions and correctly classified variants in regions that have DNAse and TF binding signals. These results suggest that missing annotation could provide a partial explanation for hard to predict regions.

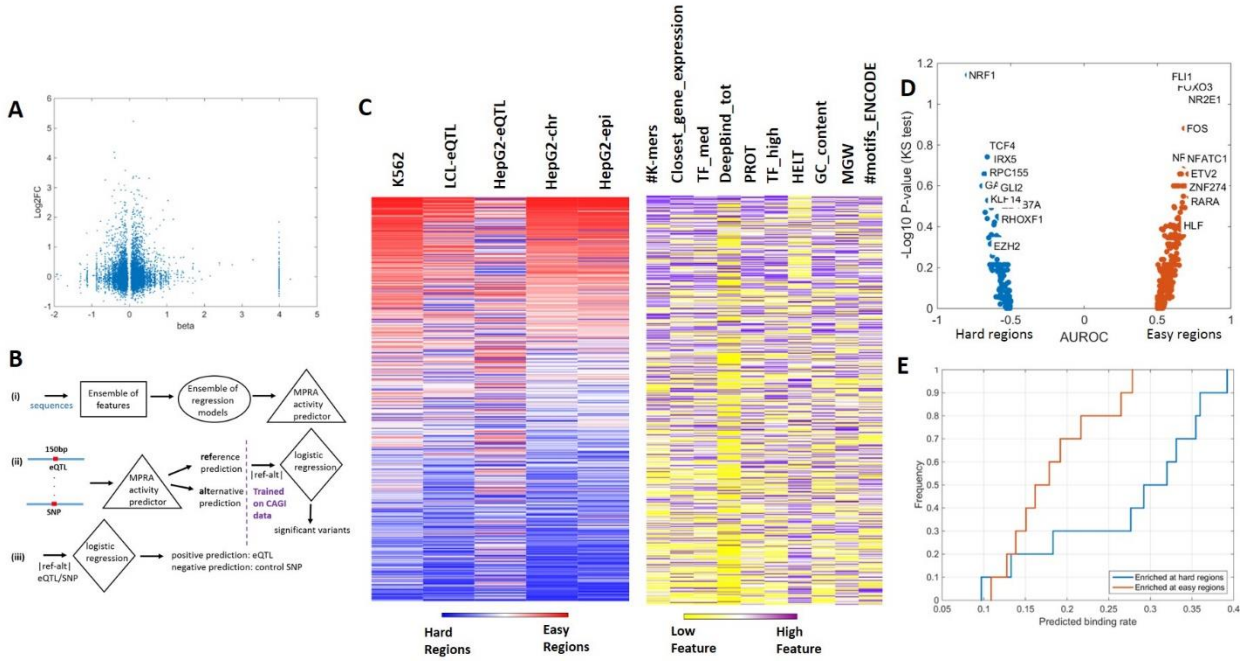


Figure 7: Regions hardness to predict. (A) Correlation between eQTL beta and MPRA activity for LCL-eQTL. (B) Flow chart for predicting regulatory variants. (C) Left panel: hardness to predict eQTL regions that are easily called using a supervised approach when using the 5 MPRA datasets for MPRA activity prediction.

Right panel: 10 most predictive features for regions hardness, the heatmap indicates the level of the feature for each one of the 500 eQTLs. (D) Contribution of individual TFs to the prediction of hard and easy regions, measured by the minus log *P* value of the ks-test for AUROC per factor. (E) Frequency of predicted binding sites across genome for TFs that are enriched in hard and easy regions.

We also noticed that the expression of the closest gene, which is very poorly correlated with MPRA activity (Figure 1) comes up as the second strongest feature for regions "hardness" (**Figure 7C**, right). This is a location specific feature, suggesting that one reason for the low accuracy when using MPRA data may be the difference in chromosomal context (i.e., either episomal, or random chromosomal integration site). When examining the prediction of regions hardness using predicted TF binding sites [11] (**Figure 7D**) we find that factors that are enriched in hard regions have higher binding rates across the genome attesting to their nonspecific nature and making them less powerful predictors (**Figure 7E**).

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 a number of 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. Finally, we examine how MPRA can be leveraged to predict expression associated variants across the endogenous genome. Our results represent, to the best of our knowledge, the first such comprehensive analysis.

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). In the MPRA setting the *cis* environment (e.g. chromatin) is altered, thus generally not cell type specific and the *trans* environment (e.g. TF binding) remains similar, hence we can still observe predictive factors that are cell type specific.

When we extend our approach to predict eQTLs across the genome, we notice that more complex regions are easier to predict and that lack of annotation might explain some of the reasons for hard predictions. We also notice that properties related to the *cis* "environment" of the sequence (e.g. the expression of the closest gene) are correlated with hardness to predict, indicating that the nature of MPRA integration (i.e. even if the MPRA is done in a chromosomal context, the integration sites are not controlled for) can have a strong effect on prediction performance.

Our work provides a comprehensive resource of annotation for thousands of endogenous sequences across the genome (**Supplementary Tables 1,4,5**). Furthermore, we demonstrate the performance of different machine learning models for MPRA activity prediction (**Supplementary Table 2**) by using publicly available tools, and will make these prediction models available for public use. Our approach can highlight functionally important regulatory regions across the genome in a cell-type agnostic fashion and can be leveraged for an efficient design of future MPRA experiments.

METHODS

Featurization - Individual features

Individual features can be categorized into several classes: **(1)** *Experimentally measured epigenetic properties*, including Transcription factor (TF) binding sites (TFBS), histone marks, chromatin accessibility (primarily by identifying DNase-hypersensitivity sites; henceforth abbreviated as DHS), and DNA- methylation. To define these features, each tested region is mapped to the reference human genome, and then queried against tracks of epigenetic properties from ENCODE [16], measured in multiple cell lines. These features include: DNase, Ctcf, Ezh2 H2az, H3k27ac, H3k27me, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me1, H3k9me3, H4k20me1, P300 - for these features we examine both cell type specific overlaps and sum of overlaps across all cell types. ChIP-seq_sum - the sum of all ChIP-seq peaks that overlap a region and ChIP-seq_sum_cell_type - the sum of cell type specific ChIP-seq peaks. ChIP-seq_sum_shuffled - the sum of randomly and equally sized chosen non-cell type specific

ChIP-seq peaks. **(2)** *Predicted epigenetic properties.* This set of features covers similar properties as the experimentally-derived ones (e.g., TFBS or DHS). However, instead of being directly measured, the properties are inferred based on the DNA sequence of the respective MPRA construct, using models trained on experimental data (e.g., protein binding microarrays for TFBS [38], or DNase-seq for DHS [16]). We use a wide array of models for prediction of epigenetic properties from sequence, from simple DNA-binding motif scoring [39] to more recent supervised learning algorithms such as *DeepBind* [11] and *DeepSea* [12]. These features include: #motifs_ENCODE—number of predicted TFBS based on the presence of DNA-binding motifs from ENCODE, min_dist_to_TSS_ENCODE—the distance between the transcription start site of the MPRA construct and the nearest motif hit, DeepBind_tot - using a neural network model trained on protein-binding microarrays. This feature was derived by marking the regions that score at the top 90% as hits for every TF, and for every region, count the number of TFs for which it has a hit. DeepSea_TF_tot - using a neural network model trained on chromatin profiling data. This feature was derived by marking the regions that score at the top 90% as hits for TF binding, and for every region, count the number of TFs for which it has a hit. TF_high, TF_med, TF_low - we anonymize TFs' identity by marking all TFs that are bound in the 90% percentile (based on DeepBind) and count how many of them are the top/middle/low expressed factors based on RNA-seq out of the top/middle/low 100 expressed genes in the corresponding cell type. minor groove width (MGW), ROLL, propeller twist (PROT), and helix twist (HELT) are DNA shape features [40] derived by taking the mean of their signal across each region. **(3)** DNA *k-mer frequencies.* Including #K-mer - the sum of 5-mer appearances in a region, #PolyA, #PolyT - length of polyA/T subsequence and GC content. **(4)** *Loci specific features*, including evolutionary conservation scores predicted by phastCons [41], closest_gene_expression - the expression of the closest gene (TPM) from RNA-seq data in the corresponding cell type, promoter/exon/intron/distal - genomic annotation indicating if a sequence intersects a promoter, exon or intron or otherwise it is annotated as distal.

<u>Featurization - Features used in prediction models</u>

For predicting MPRA activity we used the following sets of features: DeepBind - 515 predicted TF binding using a neural network model trained on protein-binding microarrays [11], DeepSea - predicted 919 TF binding, DNA accessibility, and histone modifications using a neural network model trained on chromatin profiling data [12], motifs - ~2,000 predicted motif hits [16] from simple DNA-binding motif scoring [39], K-mers - appearance of 5-mers, epigenetic - 1,095 overlaps of experimentally tested epigenetic marks in various cell types [16], across MPRA tested regions. summary (remaining summary features)- refers to a set of different features, including: polyA/T subsequences, GC content, DNA shape features: mean of predicted minor groove width (MGW), ROLL, propeller twist (PROT), and helix twist (HELT) [40], evolutionary conservation scores predicted by phastCons [41], distance to TSS - distance to the end of the MPRA construct, sum of

motif hits - sum of TF that bind to a region (based on ENCODE ChIP-seq data [16] for different factors across various cell types), anonymization of TFs - we anonymize TFs' identity by marking all TFs that are bound in the 90% percentile (based on DeepBind) and count how many of them are the top/middle/low expressed factors based on RNA-seq out of the top/middle/low 100 expressed genes in the corresponding cell type, and closest gene expression derived from RNA-seq data in the corresponding cell type.

Statistical tests

We examine the predictivity of features and accuracy of prediction models using several statistical tests. For regression task - i.e. predicting MPRA output we applied several correlation measures (Pearson, Spearman, Kendall), considering either the entire test data; variants at the top 25% of quantitative measurements; or a binning of the data. For classification task - i.e. binary prediction indicating if a region is active or not, we record the AUROC (area under receiver operating characteristic curve) and AUPRC (area under precision recall curve). For both regression and classification tasks, we applied a hypergeometric test. To better account for the binary predicted values (i.e., class 1 features), we also applied a fold enrichment test by examining the overlap between the set of true positives (active sequences) and the predicted positives. For quantitative features (classes 1, 3), the predicted positives were defined as regions with value higher than the mean. The significance of each test was evaluated by the respective statistical test (correlation *P* values for the regression tasks; Kolmogorov–Smirnov test for ROC and PR; hypergeometric *P* value for the enrichment test). All *P* values were corrected using the Benjamini–Hochberg procedure, and only associations below a FDR of 5% are presented.

Prediction models

All five regression and two classification models are used with default parameters. The ensemble of methods is implemented by considering the average of predicted outputs for the 5 regression models and the average of predicted probabilities for the 2 classification models.

TF categories when comparing two cell types

We define three categories of TFs: (1) Cell type specific factors (Figure 5 upper panel: green/blue dots respective to the cell type) are determined by a p-value threshold (0.05/#TFs) of the ks-test for AUROC per factor in each cell type, i.e. when comparing two cell types: a factor is specific to cell type 1 but not cell type 2, if the p-value is smaller than the threshold in cell type 1 but larger than the threshold in cell type 2. (2) Common factors (Figure 5 upper panel: red dots) that are significant in both cell types, i.e. p-value in both cell types is lower than the threshold. (3) Factors with non-significant contribution in both cell types (Figure 5 upper panel: black dots) i.e. p-value in both cell types is higher than the threshold.

Detecting eQTL

eQTLs from the GRASP dataset [36] were matched (Zhou and Troyanskaya, 2015) with controls from 1000 genomes project [37] with different distances from the eQTLs and similar allele frequency distribution . The specific distance bins are represented using the following 5 categories that stand for the mean distance of the control SNP from an eQTL: 360bp, 710bp, 6.3kbp, 31kbp, Random - matched controls from the whole genome without a distance constraint). We randomly chose 10,000 variants from the 6 categories (i.e. eQTLs as the positive set and each one of the distance bins as a negative set). The positive set is used against each one of the 5 negative sets for prediction (**Supplementary Figure 9A**).

The scheme for calling eQTLs (**Figure 7B**) includes: (i) train an ensemble of regression models on each one of the 5 MPRA datasets (separately) to predict MPRA activity. (ii) Use the trained model from step (i) to predict MPRA activity for the 150bp region harboring the reference/alternative alleles. This is done both for the positive set (eQTL) and their matched controls (i.e. negative set SNPs). (iii) apply a logistic regression model to distinguish between eQTLs and control SNPs. This model is trained on allelic skew measurements as defined in the CAGI eQTL causal challenge [10] (i.e. the significance of the difference between the transcriptional activity of the two alleles was evaluated using a *t*-test on the log-transformed RNA-seq/plasmid ratios across replicates with a FDR cutoff of 0.05). The input for the logistic regression model is the absolute difference between the predicted MPRA activity of the reference and the alternative alleles (predicted in step (i)). We applied different strategies of aggregation (e.g. using only the reference allele activity prediction) but this did not yield better results (data not shown).

Defining “hardness” of the region - we predicted 10,000 eQTL and 10,000 matched negative control SNPs across the genome and recorded the rank (tiedrank) of the classifier score, we then normalized the ranks by dividing by the maximum rank. We define region’s hardness as the absolute value of the difference between the label of the variant (either 0 or 1) and the normalized rank, representing how far we are from the right prediction.

REFERENCES

1. Weingarten-Gabbay, S. & Segal, E. The grammar of transcriptional regulation. *Hum. Genet.* **133,** 701–711 (2014).
2. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **23,** 1908–1915 (2013).
3. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30,** 265–270 (2012).
4. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).
5. Patwardhan, R. P. *Massively Parallel Functional Dissection of Regulatory Elements*. (2012).
6. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. Massively Parallel Reporter Assays in Cultured Mammalian Cells. *J. Vis. Exp.* (2014). doi:10.3791/51719
7. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30,** 521–530 (2012).
8. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45,** 1021–1028 (2013).
9. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077 (2013).
10. Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum. Mutat.* (2017). doi:10.1002/humu.23197
11. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).
12. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12,** 931–934 (2015).
13. Zeng, H., Hashimoto, T., Kang, D. D. & Gifford, D. K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32,** 490–496 (2016).
14. Zeng, H., Edwards, M. D., Guo, Y. & Gifford, D. K. Accurate eQTL prioritization with an ensemble-based framework. *Hum. Mutat.* (2017). doi:10.1002/humu.23198
15. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24,** 1595–1602 (2014).
16. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
17. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).
18. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28,** 817–825 (2010).

19. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165,** 1519–1529 (2016).
20. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
21. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27,** 38–52 (2017).
22. Erwin, G. D. *et al.* Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.* **10,** e1003677 (2014).
23. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67,** 301–320 (2005).
24. Breiman, L. *Mach. Learn.* **45,** 5–32 (2001).
25. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63,** 3–42 (2006).
26. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2009).
27. Ameyar, M., Wisniewska, M. & Weitzman, J. B. A role for AP-1 in apoptosis: the case for and against. *Biochimie* **85,** 747–752 (2003).
28. Safran, M. *et al.* Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.* **31,** 142–146 (2003).
29. Craig, M. P. & Sumanas, S. ETS transcription factors in embryonic vascular development. *Angiogenesis* **19,** 275–285 (2016).
30. Géraud, C. *et al.* 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 (2010).
31. Roberts, K. E. *et al.* Genetic risk factors for hepatopulmonary syndrome in patients with advanced liver disease. *Gastroenterology* **139,** 130–9.e24 (2010).
32. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).
34. Chuvpilo, S. Three NF-ATc isoforms with individual properties are expressed in lymphoid cells. *Immunol. Lett.* **56,** 349 (1997).
35. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
36. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30,** i185–94 (2014).

37. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
38. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **37,** D77–82 (2009).
39. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
40. Zhou, T. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* **41,** W56–62 (2013).
41. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).