Inferring transcriptional activation and repression activity maps in single-nucleotide resolution using deep-learning

Tom Aharon Hait
Tel Aviv University
Ran Elkon
Tel Aviv University
Ron Shamir (✉ rshamir@tau.ac.il)
Tel Aviv University

Method Article

Keywords:

Posted Date: August 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3270775/v1

Abstract

Recent computational methods for inferring cell type-specific functional regulatory elements have used sequence and epigenetic data. Active regulatory elements are characterized by open-chromatin state, and the novel experimental technique ATAC-STARR-seq couples ATAC-seq assays, which capture such genomic regions, with a functional assay (STARR-seq) to selectively examine the regulatory activity of accessible DNA. ATAC-STARR-seq may thus provide data that could improve the quality of computational inference of active enhancers and silencers. Here, we propose a novel regression-based deep learning (DL) model that utilizes such data for predicting single nucleotide activation and repression maps. We found that while models using only sequence and epigenetics data predict active enhancers with high accuracy, they generally perform poorly in predicting active silencers. In contrast, models building also on data of experimentally identified enhancers and silencers do substantially better in the identification of active silencers. Our model predicts many novel enhancers and silencers in the model lymphoblastoid cell line GM12878. Epigenetic signatures of the novel regulatory elements detected by our model resemble the ones shown by the experimentally validated enhancers and silencers in this cell line. ChIP-seq enrichment analysis in predicted novel silencers identify a few significant enriched transcriptional repressors such as SUZ12 and EZH2, which compose the PRC2 repressive complex. Intersection with GWAS data found that the novel predicted enhancers are specifically enriched for risk SNPs of the Lupus autoimmune disease. Overall, while silencers are still poorly understood, our results show that our DL-model can be used to complement the experimental results on regulatory element discovery.

Background

Regulatory elements that control transcription such as enhancers and promoters have been studied extensively over the past two decades (reviewed in (1); (2)). In contrast, silencers, which turn-off or reduce the transcription of their target gene, have received less attention, mainly because they are harder to validate experimentally. There is still no consensus on how to identify silencers. For example, two recent studies have applied different genomic screening techniques and identified 2,664 (3) and 3,001 (4) silencers in K562 cell line. However, there is no overlap between these two sets. Thus, robust characterization and annotation of functional silencers is a major genomic challenge. Furthermore, the candidate regions tested in such experiments may contain sub-regions that are interchangeably activating and repressing (5), making their detection even harder. For example, a recent study in Drosophila suggested the existence of bi-functional elements (BFEs), acting as silencers in some cell types and as enhancers in other cell types (6).

Functional enhancers are known to exhibit characteristic histone modifications such as acetylation of histone H3 at Lysine 27 (H3K27ac) and monomethylation of histone H3 at Lysine 4 (H3K4me1), and to reside within open genomic regions depleted of nucleosomes (1). In contrast, the precise epigenetic marks for functional silencers is not well characterized, and they can be located within both closed and open chromatin regions (7). Silencers can repress transcription through different modes: they can repress target genes by enhancing the establishment of a repressive chromatin structure (8) or by competition between transcriptional repressors with transcriptional activators or general transcription factors (TFs) on binding sites (BSs). For example, the BCL6 repressor competes with the STAT6 and CEBPB activators for BSs in the IL4 promoter to repress transcription (9). An additional mode of action is the recruitment of the Polycomb-Repressive Complex 2 (PRC2), which mediates trimethylation of histone H3 at Lysine 27 (H3K27me3) at promoters, making them inaccessible to activators (10, 11).

Several recent studies aimed at systematically predicting silencers on a genomic scale. The first approach implemented by Huang et al. (12) used an SVM-based model to predict silencers from DNase-Hypersensitive sites (DHSs) overlapped with H3K27me3 broad peaks (termed as H3K27me3-DHS sites) using transcription factor binding site (TFBS) maps, various epigenetic signals and gene expression (GE) as features. As large sets of experimentally identified silencers were not available at that time, this study used predicted sets of silencers for training the SVM model. Silencers were defined as H3K27me3-DHS sites that are negatively correlated with the nearest gene's expression across 25 cell types. However, the models were not trained against background sequences that are nonfunctional (i.e., sequences that neither increase nor decrease their target GE). In a follow-up study, Huang and Ovcharenko developed a sequence-based deep-learning (DL) method (13), which classifies sequences into silencer, enhancer and nonfunctional classes. However, the use of sequence information as the sole input to the model may limit the prediction quality. Here too, the model was trained on putative silencers and enhancers defined based on epigenomic marks. Validation against experimentally identified silencers in K562 cell line (3, 4) resulted with 0.47–0.48 area under the precision-recall (AUPR) curve, a substantial improvement over the 0.32 obtained by the previous SVM-based method. These results leave much room for improving silencer prediction. The above mentioned prediction methods aimed to predict a specific class (silencer or enhancer) per input sequence. However, the input sequences used in these studies were typically 1kb long and therefore may contain distinct sub-intervals with either activating or repressing effects (5). To the best of our knowledge, there are currently no methods that predict activating or repressing effects of an input candidate regulatory element in a single-nucleotide resolution.

Recently, an ATAC-STARR-Seq study identified experimentally a multitude of enhancer and silencer elements in GM12878 cell line (5), each with high-resolution contribution scores for activation or repression of transcription. These data can be used to develop more robust computational methods to predict regulatory elements (REs) and to gain additional biological insights on functional silencers.

Here, we present a novel regression-based DL model that predicts per-nucleotide activation and repression activities within candidate sequences. Our model predicts many additional enhancers and silencers, and expands the current biological knowledge of what defines functional silencers.

Results

Training on experimentally identified regulatory elements improves predictive accuracy of silencers models

Due to lack of broad sets of experimentally identified silencers, the computational models for silencers developed by Huang and Ovcharenko (13) were trained on sets of putative silencers that were defined based on their epigenomic profile rather than on experimentally detected silencer elements. The recent ATAC-STARR-Seq study by Hansen et al. provides extensive sets of identified enhancer and silencer elements in the lymphoblastoid cell line GM12878 (5). Therefore, first, we wished to compare the performance of silencer models trained on putative silencers that were defined based on epigenomic marks to the performance of models trained on experimentally identified silencers.

Following the epigenetic criteria used by Huang and Ovcharenko, we defined as the set of putative silencers in GM12878 all H3K27me3 peaks not overlapping either H3K27ac, H3K4me1 or H3K4me3 peaks in this cell line. In parallel, we defined a set of putative enhancers that are active in GM12878 as the regions of ATAC-seq peaks overlapping H3K27ac, but not H3K27me3 peaks in this cell line. We also defined a background set of regulatory elements that are non-functional in GM12878 as regions of ATAC-seq and H3K27me3 peaks randomly chosen from five other cell types that were not detected in GM12878. Overall, this epigenetic approach defined 41,548 enhancers, 24,554 silencers and 396,612 nonfunctional peaks. We applied the Convolutional Neural Network (CNN) method introduced by Huang and Ovcharenko on the GM12878 training set using 1kb sequence as the only feature, and evaluated how accurately it classified the experimentally identified elements detected by ATAC-STARR-Seq in this cell line (22,336 enhancers, 19,289 silencers and 175,088 nonfunctional ATAC-seq peaks; **Methods**). The CNN model achieved for enhancers 0.3 AUPRC, and for silencers 0.06 AUPRC (**Supplementary Figure 1**).

Next, we applied the same CNN method, but now trained the model using the sequences identified experimentally as regulatory elements by ATAC-STARR-Seq (Hansen et al.) Chromosomes 1-5, 9-22 and X constituted the **training set**. Chromosome 6 was used as a **validation set** for tuning the model's hyper-parameters. The **test set** used for evaluation of the model's performance included chromosomes 7 and 8.

The predictive performance of enhancer models trained on the experimentally identified enhancers was 0.37 AUPRC, a bit higher than the performance obtained by the enhancer models trained on putative enhancers defined based on epigenomic marks (0.3 AUPRC). In contrast, for silencers, the performance of the models trained on experimentally identified silencers was 0.77 AUPRC, dramatically higher than that obtained by the silencer model trained on REs defined by epigenomic marks (0.06 AUPRC) (**Supplementary Figure 1**). This result reflects the much better knowledge we currently have on epigenomic marks defining active enhancers compared to those that mark active silencers. Furthermore, as extensive sets of experimentally identified enhancers and silencers are available for only a limited number of cell lines, our result indicates that the availability of epigenomic profiles for canonical marks in various cell lines is sufficient for reasonable prediction of enhancers in these cells, but it does not allow accurate prediction of the landscape of active silencers.

Improved deep-learning model for prediction of enhancer and silencer elements

Next, we aimed to build a DL model for regulatory elements with improved accuracy. We reasoned that a DL model can utilize the quantitative output measured by STARR-Seq for the effect of the probed genomic intervals on transcriptional activity, rather than using discrete classes (Enhancer/Silencer/Non-functional categories) in the model learning phase. Therefore, we implemented a two-steps model as follows: Step 1 implements a regression model that predicts, in a single-nucleotide resolution, activation and repression effects in the trained cell type. Step 2 is a 3-class classification model built upon the trained regression model (**Fig. 1a**). The input to our model are 1kb sequences of ATAC-seq peaks together with epigenetic signals of DNA methylation, H3K27ac, and H3K4me1 in that interval (**Fig. 1b;** see next section for how we selected the epigenetic marks).

The regression model was built using activation and repression profiles measured for GM12878 ATAC-Seq peaks by STARR-Seq in 50-bp windows (5) (**Methods**). These windows were computationally merged to 21,125 silencers and 30,078 enhancers. We also generated an exploratory set composed of 70,937 GM12878 ATAC-seq peaks that did not overlap any silencer or enhancer identified by ATAC-STARR-Seq in this cell line. These peaks were excluded from the training phase and used in downstream analyses. We tested three different DL architectures previously used in genomic analyses: deepTACT (14), CNN (13) and ResNet (15). We also tested a simple linear regression as a baseline model. In each DL architecture, we replaced the last layer by a new dense layer that outputs 1,000 regression scores, one per position in the input sequence (**Fig. 1a**; **Methods**). Models were compared based on their classification performance in the second step.

In Step 2 we implemented a 3-category classification model by appending two dense layers to the regression network, to account for dependency between adjacent nucleotides' activation and repression levels. The first layer consists of 300 outputs, and the second, final layer, has three outputs, corresponding to the classes to be predicted: enhancer, silencer and nonfunctional. The predicted class is the one receiving the highest probability.

For the classification task, input 1kb sequences were labeled using the following scheme: (1) we scored each sequence by summing over the activation and repression levels at every nucleotide, (2) we divided the sequences into two sets: those with positive and negative sums, (3) in the positive set, the top 25th percentile were labeled as enhancers, (4) in the negative set, sequences at the bottom 25th percentile were labeled as a silencer, (5) all other sequences were labeled as nonfunctional. Overall, the 85% and 76% of the silencers and enhancers called by the original ATAC-STARR-Seq matched the labels they got by this scheme.

We again used enhancers and silencers from chromosomes 1-5, 9-22 and X for the **training set**. Elements from chromosome 6 were used as a **validation set**, and the **test set** included he elements from chromosomes 7 and 8. The three DL architectures had similar performance (**Fig. 1c**), and all performed better than the simple linear regression model. All DL models performed quite well in predicting silencers (AUPRC 0.81-0.86), and much better than the sequence based model of Huang and Ovcharenko (13) (AUPRC 0.77; **Supplementary Figure 1).** Performance of the DL models in predicting enhancers were much lower (AUPRC 0.51-0.55). This might be attributed to the fact that the observed activation levels of enhancers are not clearly distinguishable from the nonfunctional levels (**Fig. 2**). Overall, the deepTACT model performed best in predicting both enhancers and silencers. Thus, we used this model in downstream analyses.

Epigenetic markers improve prediction performance

The silencers prediction models developed by Huang and Ovcharenko used only sequence information as input. Our DL model utilizes also epigenetic data. Therefore, next, we examined whether the addition of epigenetic information improves the prediction performance. To this end, we trained the deepTACT model on sequences alone or on sequences together with combinations of additional epigenetic markers. Indeed, our result shows that adding the epigenetic data, and specifically H3K27ac and H3K4me1 signals, improved the prediction performance of our model, with more prominent improvement obtained for enhancers (AUPRC improves from 0.29 to 0.54 for enhancers and from 0.76 to 0.85 for silences) (**Supplementary Table 1**).

When plotting the average signal across sequences of predicted classes, we found that our model captures epigenetic signals that were not part of the input training data and are relevant to the activity of enhancers and silencers (**Fig. 2**). For example, high signal for the transcriptional co-activator P300 (EP300), a histone acetyltransferase known to bind active enhancers, was obtained within predicted enhancers but was markedly depleted within silencers. On the other hand, in flanking nucleosomes of predicted silencers we observed high signals for the enhancer of zeste homolog 2 (EZH2), which is part of the PRC2 complex, and for H3K27me3. In addition, predicted silencers seem to be more methylated compared to the other two classes. EZH2 can also serve as an activator (16), which could explain the high signals it obtained at the center of the predicted enhancers in the test set (**Fig. 2**).

Overall, silencers predicted by our model tend to be more methylated and more strongly marked by H3K27me3 than enhancers (**Fig. 2**). On the other hand, as expected, predicted enhancers tend to be marked by H3K27ac and H3K4me1 and bound by EP300.

Next, we set to determine which features contributed the most to the classification. For this task, we used the integrated gradients (IG) approach (17) (**Methods**), which calculates feature importance scores per input sample given their labels. The sign of these scores indicate a positive or negative correlation between the feature signal and the classification score. The magnitude of these scores indicates the contribution of the feature to the classification score. We applied this approach to input sequences in the test set given their labels. We found that enhancer classification scores were most positively correlated with H3K4me1 and H3K27ac levels followed by the DNA bases C and G, and DNA methylation features (**Fig. 3a**). The contribution of both H3K27ac and methylation is in agreement with previous findings of their bivalent role in enhancers (18). In addition, methylation is associated with GC-rich regions, and, as expected enhancers tend to be GC-rich. Interestingly, the G and C features were the only major contributors to silencer classification, with little contribution from the epigenetic marks. This could be attributed to the fact that silencers tend to be closer to TSSs compared to enhancers (mean distance:

18,782 bp vs. 52,324 bp; **Supplementary Figure 2**). Regions closer to TSSs, e.g., promoters, are highly GC-rich (19). Both enhancer and silencer classification scores were negatively correlated with A and T features. In contrary to enhancer and silencer classifications, classification scores for nonfunctional elements were most positively correlated with A and T features. Chromatin in AT-rich regions is more compacted than in GC-rich regions (20), which could explain why nonfunctional regions are AT-rich. Nonfunctional classification scores were strongly negatively correlated with H3K4me1 levels, as this marker is mostly associated with enhancer regions.

Next, we used the feature importance scores to find enriched motifs in the sequences using TF-MoDISco (21) (**Methods**). We identified one motif within silencers in the test set. This motif matched the binding motif of SP2 and SP3 TFs (using TomTom (22)) (**Fig. 3b**), which bind GC-rich elements. A richer set of 8 motifs was found within enhancers in the test set. Among the motifs, one matched Myocyte enhancer factor (MEF) TFs (**Fig. 3c**), and others matched known B-cell TFs such as: PRDM6, BCL11A, and IRF3 (**Supplementary Table 2**).

deepTACT predicts novel enhancers and silencers in GM12878

We applied the trained deepTACT model on the ATAC-seq peaks in the exploratory set (containing the set of 70,937 ATAC-seq peaks in GM12878 that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription) in order to find novel enhancers and silencers in GM12878 which were missed by the ATAC-STARR-seq experiment (**Methods**). The model predicted 3,752 novel enhancers and 518 novel silencers. The epigenetic marks on these predicted elements are similar to those obtained on the experimentally identified enhancers and silencers (**Fig. 2; Supplementary Figure 3**).

To provide further support for the functionality of these novel predictions, we examined their enrichment for eQTLs and GWAS variants. We used eQTL data from Lymphoblastoid cell lines downloaded from the GEUVADIS database (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GEUVADIS/ge/**; Methods**). Using logistic regression and accounting for the potential confounding effect of distance to nearest TSS (**Methods**), we found that the set of novel enhancers predicted by our model is significantly enriched for eQTLs (P<3.1E-24; compared to ATAC-seq peaks not predicted as enhancers/silencers). We observed no eQTL enrichment in the set of predicted silencers, possibly due to their low number (n=518). On the other hand, the sets of experimentally identified enhancers and silencers were both enriched for eQTL signal (P<8.0E-27 and P<6.7E-45, respectively).

Next, we used GWAS summary statistics for 50 diseases and traits from Groenewoud et al. (23). In each one, we kept the SNPs with p-value < 10^{-7}. When performing enrichment analysis of the SNPs in each predicted class, we found that the set of experimentally identified enhancers was enriched for systemic lupus erythematosus (SLE) risk SNPs (Q<5.2E-5; **Supplementary Fig. 4a**), an autoimmune disease involving B-cells, as well as for schizophrenia (SCZ) risk SNPs (Q<5.4E-7), in line with a study that implicated increased levels of B-cell cytokines and autoantibodies in SCZ (24). The silencers were also enriched for some diseases albeit at lower statistical significance (**Supplementary Fig. 4b**). Reassuringly, the set of novel enhancers predicted by our model was also enriched for SLE (**Fig. 4a**; Q<1.5E-5) and schizophrenia risk SNPs (Q<1.2E-3). No enrichment for GWAS risk SNPs was found within the set of predicted silencers.

Among the SLE risk SNPs in the novel enhancers is rs8052690, located within an enhancer that interacts, according to C-HiC analysis, with the promoter of the IRF8 gene (25) (**Fig. 4b**). As an another example, the SLE risk SNP rs13240595, which has ~2.5-fold enhancing effect as measured using MPRA (26), is located within a novel enhancer, which is predicted (by FOCS (27) and GeneHancer (25) enhancer-promoter maps) to interact with the promoter of the TNPO3 gene. TNPO3 was previously shown to be associated with SLE (**Supplementary Fig. 5**) (28).

Predicted novel enhancers and silencers are enriched for binding motifs of known transcriptional activators and repressors

To further support the functionality of the novel enhancers and silencers predicted by our model for GM12878 in the exploratory set, we performed motif enrichment analyses (**Methods**). Using very stringent cutoffs of q-value=1E-40 and 1.5 fold-enrichment, 54 motifs were found within the novel enhancers, including some well-established B-cell TFs: PAX5, IRF8, BCL11A and SPIB (**Supplementary Fig. 6a**). 42 (78%) of these motifs were also found among the 93 enriched TFs detected in the set of experimentally identified enhancers. Within the novel predicted silencers, we detected four enriched TFs (**Supplementary Fig. 6b**). Among them, ZBTB17 and PATZ1 were implicated as transcriptional repressors (29). These four enriched TFs were also found among the 146 enriched TF motifs detected in the set of experimentally identified silencers.

In addition to motif analysis, we also examined enrichment for physical TF binding sites in GM12878. To this goal, we downloaded all 154 available GM12878 ChIP-seq experiments from ENCODE project and analyzed their enrichment within the predicted and experimentally identified sets of enhancers and silencers. For the novel silencers, using stringent cutoffs of q-value=1E-20 and at least 10 fold-enrichment, we found four enriched proteins (**Fig. 5a**), SUZ12, HDAC6, EZH2 and NRF1. Notably, SUZ12 and EZH2 are members of the PRC2 complex, which represses transcription (30). HDAC6 is a histone deacetylate and marks epigenetic repression (31). The experimentally identified silencers were enriched for binding of 35 proteins (**Supplementary Fig. 7a**)

The predicted enhancers were enriched for 26 proteins (**Fig. 5b),** including MAX and MYC, which when in complex act as activators in B-cells (32), and IRF3, which is known to be involved in B-cell functions (33). 18 out of 26 enriched proteins (~69%) were also enriched within the experimentally identified enhancers (**Supplementary Fig. 7b**).

Discussion

Our first goal in this study was to test if a DL model trained on putative silencers labeled using epigenomic data can accurately detect experimentally identified silencers. To this end, we compared two class labeling approaches: the epigenetic approach, in which putative enhancers and silencers are labeled using epigenetic data, and the experimentally identified approach, in which enhancers and silencers are labeled using elements identified by ATAC-STARR-Seq assay. We trained a CNN model proposed by Huang and Ovcharenko on each dataset and tested the performance of the models on experimentally identified test set. While both trained CNN models performed similarly on predicting true enhancers (0.3 and 0.37 AUPRC for the models trained using the epigenetic and the experimental approaches, respectively), the silencer prediction performance of the model trained on the experimental dataset was dramatically higher (0.77 AUPRC) than that obtained by the model trained on the epigenetic dataset (0.06 AUPRC; **Supplementary Fig. 1**). These results reflect the much better knowledge that we currently have on epigenomic marks defining active enhancers compared to those defining active silencers.

Our second goal was to build a computational model that predicts activation and repression transcriptional activities at single nucleotide resolution within regulatory elements. To this end, we used the ATAC-STARR-Seq quantitative results to train a regression-based DL model combined with a classification model to classify sequences into enhancer, silencer or nonfunctional elements. We compared published DL architectures and found that deepTACT performed best in terms of AUROC and AUPRC (**Fig. 1c**). Predicted silencers harbor high levels of the H3K27me3 repressive mark, whereas predicted enhancers harbor high levels of H3K27ac and H3K4me1 activation marks (**Fig. 2**; **Supplementary Fig. 3**).

We applied the trained deepTACT model on an exploratory dataset, which included the ATAC-seq peaks in GM12878 that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription. Within this set, the model identified 3,752 and 518 novel putative enhancers and silencers, which were possibly missed by the experiment. Reassuringly, 18 of the predicted novel enhancers overlapped 42 Lupus risk SNPs, including rs13240595 Lupus risk-SNP, which was shown to have 2.5-fold enhancing effect by MPRA analysis (26). We showed that predicted enhancer sequences tend to contain significantly more eQTLs than predicted nonfunctional sequences. ChIP-seq enrichment analysis within predicted novel silencers identified four major enriched proteins: SUZ12, HDAC6, EZH2 and NRF1. SUZ12 and EZH2 form the PRC2 repressive complex known to bind silencers (**Fig. 5a**). Predicted enhancers, on the other hand, were enriched for many proteins, the majority of which are known to induce transcription (**Fig. 5b**).

Our study has several limitations. It was performed on a single cell type for which genome-wide experimentally identified enhancers and silencers are available. Additional validation would necessitate experiments in more cell types. A major future challenge is to transfer the trained model on GM12878 cell type to other cell types where activation and repression levels are not in hand. As proof of concept, we retrained only the last two dense layers in the classification step on HepG2 and K562 cancer cell lines (**Fig. 1a**). We constructed training and test sets for these cell lines using (a) enhancers detected by STARR-seq experiments done by ENCODE in these cell lines, and (b) ATAC-seq regions overlapping H3K27me3 ChIP-seq peaks as putative silencers. Our results on test sets from HepG2 and K562 achieved high AUROC and moderate AUPR scores for enhancer and silencer classifications (**Supplementary Fig. 8**). This analysis indicates a great promise for the application of transfer learning techniques for predicting REs in many cell types.

Conclusions

- Computational models trained on enhancer and silencer sequences labeled using epigenetic data generally perform poorly in predicting silencers
- Leveraging data from experimentally identified enhancers and silencers substantially improves silencer prediction accuracy
- ATAC-STARR-seq experiment might miss true enhancers and silencers. These regulatory elements can be recovered using DL models

Methods

GM12878 data preparation

101,896 GM12878 ATAC-STARR-seq peaks were obtained from (5) (GEO dataset GSE181317) and resized to 1kb around their central positions. Experimentally identified silencer (n = 21,125) and enhancer (n = 30,078) regions and their repression or activation signals, as measured by STARR-Seq in GM12878, were also taken from the same dataset. Transcriptional repression and activation signals were measured at resolution of 50 bp. ATAC-seq, H3K4me1, H3K27ac, H3K27me3 and WGBS DNA methylation signal datasets in GM12878 were downloaded from the ENCODE project (https://www.encodeproject.org/). 44,494,433 CpG sites with at least 4 mapped reads were kept. The methylation level in each CpG site is the fraction of methylated reads that cover it. CpGs with insufficient coverage were given a methylation level of -1.

The input data to our model is a 1000x7 matrix. For each of the 1000 bases, the first four features are one-hot encoding of the DNA sequence of the ATAC-STARR-seq peak, followed by nucleotide-resolution signals for DNA methylation, H3K27ac and H3K4me1. We normalized the features to mean 0 and std 1. The 1k target vector is a per-position value with a positive activation signal for enhancers, negative repression signal for silencers, and 0 otherwise.

The model was trained on data from chromosomes 1–5, 9–23. Data from chromosome 6 were used for validation of the model while tuning the hyper-parameter (the number of training epochs). Data from chromosomes 7 and 8 were held out as a test set to assess the model's performance.

For model training and testing, positive cases were ATAC-seq peaks overlapping experimentally identified enhancers or silencers. Following the approach of Huang and Ovcharenko (13), we used as negative cases ATAC-seq peaks that were detected in other five cell types, but not in GM12878. For each positive peak, six negative ones were randomly sampled from the same chromosome. Overall, our dataset contained 216,713 cases: 30,959 positive peaks and 185,754 negatives. GM12878 ATAC-seq that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription were left out from the phase of model training and testing, and were used as an exploratory set in downstream analyses.

Model implementation

Our model implementation is divided into two steps: In step 1, we implemented a deepTACT model as follows: (A) we used model architecture and hyper-parameters similar to those implemented in Li et al (14). (B) The last dense layer outputs 1,000 scores, one for each position in the input sequence, aiming to predict the activation or repression scores measured by STARR-Seq for this genomic interval. Intermediate batch normalization and Dropout layers were used to prevent overfitting. Model training was performed with the mean squared error (MSE) loss function using the 'rmsprop' optimizer. We found the number of epochs required for training the model using the MSE on the validation set. In step 2, the 1000-scores output of the last dense layer of the model in step 1 is fed into a dense layer of 300 outputs scores followed by a dense layer that outputs three scores – for predicting Enhancer, Silencer or Non-functional elements - with the softmax activation function.

Inferring enhancer and silencer intervals

Given a sequence, x, and its epigenetic signals, Step 1 of our model outputs for each position j a transcriptional activity score. The score can be positive, indicating that position j is involved in transcriptional activation (in GM12878 cell line), negative, indicating that position j is involved in transcriptional repression, or 0 (i.e., suggesting position j is non-functional). To summarize the output, we applied a 50bp sliding window with step size of 10 on the 1,000 scores the model outputs. We define a window as an Enhancer if all scores within that window are above a certain threshold (t_e). Similarly, we define a window as a Silencer if all scores within that window are below a certain threshold (t_s). We merged overlapping windows that had the same label. We selected the t_e and t_s thresholds as those yielding the maximum $F1$ score on the test set. For enhancers, the F1 score was computed by considering as positives the true activating positions in the test set, and considering as negatives - all other positions in the test set. Predicted activating positions that matched true activating positions were considered as true positives whereas unmatched predicted activating positions were considered as false positives. The same principle was applied for silencers.

The novel enhancer and silencer windows predicted (for GM12878) in the exploratory set are provided in **Supp. Table 3**.

Alternative models

We implemented three alternative models: (1) a simple linear regression implemented as a single dense layer in a DL model, (2) the CNN model of Huang and Ovcharenko (13), and (3) the ResNet-based model from Luo et al (15).

Comparison of models trained on either experimentally identified or on epigenetically called enhancers and

silencers

We took the CNN architecture as implemented by Huang and Ovcharenko (13) and used it to compare models trained either on (A) regulatory elements called based on epigenomic markers as done by Huang and Ovcharenko: (1) silencers: H3K27me3 ChIP-seq peaks not overlapping either H3K27ac, H3k4me1 or H3k4me4 ChIP-seq peaks, (2) enhancers: ATAC-seq peaks overlapping H3K27ac ChIP-seq peaks, and (3) nonfunctional: ATAC-seq peaks from five other cell types not detected in GM12878; or (B) regulatory elements experimentally identified by the ATAC-STARR-Seq assay (as described above). We measured the performance of the two models in terms of AUPRC of detection experimentally identified elements. In both approaches, only sequences (without any epigenetic signal) were provided to the CNN model as input (as done in Huang and Ovcharenko).

Feature importance scores using integrated gradients

To determine which features contribute the most to correct classification we used the integrated gradients (IG) approach (17). The main idea behind this approach is to find the contribution of input features to the prediction by calculating the integral of the model's output gradients over a straight path between a chosen 'proper baseline' input and the actual input. To do so, 50 points are sampled along the path and the output gradients are calculated for each point. Accumulating the gradients from all points defines the integrated gradients, which are used as the feature importance score. We chose a proper baseline input as follows:

$$baseline_{i,j} = \begin{cases} 0 & j \neq Methylation \\ -1 & j = Methylation \end{cases}$$

Where $1 \leq i \leq 1000$ and j is the feature type: A, C, G, T, Methylation, H3K27ac or H3K4me1. This baseline corresponds to a sequence with 0 (or NA) signal for all seven features. After computing the integrated gradients per position i and feature j, we summed them up across all positions to represent the integrated gradients of feature j. The feature importance score of each feature is the average of the integrated gradients across all inputs per class.

Identification of motifs within importance scores

We used TF-MoDISco to find recurrent motifs within subsequences with highly positive or negative importance scores (21). The tool first finds subsequences of high importance scores, aligns and clusters them, and then finds a set of recurring motif patterns. We computed the IG of the enhancer and silencer sequences in the test set. To account only for changes in the nucleotide composition we kept the epigenetic features fixed along the path between the baseline and the input. The null IG distribution used in TF-MoDISco was generated by dinucleotide shuffling the original sequences and computing their IG.

Motif finding

We applied the simple enrichment analysis (SEA) tool from the MEME suite (https://meme-suite.org/meme/tools/sea) on the inferred sequences with Human HOCOMOCO v11 PWMs (34). A

Markov model of order of 1 was chosen to model the background distribution.

ChIP-seq enrichment

We downloaded all 158 ENCODE ChIP-seq narrowpeak bed files of GM12878 cell type. For each peak file, we represented each peak by the single nucleotide position that had maximum ChIP-seq signal. Then, we computed the number of these positions that overlapped with 1kb predicted enhancers, silencers and nonfunctional sequences from the exploratory dataset. We computed the enrichment fold-change as follows:

$$FC\left(protein\right) = \frac{\frac{\#overlapsintargetset}{|targetset|}}{\frac{\#overlapsinbgset}{|bgset|}}$$

Where the target set is either the enhancers or silencers. The bg set included all sequences in the exploratory set. We used the Hypergeometric test (python 'scipy.stats.hypergeom') to evaluate the significance of the enrichment. Benjamini-Hochberg multiple testing correction was used to correct the p-values (35).

eQTL and GWAS risk SNPs enrichment

GWAS summary statistics of 50 traits were downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/) and preprocessed as described in Groenewoud et al. (23). For each trait, we retained associated variants with p-value < 1E-7. Then, similar to ChIP-seq enrichment above, we computed the overlap of the risk SNPs with the predicted enhancers and silencers from the exploratory dataset and computed the significance of the enrichment using HG test.

As for eQTL enrichment analysis, we downloaded the lymphoblastoid cell line (LCL) GEUVADIS eQTL dataset (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GEUVADIS/ge/) and computed the overlap of the eQTLs with the predicted enhancers, silencers and nonfunctional sequences in the exploratory dataset. To find whether eQTLs tend to overlap enhancers more than the nonfunctional sequences we implemented a logistic regression test: $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ where: Y_i denotes whether sequence i has an eQTL or not, X_1^i denotes whether sequence i is an enhancer or a nonfunctional sequence, and X_2^i is the distance from region i mid-position to the nearest gene TSS. We added the distances to the nearest gene as this distance may confound the association between eQTLs and genomic intervals. We used the python statsmodels.sd.Logit function to implement logistic regression and to infer significance of the coefficients. If β_1 is positive and significant then we concluded that the eQTLs are significantly enriched within the set of enhancers. Similar analysis was done for silencers versus the nonfunctional sequences.

Declarations

Availability of data and materials

Python scripts and links to download the data used in this study are available on GitHub: https://github.com/Shamir-Lab/EnhancerSilencerDL .

Competing interests

The authors declare that they have no competing interests.

Funding

The study is supported in part by Israel Science Foundation grant No. 2206/22 (to R.S.), the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics (to R.E. and R.S.), the Tel Aviv University Center for AI and Data Science (TAD, to R.E. and R.S.), Len Blavatnik and the Blavatnik Family foundation (to R.S), a joint program grant from the Cancer Biology Research Center (CBRC), Djerassi Oncology Center, Edmond J. Safra Center for Bioinformatics and Tel Aviv University Center for AI and Data Science (TAD) (to R.E.). T.A.H. is supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Authors' contributions

TAH, RE, and RS designed the research. TAH performed the analyses. All authors analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

R.E. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics, Tel Aviv University. This work was carried out in partial fulfillment of the requirements for the Ph.D. degree of T.A.H. at the Blavatnik School of Computer Science, Tel Aviv University.

References

1. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15(4):272–86.
2. Dao LTM, Galindo-albarrán AO, Castro-mondragon JA, Andrieu-soler C, Medina-rivera A, Souaid C, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Publ Gr [Internet]. 2017;49(7):1073–81. Available from: http://dx.doi.org/10.1038/ng.3884
3. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52(3):254–63.
4. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. Candidate silencer elements for the human and mouse genomes. Nat Commun. 2020;11(1):1–15.
5. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor–bound activators and silencers within chromatin-accessible regions of the human genome. Genome Res. 2022;32(8):1529–41.

6. Gisselbrecht SS, Palagi A, Kurland J V, Rogers JM, Ozadam H, Zhan Y, et al. Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol Cell. 2020;77(2):324-337.e8.
7. Zhang Y, See YX, Tergaonkar V, Fullwood MJ. Long-Distance Repression by Human Silencers: Chromatin Interactions and Phase Separation in Silencers. Cells. 2022;11(9):1–17.
8. Cai Y, Zhang Y, Loh YP, Tng JQ, Lim MC, Cao Z, et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun [Internet]. 2021;12(1). Available from: http://dx.doi.org/10.1038/s41467-021-20940-y
9. Harris MB, Mostecki J, Rothman PB. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. J Biol Chem. 2005;280(13):13114–21.
10. Ngan CY, Wong CH, Tjong H, Wang W, Goldfeder RL, Choi C, et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat Genet. 2020;52(3):264–72.
11. van Kruijsbergen I, Hontelez S, Veenstra GJC. Recruiting polycomb to chromatin. Int J Biochem Cell Biol. 2015;67:177–87.
12. Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 2019;29(4):657–67.
13. Huang D, Ovcharenko I. Enhancer-silencer transitions in the human genome. Genome Res. 2022;32(3):437–48.
14. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res [Internet]. 2019 Jun 4;47(10):e60–e60. Available from: https://doi.org/10.1093/nar/gkz167
15. Luo Z, Zhang J, Fei J, Ke S. Deep learning modeling m6A deposition reveals the importance of downstream cis-element sequences. Nat Commun. 2022;13(1):1–16.
16. Wang J, Yu X, Gong W, Liu X, Park K-S, Ma A, et al. EZH2 noncanonically binds cMyc and p300 through a cryptic transactivation domain to mediate gene activation and promote oncogenesis. Nat Cell Biol. 2022;24(3):384–99.
17. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International conference on machine learning. 2017. p. 3319–28.
18. Charlet J, Duymich CE, Lay FD, Mundbjerg K, Sørensen KD, Liang G, et al. Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. Mol Cell. 2016;62(3):422–31.
19. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 2012;22(12):2399–408.
20. Dekker J. GC-and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 2007;8:1–14.

21. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. arXiv Prepr arXiv181100416. 2018;

22. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2):1–9.

23. Groenewoud D, Shye A, Elkon R. Incorporating regulatory interactions into gene-set analyses for GWAS data: A controlled analysis with the MAGMA tool. PLOS Comput Biol. 2022;18(3):e1009908.

24. van Mierlo HC, Broen JCA, Kahn RS, de Witte LD. B-cells and schizophrenia: A promising link or a finding lost in translation? Brain Behav Immun. 2019;81:52–62.

25. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017.

26. Lu X, Chen X, Forney C, Donmez O, Miller D, Parameswaran S, et al. Global discovery of lupus genetic risk variant allelic enhancer activity. Nat Commun. 2021;12(1):1611.

27. Hait TA, Amar D, Shamir R, Elkon R. FOCS : a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. 2018;19(1):59.

28. Kottyan LC, Zoller EE, Bene J, Lu X, Kelly JA, Rupert AM, et al. The IRF5–TNPO3 association with systemic lupus erythematosus has two components that other autoimmune disorders variably share. Hum Mol Genet. 2015;24(2):582–96.

29. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinforma. 2016;54(1):1–30.

30. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011;469(7330):343–9.

31. Grozinger CM, Hassig CA, Schreiber SL. Three proteins define a class of human histone deacetylases related to yeast Hda1p. Proc Natl Acad Sci. 1999;96(9):4868–73.

32. Pérez-Olivares M, Trento A, Rodriguez-Acebes S, González-Acosta D, Fernández-Antorán D, Román-García S, et al. Functional interplay between c-Myc and Max in B lymphocyte differentiation. EMBO Rep. 2018;19(10):e45770.

33. Oganesyan G, Saha SK, Pietras EM, Guo B, Miyahira AK, Zarnegar B, et al. IRF3-dependent type I interferon response in B cells regulates CpG-mediated antibody production. J Biol Chem. 2008;283(2):802–8.

34. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs. bioRxiv. 2021;

35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;289–300.

Figures

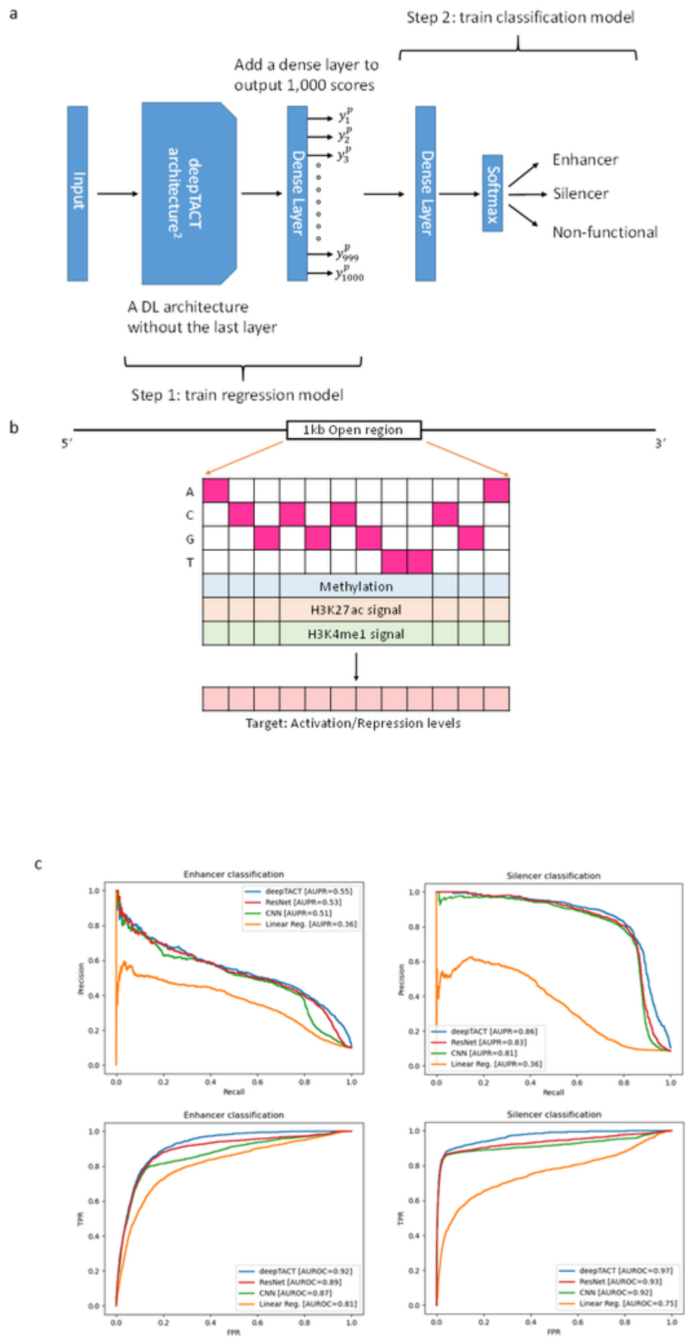

Figure 1

Model implementation and comparison. (a) Model architecture. (b) Schematic figure of the input and output structure. (c) Performance of the models (**Methods**).

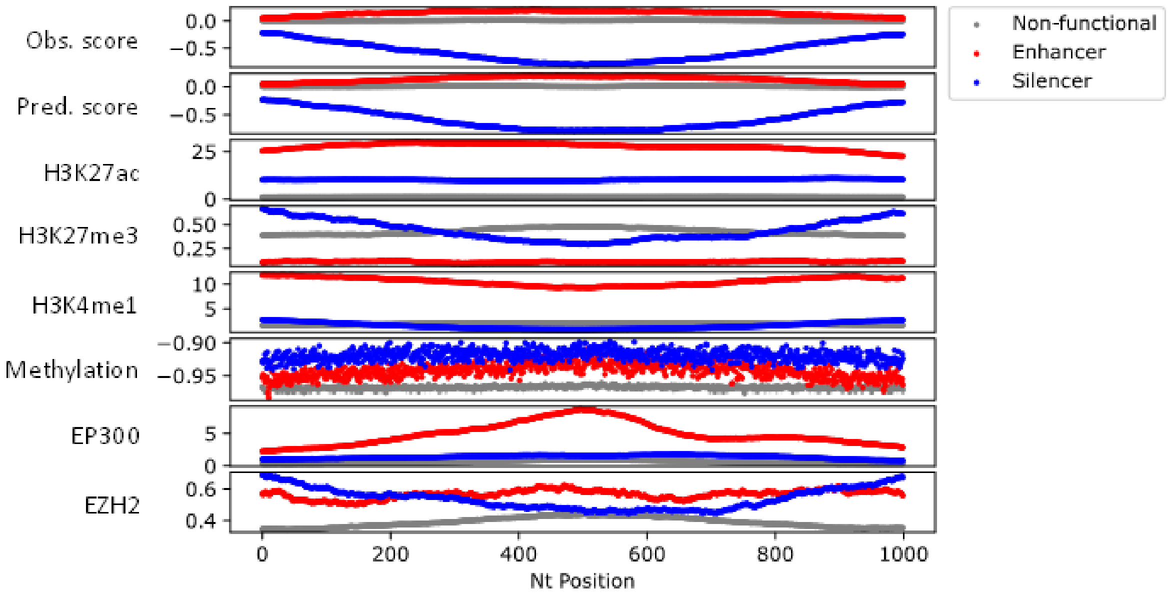

Figure 2

Summary of epigenetic markers in the test set. Top to bottom: observed scores (as measured by STARR-Seq), predicted scores (output of Step1 – the regression model), H3K27ac, H3K27me3, H3K27me1, Methylation, EP300 and EZH2. Predicted enhancers, silencers and nonfunctional are marked by red, blue and grey colors, respectively. In each predicted class and each track, the average signal per position in the 1kb sequences is shown. In b, the grey curve overlaps the blue curve for H3K27ac and the red curve for the EZH2.

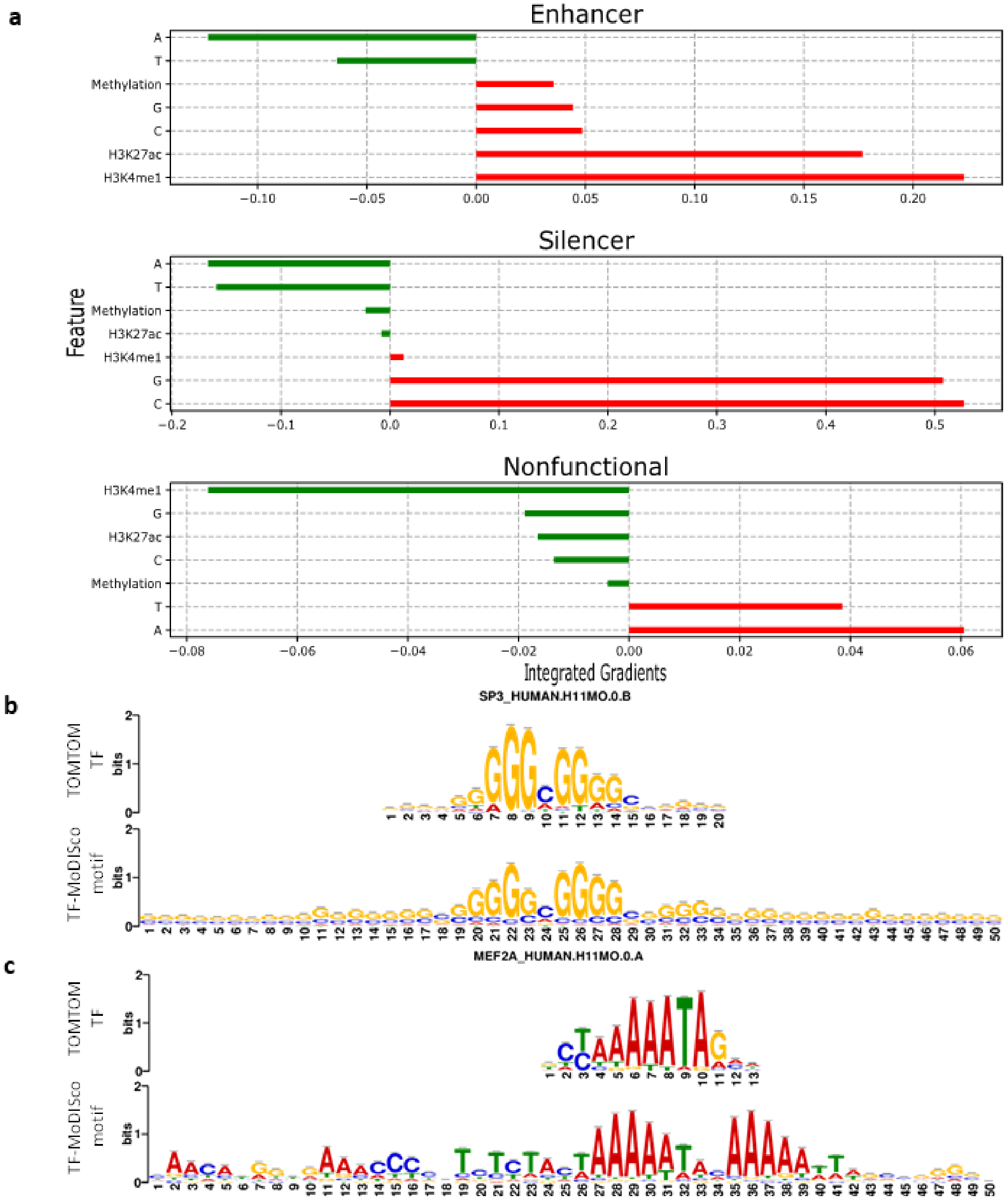

Figure 3

Feature importance scores computed for each class on the test set. (a) We used the integrated gradients approach to assign feature importance scores to the sequences per class: enhancer (top), silencer (middle) and nonfunctional (bottom). Positive or negative importance scores reflect a positive or negative correlation between the feature and the classification score, respectively. The magnitude of these scores measures the contribution of the feature to the classification score. (b) The top enriched motif in silencers as computed by TF-MoDISco and the corresponding known TF matched by TomTom. (c) Same as (b) for enhancers.

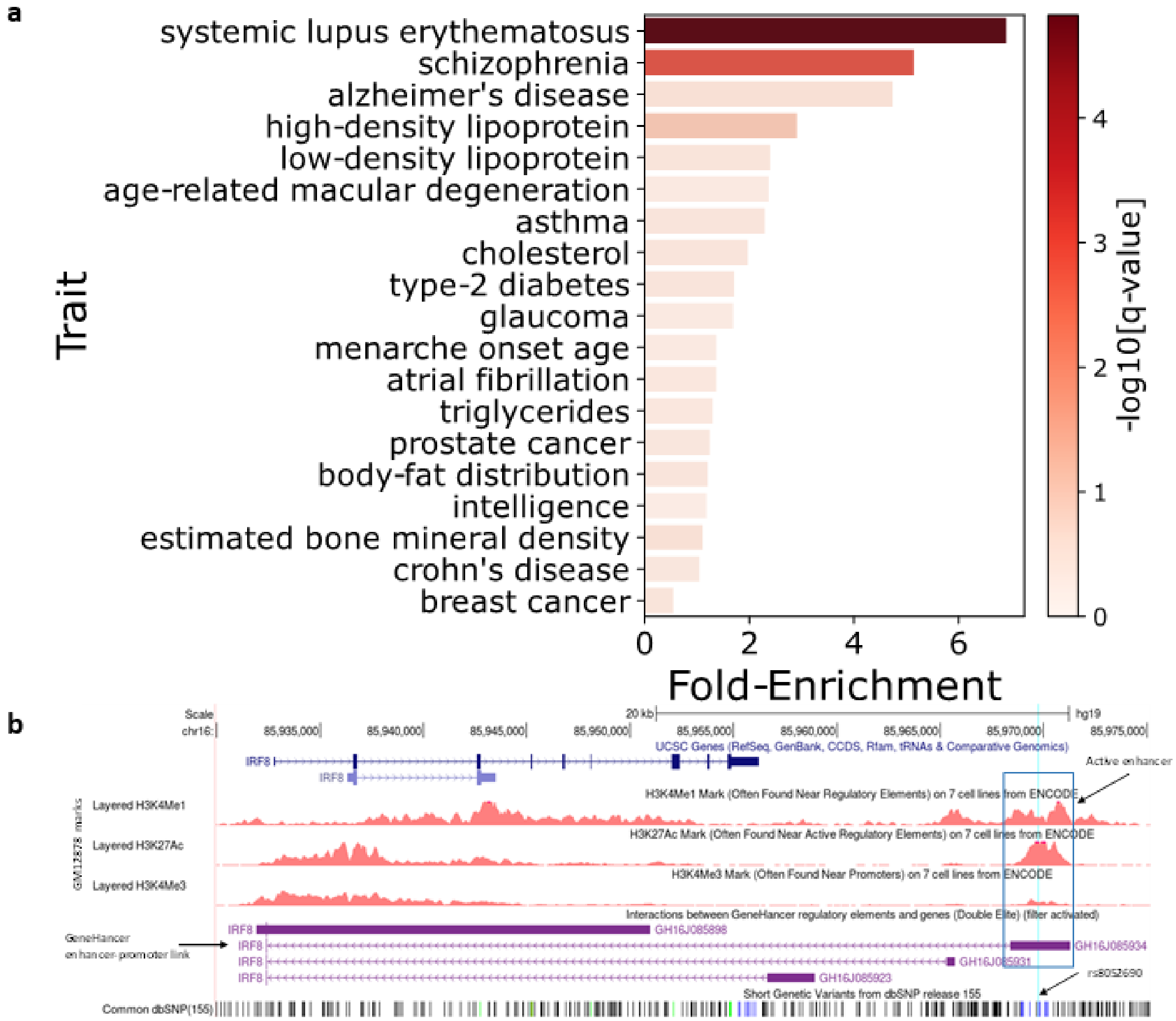

Figure 4

Enrichment of GWAS risk SNPs within predicted enhancers. (a) Enrichment for GWAS SNPs. Traits with at least one risk SNP overlapping an element in the exploratory set are shown. q-values are FDR-corrected Hypergeometic test p-values. (b) UCSC genome browser tracks of SLE risk SNP, rs8052690 (marked in arrow), falling within a predicted active enhancer that physically interacts with the promoter of IRF8.

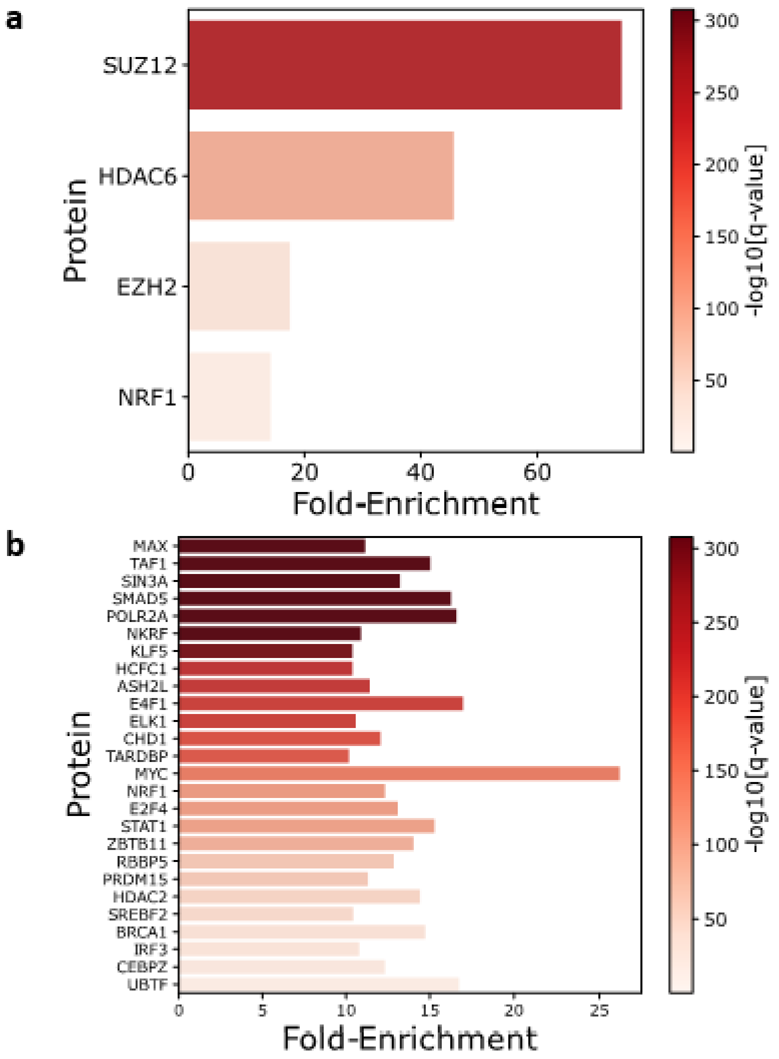

Figure 5

ChIP-seq enrichment analysis in predicted enhancers and silencers detected in the exploratory set. (a) Silencers. (b) Enhancers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.docx
- TableS2.xlsx
- TableS3.xlsx