A framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation across organisms

Anurag Sethi[1,2,†], Mengting Gu[1,†], Emrah Gumusgoz[6], Landon Chan[3], Koon-Kiu Yan[1,2], Joel Rozowsky[1,2], Iros Barozzi[7], Veena Afzal[7], Jennifer Akiyama[7], Ingrid Plajzer-Frick[7], Catherine Pickle[7], Momoe Kato[7], Tyler Garvin[7], Quan Pham[7], Anne Harrington[7], Brandon Mannion[7], Elizabeth Lee[7], Yoko Fukuda-Yuzawa[7], Axel Visel[7], Diane E. Dickel[7], Kevin Yip[4], Richard Sutton[6], Len A. Pennacchio[7] and Mark Gerstein[1,2,5]

Deleted: Kevin Yip[4],

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America
[2] Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America
[3] School of Medicine, The Chinese University of Hong Kong, China
[4] Department of Computer Science, The Chinese University of Hong Kong, China
[5] Department of Computer Science, Yale University, New Haven, Connecticut, United States of America
[6] Department of Internal Medicine, Section of Infectious Diseases, Yale University School of Medicine, New Haven, Connecticut, United States of America
[7] Functional Genomics Department, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

Abstract

Enhancers are important noncoding elements, but they have been traditionally hard to characterize experimentally. Only a few mammalian enhancers have been validated, making it difficult to train statistical models for their identification properly. Instead, postulated patterns of genomic features were used heuristically for identification. The development of massively parallel assays allows the characterization of large numbers of enhancers for the first time. Here, we develop a framework that uses them to create shape-matching filters based on enhancer-associated meta-profiles of epigenetic features. These features are combined with supervised machine learning algorithms (i.e., SVMs) to predict enhancers. We demonstrated that our model can be applied to predict enhancers in mammalian species (eg, mouse and human). The predictions are comprehensively validated using a combination of *in vivo* and *in vitro* assays (133 mouse transgenic enhancer assays in 6 different tissues and 25 human H1 hESC transduction-based reporter assays). The validation results confirm that our model can accurately predict enhancers in different species without re-parameterization. Finally, we predict enhancers in cell lines with many transcription-factor binding sites. This highlights distinct differences between the type of binding at enhancers and promoters, enabling the construction of a secondary model discriminating between these two.

Introduction

Enhancers are gene regulatory elements that activate expression of target genes from a distance [1]. Enhancers are turned on in a space and time-dependent manner contributing to the formation of a large assortment of cell-types with different morphologies and functions even though each cell in an organism contains a nearly identical genome [2-4]. Moreover, changes in the sequences of regulatory elements are thought to play a significant role in the evolution of species[5-9]. Understanding enhancer function and evolution is currently an area of great interest because variants within distal regulatory elements are also associated with various traits and diseases during genome-wide association studies [10-12]. However, the vast majority of enhancers and their spatiotemporal activities remain unknown because it is not easy to predict their activity based on DNA sequence or chromatin state [13, 14].

FURTHERMORE,

Traditionally, the regulatory activity of enhancers and promoters were experimentally validated in a non-native context using low throughput heterologous reporter constructs leading to a small number of validated enhancers that function in the same mammalian cell-type [15, 16]. ~~In addition to the small numbers~~, the validated enhancers were typically selected based on conserved noncoding regions [17] with particular patterns of chromatin [18], transcription-factor binding, [19] or noncoding transcription [20]. The small number and biases within the validated enhancers make them inappropriate for parameterizing tissue-specific enhancer prediction models [16]. As a result, most theoretical methods to predict enhancers could not optimally parameterize their models using a gold-standard set of functional elements. Instead, most of these models were parameterized based on certain heuristic features associated with enhancers, which were then utilized to predict enhancers [19, 21-30]. For example, two widely used methods for predicting enhancers were based on the fact that these elements are expected to contain a cluster of transcription factor binding sites [24] and their activity is often correlated with an enrichment of particular post-translational modifications on histone proteins [27, 30]. These predictions could not be comprehensively assessed as few putative enhancers could be validated experimentally due to the low throughput of validation assays and it remains challenging to assess the performance of different methods for enhancer prediction.

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

In recent times, due to the advent of next-generation sequencing, a number of transfection and transduction-based assays were developed to experimentally test the regulatory activity of thousands of regions simultaneously in a massively parallel fashion [31-37]. In these experiments, several plasmids that each contain a single core promoter upstream of a luciferase or GFP gene are transfected or transduced into cells. These plasmids are used to test the regulatory activity of different regions by placing one region within the screening vector in each plasmid as differences in the gene's expression occur due to the differences in the activity of the tested region. STARR-seq was one such massively parallel reporter assay (MPRA) that was used to test the regulatory activity of the *Drosophila* genome by inserting candidate fragments from the genome within the 3' untranslated region of the luciferase gene. STARR-seq identified thousands of cell-type specific enhancers and promoters within the *Drosophila* genome [31, 38]. MPRAs have confirmed that active enhancers and promoters tend to be depleted of histone proteins and contain accessible DNA on which various transcription factors and cofactors bind [39, 40]. These regulatory regions also tend to be flanked by nucleosomes that contain histone proteins with certain characteristic post-translational

2 modifications. These attributes lead to an enriched peak-trough-peak (“double peak”) signal in different ChIP-Seq experiments for various histone modifications such as acetylation on H3K27 and methylations on H3K4. The troughs in the double peak ChIP-seq signal represent the accessible DNA that leads to a peak in the DNase-I hypersensitivity (DHS) at the enhancers [41]. However, the optimal method to combine information from multiple epigenetic marks to make cell-type specific regulatory predictions remains unknown. For the first time, using data from several MPRAs, we have the ability to properly train our models based on a large number of experimentally validated enhancers and test the performance of different models for enhancer prediction using cross validation.

Formatted: Highlight

Here we develop a framework for making supervised enhancer prediction models using MPRA datasets. We make use of all published data resources to provide a comprehensive model for enhancer prediction that can be applied across different contexts (i.e., different species and tissue types); we validate our model in a variety of different contexts. In particular, we utilized extensive datasets from STARR-seq experiments performed on *Drosophila* cell lines to create and parameterize our model. Unlike previous prediction methods that focused on the enrichment (or signal) of different epigenetic datasets, we developed a method to also take into account the enhancer-associated pattern within different epigenetic signals. As the epigenetic signal around each enhancer is noisy, we aggregated the signal around thousands of enhancers identified using MPRAs to increase signal-to-noise ratio, and identified the shape associated with active regulatory regions. Previous ENCODE and modENCODE efforts showed that the chromatin modifications on active promoters and enhancers were conserved across higher eukaryotes [42-48]. The signal of different chromatin modifications upstream of a gene have been used to create a universal model for predicting its expression and the parameters of the model were transferable across humans, flies, and worm. Here, we further explored this conservation of epigenetic signal shapes for constructing simple-to-use transferrable statistical models with six parameters that were used to predict enhancers and promoters in diverse eukaryotic species including fly, mouse, and human. We showed that the enhancer predictions from our transferrable model was comparable to the prediction accuracy of species-specific models.

Working across organisms also allowed us to take advantage of different assays to validate our predictions in a robust fashion using multiple experimental approaches. In the first stage, we predicted enhancers in six different embryonic mouse tissues and tested the activity of these predictions *in vivo* with transgenic mouse assays. Due to the obvious ethical considerations of performing such transgenic assays in human embryos, we then proceeded to test the activity of these elements in a human cell-line *in vitro*.

✓

H1-hESC is a highly studied human cell-line in which a comprehensive set of transcription factor (TF) binding experiments are available. After validating our predictions, the many TFs provided us with the opportunity to differentiate between the enhancers and promoters. The pattern of TF and co-TF binding at active enhancers is much more heterogeneous than the corresponding patterns on promoters, which can be used to distinguish enhancers from promoters with high accuracy. Thus, our methods provide a framework that utilizes different epigenetic genomics datasets to predict active regulatory regions in a cell-type specific manner. Further functional genomics datasets can be utilized to identify key TFs associated with active regulatory regions within these cell types.

Results

Aggregation of epigenetic signal (in *Drosophila*) to create metaprofile:

We developed a framework to predict active regulatory elements using the epigenetic signal patterns associated with experimentally validated promoters and enhancers [31]. We aggregated the signal of histone modifications on MPRA peaks to remove noise in the signal and created a metaprofile of the double peak signals of histone modifications flanking enhancers and promoters. MPRA peaks typically consist of a mixture of enhancers and promoters, and at this stage, we do not differentiate between the two sets of regulatory elements. As STARR-seq quantifies enhancer activity in an episomal fashion, not all STARR-seq peaks would be active in the native chromatin environment. Arnold C. et al showed that the STARR-seq peaks that occur in enriched DNase hypersensitivity or H3K27ac modifications tend to be near active genes while other STARR-seq peaks tend to be associated with enrichment of repressive marks such as H3K27me3. Hence, we took the overlap of the STARR-seq enhancers with H3K27ac and/or DHS peaks to get a high confident set of enhancers that are active in vivo, based on which the metaprofiles were created. These metaprofiles were then utilized in a pattern recognition algorithm for predicting active regulatory elements in a cell-type specific manner.

WITH

UFRC

!

Deleted: promoters and enhancers

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 12 pt

The STARR-seq studies on *Drosophila* cell-lines provide the most comprehensive MPRA datasets as the whole genome was tested for regulatory activity within these assays and these assays were performed with multiple core promoters [31, 49]. Hence, we chose to create metaprofiles using the histone modification H3K27ac at active STARR-seq peaks (see Figure 1 and Methods) identified within the *Drosophila* S2 cell-line. Approximately 70% of the active STARR-seq peaks contain an easily identifiable double peak pattern even though there is a lot of variability in the distance between the two maxima of the double peak in the ChIP-chip signal (Figure S1). While the minimum tends to occur in the center of these two maxima on average, the distance between the two maxima in the double peaks can vary between 300 and 1100 base pairs. During aggregation, we aligned the two maxima in the H3K27ac signal across different STARR-seq peaks, followed by interpolation and smoothening the signal before calculating the average metaprofile. In addition, an optional flipping step was performed to maintain the asymmetry in the underlying H3K27ac double peak because it may be associated with the directionality of transcription [50]. We also calculated the dependent metaprofiles for thirty other histone marks and DHS signal by applying the same set of transformations to these datasets. The metaprofile for the histone marks associated with active regulatory regions were also double peak signals, and the maxima across different histone modification signals tended to align with each other on average (Figure S2). This indicates that a large number of histone modifications tend to simultaneously co-occur on the nucleosomes flanking an active enhancer or promoter. In contrast, as expected, the DHS signal displayed a single peak at the center of the H3K27ac double peak (Figure 1). In addition, repressive marks such as H3K27me3 were depleted in these regions, and the metaprofile for these regions did not contain a double peak signal (Figure S2).

Deleted: (cite31, 50)

Formatted: Highlight

Match of a metaprofile is predictive of regulatory activity:

We evaluated whether these metaprofiles can be utilized to predict active promoters and enhancers using matched filters, a well-established algorithm in template recognition. A matched filter is the optimal pattern recognition algorithm that uses a shape-matching filter to recognize the occurrence of a template in the presence of stochastic noise [51]. We evaluated whether the occurrence of the epigenetic metaprofiles identified for the histone marks and DHS can be used to predict active enhancers and promoters using receiver operating characteristic (ROC) and precision-recall (PR) curves. PR curves are particularly useful to assess the performance of classifiers in skewed or imbalanced data sets in which one of the classes is observed much more frequently compared to the other class, as it plots the fraction of true positives among all predicted positives. If the area under a PR curve is higher, the corresponding model has a low false discovery rate and can easily distinguish between the positives from the negatives. On the other hand, in skewed datasets, the area under ROC curves could be high even when the FDR is high even. This is because, in these cases, even if a small fraction of negatives are predicted to be positive by the model, the false discovery rate can be high as the total number of true positives are much smaller than the total number of true negatives [52]. The matched filter score is higher in genomic regions where the template pattern occurs in the corresponding signal track while it is low when only noise is present in the signal (Figure 1). Due to the aforementioned variability of the distance between the double peaks, we allow the widths of scanned regions to vary between 300-1100 basepairs (at steps of 25 basepairs). A single H3K27ac metaprofile was applied to match different width and the highest score was used to rate the regulatory potential of this region (see Methods). The dependent profiles are subsequently used on the region of the same width to score the corresponding genomic tracks.

TRANS

CIRC

We used 10-fold cross validation to assess the performance of matched filters for individual histone marks to predict active STARR-seq peaks. In Figure 2, we observe that the H3K27ac matched filter is the single most accurate feature for predicting active regulatory regions (AUROC=0.92, AUPR=0.72) identified using STARR-seq. This is consistent with the literature as H3K27ac enriched peaks are often used to predict active promoters and enhancers [23, 53, 54]. In general, several histone acetylations (H3K27ac, H3K9ac, H4K12ac, H2BK5ac, H4K8ac, H4K5ac, H3K18ac) marks as well as the H1, H3K4me2, and DHS are the most accurate prediction features (Table S1) because the matched filter scores for these features are higher on the STARR-seq peaks. The degree to which the matched filter scores for promoters and enhancers are higher than the matched filter scores for the rest of the genome is a measure of the signal to noise ratio for regulatory region prediction in the corresponding feature's genomic track. The larger the separation between positives and negatives, the greater the accuracy of the corresponding matched filter for predicting active regulatory regions. Interestingly, the distribution of matched filter scores for STARR-seq peaks are unimodal for each histone mark except for H3K4me1, H3K4me3, and H2Av, which are bimodal (Figure S3). We also show that the matched filter scores are more accurate for predicting active STARR-seq peaks than the enrichment of signal alone as they outperform histone peak calling on ROC and PR curves (Figure S4).

While a single STARR-seq experiment identifies thousands of active regulatory regions, these regions display core-promoter specificity, and different sets of enhancers are identified when different core promoters are used in the same cell-type [55-59]. As we wanted to create a framework to predict all the enhancers and promoters active in a particular cell type, we combined the peaks identified from multiple STARR-seq

Deleted: Due to the aforementioned variability in the double peak pattern, the H3K27ac signal track is scanned with multiple matched filters with templates that vary in width between the two maxima in the double peak and the highest matched filter score with these matched filters is used to rate the regulatory potential of this region (see Methods). The dependent profiles are then used on the same region with the matched filter to score the corresponding genomic tracks.

Formatted: Highlight

DEFEND HERE ?

experiments in the S2 cell-type and reassessed the performance of the matched filters at predicting these regulatory regions. Merging the STARR-seq peaks from multiple core promoters in the S2 cell-type leads to higher AUROC and AUPR for the matched filters from most histone marks (Figure 2 and Table S2).

Machine learning can combine matched filter scores from different epigenetic features

We built an integrated model with combined matched filter scores of the most informative epigenetics marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and DHS) associated with active regulatory regions using a linear SVM ~~[54]~~ [60]. The selection of these six features is based on their matched filter score performance, their importance in the integrated model and the data availability (See Methods). Particularly, the combination of these six features allows the integrated model be applied to a variety of cell lines and tissues, as many relevant ChIP-seq and DNase experiments have been performed by the Roadmap Epigenomics Mapping [61] and the ENCODE [62] Consortia in a wide variety of samples. We also assessed the performance of other statistical approaches including a nonlinear SVM for combining the features. While all these approaches performed similarly (Figure S5), a linear SVM is used in our framework for its better interpretability.

AS A SINGLE FEAT. MF

AGREE MORE

During integration, the normalized matched filter score for each epigenetic feature in a particular region is scaled by its optimized weight and added together to form a discriminant function. The sign of the discriminant function is then used to predict whether the region is regulatory. The features with large positive and negative weights are predicted to be important for discriminating regulatory from non-regulatory regions. The optimized weights can also be used to measure the amount of non-redundant information added by each feature in the integrated model. According to the model, the acetylations (H3K27ac and H3K9ac) are the most important feature for predicting active regulatory regions. The DHS matched filter performed well as an individual feature (AUPR in Figure 2) to predict enhancers, but had a lower weight among the six features likely due to the fact that the information in DHS is redundant with the information contained within the histone mark, eg. the DHS peaks usually occur at the trough region between two maxima in the histone signal. Despite the redundancy, combination of the DHS and histone signals is more predictive of regulatory activity as the reinforcing signals are strengthened compared to the uncorrelated noise in each signal track. The integrated model, as expected, achieved a higher accuracy than the individual matched filter scores (Figure 2), as they can leverage information from multiple epigenetic marks. We also trained a 6-parameter SVM model using STARR-seq data in BG3 cell-line. The model is highly accurate at predicting active enhancers and promoters in the S2-cell line (Figure S6), indicating our framework of combining epigenetic features with a linear SVM model to predict enhancers is applicable across species of great evolutionary distance.

MORE

THE PREDICTION AS

NO

To assess the information contained in other epigenetic marks, we combined the matched filters from all 30 measured histone marks along with the DHS matched filter in separate statistical models (Figure S7) and these models displayed higher accuracy (AUROC=0.97, AUPR=0.93 for SVM model with multiple core promoters) than the 6 feature model presented in Figure 2. The feature weights in this model indicated that H3K27ac contains the most information regarding the activity of regulatory regions. However, we found that a few other acetylations such as H2BK5ac, H4ac, and H4K12ac

Formatted: Strikethrough, Highlight

Formatted: Not Strikethrough

Formatted: Not Strikethrough

Formatted: Highlight

Deleted: T

Deleted: selection

Deleted: ensures that

Deleted: can

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial, Not Bold

Formatted: Font:(Default) Arial, Not Bold

Formatted: Font:Not Bold

Deleted: The DHS matched filter performed well as an individual feature (AUPR in Figure 2) to predict enhancers and can be highly predictive of regulatory activity in combination with other marks such as H3K27ac (Moore et al., in review). However, in the integrated model, the information in DHS is redundant with the information contained within the five histone marks as indicated by the fact that it has the lowest weight among the six features in the integrated model.

Formatted: Highlight

Comment [MG1]: Across different cell lines

Formatted: Highlight

Formatted: Font:(Default) Times New Roman, (Asian) Times New Roman, 12 pt contain additional non-redundant information regarding the activity of these regulatory regions and might improve the accuracy of promoter and enhancer prediction from machine learning models.

To evaluate the impact of the training sample size on model performance, we did a saturation analysis where we down sampled the training data to different levels of fractions and evaluated the model performance on the remaining data. For each fraction level, we did a 10-fold cross-validation (see methods) and then took the average of the ten outputs. The result shows that the average AUPR increases with increasing size of training data, and it starts to saturate for our SVM model with 80%-90% of the experimental data for training. In contrast to that, the average AUROC remain comparable with varying training size, but the performance variances decrease with increasing training data size. This indicates that a 5-fold cross validation might be sufficient with this size of data, as a 5-fold cross validation uses 80% of the data for training and the remaining 20% of the data for testing. In fact, even a 2-fold cross validation could work as the AUPR is close to saturation with 50% of the data for training.

Formatted: Font:(Default) Arial, 11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:(Default) Arial, 11 pt

Deleted:

Distinct epigenetic signals associated with promoters and enhancers

We proceeded to create individual metaprofiles and machine learning models for the two classes of regulatory activators – promoters (or proximal) and enhancers (or distal). We divided all the active STARR-seq peaks into promoters or enhancers based on their distance to the closest transcription start site (TSS) to delineate their likely function in the native context. Due to the conservative distance metric used in this study (1kb upstream and downstream of TSS in *Drosophila* genome), the enhancers are regulatory elements that are not close to any known TSS and could be considered to enhance gene transcription from a distance. However, a few of the promoters may also regulate distal genes in addition to their promoter activity. We then created metaprofiles of the different epigenetic marks on the promoters and enhancers and assessed the performance of the matched filters for predicting active regulatory regions within each category (Figure 3). The highest matched filter scores are typically observed on promoters, and the matched filters for each of the six features tended to perform better for promoter prediction. The H3K27ac matched filter continues to outperform other epigenetic marks for predicting active promoters and enhancers. In addition, the DHS, H3K9ac, and H3K4me2 matched filters also performed reasonably for promoter and enhancer prediction. Similar to previous studies [63, 64], we observed that the H3K4me1 metaprofile performs better for predicting enhancers while it is close to random for predicting promoters. In contrast, the H3K4me3 metaprofile can be utilized to predict promoters and not enhancers. The histogram for matched filter scores shows that H3K4me1 matched filter score is higher near enhancers while the H3K4me3 matched filter score tends to be higher near promoters (Figure S8). The mixture of these two populations lead to bimodal distributions for H3K4me1 and H3K4me3 matched filter scores when calculated over all regulatory regions (Figure S3).

We created different integrated models to learn the combination of features associated with promoters and enhancers respectively. These integrated models outperformed the individual matched filters at predicting active enhancers and promoters (Figures 3 and S9). In addition, the weights of the individual features identified the difference in roles of the H3K4me1 and H3K4me3 matched filter scores at discriminating active promoters and enhancers from inactive regions in the genome. The promoter-based (enhancer-based) model performed much more poorly at predicting enhancers (promoters) indicating the unique properties of these regions (Figures S10 and S11). We also created two integrated models utilizing matched filter scores of all thirty histone marks as features for predicting enhancers and promoters. The additional histone marks provided independent information regarding the activity of promoters and enhancers as these features increased the accuracy of these models (Figure S12). The weights of different features indicate that H2BK5ac again displays the most independent information for accurately predicting active enhancers and promoters. We observe similar trends and accuracy with several different machine learning methods (Figures S9 and S12). To investigate in the impact of different distance metrics used to segregate enhancers and promoters, we repeated our analysis with different distance CUTOFFS ~~metrics~~ (0.5kb, 1.5kb, 2.0kb and 2.5kb). While the accuracy as measured by the AUROC of different features and the integrated model slightly reduces as the distance ~~cutoff~~ increases, the importance of each feature in the integrated model as measured by the GINI score remains similar (Figure SXX).

Deleted:

Application of STARR-seq model to predict enhancers in mammalian species

One of the important findings of previous ENCODE and model organism ENCODE efforts is the conservation of chromatin marks close to regulatory elements across hundreds of millions of years of evolution [42-48]. The relationship of chromatin marks to gene expression was very similar, for instance, in worms, flies, mice and human, so much that one could build a statistical model relating chromatin modification to gene expression that would work without re-parameterization across different organisms. This motivated us to apply our well-parameterized model based on the STARR-seq data from flies to mammalian systems -- eg. mouse and human -- and test our model performance.

We started BY MAKING with genome-wide predictions of regulatory regions in mouse. Tissue-specific epigenetic signals were processed and USED FOR applied to our model to account for the tissue specificity of enhancers. Predictions are made in six different tissues (forebrain, midbrain, hindbrain, limb, heart and neural tube) at mouse e11.5 stage (Genome-wide predictions are available through our website at https://goo.gl/E8fLNN). These tissues are selected as their epigenetic signals are highly studied in mouse ENCODE, providing us with a rich source of raw data that can be utilized for making enhancer and promoter predictions. In addition, the VISTA database contains close to 100 validated enhancers that can be used for test for each of these tissues. Using our model, we predicted 31K to 39K regulatory regions in individual tissues in mouse, with each region ranging from 300bp to 1100bp. Notably, a consistent proportion of two-thirds (66%~70%) of these predicted regulatory regions are distal regulatory elements for all six tissues, with the other one-third (30%~34%) being proximal regulators (Table S3). These numbers agree with a previous enhancer evolution study [8], and suggest that the amount of enhancers and promoters are likely comparable in different tissues.

Similarly, we did genome wide prediction of regulatory regions in ENCODE top tier human cell lines, including H1-hESC, GM12878, AS 49? K562, HepG2 and MCF-7 (all available through our website). For each cell line, we utilized the 6-parameter integrated model to predict active enhancers and promoters based on the epigenetic datasets measured by the ENCODE consortium [62]. In H1-hESC, for example, we predicted 43463 active regulatory regions, of which 22828 (52.5%) are within 2kb of the TSS and are labeled as promoters. A large proportion of the predicted enhancers are found in the introns (30.41%) and intergenic regions (13.93%) (Figure S13). The predicted promoters and enhancers are significantly closer to active genes than might be expected randomly (Figure S14).

Whole genome STARR-seq enables proper training of enhancer prediction model

We next tried to evaluate how well the STARR-seq model did on predicting mammalian enhancers. Particularly, we want to compare the current mouse enhancer predictions with predictions from models directly trained on mouse data. The relatively large number of known mouse enhancers from VISTA database enabled us to parameterize a model in a same way as what we did with the *Drosophila* STARR-seq data. However, the VISTA database is not nearly at the same scale as the STARR-seq dataset. In total, we pulled together 1253 tissue specific positive regions and 8631 tissue specific negative regions from the assays.

With VISTA database, we trained four models based on four sets of available E11.5 mouse tissue-specific enhancers (hindbrain, limb, midbrain and neural tube), and assessed them using 10-fold cross-validation respectively. (There are no DHS data available for E11.5 forebrain and heart thus these two tissues are excluded for fair comparison). The average AUROC value is compared to the AUROC of testing STARR-seq trained model on the same VISTA enhancer data. Despite the significantly unbalanced negative to positive ratios of mouse enhancers in the database, the 6-parameter integrative SVM models learned using balanced *Drosophila* STARR-seq data were highly accurate at predicting active enhancers and promoters in mouse (Figure S15 A). The cross-validated mouse model, while it did well, performed no better on predicting mouse tissue specific enhancers. We found that the best performing one among the mouse models is for tissue midbrain, likely due to the fact that the number of validated midbrain enhancers is the largest. To construct a larger training sample for mouse, we pooled together the normalized z-scores of matched filter scores for six epigenetic signals of all four tissues, and parameterized a model using this larger set of data. Again, we observed that the original model trained with *Drosophila* STARR-seq data performed equally well on predicting mouse enhancers and much better in predicting fly enhancers (Figure S15 B). Overall, the result suggests that using the larger and more comprehensive STARR-seq data set for parameter tuning was superior to using the smaller mouse data set, even on mouse.

COULD WE USE TO ARGUE AGAINST K27ac bias

CHECK

Given the above overall statistical evaluations, we are confident in the STARR-seq parameterized model. We then set out to do targeted unbiased validations of the mammalian enhancers predicted, which is described in the next two sections.

Validation in vivo in Mouse

To test the activity of predicted mouse enhancers in vivo, we performed transgenic mouse enhancer assay in e11.5 mice for 133 regions in heart and forebrain, including 102 regions selected based on the H3K27ac signals rank of corresponding mouse tissues, and 31 regions selected by an ensemble approach from human homolog sequences. For each tested candidate, a read out of activity across the entire embryo is collected. The number of transgenic mice that showed the pattern for each tissue is also recorded for reproducibility check (See Methods and Supplement Table S4, S5). In

Deleted: (See Methods and Supplement Table S4, S5) addition, we obtained another set of transgenic mouse enhancer assay results from ENCODE Phase III Encyclopedia (Moore et al., in review), which assessed 151 regions in mouse e11.5 hindbrain, midbrain and limb. The combined results from these two large sets of validations, as well as any previously tested tissue-specific e11.5 enhancers from VISTA database, allow us to comprehensively evaluate our enhancer predictions in all six e11.5 mouse tissues.

Among the first 102 tested regions, 62 are selected based on forebrain H3K27ac signal rank, with 20, 22, 20 regions being in the top, middle and bottom rank respectively. Another 40 regions are selected by heart H3K27ac signal rank with half of them coming from the top rank and the other half coming from the middle rank. The bottom ranked regions were skipped because the activity of middle ranked regions dropped off so much. Consistently, the observed active rate of assessed regions decreases from top tier to bottom tier. For the other 31 human homolog sequences, 12.9% and 9.7% of the assessed regions are active in heart and forebrain respectively. The lower active rate is likely due to the fact that these human sequences are less well behaved in mouse tissues compared to their original native environment.

Deleted: The validation result suggested a great prediction accuracy of our model: 61% predicted active rate versus 70% observed active rate for top tier, 45% predicted active rate versus 32% observed active rate for middle tier, and 34% predicted active rate versus 35% observed active rate for bottom tier in forebrain, etc.

Formatted: Highlight

We evaluated the predictability of our matched filter model for each individual histone marks and DHS, as well as the integrated SVM model (Figure 4). For each tissue, our model ranks all the tested candidate elements with their predicted activity in this tissue using either individual feature or the integrated SVM model. Then the label of each element from experiment read out is used to assess the predictions with ROC and PR curve. One average, the integrated model trained with drosophila STARR-seq data achieves an AUROC of 0.80 and an AUPR of 0.37 for tissue-specific enhancer predictions in mouse (Figure 4A). Unlike AUROC, where the baseline is always 0.50, AUPR is more sensitive to the positive to negative ratio, with a baseline being just the positive rate. Since the positive rate from the experiment varies from 8.8% 17.6% among the tissues, the AUPR has a larger variance compared the AUROC.

2

Consistent with previous findings from STARR-seq data, when we assess each histone modification signals independently in mice, H3K27ac signal remains best performed histone marks for predicting enhancers. In addition, the DHS signal also performs well as an independent source, as it likely shares some common information with H3K27ac. The integrated model performs similar with the highest prediction feature in each tissue. This is likely due to the fact that the model is trained entirely with drosophila matched filter scores and might not be best optimized in the mammalian systems. We believe that the integrated model would achive better performance when applying our framework directly to mouse tissue STARR-seq dataset when it becomes available.

Deleted:

Deleted: is

Deleted: the single

Deleted: ,

Deleted: while

FLY2

We also did similar evaluation using the regulatory elements identified by the transduction-based FIREWACh assay in mouse embryonic stem cells (mESC) [36]. With the same metaprofiles, the predictions are based on epigenetic signals of mESC available from ENCODE website. Again, we observe similar results for individual histone marks and combined SVM model (Figure S16). As the *in vivo* and FIREWACh assays utilized a single core promoter to validate regulatory regions, the performance of the different models in Figures 4 and S16 are probably underestimated.

Deleted: The integrated model, as expected, achieves higher predictability than individual histone marks.

Deleted: then

DISC.

2Y

Validation in human cell lines

We proceeded to validate our STARR-seq based model for predicting human enhancers using a cell-based transduction assay. A third generation, self-inactivating HIV-1 based vector system in which the eGFP reporter was driven by the DNA element of interest was used to test putative enhancers after stable transduction of various cell lines, including H1 human embryonic stem cells (hESC) (Figure 5). The predicted enhancers, ranging from 650 to 2500 bp, were PCR amplified from human genomic DNA and inserted immediately upstream of a basal Oct-4 promoter of 142 bp. Each putative enhancer was tested in all four cell lines in replicates for both forward and reverse orientation. For controls experiments, VSV G-pseudotyped vector supernatants from each were prepared by co-transfection of 293T cells. These were used to transduce the same cell lines, with empty vector and FG12 vector serving as negative and positive controls respectively. Note that the empty vector did have the basal Oct-4 promoter along with the IRES-eGFP cassette. Putative enhancer activity was assessed by flow cytometric readout of eGFP expression 48-72 h post-transduction, normalized to the negative control.

Deleted: n in vitro

Deleted: validate

Deleted: just

Deleted: (a housekeeping promoter is used so that the activity of the putative enhancers should be similar across different cell lines)

Deleted: ,

Deleted: and t

Deleted: various

Deleted: ,

Formatted: Font:Arial, 11 pt, Font color: Black

A total of 25 predicted intergenic enhancers were selected for validation (Supplementary Table S6). These predictions were chosen at random to ensure that they truly represented the whole spectrum of predicted enhancers and not just the top tier of predicted enhancers. Of these 25 putative enhancers, 23 were successfully PCR-amplified and cloned into the HIV vector in both directions. To measure the distribution of gene expression in the absence of enhancer, we also amplified and cloned 25 non-repetitive elements with similar length distribution that were predicted to be inactive into the same SIN HIV vector. All positive and negative DNA elements were transduced and tested for activity in both forward and reverse orientations since enhancers are thought to function in an orientation-independent manner. Functional testing was performed in HOS, TZMBL, and A549 cells in addition to H1 hESC.

Deleted: randomly

Deleted: randomly

Deleted: se

Deleted:

Deleted: using

Deleted: strand

Deleted: line

Deleted: -

Deleted: s

Insertion of twelve of the 23 putative enhancers into the HIV vector resulted in a significant increase in eGFP expression (P-value < 0.05 over the distribution of gene expression for negative elements) in the H1 hESCs (Supplementary Table S7). While most of the positive enhancers displayed a significant increase in gene expression irrespective of their orientation, a few elements showed significantly higher levels of gene expression in one of the orientations. In contrast, the negatives displayed much lower levels of gene expression typically (Figure 5 and Supplementary Figure S17). In addition, most of these elements increased gene expression of eGFP in the four different cell lines even though some of the elements were preferentially active in one of the cell lines. Overall, 16 of the 23 tested predictions displayed a statistically significant increase in gene expression of the reporter gene in at least one of the cell lines (Supplementary Table S7 and Supplementary Figure S17). Given the promoter specificity of enhancers in such assays, we would anticipate that some of the elements that could not be validated in this particular vector would function as enhancers in a more natural biological context, ie, with the cognate promoter.

Deleted: -

Formatted: Highlight

Comparison against other computational methods

To further assess the performance of our model, we made comparisons against other published methods based on the same experimental results. We first did the comparison

Formatted: Normal (Web)

with ChromHMM[65], a well known method to ~~impute chromatin status~~ segment the genome based on chromatin features. Our integrated model outperforms ChromHMM in all four tissues, with an AUROC value of 0.74 in hindbrain (versus ChromHMM 0.69), and 0.78 in limb (versus ChromHMM 0.75), etc (Figure S18). In addition to the comparison with unsupervised segmentation based methods, we also compared with other published enhancer prediction tools, including CSIANN, a neural network based approach[66]; DELTA, an ensemble model integrating different histone modifications[67]; RFECS, a random forest model based on histone modifications[63], and REPTILE, a more recent published method that integrates histone modifications and whole genome bisulfite sequencing data[68]. We used the mouse experiment data published in REPTILE for the comparison, and we assessed the performance of our method compared to the four published methods mentioned above for all four mouse tissues with experimental data, ChIP-seq data and DNase data available. In 3 out of 4 tissues (hindbrain, limb and neural tube), our method has the highest AUROC as shown in supplementary figure SXX. In midbrain, the AUROC for our prediction is slightly lower than REPTILE and RFECS, possibly due to the data quality of the DNase experiment performed in midbrain. (The DNase experiment of mouse E11.5 stage midbrain is marked as “low SPOT score” in ENCODE, where SPOT stands for Signal Portion of Tag. We found that while 75% to 81% of the genome regions has DNase signals in the other three tissues, only 52% of the genome regions show DNase signal in the experiment in midbrain). It is also worth noting that our model is trained using the drosophila STARR-seq data whereas the other methods were trained directly with mouse data. We believe that our method would have better performance if mouse STARR-seq data could be applied for training in our framework.

In human we did not have an extensive amount of validated enhancer data. We checked the overlap of our predicted enhancers with the enhancer predictions from two popular algorithms in human cells, eg, chromHMM [65] and SegWay [27]. We observe that a majority of the predicted enhancers and promoters are also predicted to be enhancers and promoters by chromHMM and SegWay respectively (Figures S16 to S20). In addition, we compared our cell-type specific enhancer predictions with the integrative annotation of ChromHMM and Segway (provided by Hoffman. et al) using CAGE-defined enhancers from FANTOM5 Atlas. The FANTOM5 Atlas has included three human cell lines from ENCODE project with enhancer predictions from both methods: GM12878, K562 and HepG2. We found that the percentage of overlap for our predicted enhancers is more than three times higher than that of the combined ChromHMM and Segway enhancers in each of these cell lines. Despite the fact that our framework predicted a smaller number of enhancers, the exact number of overlap is still higher for our predictions. Around 40% of the CAGE-defined enhancers overlap with our predicted enhancers, while 23% to 34% overlap with the enhancers predicted by integrative ENCODE annotation method (Figure SX). We also compared the predicted enhancers from our model with their promoter annotations using FANTOM5 promoter sets. Again, the promoters predicted in our model has a higher fraction of overlap with the FANTOM promoters (Figure SXX).

Similar to the comparison in mouse tissues, we seek to compare with other published enhancer predictions in human cell lines with FANTOM5 experiment data. Since we didn’t find the promoter predictions of these methods, we compared our predicted

Formatted: Strikethrough

Formatted: Not Strikethrough

Deleted: .

Formatted: Font:11 pt

Deleted: [66][60]

Formatted: Font:(Default) -webkit-standard, (Asian) +Theme Body Asian (宋体), 11 pt, Font color: Black

Deleted: In addition to the comparison with unsupervised segmentation based algorithms, we show that our methods outperform other supervised models such as Capra, et al. [66] where the model was trained using VISTA enhancer database (see Supplement).

Formatted: Highlight

Formatted: Normal (Web), Justified

Deleted: Instead, we compared

Deleted: to

Formatted: Font:11 pt

Deleted:

Formatted: Font:11 pt

Formatted: Font:11 pt, Highlight

Formatted: Font:11 pt

Formatted: Font:-webkit-standard, 12 pt, Font color: Black

Formatted: Font:11 pt

enhancers with enhancer predictions from CSI-ANN, DEEP and RFECS, in addition to the integrative ENCODE annotation. As most of these methods don't have published enhancer predictions for GM12878 and HepG2, the comparison was done in the K562 cell line. We find that our predicted K562 enhancers has a similar fraction of overlap with FANTOM5 enhancers compared to that of CSI-ANN, but the fraction is more than twice as high as that of DEEP and RFECS (Figure SXXX). Thus, we believe that our ~~drosophila~~ STARR-seq based enhancer prediction model performs well in the mammalian systems.

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt

Deleted:

Integrative analysis in human cell-lines: Different Transcription Factors bind to enhancers and promoters

We further studied the differences in TF binding at promoters and enhancers (Figure 6 and Figure S23). We focused on the human H1-hESC cell line as there is large amount of functional genomic assays from the ENCODE [62] and Roadmap Epigenomics Mapping Consortium [61] within these cell lines. Together, the consortia have generated ChIP-Seq data for 60 transcription related factors in H1-hESC cell line, including a few chromatin remodelers and histone modification enzymes. Collectively we call all these transcription related factors "TF"s for simplicity.

We show that the patterns of TF binding within regulatory regions can be utilized in a logistic regression model to distinguish active enhancers from promoters with high accuracy (AUPR = 0.89, AUROC = 0.87) (Figure 6). We were also able to identify the most important features that distinguish promoters from enhancers. In addition to TATA-box associated factors such as TAF1, TAF7, and TBP, the RNA polymerase-II binding patterns as well as chromatin remodelers such as KDM5A and PHF8 are some of the most important factors that distinguish promoters from enhancers in H1-hESC. This provides a framework that can be utilized to identify the most important TFs associated with active enhancers and promoters in each cell-type.

We found that while most promoters and enhancers contain multiple TF binding sites, the pattern of TF binding at promoters is different from that at enhancers and that TF-binding at enhancers displays more heterogeneity: more than 70% of the promoters bind to the same set of 2-3 sequence-specific TFs, which is not observed for enhancers (Figure 6C and S24). For example, the majority of the promoters contain peaks for several TATA-associated factors (TAF1, TAF7, and TBP). These TF co-associations could lead to mechanistic insights of cooperativity between TFs. Similarly, CTCF and ZNF143 may function cooperatively as they are observed to co-occur frequently at distal regulatory regions, consistent with previous report [69]. To check if the STARR-seq based enhancer predictions have different TF binding patterns, we compared the fraction of TF occupancy of our predicted enhancers with that of RFECS, which is shown to have similar or better TF binding patterns with other methods like CSI-ANN [63]. To make the comparison, we use the same H1hESC DNase peaks, p300 ChIP-seq peaks and 3 other TFs (NANOG, OCT4, SOX2) binding sites they provided and the same 2.5kb frame distance. We show that the TF binding patterns of these two sets of predicted enhancers are very similar (Figure SX). Notably, while RFECS took p300

Deleted: T

Deleted: also

Deleted: For example,

Deleted: s

Deleted: to a previous study

Deleted: , CTCF and ZNF143 may function cooperatively as they are observed to co-occur frequently at distal regulatory regions in this study.

?

NEW PARA
INTRO
IT ...

binding regions as positive training sets, only 25% or less of predicted enhancers were within 2.5kb of any p300 binding sites, and this is consistent across different methods [63]. Overall, the high heterogeneity associated with enhancer TF-binding is consistent with the absence of a sequence code (or grammar) which can be utilized to identify active enhancers on a genome-wide fashion.

Deleted: [19]

Deleted: easily

Discussion

In this paper, we have developed a framework using transferable supervised machine learning models trained on regulatory regions identified by MPRAs to accurately predict active enhancers in a cell-type specific manner. Current, most existing methods were parameterized (not properly “trained”) on regions that had various features associated with promoters and enhancers and only a small number of these regions were typically tested for regulatory activity experimentally in an *ad hoc* manner [19, 21-30]. The rich amount of whole genome STARR-seq experiments [31] can now establish the characteristic pattern flanking active regulatory regions within certain histone modifications. This motivated us to train a shape-matching and filtering model that can be used to identify these patterns within the shape of the ChIP-seq signals. As the chromatin marks and epigenetic profiles associated with active regulatory regions are highly conserved among organisms [42-48], we showed that a well parameterized model in one model organism can be transferred to another with high prediction accuracy.

In the model, we compared close to 30 epigenetic signals for their ability to predict regulatory elements individually. The H3K27ac matched filter remains the single most important feature for predicting active regions while H3K4me1 and H3K4me3 are shown to distinguish promoters and enhancers. We characterized the amount of redundant information within the metaprofile of different epigenetic features and showed that the ChIP-seq signals of H2BK5ac, H4ac and H2A provide independent information that helps to improve the accuracy of promoter and enhancer predictions. In addition to these 30-feature models, we also provide a simple to use six-parameter SVM model for combining H3K27ac, H3K9ac, H3K4me1, H3K4me2, H3K4me3, and DHS to predict active promoters and enhancers in a cell-type specific manner. These six histone marks have been measured for a number of different tissues and cell-types by the Roadmap Epigenomics Mapping [39], the ENCODE [62], and the modENCODE Consortia [70]. Based on these signals, our model could be applied in a tissue and cell-type specific fashion in other organisms like mouse and human. We trained our models with datasets from different species and demonstrated that the high-quality STARR-seq data from *Drosophila* is sufficient to train a well transferable model. We also compared our result with chromHMM [65] and SegWay [27] predictions and observed the majority of them overlap (Figure S16 to S19).

To avoid potential biases, we chose to validate our model using multiple regulatory assays including *in vivo* transgenic assays and in vitro transductions assays, in which the predicted region is tested for regulatory activity in the native chromatin environment. The transgenic assays are performed in E11.5 mice for 133 regions of three rank tiers predicted active in mouse heart and forebrain. The experiment is supplemented by another set of 151 assayed regions predicted active in mouse hindbrain, midbrain and limb in ENCODE Phase III Encyclopedia (Moore et al., in review). Together with other validated regulatory regions from VISTA database, we were able to comprehensively validate our tissue-specific predictions in six different tissues in mouse. As we show in figure 4, the H3K27ac and DHS signals continue to be the highest predictive signals in mouse. We also did a ~~similar evaluation with publicly available FIREWACh~~ assay data [36] in mouse, and the results are consistent. Taken together, we showed that the matched filter model is transferable with high accuracy in predicting active enhancers in mouse tissues.

LINK EARLIER

The human cell-line specific regulatory elements predictions are validated through *in vitro* transduction assays in human H1-hESC cells. The majority of the predicted elements displayed a significant increase in expression of the reporter gene, further confirming the predictability of our model in mammalian organisms. ¶ H1-hESC is a highly studied cell line, allowing us to analyze the differences in the patterns of TF binding at proximal and distal regulatory regions. The TF binding and co-binding patterns at enhancers are much more heterogeneous than that at promoters. This heterogeneity in TF binding patterns makes it more difficult to predict enhancers due to the absence of obvious sequence patterns in distal regulatory regions. However, we were able to create accurate machine learning models that can distinguish proximal promoter regions from distal enhancers based on the patterns of TF ChIP-seq peaks within these regulatory regions. The conservation of the epigenetic underpinnings underlying active regulatory regions sets the stage for our method to study the evolution of tissue-specific enhancers and their genomic properties across different eukaryotic species.

RELATE TO... SLIGHTLY

Our results echo to the previous findings that the epigenetic profiles associated with active enhancers and promoters are highly conserved in evolution [42-48]. Therefore, our model of integrating shape-matching epigenetic scores using *Drosophila* STARR-seq enhancers can be applied to predict on a variety of tissues and cell lines in other species. In the cross-comparison, we show that the six-parameter integrated model trained in STARR-seq data performs equally well at predicting mouse tissue enhancers with a model trained in VISTA mouse enhancer data. This highlights the advantage of modeling based on a comprehensive genome-wide experimental assay. In the future, we expect that more extensive whole-genome STARR-seq dataset will become available on mammalian systems. It could thus be advantageous to re-train the matched filter model on the state-of-art datasets. With the set up of our framework, re-training the model with newly generated datasets should be straightforward. We envision that our framework would benefit from these datasets and generate more comprehensive regulatory element annotations across different eukaryotic species.

Implementation: source code and datasets

We have implemented our methods in Python. The source code is available at the website https://goo.gl/E8fLNN. A dockerized image is also provided for download at this site.

The datasets and output annotations referenced in the paper are available in the supplement and on the website. In particular, the transgenic mouse reporter assay result is shown in Table S4 and Table S5, and these results are also made available in VISTA Enhancer Browser. Please refer to the supplement for more details.

Acknowledgement:

M.G. were supported by the NIH grant HG009446-01. A.V. and L.A.P. were supported by NHLBI grant R24HL123879, and NHGRI grants R01HG003988, and U54HG006997, and UM1HG009421 where research was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California.

References:

1. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.
2. Ong, C.T. and V.G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression.* Nat Rev Genet, 2011. **12**(4): p. 283-93.
3. Woolfe, A., et al., *Highly conserved non-coding sequences are associated with vertebrate development.* PLoS Biol, 2005. **3**(1): p. e7.
4. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.
5. Cotney, J., et al., *The evolution of lineage-specific regulatory activities in the human embryonic limb.* Cell, 2013. **154**(1): p. 185-96.
6. Degner, J.F., et al., *DNase I sensitivity QTLs are a major determinant of human expression variation.* Nature, 2012. **482**(7385): p. 390-4.
7. Shibata, Y., et al., *Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection.* PLoS Genet, 2012. **8**(6): p. e1002789.
8. Villar, D., et al., *Enhancer evolution across 20 mammalian species.* Cell, 2015. **160**(3): p. 554-66.
9. Xiao, S., et al., *Comparative epigenomic annotation of regulatory DNA.* Cell, 2012. **149**(6): p. 1381-92.
10. Wray, G.A., *The evolutionary significance of cis-regulatory mutations.* Nat Rev Genet, 2007. **8**(3): p. 206-16.
11. Corradin, O. and P.C. Scacheri, *Enhancer variants: evaluating functions in common disease.* Genome Med, 2014. **6**(10): p. 85.
12. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.* Am J Hum Genet, 2014. **95**(5): p. 535-52.
13. Slattery, M., et al., *Absence of a simple code: how transcription factors read the genome.* Trends Biochem Sci, 2014. **39**(9): p. 381-99.
14. Levo, M., et al., *Unraveling determinants of transcription factor binding outside the core binding site.* Genome Res, 2015. **25**(7): p. 1018-29.
15. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nat Rev Genet, 2013. **14**(4): p. 288-95.
16. Erwin, G.D., et al., *Integrating diverse datasets improves developmental enhancer prediction.* PLoS Comput Biol, 2014. **10**(6): p. e1003677.
17. Pennacchio, L.A., et al., *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502.
18. Nord, A.S., et al., *Rapid and pervasive changes in genome-wide enhancer usage during mammalian development.* Cell, 2013. **155**(7): p. 1521-31.
19. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.
20. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-61.
21. Narlikar, L., et al., *Genome-wide discovery of human heart enhancers.* Genome Res, 2010. **20**(3): p. 381-92.

22. Visel, A., et al., *Ultraconservation identifies a small subset of extremely constrained developmental enhancers.* Nat Genet, 2008. **40**(2): p. 158-60.
23. Bonn, S., et al., *Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.* Nat Genet, 2012. **44**(2): p. 148-56.
24. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.
25. Ghandi, M., et al., *Enhanced regulatory sequence prediction using gapped k-mer features.* PLoS Comput Biol, 2014. **10**(7): p. e1003711.
26. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.
27. Hoffman, M.M., et al., *Unsupervised pattern discovery in human chromatin structure through genomic segmentation.* Nat Methods, 2012. **9**(5): p. 473-6.
28. Kharchenko, P.V., et al., *Comprehensive analysis of the chromatin landscape in Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 480-5.
29. He, H.H., et al., *Nucleosome dynamics define transcriptional enhancers.* Nat Genet, 2010. **42**(4): p. 343-7.
30. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
31. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
32. Dickel, D.E., et al., *Function-based identification of mammalian enhancers using site-specific integration.* Nat Methods, 2014. **11**(5): p. 566-71.
33. Gisselbrecht, S.S., et al., *Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos.* Nat Methods, 2013. **10**(8): p. 774-80.
34. Kwasnieski, J.C., et al., *High-throughput functional testing of ENCODE segmentation predictions.* Genome Res, 2014. **24**(10): p. 1595-602.
35. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.
36. Murtha, M., et al., *FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.* Nat Methods, 2014. **11**(5): p. 559-65.
37. Patwardhan, R.P., et al., *Massively parallel functional dissection of mammalian enhancers in vivo.* Nat Biotechnol, 2012. **30**(3): p. 265-70.
38. Yanez-Cuna, J.O., et al., *Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features.* Genome Res, 2014. **24**(7): p. 1147-56.
39. Shlyueva, D., G. Stampfel, and A. Stark, *Transcriptional enhancers: from properties to genome-wide predictions.* Nat Rev Genet, 2014. **15**(4): p. 272-86.
40. Maston, G.A., et al., *Characterization of enhancer function from genome-wide analyses.* Annu Rev Genomics Hum Genet, 2012. **13**: p. 29-57.
41. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.

42. Yue, F., et al., *A comparative encyclopedia of DNA elements in the mouse genome.* Nature, 2014. **515**(7527): p. 355-64.
43. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.
44. Dong, X., et al., *Modeling gene expression using chromatin features in various cellular contexts.* Genome Biol, 2012. **13**(9): p. R53.
45. Cheng, C. and M. Gerstein, *Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.* Nucleic Acids Research, 2012. **40**(2): p. 553-568.
46. Cheng, Y., et al., *Principles of regulatory information conservation between mouse and human.* Nature, 2014. **515**(7527): p. 371-+.
47. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.
48. Gjoneska, E., et al., *Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease.* Nature, 2015. **518**(7539): p. 365-9.
49. Ablikim, M., et al., *Analysis of D+ -> (K)over-bar(0)e(+)nu(e) and D+ -> pi(0)e(+)nu(e) semileptonic decays.* Physical Review D, 2017. **96**(1).
50. Kundaje, A., et al., *Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements.* Genome Res, 2012. **22**(9): p. 1735-47.
51. Kumar, V.B.V.K., A. Mahalanobis, and R.D. Juday, *Correlation Pattern Recognition*. 2005.
52. Davis, J. and M. Goadrich, *The Relationship Between Precision-Recall and ROC Curves.* Proceedings of the 23rd international conference on Machine Learning, 2006: p. 233-240.
53. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
54. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.
55. Butler, J.E. and J.T. Kadonaga, *Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs.* Genes Dev, 2001. **15**(19): p. 2515-9.
56. Li, X. and M. Noll, *Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo.* EMBO J, 1994. **13**(2): p. 400-6.
57. Merli, C., et al., *Promoter specificity mediates the independent regulation of neighboring genes.* Genes Dev, 1996. **10**(10): p. 1260-70.
58. Ohtsuki, S., M. Levine, and H.N. Cai, *Different core promoters possess distinct regulatory activities in the Drosophila embryo.* Genes Dev, 1998. **12**(4): p. 547-56.
59. Zabidi, M.A., et al., *Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation.* Nature, 2015. **518**(7540): p. 556-9.
60. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121--167.
61. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

62. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
63. Rajagopal, N., et al., *RFECS: a random-forest based algorithm for enhancer identification from chromatin state.* PLoS Comput Biol, 2013. **9**(3): p. e1002968.
64. Koch, C.M., et al., *The landscape of histone modifications across 1% of the human genome in five human cell lines.* Genome Res, 2007. **17**(6): p. 691-707.
65. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Methods, 2012. **9**(3): p. 215-6.
66. Firpi, H.A., D. Ucar, and K. Tan, *Discover regulatory DNA elements using chromatin signatures and artificial neural network.* Bioinformatics, 2010. **26**(13): p. 1579-1586.
67. Lu, Y.M., et al., *DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications.* Plos One, 2015. **10**(6).
68. He, Y.P., et al., *Improved regulatory element prediction based on tissue-specific local epigenomic signatures.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(9): p. E1633-E1640.
69. Bailey, S.D., et al., *ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.* Nat Commun, 2015. **2**: p. 6186.
70. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97.

Figures and Captions

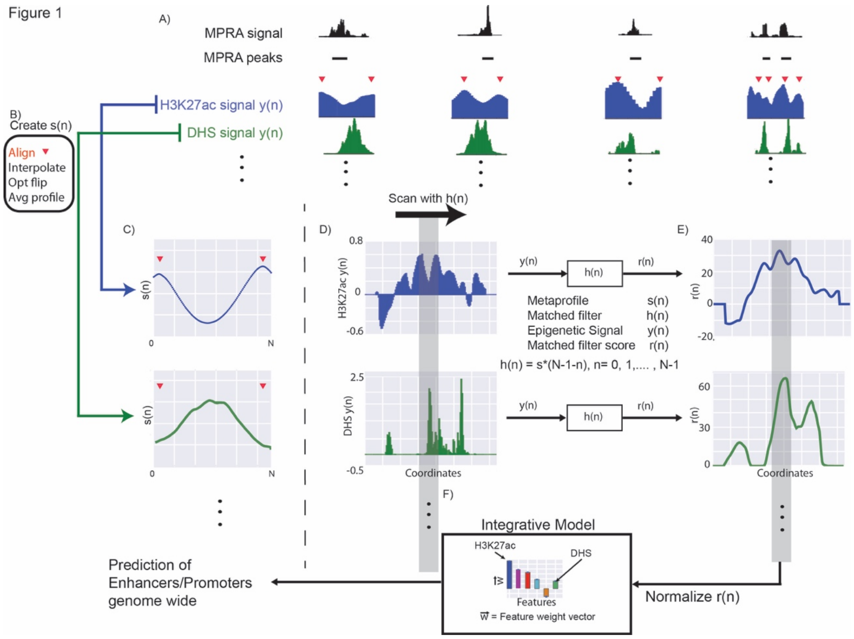
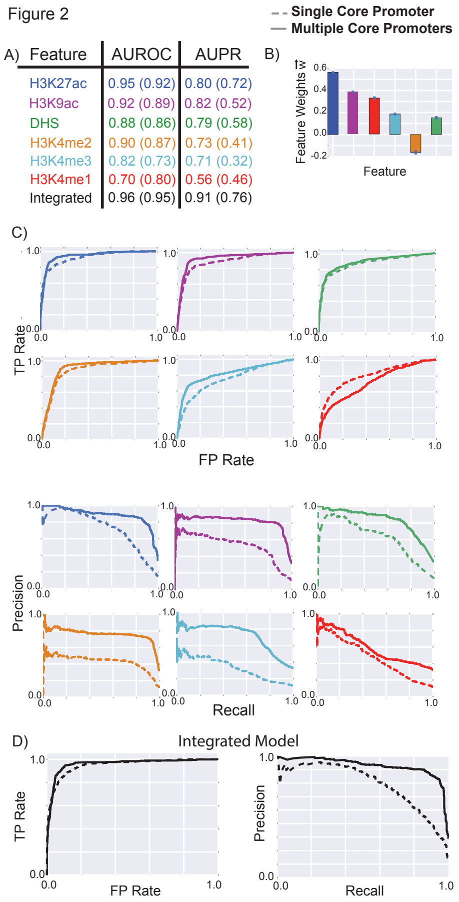


Figure 1: Creation of metaprofile. A) We identified the “double peak” pattern in the H3K27ac signal close to STARR-seq peaks. The red triangles denote the position of the two maxima in the double peak. B) We aggregated the H3K27ac signal around these regions after aligning the flanking maxima, using interpolation and smoothing on the H3K27ac signal, and averaged the signal across different MPRA peaks to create the metaprofile in C). The exact same operations can be performed on other histone signals and DHS to create metaprofiles in other dependent epigenetic signals. D) Matched filters can be used to scan the histone and/or DHS datasets to identify the occurrence of the corresponding pattern in the genome. E) The matched filter scores are high in regions where the profile occurs (grey region shows an example) and it is low when only noise is present in the data. The individual matched filter scores from different epigenetic

datasets can be combined using integrated model in F) to predict active promoters and enhancers in a genome wide fashion.

Figure 2



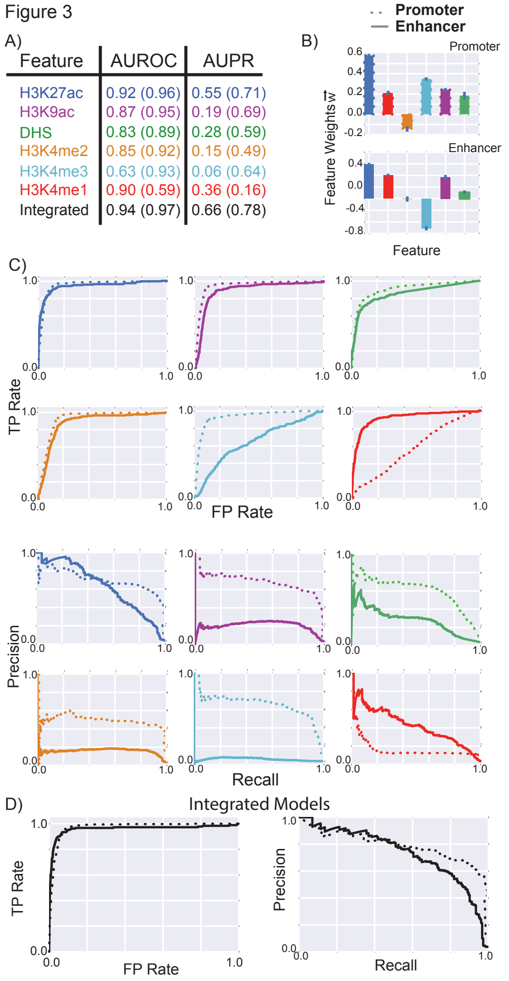
A)

Feature	AUROC	AUPR
H3K27ac	0.95 (0.92)	0.80 (0.72)
H3K9ac	0.92 (0.89)	0.82 (0.52)
DHS	0.88 (0.86)	0.79 (0.58)
H3K4me2	0.90 (0.87)	0.73 (0.41)
H3K4me3	0.82 (0.73)	0.71 (0.32)
H3K4me1	0.70 (0.80)	0.56 (0.46)
Integrated	0.96 (0.95)	0.91 (0.76)

Figure 2: Performance of matched filters and integrated models for predicting MPRA peaks. The performance of the matched filters of different epigenetic marks and the integrated model for predicting all STARR-seq peaks is compared here using 10-fold cross validation. A) The area under the receiver-operating characteristic (AUROC) and the precision-recall (AUPR) curves are used to measure the accuracy of different matched filters and the integrated model. B) The weights of the different features in the integrated model are shown and these weights may be used as a proxy for the importance of each feature in the integrated model. C) The individual ROC and PR curves for each matched filter and the integrated model are shown. The performance of these features and the integrated model for predicting the STARR-seq peaks using multiple core promoters and single core promoter are compared. The numbers within the parentheses in A) refer to the AUROC and AUPR for predicting the peaks using a single

STARR-seq core promoter while the numbers outside the parentheses refers to the performance of the model for predicting peaks from multiple core promoters.

Figure 3



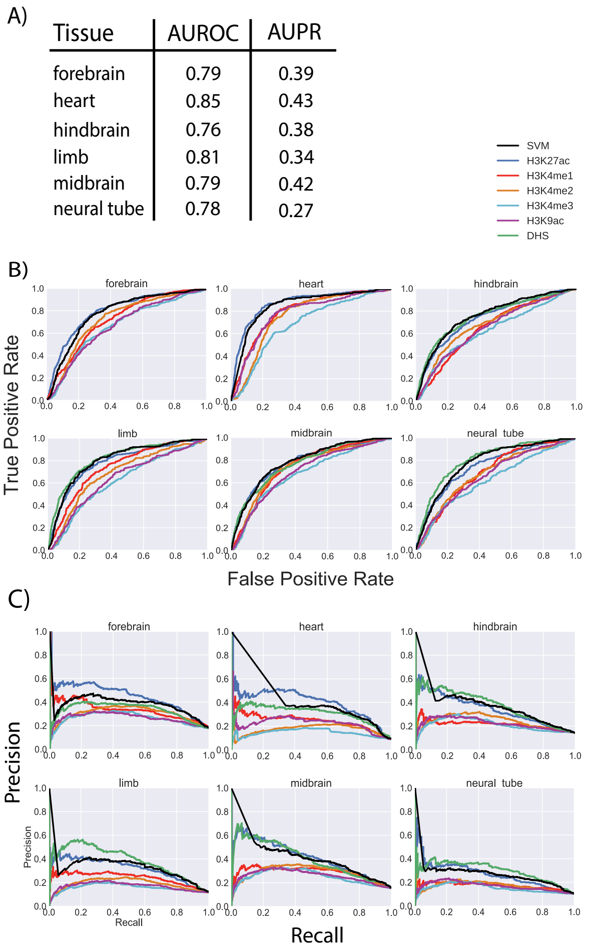
A)

Feature	AUROC	AUPR
H3K27ac	0.92 (0.96)	0.55 (0.71)
H3K9ac	0.87 (0.95)	0.19 (0.69)
DHS	0.83 (0.89)	0.28 (0.59)
H3K4me2	0.85 (0.92)	0.15 (0.49)
H3K4me3	0.63 (0.93)	0.06 (0.64)
H3K4me1	0.90 (0.59)	0.36 (0.16)
Integrated	0.94 (0.97)	0.66 (0.78)

Figure 3: Performance of matched filters and integrated models for predicting promoters and enhancers. The performance of the matched filters of different epigenetic marks and the integrated model for predicting active promoters and enhancers are compared here using 10-fold cross validation. A) The numbers within parentheses refer to the AUROC and AUPR for predicting promoters while the numbers outside parentheses refer the performance of the models for predicting enhancers. B) The weights of the different features in the integrated models for promoter and enhancer prediction are shown. C) The individual ROC and PR curves for each matched filter and the integrated model are shown. The performance of these features and the integrated

model for predicting the active promoters and enhancers using multiple core promoters are compared.

Figure 4



A)

Tissue	AUROC	AUPR
forebrain	0.79	0.39
heart	0.85	0.43
hindbrain	0.76	0.38
limb	0.81	0.34
midbrain	0.79	0.42
neural tube	0.78	0.27

Figure 4: Performance of matched filters and integrated model for predicting active enhancers in mice. The performance of the *Drosophila* STARR-seq based matched filters and the integrated model for predicting active enhancers identified by transgenic mouse enhancer assays in 6 different tissues of E11.5 mice. A) The AUROC and AUPR for the integrated SVM model in 6 tissues. The weights of the different features in the integrated model is the same as the weights shown in Figure 3 for enhancers. B) The individual ROC curves of each feature and the integrated SVM model for each tissue. C) The individual PR curves of each feature and the integrated SVM model for each tissue.

Deleted: Conservation of epigenetic features.

Deleted: at

Deleted: in

Deleted: Average

Deleted: predicting enhancers by different features and by the

Deleted: and the integrated

Deleted: are shown

Deleted: are shown

Figure 5

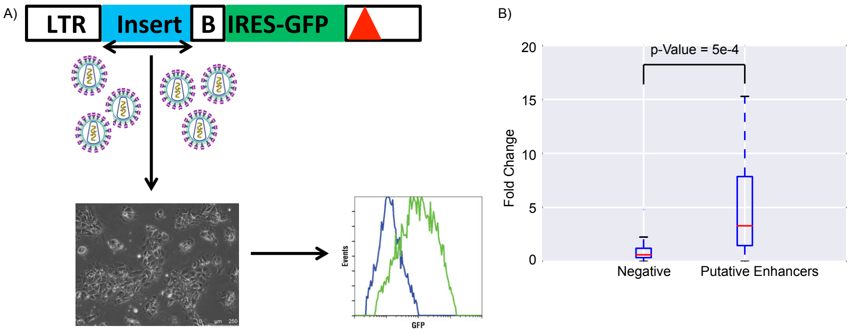


Figure 5: Enhancer Validation Experiments. A) Schematic of the enhancer validation experiment flow. At top is the third generation HIV-based self-inactivating vector (deletion in 3’ LTR indicated by red triangle), with PCR-amplified test DNA (blue, two-headed arrow indicates fragment cloned in both orientations) inserted at 5’ of a basal (B) Oct4 promoter driving IRES-eGFP (green). Vector supernatant was prepared by plasmid co-transfection of 293T cells. Targeted cells are tranduced and then analyzed by flow cytometry a few days later. Shown below is the expected post-transduction structure of the SIN HIV vector, with a duplication of the 3’ LTR deletion rendering both LTRs non-functional B) Fold change of gene expression of eGFP is compared between negative elements and putative enhancers chosen at random, with p-value measured by Wilcoxon signed-rank test.

Deleted: A s

Deleted: scheme is show

Deleted: was

Deleted: ,

Deleted: just

Deleted:

Deleted: and used to transduce cellular targets and

Deleted: The f

Deleted: for experiments

Deleted: . The

Deleted: V

Deleted: of the difference in activity is

Deleted: using a

Deleted:

Figure 6

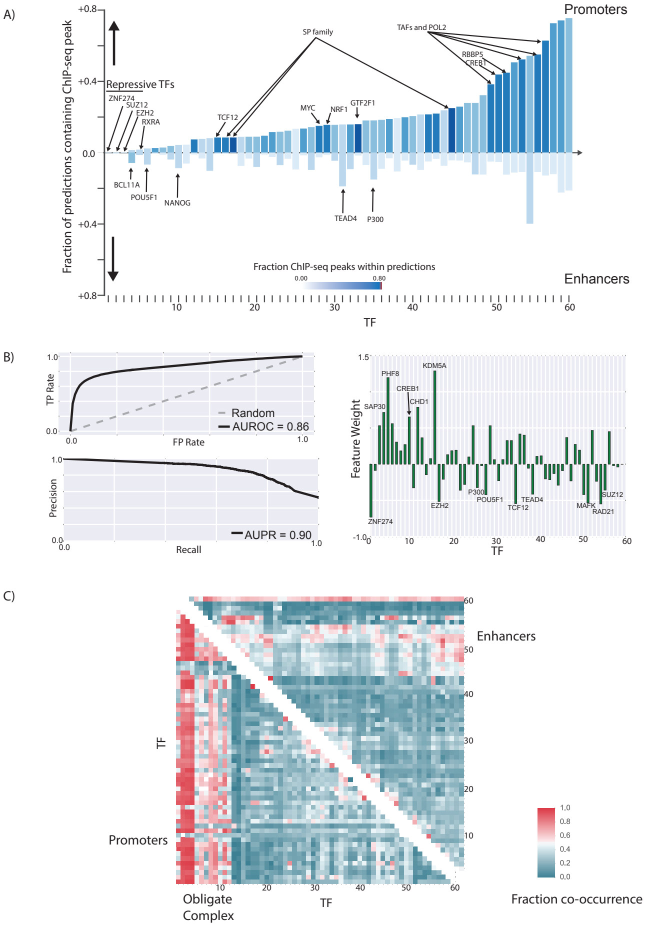


Figure 6: Differences in TF binding patterns at enhancers and promoters. A) The fraction of predicted promoters and enhancers that overlap with ENCODE ChIP-seq peaks for different TFs in H1-hESC are shown. The names of all TFs in the figure can be viewed in Figure S20. B) The AUROC and AUPR for a logistic regression model created using the pattern of TF binding at each regulatory region to distinguish enhancers from promoters are shown. The weight of each feature in the logistic regression model can be used to identify the most important TFs that distinguish enhancers from promoters. C) The patterns of TF co-binding at active promoters and enhancers are shown. The names of all the TFs in this graph can be viewed in Figure S21.