Enhancer prediction using pattern recognition of epigenetic signals -> MPRA training

Abstract

Enhancers are an important category of tissue-specific noncoding functional elements, whose activity is often associated with changes in gene expression across different tissues. Unfortunately, until recently, enhancers were difficult to characterize experimentally and only a small number of tissue-specific mammalian enhancers were rigorously validated. Hence, for predicting enhancers, it was difficult to properly train statistical models based on experimentally validated enhancers. Instead, in more heuristic fashion, the presence of genomic features associated with enhancers was used to predict them. 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 and their activity is often correlated with an enrichment of certain post-translational modifications on histone proteins. Recently, a large number of massively parallel assays for characterizing enhancers were developed to identify thousands of cell-type specific enhancers. We use the output of these assays to train and test a statistical model for predicting enhancers. The MPRAs have established that a characteristic double peak pattern in histone marks is associated with active enhancers. This motivated us to develop linear filters to identify the occurrence of promoter and enhancer-associated patterns in different epigenetic signals. We then combine these filters using simple linear models, which allow us to predict enhancers in a cell-type specific manner, and we show that our models can be transferred without change between various cell lines and even between different organisms. The conservation of these enhancer-associated features allows us to characterize them on a large scale across many tissues and cell lines. Our model also allows us to characterize enhancers in cell lines with many experimentally measured transcription factor binding sites, and this, in turn, highlights distinct differences between the type of transcription factor binding at enhancers and promoters, enabling the construction of a secondary model that better discriminates between these two active regulatory regions.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: within signal

Anurag Sethi 5/9/2016 12:07 PM
Deleted: datasets

Anurag Sethi 5/9/2016 12:07 PM
Deleted: Now, there are

Anurag Sethi 5/9/2016 12:07 PM
Deleted: We initially

Anurag Sethi 5/9/2016 12:07 PM
Deleted:

Anurag Sethi 5/9/2016 12:07 PM
Deleted: with few parameters. This characterizes

Anurag Sethi 5/9/2016 12:07 PM
Deleted: model

Anurag Sethi 5/9/2016 12:07 PM
Deleted: This

Anurag Sethi 5/9/2016 12:07 PM
Deleted: enhancers

Anurag Sethi 5/9/2016 12:07 PM
Formatted: Font color: Custom Color(RGB(33,33,33)), Pattern: Clear (White)

Anurag Sethi 5/9/2016 12:07 PM
Deleted:

Introduction

Enhancers are gene regulatory elements that activate expression of target genes from a distance \cite{}. 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 nearly identical genome \cite{}. Moreover, changes in the sequences of regulatory elements are thought to play a significant role in the evolution of species \cite{}. 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 \cite{}. 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 \cite{}.

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 \cite{}. In addition to the small numbers, the validated enhancers were typically biased towards conserved noncoding regions \cite{} with particular patterns of chromatin or transcription factor binding \cite{} making these validated enhancers inappropriate for training supervised machine learning models of enhancers. 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 trained based on certain heuristic features associated with enhancers, which were then utilized to predict enhancers. A small number of the predicted enhancers were then validated experimentally to test the accuracy of these predictions. However, as very few enhancers had been experimentally validated, 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 \cite{}. 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 them near the core promoter as differences in the gene's expression occur due to the differences in the activity of the tested region. STARR-seq was one such MPRA that was used to test the regulatory activity of the fly genome in several cell-types \cite{} 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 \cite{}. These regulatory regions also tend to be flanked by nucleosomes that contain histone proteins with certain characteristic post-translational modifications \cite{}. These characteristics lead to an enriched "double peak" signal in different ChIP-Seq experiments for various histone modifications such as acetylation on H3K27 and methylations on H3K4 \cite{}. The troughs in the double peak ChIP-seq signal represent the accessible DNA that leads to a peak in the DNase-I hypersensitivity at the enhancer \cite{}. However, we are still unsure about the optimal method to combine information from multiple epigenetic marks to make cell-type specific regulatory predictions. For the first time, using data from massively parallel reporter assays (MPRAs), we have the

Anurag Sethi 5/9/2016 12:07 PM
Deleted: leading

Anurag Sethi 5/9/2016 12:07 PM
Deleted: is

Anurag Sethi 5/9/2016 12:07 PM
Moved down [1]: enhancers and promoters tend to be depleted of histone proteins and contain accessible DNA on which various transcription factors and cofactors bind \cite{}. These regulatory regions also tend to be flanked by nucleosomes that contain histone proteins with certain characteristic post-translational modifications \cite{}.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: Active

Anurag Sethi 5/9/2016 12:07 PM
Deleted: These characteristics lead to an enriched "double peak" signal containing troughs on regulatory regions within different ChIP-Seq experiments for various histone modifications such as acetylation on H3K27 and methylations on H3K4 \cite{}. Hence, conservation, TF binding motifs, TF binding sites, as well as enrichment of epigenetic marks have each been used to train models for enhancer prediction \cite{}.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: . ... [1]

Anurag Sethi 5/9/2016 12:07 PM
Deleted: and we are still unsure about the optimal method to combine information from multiple chromatin marks to make cell-type specific regulatory predictions.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: up to a hundred thousand

Anurag Sethi 5/9/2016 12:07 PM
Deleted: \cite{}.

Anurag Sethi 5/9/2016 12:07 PM
Moved (insertion) [1]

Anurag Sethi 5/9/2016 12:07 PM
Deleted: able to identify thousands of cell-type specific enhancers.

ability to properly train our models based on a large number of experimentally validated enhancers and assess the performance of different models for enhancer prediction.

We developed a new supervised machine-learning method trained from large numbers of experimentally active regulatory regions 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 to identify 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 like fly, mouse, and human. We applied these models to predict active enhancers and promoters in the 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.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: enhancer

Anurag Sethi 5/9/2016 12:07 PM
Deleted: focus

Anurag Sethi 5/9/2016 12:07 PM
Deleted: signal

Anurag Sethi 5/9/2016 12:07 PM
Deleted: datasets associated with active regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: and identified

Anurag Sethi 5/9/2016 12:07 PM
Deleted: signal

Anurag Sethi 5/9/2016 12:07 PM
Deleted: machine learning

Anurag Sethi 5/9/2016 12:07 PM
Deleted:

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 \cite{}. 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. These metaprofiles were then utilized in a pattern recognition algorithm for predicting active promoters and enhancers in a cell-type specific manner.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory region

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 \cite{}. We also calculated the

Anurag Sethi 5/9/2016 12:07 PM
Deleted: There

Anurag Sethi 5/9/2016 12:07 PM
Deleted: Finally, we

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). 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 linear filter to recognize the occurrence of a template in the presence of stochastic noise \cite{}. 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 class. 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 \cite{}. The matched filter score is higher in genomic regions where the template pattern occurs in the corresponding signal track while the matched filter score 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 genomics tracks.

We used 10-fold cross validation to assess the performance of matched filters for individual histone marks to predict active promoters and enhancers identified in a STARR-seq experiment. 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 \cite{}. 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 for predicting promoters and enhancers (see Figure 2 and Table S1) because the matched filter scores for these regions on these marks are higher than the matched filter scores for non-regulatory regions (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 Gaussian for each histone mark except for a bimodal distribution for the H3K4me1, H3K4me3, and H2Av matched

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: they are more sensitive

Anurag Sethi 5/9/2016 12:07 PM
Deleted: positives and can highlight

Anurag Sethi 5/9/2016 12:07 PM
Deleted: differences

Anurag Sethi 5/9/2016 12:07 PM
Deleted: the

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: active regulatory

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

filter scores (Figure S3). We also show that the matched filter scores are more accurate for predicting active enhancers and promoters 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 \cite{}. 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 associated with active regulatory regions by the Roadmap Epigenomics Mapping \cite{} and the ENCODE \cite{} Consortia using a linear SVM \cite{} and the integrated model achieved a higher accuracy than the individual matched filters (Figure 2). 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 inegrated model. We utilized several other machine learning algorithms 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 a separate SVM model (Figure S6) and this model displayed higher accuracy 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, H4acTetra, and

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: the second best

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 marks 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 \cite{}, we observed that the H3K4me1 metaprofile performs better for predicting enhancers while it is close to random for predicting promoters. Similarly, 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 is 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:

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory region

Anurag Sethi 5/9/2016 12:07 PM
Deleted: or

Anurag Sethi 5/9/2016 12:07 PM
Deleted:).

Anurag Sethi 5/9/2016 12:07 PM
Deleted: peforms

Anurag Sethi 5/9/2016 12:07 PM
Deleted: .

Anurag Sethi 5/9/2016 12:07 PM
Deleted: six-parameter SVM

Anurag Sethi 5/9/2016 12:07 PM
Deleted: .

Anurag Sethi 5/9/2016 12:07 PM
Deleted: Figure S8).

Anurag Sethi 5/9/2016 12:07 PM
Deleted: S10 and S11

Anurag Sethi 5/9/2016 12:07 PM
Deleted: -S11

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) \cite{}. In addition, as these regulatory regions were identified using a single core promoter in FIREWACh, the performance of the different models is probably underestimated similar to Figure 2.

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. We also show that the metaprofile 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). 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) and the BG3 cell-line (Figure S12). 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. In addition, it also enables to use the metaprofiles and statistical models learned using STARR-seq data in fly to predict enhancers in higher eukaryotes.

Different Transcription Factors bind to enhancers and promoters

We utilized the 6 parameter integrated model to predict active enhancers and promoters in the H1-human embryonic stem cell (hESC) based on the epigenetic datasets measured by the ENCODE consortium. Using these models, we predicted 43463 active regulatory regions, of which 22828 are within 2kb of the TSS and are labeled as promoters. A large proportion of the predicted enhancers are found in the introns and intergenetic regions (Figure S13). The predicted promoters and enhancers are significantly closer to active genes than might be expected randomly (Figure S14).

We further studied the differences in TF binding at promoters and enhancers (Figure 5 and Figure S15). (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.) Most promoters and enhancers contain multiple TF-binding sites. However, the TF-binding of enhancers is more heterogeneous than promoters: 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 5, 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

Anurag Sethi 5/9/2016 12:07 PM
Deleted: are

Anurag Sethi 5/9/2016 12:07 PM
Deleted: marks

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: matched filter

Anurag Sethi 5/9/2016 12:07 PM
Deleted: from S2

Anurag Sethi 5/9/2016 12:07 PM
Deleted: BG3

Anurag Sethi 5/9/2016 12:07 PM
Deleted: of fly

Anurag Sethi 5/9/2016 12:07 PM
Deleted: could be utilized

Anurag Sethi 5/9/2016 12:07 PM
Deleted: SVM

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: In addition, the ChIP-seq peaks of a few TFs are mostly within the predicted enhancers and promoters.

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 5A, 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 TF that bind 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 5C and S16. These TF co-associations could lead to mechanistic insights of cooperativity between TFs. For example, similar to a previous study \cite{}, 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 works. Previously, most methods were 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. These 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 \cite{}. 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.

Using regulatory regions identified in MPRAs for training enhancer prediction models remains an area of concern because the sensitivity and selectivity of these assays remains questionable. A majority of these MPRAs test the regulatory activity of different regions by assessing its ability to induce gene expression in a plasmid after transfecting it into a cell-type of interest \cite{}. 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 \cite{}. 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,

Anurag Sethi 5/9/2016 12:07 PM
Deleted: So

Anurag Sethi 5/9/2016 12:07 PM
Deleted: we've just shown

Anurag Sethi 5/9/2016 12:07 PM
Deleted: set

Anurag Sethi 5/9/2016 12:07 PM
Deleted: that bind up

Anurag Sethi 5/9/2016 12:07 PM
Deleted: very

Anurag Sethi 5/9/2016 12:07 PM
Deleted: than bind

Anurag Sethi 5/9/2016 12:07 PM
Deleted: is

Anurag Sethi 5/9/2016 12:07 PM
Deleted: heterogeneous than TF-binding at promoters

Anurag Sethi 5/9/2016 12:07 PM
Deleted: Figure 5.

Anurag Sethi 5/9/2016 12:07 PM
Formatted: Font:Bold

Anurag Sethi 5/9/2016 12:07 PM
Deleted: . ... [2]

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: that

Anurag Sethi 5/9/2016 12:07 PM
Deleted: . Only

Anurag Sethi 5/9/2016 12:07 PM
Deleted: experimentally and the precision/recall of these different features

Anurag Sethi 5/9/2016 12:07 PM
Deleted: region prediction remained unknown.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: We created

Anurag Sethi 5/9/2016 12:07 PM
Deleted: The validity of the

Anurag Sethi 5/9/2016 12:07 PM
Deleted: using massively parallel ... [3]

Anurag Sethi 5/9/2016 12:07 PM
Deleted: machine learning

Anurag Sethi 5/9/2016 12:07 PM
Deleted: that predict enhancers

Anurag Sethi 5/9/2016 12:07 PM
Deleted: controversial as

Anurag Sethi 5/9/2016 12:07 PM
Deleted: are assumed to be

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 different 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, H4acTetra, 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 \cite{}, the ENCODE \cite{}, and the modENCODE Consortium \cite{}.

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 \cite{}. As the transcription factors that bind to each regulatory region are thought to play a key role in core-promoter specificity \cite{}, 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-lines. 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. 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.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: are

Anurag Sethi 5/9/2016 12:07 PM
Deleted: signals

Anurag Sethi 5/9/2016 12:07 PM
Deleted: shows for the first time

Anurag Sethi 5/9/2016 12:07 PM
Deleted: in

Anurag Sethi 5/9/2016 12:07 PM
Deleted: SVM

Anurag Sethi 5/9/2016 12:07 PM
Deleted: linear

Anurag Sethi 5/9/2016 12:07 PM
Deleted: regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: The

Anurag Sethi 5/9/2016 12:07 PM
Deleted: as shown in this work.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: utilized

Anurag Sethi 5/9/2016 12:07 PM
Deleted: think

Anurag Sethi 5/9/2016 12:07 PM
Deleted: could

Anurag Sethi 5/9/2016 12:07 PM
Deleted: they are

Anurag Sethi 5/9/2016 12:07 PM
Deleted: experimentally

Anurag Sethi 5/9/2016 12:07 PM
Deleted: On the other hand, the performance of machine learning

Anurag Sethi 5/9/2016 12:07 PM
Deleted: that are trained

Anurag Sethi 5/9/2016 12:07 PM
Deleted: and

Anurag Sethi 5/9/2016 12:07 PM
Deleted: may be underestimated when utilizing data from a single core promoter as shown here in Figure 2. On

Anurag Sethi 5/9/2016 12:07 PM
Deleted: could

Anurag Sethi 5/9/2016 12:07 PM
Deleted: We also analyzed

Anurag Sethi 5/9/2016 12:07 PM
Deleted: distal regulatory regions

Anurag Sethi 5/9/2016 12:07 PM
Deleted: proximal regulatory regions.

Anurag Sethi 5/9/2016 12:07 PM
Deleted: distal regulatory regions

Figure Captions

Anurag Sethi 5/9/2016 12:07 PM
Deleted:

Figure 1: Creation of metaprofile. A) We identified the "double peak" pattern in the H3K27ac signal close to STARR-seq peaks. 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 4: 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 S15. B) The AUROC and AUPR for a logistic regression model created from 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 S16.