Supporting Information for Matched Filter Paper

Methods

Creation of Metaprofile:

We utilized the smoothed histone signal tracks provided for the S2 cell-line by the modENCODE consortium \cite{} to aggregate the corresponding histone signals around the STARR-seq peaks \cite{}. This aggregation was performed to remove noise before using the metaprofile *s(n)* for identifying active regulatory regions in the genome. The genome-wide profile for open chromatin (DNase-seq or DHS) for the S2 cell-line was calculated based on the experiments by the Stark lab \cite{}. To create the smoothened metaprofile, we utilized the H3K27ac signals of “double peak” regions around active STARR-seq peaks in the S2 cell-line \cite{}. The active regulatory regions are assumed to be STARR-seq peaks that occur on open DHS regions in the genome \cite{}. In this study, we chose all the STARR-seq peaks that overlap with DHS or H3K27ac peaks to be active regulatory regions in the genome.

To identify double peak regions, we initially identified the minimum in the H3K27ac signal track closest to the middle of the STARR-seq peaks. A minimum is accepted if it has the lowest signal within a 100 base pair region in the H3K27ac signal track. Then we proceed to identify the flanking maxima (both sides of the minimum) within a total of 2-kilo base pair region of the STARR-seq peak (1kb on each direction from the center of the STARR-seq peak). These maxima are accepted only if they have the highest signal within a 100 base pair region in the H3K27ac signal track. Approximately 70% of the active STARR-seq peaks contained an identifiable double peak within the H3K27ac signal.

After identifying the double peaks surrounding STARR-seq peaks, we aggregated the signal after aligning the maxima flanking the regulatory region. The signal track is interpolated with a cubic spline fit so that the signal track contains equal number of points for each double peak region. The aggregated signal tracks are averaged to create the metaprofile for the double peak regions. While the signal tracks are aggregated based on identifying the double peak regions in the H3K27ac signal track, the same set of operations can be performed with any histone mark expected to have the double peak pattern flanking regulatory regions.

In addition, while creating the metaprofile for H3K27ac signal close to active STARR-seq peaks, we also performed the same set of transformations on other dependent epigenomic datasets (other histone marks and/or DHS signal). In this study (Figures 1 and S2), the dependent profiles for all other epigenetic datasets are calculated by averaging the corresponding signal based on identifying double peak regions within H3K27ac signal.

Matched Filter Algorithm:

The epigenetic signal at enhancers and promoters can be approximated as the linear superposition of background noise and the metaprofile *s(n)* learned in Figure 1 (Figure S2) for the corresponding experimental dataset. The matched filter *h(n)* is used to scan the epigenetic signal to identify the occurrence of the metaprofile pattern within different regions of the genome. The matched filter process is equivalent to the computation of the cross correlation between the signal *y(n)* and the reverse of the transformed metaprofile template *s*(N-n)* (where *N* is the total number of points in the template). In other words:

$$r(n) = \sum_{i=1}^{N} y(i) * h(i)$$

where *h(i)* is the matched filter and can be written as:

$$h(i) = s^{*}(N - i)$$

As shown in Figure S1, there is a large amount of variability in the span (distance between the two peaks in the histone signal) of the regulatory region in the epigenetic signal. As a result, we scan the genome with the matched filter scanning different spans of the genome (distance between the two peaks allowed to vary between 300 and 1100 base pairs) and take the highest score as the matched filter score for that region. The matched filter is the filter that recognizes any given template in the presence of noise in a signal with the highest signal-to-noise ratio. In the presence of white noise alone, the matched filter score is low and follows a Gaussian distribution (negatives). The presence of the metaprofile within the signal leads to higher matched filter scores for positives.

Statistical Learning Models

The matched filter scores for negatives for different histone marks are unimodal that can be fit using separate Gaussian distributions. The Z-scores of matched filter scores with respect to the negatives (random regions of genome) are used as input features for training different statistical learning models. In the main text, we discuss our results of the Support Vector Machine (SVM) model, which is one of the most versatile and successful binary classifiers \cite{}. We utilized a linear kernel to distinguish between the positives and negatives. The linear SVM identifies a decision boundary that maximally discriminates the epigenetic features of regulatory regions from random regions of the genome in the SVM feature vector space. In the Supporting Information, we also present results for Ridge Regression, Random Forest, and Gaussian Naïve Bayes models and the accuracy of different models are comparable. We use scikit-learn \cite{} for training and assessing the performance of all the machine learning models.

Assessing the Models:

In order to assess the accuracy of matched filter for predicting enhancers and promoters, we used 10-fold cross validation. In Figure 2, the positives are defined as the active peaks (intersecting with DHS or H3K27ac peaks) from a single STARR-seq experiment (singe core promoter) or the union of active peaks from multiple STARR-seq experiments (multiple core promoters). The negatives are randomly chosen regions in the genome with H3K27ac signal that had the same width distribution as the distribution of distance between double peaks near STARR-seq peaks (shown in Figure S1). We typically chose between 5 to 10x number of negatives as compared to number of positives in Figures 2, 3, and 4 as the number of enhancers and promoters in the genome (positives) are far lesser than the number of negatives. During 10-fold cross validation, the positives and negatives are randomly divided in to 10 groups each. Nine of the 10 groups are randomly combined to train the model and the predictions are tested on the 10th group. To evaluate the performance of trained classifiers, we performed 10-fold cross-validation on the training data and quantified our results with area under ROC, and area under precision-recall (AUPR) curves. In Figure 3, the active promoters are defined as active STARR-seq peaks (multiple core promoter) within 1 kb of TSS (Ensembl release 78) while enhancers were active STARR-seq peaks more than 1kb from any TSS in *Drosophila melanogaster*. While calculating the matched filter for positives and negatives, we considered the best scoring matched filter score after padding each region to 1.5kb width. In Figure 4, the promoters are defined as FIREWACh peaks within 2 kb of TSS (GENCODE release vM4) while enhancers were FIREWACh peaks more than 2kb from any TSS. The FIREWACh assay is performed in a transduction assay and was based on ChIP-seq peaks of a few key TFs. Hence, we did not split the FIREWACh peaks in to active and poised enhancers and promoters.

H1-hESC whole genome prediction

To predict enhancers and promoters on the whole genome, we utilized the 6 parameter machine learning model shown in Figure 2. The H3K27ac matched filter was utilized with 5% FDR (with respect to negative Gaussian model) to set the threshold for the machine learning model. There were 43463 active regulatory regions predicted in the human genome (< 2% of genome). All regions within 2kb of TSS were annotated as promoters while active regulatory regions that were more than 2kb from TSS were annotated as enhancers. The distribution of the expression of closest gene (GENCODE v19 TSS) from ENCODE RNA-seq dataset for H1-hESC was compared to the expression of all genes from H1-hESC. The Wilcoxon test was used to measure the significance of changes in gene expression.

H1-hESC TF binding

To measure the differences in TF binding and co-binding patterns at promoters and enhancers, we overlapped the ChIP-seq peak from ENCODE with our predicted enhancers and promoters using intersectBed. The two regions were considered to be overlapping if at least 25% of the ChIP-seq peak was overlapping with the predicted enhancer or promoter.