Using pattern recognition of epigenetic signals for supervised enhancer prediction

Anurag Sethi[1,2], Mengting Gu[1], Emrah Gumusgoz[6], Landon Chan[3], Koon-Kiu Yan[1,2], Kevin Yip[4], Joel Rozowsky[1,2], Richard Sutton[6], and Mark Gerstein[1,2,5]

[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 Hong Kong, China

[4] Department of Computer Science, The Chinese University 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

Abstract

Enhancers are important noncoding elements. Unfortunately, until recently, they were difficult to characterize experimentally, and only a few mammalian enhancers were validated, making it difficult to properly train statistical models for their identification. Instead, postulated patterns of genomic features were used heuristically for identification. Recently, a large number of massively parallel assays for characterizing enhancers have been developed. Here, we use them to create shape-matching filters based on enhancer-associated metaprofiles in epigenetic features. We then combine different features with simple, linear models and predict enhancers in a supervised fashion. By cross-validating and testing our models, we show that they can be transferred without re-parameterization between cell lines and even between organisms. Finally, we predict enhancers in cell lines with many transcription-factor binding sites and validate these enhancers experimentally. In turn, this highlights distinct differences between the type of binding at enhancers and promoters, enabling the construction of a secondary model discriminating between these two.

Significance Statement

Enhancers are import regulatory elements in the genome. The distance between the enhancer and its regulating genes varies between several kilobytes to megabytes, making it hard annotate enhancer region both experimentally and computationally. Here we demonstrate that by integrating epigenetic features with supervised machine learning models, we can achieve high accuracy of enhancer prediction. The match filter tool providing a general framework to identify enhancers across cell lines.

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].

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 of the widest 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 certain post-translational modifications on histone proteins [27, 30]. These predictions were not rigorously assessed as very few putative enhancers could be validated experimentally and it remains challenging to assess the performance of different methods for enhancer prediction.

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 contains 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 near the core promoter 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 fly genome in several cell-types [31, 38] and was used to identify thousands of cell-type specific enhancers and promoters. 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 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 enhancer [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.

We developed a new supervised machine-learning method that was trained and tested on large number of experimentally active regulatory regions identified in MPRAs to accurately predict active enhancers and promoters in a cell-type specific manner. 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 the signal-to-noise ratio and identified the shape associated with active regulatory regions. The epigenetic signal shapes associated with promoters and enhancers are conserved across millions of years of evolution and these models can be used to predict enhancers and promoters in different cell-types and tissues and across diverse eukaryotic species. We further created simple to use transferrable statistical models with six parameters that can be used to predict enhancers and promoters in several eukaryotic species including fly, mouse, and human. We applied these models to predict active enhancers and promoters in the H1-human embryonic stem cell (H1-hESC), a highly studied human cell-line in the ENCODE datasets. These analyses show that the pattern of transcription factor (TF) binding and co-binding varies between enhancers and promoters. The pattern of TF and co-TF binding at active enhancers is much more heterogeneous than the corresponding patterns on promoters. The pattern of TF binding 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 and then utilizes further functional genomics datasets to identify key TFs associated with active regulatory regions within these cell-types.

Results

Aggregation of epigenetic signal to create metaprofile:

We developed a framework to predict activating regulatory elements utilizing 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. These metaprofiles were then utilized in a pattern recognition algorithm for predicting active promoters and enhancers in a cell-type specific manner.

These metaprofiles were initially created using the histone modification H3K27ac at active STARR-seq peaks (see Figure 1 and Methods) identified in the S2 cell-line of fly. 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). Even though 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 [42]. For the first time, 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).

Occurrence of 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 [43]. 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. The 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 as compared to the other. On these imbalanced data sets, PR curves are useful alternative to ROC curves as the precision is directly related to the false detection ratio at different thresholds. The PR curve highlights differences in performance of different models even when their ROC curves remain comparable [44]. 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 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.

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, 45, 46]. In general, several histone acetylation (H3K27ac, H3K9ac, H4K12ac, H2BK5ac, H4K8ac, H4K5ac, H3K18ac) marks as well as the H1, H3K4me2, and DHS matched filters are the most accurate marks (see Figure 2 and

Table S1) because the matched filter scores for these regions on these marks are higher for STARR-seq peaks (Figure S3). 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 and 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 enrichment of signal alone as they outperform the histone peaks 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 [47-51]. 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 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).

Machine learning can combine matched filter scores from different epigenetic features:

We combined the normalized matched filter scores (see Methods) from six different epigenetic marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and DHS) associated with active regulatory regions by the Roadmap Epigenomics Mapping [52] and the ENCODE [53] Consortia using a linear SVM [54] and the integrated model achieved a higher accuracy than the individual matched filter scores (Figure 2). We also assessed the performance of other statistical approaches for combining the features (including non-linear models) in Figure S6 and all these models performed similarly. By using only six features, we ensure that our model is capable of being applied to many cell-lines and tissues on which the relevant experiments have been performed. These models are trained to learn the patterns in the matched filter scores for different epigenetic marks within experimentally verified regulatory regions and we chose these marks as we wanted to assess the applicability of these machine learning models to predict active enhancers and promoters across different cell-types and species. As expected, the integrated models outperformed the individual matched filter scores, as they are able to leverage information from multiple epigenetic marks. In addition, the six-parameter integrated model displayed higher accuracy after combining the peaks identified using different core promoters. In the integrated model, the normalized matched filter score for each epigenetic feature in a particular region is scaled by its optimized weight and added together to form the 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 regions from non-regulatory regions in such models. They 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 from inactive regions. While the DHS matched filter performed well as an individual feature (AUPR in Figure 2), the information in DHS is redundant with the information in the histone marks as indicated by the fact that it has the lowest weight among the six features in the integrated model. We compared several other machine learning algorithms including nonlinear SVM (results not shown) to combine the machine learning models and found that they all displayed nearly similar accuracy and similar features were more important across these different models (Figure S5).

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 S6) and these model 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 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 (Figure S6).

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 fly), the enhancers are regulatory elements that are not close to any known TSS even though a few of the promoters may actually function as enhancers. 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 (Figure 3). In addition, the DHS, H3K9ac, and H3K4me2 matched filters also performed reasonably for promoter and enhancer prediction. Similar to previous studies [55, 56], 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 show that H3K4me1 matched filter score is higher near enhancers while the H3K4me3 matched filter score tends to be higher near promoters (Figure S7). 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 two different integrated models to learn the combination of features associated with promoters and enhancers. These integrated models outperformed the individual matched filters at predicting active enhancers and promoters (Figures 3 and S8). 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 for 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 S9). The weights of different features indicate that H2BK5ac again displays the most independent information for accurately predicting active enhancers and promoters (Figures S9). We observe similar trends and accuracy with several different machine learning models (Figures S8 and S9).

The epigenetic underpinnings of active regulatory regions are highly conserved in evolution:

In order to assess the transferability of these metaprofiles and machine learning models for predicting regulatory regions in other tissues and cell-types, we assessed the accuracy of these models for predicting regulatory elements identified using the transduction-based FIREWACh assay in mouse embryonic stem cells (mESC) [36]. The metaprofiles for individual histone marks learned using active promoters and enhancers identified with the STARR-seq assay in the S2 cell-line were used with matched filters to predict the regulatory activity of different regions in mESC based on the epigenetic signals in mESC (Figure 4). The matched filters for individual histone marks displayed similar accuracy for predicting enhancers and promoters in mESC as in the original S2 cell-line. In addition, the 6-parameter SVM models learned using STARR-seq data in S2 cell-line were also highly accurate at predicting active enhancers and promoters in mouse (Figure 4).

This indicates that the epigenetic profiles associated with active enhancers and promoters are conserved over 600 million years of evolution underscoring the importance of such epigenetic modifications in maintaining the regulatory role of enhancers and promoters across different cell-types and species. As these regulatory regions were identified using a single core promoter in FIREWACh, the performance of the different models in Figure 4 is probably underestimated. The accuracy of these models enables us to use the metaprofiles and statistical models learned using STARR-seq data in fly to predict enhancers in different cell-lines and eukaryotic species. Consistent with this, the metaprofile and machine learning models learned using STARR-seq experiment in BG3 cell-line (fly) can be utilized to predict active promoters and enhancers in the S2 cell-line (Figure S12).

Validation of Enhancer Prediction Models

The ENCODE consortium has 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 utilized the 6 parameter integrated model to predict active enhancers and promoters in the hESC cell-line based on the epigenetic datasets measured by the ENCODE consortium. This provides us with a system to validate our enhancer prediction model as well as to study the patterns of TF binding within enhancers and promoters. Using these models, 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). By comparing the matched filter predicted enhancers and promoters with chromatin states predicted by chromHMM [30] 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 S15 to S18).

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 validate putative enhancers after stable transduction of various cell lines, including H1 hESC (Figure 5). The predicted enhancers, ranging from 650 to 2500 bp, were PCR amplified from human genomic DNA and inserted just upstream of a basal Oct-4 promoter of 142 bp (a housekeeping promoter is used so that the activity of the putative enhancers should be similar across different cell lines). VSV G-pseudotyped vector supernatants from each were prepared by co-transfection of 293T cells, and these were used to transduce the various cell lines, with empty vector and FG12 vector serving as negative and positive controls, respectively. Putative enhancer activity was assessed by flow cytometric readout of eGFP expression 48-72 h post-transduction, normalized to the negative control.

A total of 25 predicted intergenic enhancers were randomly selected for validation (Supplementary Table S3). These predictions were chosen randomly to ensure that these 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 amplified and cloned into the HIV vector. 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 using the same HIV vector. All positive and negative DNA elements were transduced and tested for activity in both forward and reverse strand orientations since enhancers are thought to function in an orientation-independent manner. Functional testing was performed in HOS, TZMBL, and A549 cell lines in addition to H1-hESCs.

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 distribution of gene expression for negative elements) in the H1-hESCs (Supplementary Table S3). While most of the positive enhancers displayed a significant increase in gene expression irrespective of their orientation during orientation, a few elements showed significantly higher levels of gene expression in one of the orientations (Supplementary Table S4). In contrast, the negatives displayed much lower levels of gene expression typically (Figure 5 and Supplementary Figure S19). In addition, most of these elements increased gene expression of GFP 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 statistically significant increase in gene expression of the reporter gene in at least one of the cell lines (Supplementary Table S3 and Supplementary Figure S19). 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.

Different Transcription Factors bind to enhancers and promoters

We further studied the differences in TF binding at promoters and enhancers (Figure 6 and Figure S20). Most promoters and enhancers contain multiple TF-binding sites. However, the TF-binding of enhancers is more heterogeneous than promoters: in particular, more than 70% of the promoters bind to the same set of 2-3 sequence-specific TFs, which is not observed for enhancers. The majority of the promoters also contain peaks for several TATA-associated factors (TAF1, TAF7, and TBP). 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 easily identify active enhancers on a genome-wide fashion.

In Figure 6, 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). 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 KDM4A 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.

In Figure 6A, we show that the pattern of TF binding at promoters is different from that at enhancers and TF-binding at enhancers displaying more heterogeneity. As the set of TFs binding promoters is fairly uniform, the same pairs of TF also tend to bind together on promoters. In contrast, for enhancers, the patterns of TF co-binding is much more heterogeneous and different enhancers tend to contain different TF-pairs. This can be observed in the patterns of TF co-binding in Figures 6C and S21. These TF co-associations could lead to mechanistic insights of cooperativity between TFs. For example, similar to a previous study [57], CTCF and ZNF143 may function cooperatively as they are observed to co-occur frequently at distal regulatory regions in this study.

Discussion

Our ability to accurately predict active enhancers in a cell-type specific manner using transferable supervised machine learning models that were trained based on regulatory regions identified using new NGS-enabled MPRAs distinguishes our method from previous enhancer prediction methods. Currently, most existing methods were parameterized (not properly “trained”) with 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. The MPRAs were able to firmly establish that certain histone modifications occur on nucleosomes flanking active regulatory regions leading to the formation characteristic double peak pattern within the ChIP-signal [39]. This motivated us to create matched filter models that were able to identify these patterns within the shape of the ChIP-signal in the presence of stochastic noise with the highest signal to noise ratio. Furthermore, we were able to combine the matched filter scores from different epigenetic features using simple transferrable linear SVM models and learned the most informative epigenetic features for regulatory region predictions.

The sensitivity and selectivity of various MPRAs is currently a matter of debate. A majority of these MPRAs test the regulatory activity of different regions by assessing their ability to induce gene expression in a plasmid after transfecting it into a cell-type of interest [31]. Such assays may not recapitulate the native chromatin environment found in chromosomes, which may be necessary for assessing whether the regulatory region is active in its genomic environment.

Here, we show for the first time, that the patterns in the epigenetic signals associated with active enhancers identified using a transfection-based assay (STARR-seq) can be utilized to predict the activity of enhancers in a transduction-based assay (FIREWACh). During the FIREWACh assay, random nucleosome-free regions in mESC were captured and assayed for regulatory activity of the GFP gene by utilizing a lentiviral plasmid vector and inserted (or transduced) these vectors into the chromosome in mESC cells. As the FIREWACh assay tests the regulatory activity of enhancers after transduction, we assume that these regions were tested in their native chromatin environment and transduction-based assays form a more stringent test for regulatory activity. However, due to the shorter length of the tested region (< 300 bp) and the single core promoter used in the FIREWACh assay, we think that the accuracy of the statistical models in Figure 4 is underestimated.

We were able to assess the accuracy of different epigenetic metaprofiles for predicting regulatory activity using our statistical models. While different acetylation modifications are associated with active regions of the genome, we were able to compare close to 30 histone marks for enhancer and promoter predictions. The H3K27ac matched filter remains the single most important feature for predicting active regulatory regions while H3K4me1 and H3K4me3 are known to distinguish promoters from enhancers. However, our analysis characterizes the amount of redundancy in information within the metaprofile of different epigenetic features for predicting active regulatory regions and shows that ChIP-experiments of H2BK5ac, H4ac, and H2A variants could also produce independent information that can 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. We also showed that the metaprofiles and the combination of epigenetic marks associated with active regulatory regions are highly conserved in evolution making these models highly transferable. These six histone marks have been measured for a number of different tissues and cell-types by the Roadmap Epigenomics Mapping Consortium [39], the ENCODE [53], and the modENCODE Consortium [58]. The enhancers predicted using our machine learning models were experimentally validated in human cell lines.

One aspect that is discussed less frequently is the effect of core promoter on enhancer and promoter prediction. MPRAs show that the regulatory activity of enhancers and promoters in a regulatory assay depends on the core promoter used during the experiment [51]. As the transcription factors that bind to each regulatory region are thought to play a key role in core-promoter specificity [47, 51], we suspect that machine learning models that contain sequence or motif-based features may be biased towards certain transcription factor binding sites when trained with regulatory regions identified using a single-core promoter. To avoid such biases, it would be more appropriate to train models with sequence-based features when the validation experiments are performed with multiple core promoters. In the absence of validation data with multiple core promoters, it may be more suitable to train models using epigenetic features as such models contain no sequence-based information. In comparing the predictions from such models with experiments using a single core promoter, some of the strongest predictions may be mislabeled as negatives even though they contain some regulatory activity leading to a lower accuracy estimate as shown in Figure 2.

As the epigenetic profiles and statistical models learned in this study are transferable across different cell-lines and species, we are able to apply these models to predict active enhancers and promoters in different cell-types. We applied these models to predict enhancers and promoters in H1-hESC, a highly studied ENCODE cell-line. This allowed 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 is much more heterogeneous than that at promoters. We think that this heterogeneity in TF binding patterns makes it much more difficult to predict enhancers due to the absence of obvious sequence patterns in distal regulatory regions. However, we were also able to create highly accurate machine learning models that are able to 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.

Figure 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: 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: 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: Conservation of epigenetic features. The performance of the fly-based matched filters and the integrated model for predicting active promoters and enhancers in mouse embryonic stem cells identified using FIREWACh. A Similar to Figure 3, 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 identified using FIREWACh are shown.

Figure 5: Enhancer Validation Experiments. A) A schematic of the enhancer validation scheme is show**.** At top is 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 was cloned in both orientations), inserted just 5' of a basal (B) Oct4 promoter driving IRES-eGFP (green). Vector supernatant was prepared by plasmid co-transfection of 293T cells and used to transduce cellular targets and analyzed by flow cytometry a few days later. B) The fold change of gene expression of eGFP is compared between negative elements and putative enhancers chosen for experiments. The p-Value of the difference in activity is measured using a Wilcoxon signed-rank test.

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 S19. 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 S20.

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. Kundaje, A., et al., *Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements.* Genome Res, 2012. **22**(9): p. 1735-47.
43. Kumar, V.B.V.K., A. Mahalanobis, and R.D. Juday, *Correlation Pattern Recognition*. 2005.
44. 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.
45. 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.
46. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.
47. 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.
48. 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.
49. Merli, C., et al., *Promoter specificity mediates the independent regulation of neighboring genes.* Genes Dev, 1996. **10**(10): p. 1260-70.
50. 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.
51. Zabidi, M.A., et al., *Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation.* Nature, 2015. **518**(7540): p. 556-9.
52. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
53. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
54. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121--167.
55. Rajagopal, N., et al., *RFECS: a random-forest based algorithm for enhancer identification from chromatin state.* PLoS Comput Biol, 2013. **9**(3): p. e1002968.
56. 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.
57. Bailey, S.D., et al., *ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.* Nat Commun, 2015. **2**: p. 6186.
58. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97.