nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-0907-8

Check for updates

Supervised enhancer prediction with epigenetic pattern recognition and targeted validation

Anurag Sethi[1,8], Mengting Gu[2,3,8], Emrah Gumusgoz[4], Landon Chan[5], Koon-Kiu Yan[1], Joel Rozowsky[1], Iros Barozzi[6], Veena Afzal[6], Jennifer A. Akiyama[6], Ingrid Plajzer-Frick[6], Chengfei Yan[1], Catherine S. Novak[6], Momoe Kato[6], Tyler H. Garvin[6], Quan Pham[6], Anne Harrington[6], Brandon J. Mannion[6], Elizabeth A. Lee[6], Yoko Fukuda-Yuzawa[6], Axel Visel[6], Diane E. Dickel[6], Kevin Y. Yip[7], Richard Sutton[4], Len A. Pennacchio[6] and Mark Gerstein[1,2,3] ✉

Enhancers are important non-coding elements, but they have traditionally been hard to characterize experimentally. The development of massively parallel assays allows the characterization of large numbers of enhancers for the first time. Here, we developed a framework using *Drosophila* STARR-seq to create shape-matching filters based on meta-profiles of epigenetic features. We integrated these features with supervised machine-learning algorithms to predict enhancers. We further demonstrated that our model could be transferred to predict enhancers in mammals. We comprehensively validated the predictions using a combination of in vivo and in vitro approaches, involving transgenic assays in mice and transduction-based reporter assays in human cell lines (153 enhancers in total). The results confirmed that our model can accurately predict enhancers in different species without re-parameterization. Finally, we examined the transcription factor binding patterns at predicted enhancers versus promoters. We demonstrated that these patterns enable the construction of a secondary model that effectively distinguishes enhancers and promoters.

Enhancers are gene regulatory elements that activate expression of target genes from a distance[1]. The vast majority of enhancers and their spatiotemporal activities remain unknown[2,3]. Understanding enhancer function and evolution is currently an area of great interest because many variants within distal regulatory elements also have been associated with various traits and diseases during genome-wide association studies[4–6]. Traditionally, regulatory activities of enhancers were experimentally validated using heterologous reporter constructs, which has led to a relatively small number of enhancers that are functionally validated in several selected mammalian-cell types[7,8]. These validated enhancers are typically in conserved non-coding regions[9,10] with particular patterns of chromatin[11], transcription factor (TF) binding[12] or non-coding transcription[13]. When complex computational methods for predicting tissue- or cell-line-specific enhancers were trained on these validated enhancers, they could be susceptible to potential biases and were difficult to generalize to other tissues or species, as the number of training data were usually not sufficient. Some published methods also featured models trained on the basis of TF-binding sites[12,14–16]. The TF-binding sites provide a larger dataset for training. However, most enhancers do not bind to one or a small group of TFs. In addition, it has remained challenging to assess the performance of different methods for enhancer prediction with a limited number of putative enhancers being validated.

The development of self-transcribing active regulatory region sequencing (STARR-seq) has made it possible to quantitatively assess the activity of millions of candidate enhancers across entire genomes[17]. In these experiments, plasmids that each contain a potential enhancer element downstream of a green fluorescent protein (GFP) gene are transfected into target cells. The differences in the activities of the tested regions are reflected by quantifying the levels of the resulting reporter transcripts through sequencing. STARR-seq confirmed previous findings that active enhancers and promoters are usually located at open chromatin regions where various TFs and cofactors bind[18–20]. In addition, it confirmed that the regulatory regions are often flanked by nucleosomes that contain histone proteins with certain characteristic post-translational modifications, such as histone H3 acetyl K27 (H3K27ac)[21]. These attributes lead to an enriched peak-trough-peak ('double peak') signal, which has been observed in previous studies[22]. Recently, similar epigenetic patterns were repeatedly observed close to regulatory regions identified in a number of massively parallel reporter assays[23,24].

We developed a method to take into account the specific enhancer-associated pattern within different epigenetic signals. Previous Encyclopedia of DNA Elements (ENCODE) and modENCODE efforts showed that the chromatin modifications on active promoters and enhancers are conserved across higher eukaryotes[25–31]. We further explored this conservation of epigenetic signal shapes for constructing simple-to-use transferrable statistical models using six epigenetic marks to predict enhancers and promoters in different eukaryotic species, including fly, mouse and human.

Working on different organisms allowed us to take advantage of different assays to validate our predictions in a robust fashion using

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [3]Department of Computer Science, Yale University, New Haven, CT, USA. [4]Department of Internal Medicine, Section of Infectious Diseases, Yale University School of Medicine, New Haven, CT, USA. [5]School of Medicine, The Chinese University of Hong Kong, Hong Kong, China. [6]Functional Genomics Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Department of Computer Science, The Chinese University of Hong Kong, Hong Kong, China. [8]These authors contributed equally: Anurag Sethi, Mengting Gu. ✉e-mail: mark@gersteinlab.org

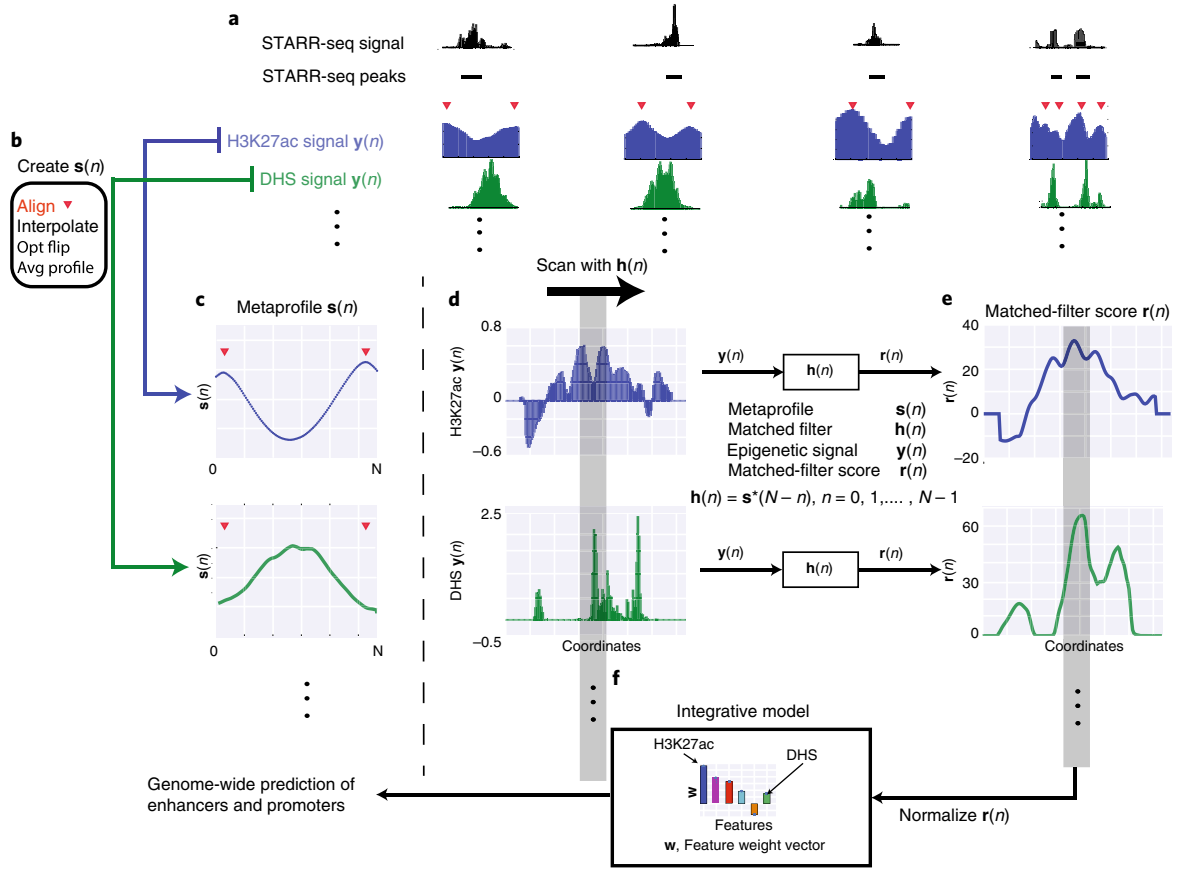


Fig. 1 | Flowchart of the matched-filter model. a, We identified the double-peak pattern in the H3K27ac signal close to STARR-seq peaks. The red triangles denote the position of the two maxima in the double peak. **b**, We aggregated the H3K27ac signal around these regions after aligning the flanking maxima, using interpolation and smoothing on the H3K27ac signal, and averaged the signal across different STARR-seq peaks to create the metaprofile in **c**. The same operations were performed on other histone signals and DHS to create metaprofiles in other dependent epigenetic signals. **d**, Matched filters were used to scan the histone and/or DHS datasets to identify the occurrence of the corresponding pattern in the genome. We use **s** to denote the metaprofile, **h** for the filter, **y** for the epigenetic signal and **r** for the matched-filter score. $\mathbf{s}^*(N-n)$ is the complex conjugate of the flipped metaprofile **s**, where N is the length of the metaprofile. **e**, The matched-filter scores are high in regions where the profile occurs (gray region shows an example), but are low when only noise is present in the data. **f**, The individual matched-filter scores from different epigenetic datasets were combined using integrated model to predict active promoters and enhancers in a genome-wide fashion.

multiple experimental approaches. In the first stage, we predicted enhancers in six embryonic mouse tissues and tested the activity of these predictions in vivo with transgenic mouse assays. We then proceeded to test the activity of these elements in vitro in human cell lines, such as H1 human embryonic stem cells (H1-hESCs), an extensively studied and well-characterized cell line. We showed that the enhancer predictions from our transferrable model are comparable to the prediction accuracy of species-specific models.

Results

Aggregation of epigenetic signals in *Drosophila* to create metaprofiles. We developed a framework to predict active regulatory elements using the epigenetic signal patterns associated with experimentally validated promoters and enhancers (Fig. 1). The STARR-seq studies on *Drosophila* cell lines provide the most comprehensive datasets as they were performed genome-wide and with multiple core promoters[17,32]. These peaks typically consist of a mixture of enhancers and promoters. At this stage, we did not differentiate between the two sets of regulatory elements. As STARR-seq quantifies enhancer activity in an episomal fashion, not all peaks would be active in the native chromatin environment. Arnold and colleagues[17] showed that the STARR-seq peaks that occur with enriched DNase hypersensitivity or H3K27ac modifications tend to be associated with active genes, whereas other STARR-seq peaks tend to be associated with enrichment of repressive marks, such as H3 trimethylated at K27 (H3K27me3). Hence, we took the overlap of the STARR-seq enhancers with H3K27ac and/or DHS peaks to get a high-confidence set of enhancers that are active in vivo, and based on these, we created representative metaprofiles for each histone modification and DNase signal, respectively. During aggregation, we first aligned the two maxima in the H3K27ac signal across active STARR-seq peaks, followed by interpolation of the signal before calculating the average to generate the metaprofile. Then, we calculated the dependent metaprofiles for other histone marks following the same procedure (Fig. 1).

Match of a metaprofile is predictive of regulatory activity. To calculate the matched-filter scores, we first smoothed the input signal track for each epigenetic mark. Then, we scanned the H3K27ac signal track to find each pair of local maximum points between 300 and 1,100 base pairs (bp). Due to the variability of the distance between the double peaks, we interpolated each double-peak region before convolving it with the filter to get an initial score (Extended Data Fig. 1). If there were multiple overlapping double-peak regions, we used the highest score within a 1,500-bp region as the prediction for the regulatory potential. We then calculated the matched-filter

scores for other epigenetic marks on the basis of those same double-peak regions (Methods).

We calculated the matched-filter score for all 30 epigenetic-modification signals available in the *Drosophila* cell lines on STARR-seq peaks and a negative control set (Extended Data Fig. 2). The negatives were randomly chosen regions in the genome that were not STARR-seq peaks and that had the same length distribution as the enhancers from STARR-seq ('Model assessment' in Methods). Interestingly, the distribution of matched-filter scores for STARR-seq peaks was unimodal for each histone mark except for H3K4me1, H3K4me3 and H2Av, which had bimodal distributions. We looked at the degree to which the matched-filter scores for promoters and enhancers were higher than the matched-filter scores for the rest of the genome (Extended Data Fig. 2), as this is a measure of the signal-to-noise ratio for prediction of regulatory regions. We observed that the H3K27ac matched-filter score was the most accurate feature for predicting active regulatory regions identified using STARR-seq (Supplementary Table 1), consistent with previous studies[21,33,34]. In addition, several histone acetylation marks, as well as H1 and H3K4 methylations and DHS, were also accurate prediction features, whereas other histone marks, such as H3K79m1 and H4K20me1, were not well suited, as their matched-filter scores for positive regions and negative regions were not distinguishable.

To quantitatively evaluate whether the occurrence of the epigenetic metaprofiles could be used to predict active enhancers and promoters, we did a tenfold cross-validation assessing the average area under the receiver operating characteristic (AUROC) and area under the precision–recall (AUPR) curves. Comparing the matched-filter result with the peak-calling result, we found that the AUROC and AUPR of the matched-filter scores for different histone modifications were higher than those of the peaks of corresponding histone marks (Fig. 2), suggesting that the matched-filter score is more accurate in predicting active STARR-seq peaks than the simple enrichment of the signals.

Integration of matched-filter scores of multiple epigenetic features. We first combined the matched-filter scores from all 30 measured histone marks along with the DHS in statistical models, such as random forest and support vector machine (SVM) (Extended Data Fig. 3). We evaluated the performance of the integrated model using tenfold cross-validation. For each fold of validation, 90% of the positives and negatives were used to build metaprofiles for each epigenetic mark, generate matched-filter scores and train the integrative model. The remaining 10% of data were used to test model accuracy. The integrated models with 30 epigenetic features displayed high accuracy (average AUROC = 0.97 and AUPR = 0.93 for SVM model with multiple core promoters). We obtained the feature coefficients or Gini scores of each epigenetic mark from the integrated models.

We then built an integrated model with combined matched-filter scores of six commonly available and discriminative epigenetic marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac and DHS) associated with active regulatory regions using a linear SVM[35]. The selection of these six features was based on their matched-filter score performance, their importance in the integrated model and data availability ('Feature selection' in Methods). We then assessed the performance of different statistical approaches, including random forest, ridge regression, Naive Bayes and SVM to combine the features. While all these approaches performed similarly (Extended

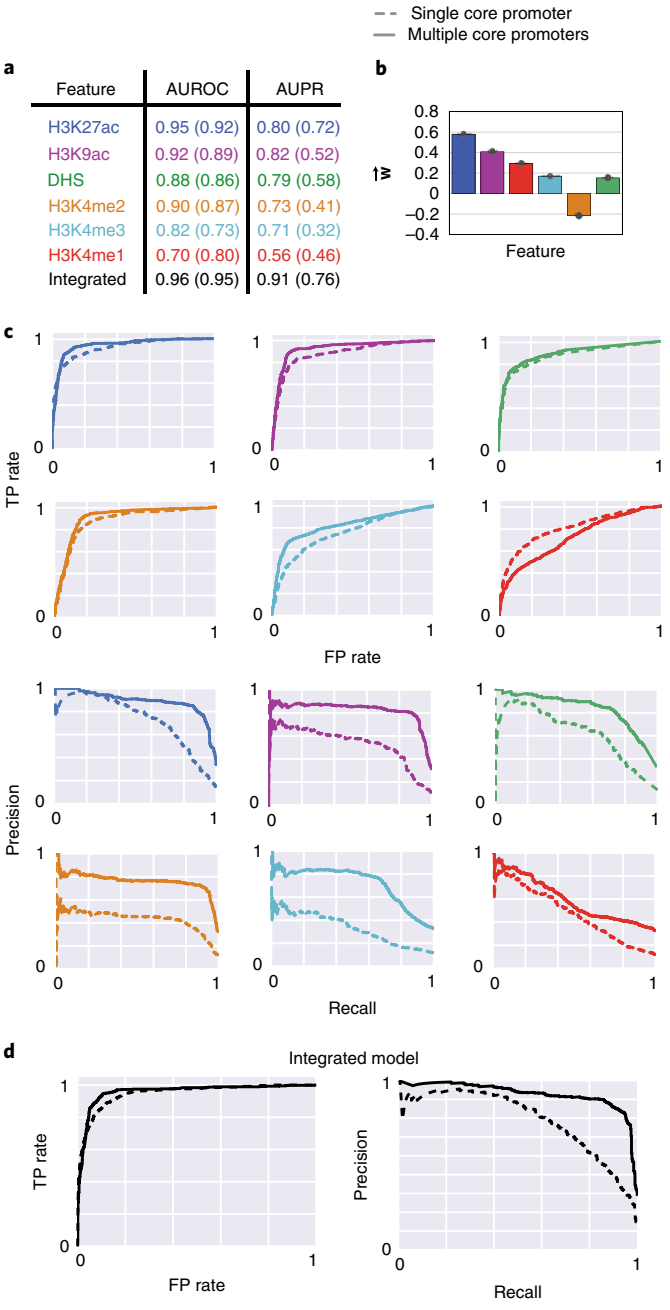


Fig. 2 | Performance of matched filters and integrated models for predicting STARR-seq peaks, compared with that of peak-based models. The performance of the matched filters of different epigenetic marks and the integrated model for predicting all STARR-seq peaks was compared using tenfold cross-validation. **a**, The AUROC and AUPR curves were used to measure the accuracy of different matched filters and the integrated model. **b**, Weights of the different features in the integrated model are plotted; the mean value is displayed in the bar plot, and the error bars show the s.d. of feature weights measured by tenfold cross-validation. These weights may be used as a proxy for the importance of each feature in the integrated model. **c**,**d**, 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 a single core promoter were compared with the performance of peak-based models. The colored numbers within the parentheses in **a** refer to the AUROC and AUPR for predicting the peaks using a single STARR-seq core promoter; the colored numbers outside the parentheses refer to the performance of the model for predicting peaks from multiple core promoters; the gray numbers in the parentheses refer to the performance of the peak-based models. TP rate, true-positive rate; FP rate, false-positive rate.

Data Fig. 4), we used a linear SVM in our framework because its performances were the most stable in cross-validations.

We found that the simplified SVM model had a high performance similar to that of the full SVM model using all 30 epigenetic marks, with an AUROC of 0.96 (0.97 in the full model) and an AUPR of 0.91 (0.93 in the full model). We also trained an SVM model using all STARR-seq peaks (including those with no DHS and H3K27ac signals) with the same six features. We found that H3K27ac still had the highest Gini score in random forest, albeit with a slightly smaller coefficient in SVM (Supplementary Fig. 2). In general, the integrated model trained on the six features achieved good performance upon cross-validation, and this set of input features allowed the integrated model to be applied to a variety of cell lines and tissues, as many relevant chromatin immunoprecipitation with sequencing (ChIP–seq) and DNase experiments have been performed by the Roadmap Epigenomics Mapping[36] and the ENCODE[37] Consortium in a wide variety of samples.

Distinct epigenetic signals associated with promoters and enhancers. We created individual metaprofiles and machine-learning models for the two classes of regulatory activators—promoters (or proximal) and enhancers (or distal). We assessed the performance of the matched filters for predicting active regulatory regions within each category (Fig. 3). We also combined the peaks identified from multiple STARR-seq experiments of S2 cells and reassessed the performance of the matched filters at predicting promoters and enhancers, respectively. Merging the STARR-seq peaks from multiple core promoters led to higher AUROC and AUPR scores for the matched filters of most histone marks (Supplementary Table 2). The highest matched-filter scores were typically observed on promoters, and the matched filters for each of the six features tended to perform better for promoter prediction. We observed, similar to what was found in previous studies[38,39], that the H3K4me1 metaprofile was very predictive for enhancers but was close to random for predicting promoters. In contrast, the H3K4me3 metaprofile could be utilized to predict promoters and not enhancers. The histogram for matched-filter scores showed that the H3K4me1 matched-filter score was higher near enhancers, whereas the H3K4me3 matched-filter score tended to be higher near promoters. The mixture of these two populations led to bimodal distributions for H3K4me1 and H3K4me3 matched-filter scores when calculated over all regulatory regions (Extended Data Fig. 2).

We again trained different statistical models to learn the combination of features associated with promoters and enhancers, respectively. These integrated models outperformed the individual matched filters at predicting active enhancers and promoters (Fig. 3 and Extended Data Fig. 5). In addition, the weights of the individual features identified the difference in the roles of H3K4me1 and H3K4me3 matched-filter scores at discriminating active promoters and enhancers from inactive regions in the genome. The trained promoter-specific model has a high weight for H3K4me3, which is considered a marker for promoters[33], but a lower weight for H3K4me1, which is considered a marker for enhancers[33]. This result is reversed in the enhancer-specific model, indicating the unique features that were captured for different identification tasks (Supplementary Figs. 4 and 5). We also created two integrated models utilizing matched-filter scores of all 30 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 (Extended Data Fig. 6).

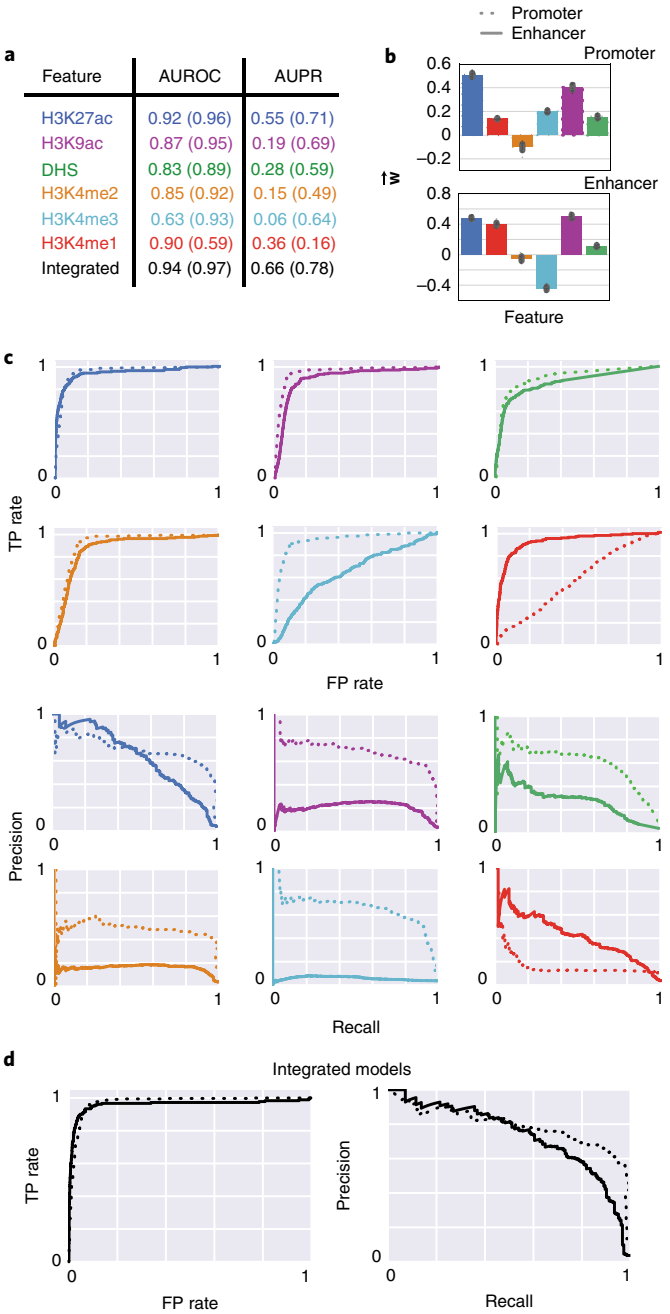


Fig. 3 | Performance of matched filters and integrated models for predicting promoters and enhancers. The performance of the matched filters of different epigenetic marks and the integrated model for predicting active promoters and enhancers were compared using tenfold cross-validation. **a**, The numbers within parentheses refer to the AUROC and AUPR for predicting promoters; the numbers outside the parentheses refer the performance of the models for predicting enhancers. **b**, Weights of the different features in the integrated models for promoter and enhancer prediction are plotted; the mean value is displayed in the bar plot, and the error bars show the s.d. of feature weights measured by tenfold cross-validation. **c**,**d**, The ROC and PR curves for each matched filter and the integrated model are shown. The performance of these features and that of the integrated model for predicting the active promoters and enhancers using multiple core promoters were compared.

Application of the STARR-seq model to predict enhancers in mammalian species. One of the important findings of previous ENCODE and model-organism ENCODE efforts was the conserved patterns of chromatin marks close to regulatory elements across hundreds of millions of years of evolution[25–31]. The relationship of chromatin marks to gene expression was very similar, for instance, in worms, flies, mice and humans. Therefore, it is possible

to build a statistical model relating chromatin modification to gene expression that would work without re-parameterization for different organisms. This motivated us to transfer our well-parameterized model based on the STARR-seq data from flies to mammalian systems (for example, mouse and human) and to test our model's performance.

We started by making genome-wide predictions of mouse regulatory regions. Predictions were made in six different tissues (forebrain, midbrain, hindbrain, limb, heart and neural tube) at embryonic day 11.5 (e11.5) (predictions are available on our website at http://matchedfilter.gersteinlab.org). Using our model, we predicted 31,000 to 39,000 regulatory regions in individual tissues in mouse, with each region ranging from 300 bp to 1,100 bp. Similarly, we performed a genome-wide prediction of regulatory regions in the ENCODE top-tier human cell lines, including H1-hESC, GM12878, K562, HepG2, A549 and MCF-7. In H1-hESC, for example, we predicted 43,463 active regulatory regions, of which 22,828 (52.5%) were within 2 kilobase pairs of the transcription start site and were labeled as promoters. Most of the predicted regulatory regions were also present near active genes (Extended Data Fig. 7).

Validation in vivo in mice. To test the activity of predicted mouse enhancers in vivo, we performed transgenic mouse enhancer assays (Extended Data Fig. 8) in e11.5 mice for 133 regions, including 102 regions selected on the basis of the H3K27ac-signal rank of the corresponding mouse tissues, and another 31 regions selected by an ensemble approach from human homolog sequences (Supplementary Tables 4–9). In addition, we included other published transgenic mouse experiments from the VISTA database for validation. In total, we had 1,253 positive regions and 8,631 negative regions pulled together from different tissues. This large set of validated enhancers allowed us to comprehensively evaluate the predictability of the matched-filter scores of each epigenetic mark, as well as the integrated SVM model (Fig. 4). On average, the integrated model trained with *Drosophila* STARR-seq data achieved an AUROC of 0.8. We did a similar evaluation with publicly available FIREWACh assay data[40] from mice, and the outcome was in accordance with our other results (Extended Data Fig. 9). For comparison, we trained an integrated model based directly on the validated mouse enhancers. We observed a similar prediction accuracy upon cross-validation (Extended Data Fig. 10).

Validation in human cell lines. We validated our STARR-seq-based model for predicting human enhancers using a cell-line-based transduction assay (Supplementary Methods). We randomly selected 20 predicted intergenic enhancers for validation. Insertion of 11 of the putative enhancers into the HIV vector resulted in a significant increase in enhanced GFP expression ($P < 0.05$ for both directions) in H1-hESCs (Supplementary Table 10 and Supplementary Data 1). The positive enhancers displayed a significant increase in gene expressions in both orientations. In contrast, the negatives displayed much lower levels of gene expression (Extended Data Fig. 11). The activity of these tested enhancers also showed cell-type specificity. More than half of the predicted enhancers show activity in H1-hESCs (Extended Data Fig. 12), but less in A549 and TZM-bl cells, which are derived from tumor cells (Supplementary Table 10). Overall, 16 of the 20 tested predictions displayed a statistically significant increase in gene expression of the reporter gene in at least one of the cell lines. Given the promoter specificity of enhancers in such assays, we anticipate that some of the elements that could not be validated in this particular vector would function as enhancers in a more natural biological context (for example, with the cognate promoter or in the absence of surrounding HIV vector sequences).

TFs exhibit different binding patterns at enhancers and promoters. We further studied the differences in TF binding at promoters and enhancers (Fig. 5). We focused on the human H1-hESC cell line, as there is a large amount of functional genomic assays from the ENCODE[37] and Roadmap Epigenomics Mapping Consortium[36] of this cell line. Together, the consortia have generated ChIP–seq data for 60 transcription-related factors in the H1-hESC cell line, including a few chromatin remodelers and histone-modification enzymes. Collectively, we call these transcription-related factors 'TFs' for simplicity.

We showed that the patterns of TF binding within regulatory regions could be utilized in a logistic-regression model to distinguish active enhancers from promoters with high accuracy (AUPR = 0.90, AUROC = 0.87) (Fig. 5). We were also able to identify the most important features that distinguish promoters from enhancers. In addition to TATA-box-associated factors such as TAF1, TAF7, and TBP, the RNA-polymerase-II binding patterns as well as those for chromatin remodelers such as KDM5A and PHF8 are some of the most important factors that distinguish promoters from enhancers in H1-hESCs. This provides a framework that can be utilized to identify the most important TFs associated with active enhancers and promoters in each cell type.

We found that although most promoters and enhancers contain multiple TF-binding sites, the pattern of TF binding at promoters was different than that at enhancers, and that TF binding at enhancers displayed more heterogeneity: more than 70% of the promoters bound to the same set of 2 or 3 sequence-specific TFs, which was not observed for enhancers (Fig. 5c). The majority of the promoters contained peaks for several TATA-associated factors (TAF1, TAF7 and TBP). These TF coassociations could lead to mechanistic insights of cooperativity between TFs. Similarly, CTCF and ZNF143 may function cooperatively as they are observed to co-occur frequently at distal regulatory regions, consistent with a previous report[41].

Discussion

In this study, we developed a framework using transferable supervised machine-learning models trained on regulatory regions identified by STARR-seq to accurately predict active enhancers in a cell-type-specific manner. The rich amount of whole-genome STARR-seq experiments established the characteristic pattern flanking active regulatory regions within certain histone modifications[17]. This motivated us to train a shape-matching and filtering model that could be used to identify these patterns in the ChIP–seq signals. As the chromatin marks and epigenetic profiles associated with active regulatory regions are highly conserved among organisms[25–31], we showed that a well-parameterized model in one model organism can be transferred to another with high prediction accuracy.

While STARR-seq provides a genome-wide unbiased test of the enhancer activity of putative sequences, it is intrinsically episomal and thus does not completely reveal the enhancer activity in the native chromatin environment. Selecting for chromosomally active enhancers using H3K27ac and DHS could introduce subtle biases in model training. To address this issue, we employed very different experimental techniques and provided orthogonal validations. This included in vivo transgenic assays and in vitro transduction assays, in which the predicted regions were tested for regulatory activity in the native chromatin environment. With these orthogonal validations, we were able to comprehensively assess our tissue-specific predictions in six tissues in mice. With multiple comparisons to other published methods trained directly on mouse data, we have shown that the matched-filter model is transferable with high accuracy in predicting active enhancers in mouse tissues. The in vitro transduction assays were performed in H1-hESCs and three other human cell lines to validate the human regulatory-element predictions. The majority of the predicted elements displayed a substantial increase in expression of the reporter gene, further confirming the capability of our model to make enhancer predictions in mammals.

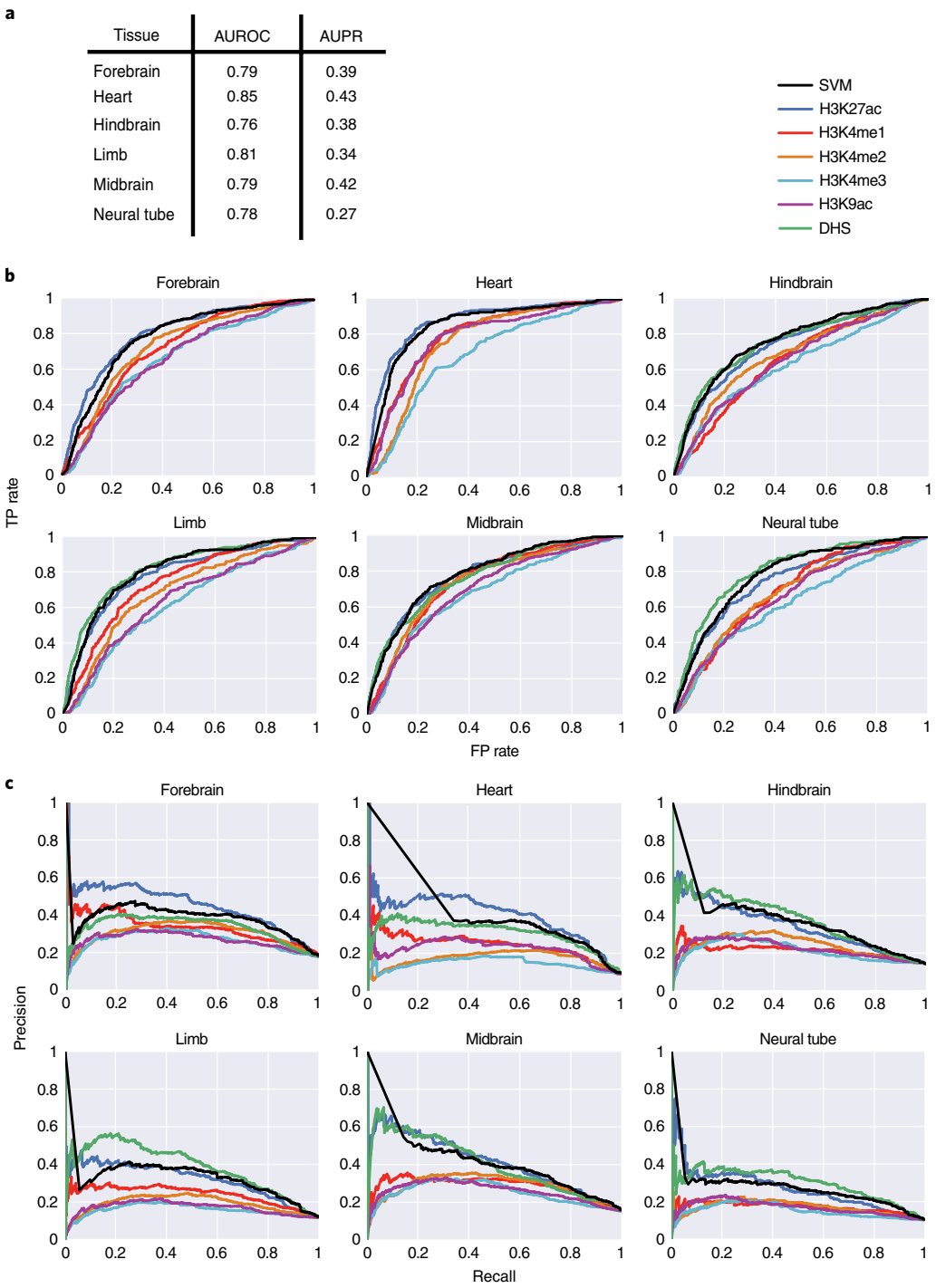


Fig. 4 | Performance of matched filters and integrated model for predicting active enhancers in mice. The performance of the *Drosophila* STARR-seq-based matched filters and the integrated model for predicting active enhancers identified by enhancer assays in six different tissues of e11.5 transgenic mice. **a**, The AUROC and AUPR are shown for the integrated SVM model in six tissues. The weights of the different features in the integrated model are the same as the weights shown in Fig. 3 for enhancers. **b**, The individual ROC curves of each feature and the integrated SVM model for each tissue are shown. **c**, The individual PR curves of each feature and the integrated SVM model for each tissue are shown.

Our predictions depend on the availability of high-quality histone ChIP–seq datasets in the relevant tissue or cell type of interest. It may be impossible to produce such high-quality datasets for different human tissues and developmental timelines.

Recently, genome-wide STARR-seq has been applied to mammalian systems such as HeLa-S3 cells[42]. In the future, we expect that more extensive whole-genome STARR-seq datasets will become available in mammalian systems. It could be advantageous to re-train the matched filter model on state-of-the-art datasets. With the set-up of our framework, re-training the model with newly generated datasets should be straightforward. We envision that our framework would benefit from these datasets and generate more comprehensive regulatory-element annotations across eukaryotic species.

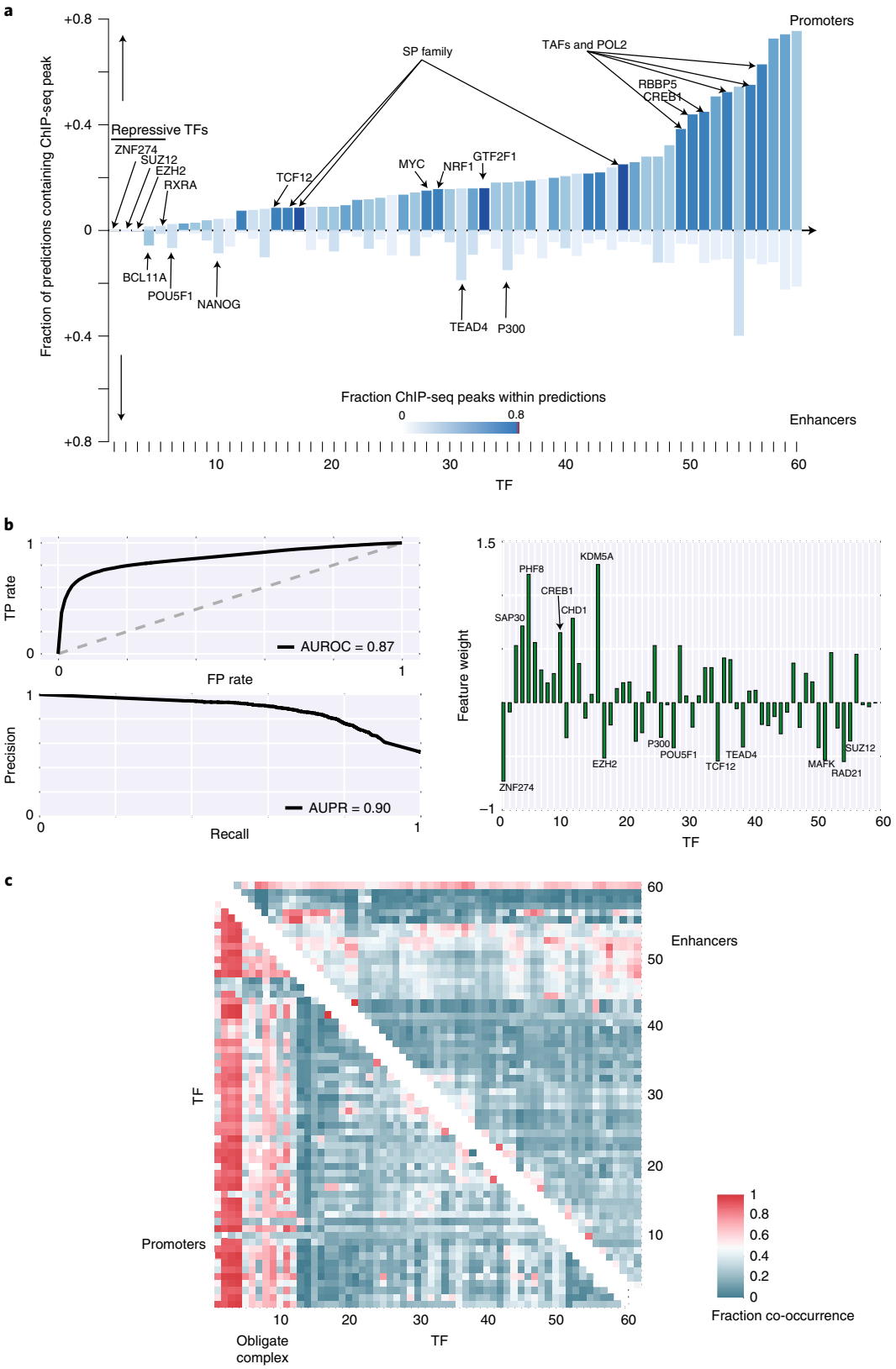


Fig. 5 | Differences in TF-binding patterns at enhancers and promoters. a, The fraction of predicted promoters and enhancers that overlap with ENCODE ChIP–seq peaks for different TFs in H1-hESC are shown. The names of all TFs in the figure can be viewed in Supplementary Fig. 11. **b**, The AUROC and AUPR for a logistic-regression model created using the pattern of TF binding at each regulatory region to distinguish enhancers from promoters are shown. The weight of each feature in the logistic-regression model could be used to identify the most important TFs that distinguish enhancers from promoters. **c**, The patterns of TF cobinding at active promoters and enhancers are shown. The TFs that co-occur at promoter regions tend to form obligate complexes. The names of all the TFs in this graph can be viewed in Supplementary Fig. 12.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0907-8.

Received: 25 September 2017; Accepted: 18 June 2020;
Published online: 29 July 2020

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
2. Levo, M. et al. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25**, 1018–1029 (2015).
3. Slattery, M. et al. Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
4. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6**, 85 (2014).
5. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
6. Wray, G. A. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.***8**, 206–216 (2007).
7. Erwin, G. D. et al. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.* **10**, e1003677 (2014).
8. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet* **14**, 288–295 (2013).
9. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
10. Visel, A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet* **40**, 158–160 (2008).
11. Nord, A. S. et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521–1531 (2013).
12. Visel, A. et al. ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
13. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
14. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol.* **10**, e1003711 (2014).
15. Narlikar, L. et al. Genome-wide discovery of human heart enhancers. *Genome Res.* **20**, 381–392 (2010).
16. Yip, K. Y. et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
17. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
18. Maston, G. A., Landt, S. G., Snyder, M. & Green, M. R. Characterization of enhancer function from genome-wide analyses. *Annu. Rev. Genomics Hum. Genet.* **13**, 29–57 (2012).
19. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet* **15**, 272–286 (2014).
20. Yanez-Cuna, J. O. et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014).
21. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
22. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
23. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet* **21**, 292–310 (2020).
24. Liu, Y. et al. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.* **18**, 219 (2017).
25. Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).
26. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* **40**, 553–568 (2012).
27. Cheng, Y. et al. Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371 (2014).
28. Dong, X. et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
29. Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
30. Gjoneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
31. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
32. Zabidi, M. A. et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
33. Cotney, J. et al. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* **22**, 1069–1080 (2012).
34. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
35. Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167 (1998).
36. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
37. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
38. Koch, C. M. et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* **17**, 691–707 (2007).
39. Rajagopal, N. et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.* **9**, e1002968 (2013).
40. Murtha, M. et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods* **11**, 559–565 (2014).
41. Bailey, S. D. et al. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).
42. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).

Methods

Creation of metaprofile. A metaprofile is a template used to estimate the signal distribution on active enhancers for one epigenetic signal. We evaluated whether we could utilize the metaprofiles to predict active promoters and enhancers using matched filters (Fig. 1). Matched filter is a well-established pattern recognition algorithm that uses a shape-matching filter to recognize the occurrence of a template in the presence of stochastic noise[43]. We started with creating the metaprofiles, which we generally denote as $\mathbf{s}(n)$, based on experimentally validated active enhancers. The STARR-seq studies on *Drosophila* cell lines provide the most comprehensive datasets, as they were performed genome wide and with multiple core promoters[17,32]. These peaks typically consist of a mixture of enhancers and promoters. At this stage, we did not differentiate between the two sets of regulatory elements. As STARR-seq quantifies enhancer activity in an episomal fashion, not all peaks would be active in the native chromatin environment. Arnold and colleagues[17] showed that the STARR-seq peaks that occur with enriched DNase hypersensitivity or H3K27ac modifications tend to be associated with active genes, whereas other STARR-seq peaks tend to be associated with an enrichment of repressive marks such as H3K27me3. Hence, we took the overlap of the STARR-seq enhancers with H3K27ac and/or DHS peaks to get a high-confidence set of enhancers that are active in vivo, upon which we created representative metaprofiles for each histone modification and DNase signal, respectively.

We utilized the smoothed histone signal tracks for the *Drosophila* S2 cell line provided by the modENCODE consortium[44] to create metaprofiles for ChIP–seq signals. The genome-wide profile for open chromatin (DNase-seq or DHS) for the S2 cell line was calculated on the basis of experiments by the Stark lab[17]. To create the metaprofiles, we aligned active STARR-seq peaks with identifiable double-peak patterns of the H3K27ac signal and aggregated the signals in the S2 cell line (Fig. 1b). Aggregation of signals over a large number of enhancers reduced the noise in the metaprofiles. 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 was accepted if it had the lowest signal within a 100-bp region in the H3K27ac signal track. We then proceeded to identify the flanking maxima (both sides of the minimum) within a total of 2-kilobase-pair (2-kb) region around the STARR-seq peak (1 kb in each direction from the center of the STARR-seq peak). These maxima were accepted only if they had the highest signal within a 100-bp region in the H3K27ac signal track.

Approximately 70% of the active STARR-seq peaks contained an identifiable double peak within the H3K27ac signal, although there was variability in the distance between the 2 maxima of the double peak in the ChIP–chip signal (Extended Data Fig. 1a). While the minimum tended to occur in the center of these 2 maxima on average, the distance between the 2 maxima in the double peaks varied between 300 and 1,100 bp. During aggregation, we first aligned the two maxima in the H3K27ac signal across active STARR-seq peaks. We then interpolated the signal with a cubic-spline fit so that the signal track contained an equal number of points for each double-peak region. All interpolation and smoothing steps were performed using the scipy module in Python. The aggregated signal tracks were averaged to create the metaprofile for the double-peak regions. While the signal tracks were aggregated on the basis of identifying the double-peak regions in the H3K27ac signal track, the same set of operations could be performed with any epigenetic mark expected to have the double-peak pattern flanking regulatory regions.

We calculated the metaprofiles of ~30 other epigenomic datasets (histone marks and DHS signal). These metaprofiles were calculated by aggregating the corresponding ChIP–seq or DHS signals based on the same regions where H3K27ac double peaks were identified, so the matched-filter scores of each epigenetic mark were calculated on the same regions in the integrated model. We observed that the metaprofiles for some epigenetic marks also showed a double-peak pattern, and the maxima across different histone-modification signals tended to align with each other on average, likely because these epigenetic marks flank enhancers in a similar pattern as H3K27ac does (Extended Data Fig. 1). This indicates that a large number of histone modifications would simultaneously co-occur on the nucleosomes flanking an active enhancer or promoter. In contrast, the repressive histone marks did not contain a double-peak pattern, so they did not have the same epigenetic template associated with enhancers. The DHS signal, as expected, displayed a single peak at the center of the H3K27ac double peak.

Matched-filter algorithm. The epigenetic signal at enhancers and promoters can be approximated as the linear superposition of background noise and the metaprofile $\mathbf{s}(n)$ learned in Fig. 1. To identify the occurrence of the metaprofile with the presence of noise, we adopted the canonical signaling processing method known as matched filter. The matched-filter process convolves the signal $\mathbf{y}(n)$ with the filter $\mathbf{h}(n)$. Before calculating the matched-filter score, the signal was interpolated to ensure that the scanned region contained the same number of points as the metaprofile:

$$\mathbf{r}(n) = (\mathbf{y} \star \mathbf{h})(n) = \sum_{i=n-N}^{n} \mathbf{y}(i)\mathbf{h}(n-i)$$

where * stands for convolution and $\mathbf{r}(n)$ is the resulting matched-filter score. The matched filter is defined as the conjugated reverse of the metaprofile template:

$$\mathbf{h}(x) = \mathbf{s}^*(N-x)$$

where N is the total number of points in the template and * denotes the complex conjugate.

There was 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 (Extended Data Fig. 1). As a result, we scanned different spans of the genome with the matched filter (distance between the 2 peaks were allowed to vary between 300 and 1,100 bp) and took the highest score as the matched-filter score for that region. Matched filter recognizes the given template in a signal in the presence of noise with the highest signal-to-noise ratio[43]. At positive regions, the presence of the metaprofile within the signal leads to high matched-filter scores. At background regions where the signal is mostly comprised of noise, the matched-filter score is low.

Statistical learning models. We built an integrated model to include matched filter scores from multiple epigenetic signals for more accurate enhancer prediction. The matched filter scores from each epigenetic signal are first normalized. The distribution of matched-filter scores in random negative regions for a particular histone mark is approximately Gaussian and it represents the background distribution in the genome. The z-scores of matched-filter scores with respect to the negatives (random regions of genome) were used as input features for training different statistical learning models. The z-score of the matched-filter score is defined as

$$z = \frac{r-\mu}{\sigma}$$

where r is the matched-filter scores, and μ *and* σ are the mean and s.d. of the Gaussian fit to the matched-filter scores for random regions in genome.

We have tested different statistical learning models, including the SVM[45], ridge regression[46], random forest[47] and Gaussian Naive Bayes[48] models. For SVM, we utilized a linear kernel to distinguish between positives and negatives. The linear SVM identifies a decision boundary that maximally separates the regulatory regions and the random regions of the genome from the decision boundary. Ridge regression is a linear-regression technique that prevents overfitting by penalizing large weights for each feature. Random forest is an ensemble learning method that operates by constructing a large number of decision trees and outputting the mean prediction of different decision trees. We used thousand trees for creating our enhancer and promoter prediction models. The naïve Bayes classifier is a family of simple probabilistic classifiers that assumes that all the features are independent of one another. We used scikit-learn[49] with default parameters for training and assessing the performance of all the statistical models. In the main text, we discussed the results of the support vector machine (SVM) model, which showed high performance, and low variance in performance upon cross-validation.

Feature selection. We selected the features to use in our framework by assessing their individual performance with matched filter, their importance in the integrative model, and their general data availability in mammalian systems. Specifically, the ability to distinguish enhancers from negative regions of each feature is shown in Extended Data Fig. 2 and Supplementary Table 1. We found that some histone marks, such as H3K27ac, give very different score distributions for the enhancer regions and random regions, whereas others, such as H3K79m1 and H4K20me1, have indistinguishable score distributions on these two categories of region.

For the importance of each feature in the integrative model, we trained an SVM model, a random forest mode, and a ridge regression using all 30 epigenetic marks, and assessed the importance of each feature using their feature coefficient or Gini score. Among these 30 features, H3K27ac, H3K4me1, H3K4me3 and H3K9ac showed high feature coefficients or a high GINI score in all three models; DHS and H3K4me2 had high GINI scores and were also widely used in previous literature to identify promoters and enhancers. In contrast, other histone marks, such as the repressive mark H3K27me3, show little contribution to the integrated model, as indicated by the Gini score and the feature coefficients.

Finally, as the 30 histone marks we tested were from *Drosophila* experimental data, many of them were unavailable in even top-tier tissues and cell lines for mouse and human. For example, H2BKac performed well in matched filter, and had a very high feature coefficient in each model, but the ChIP–seq experiment data are generally unavailable in mammalian cell lines. As our goal was to build a model with broad applicability across organisms, we decided to not include these epigenetic marks (for example, H2BK5ac, H4ac and H4K12ac) for now, but if more study is done on these histone marks in the future, we can easily include them in our framework. After filtering, we found six features that satisfied all three above criteria, namely, H3K27ac, H3K4me1, H3K4me1, H3K4me3, H3K9ac and DHS. Integrating these 6 features in the linear SVM model yielded a high performance (AUROC of 0.96, AUPR of 0.91) similar to that of the complete SVM model using all 30 epigenetic marks (AUROC of 0.97, AUPR of 0.93). We subsequently tested

the performance of this simplified model in *Drosophila* cells, mouse tissue and human cells.

In the six-feature model, the DHS signal has lower weight than the other five features (Fig. 2). It should be noted that the matched filter on DHS signal performed well on its own. The lower weight is likely due to the fact that the information in DHS is redundant with the information contained within the histone mark (for example, the DHS peaks usually occur at the trough region between two maxima in the histone signal). Despite the redundancy, the combination of the DHS and histone signals was more predictive of regulatory activity because the reinforcing signals strengthened the prediction as compared with the uncorrelated noise.

Model assessment. In order to assess the accuracy of the matched-filter model for predicting enhancers and promoters, we used a tenfold cross-validation. The STARR-seq positives and negatives were randomly divided into ten groups. For each fold of cross-validation on a single histone mark, the profiles were created with 90% of the STARR-seq positives, and the remaining 10% of the positives were used for testing the accuracy of the model. Similarly, In the integrative SVM model, the SVM was trained on 90% of the data in each fold of cross-validation, whereas the remaining 10% of the positives were used to test accuracy.

We quantified our model performance with AUROC and AUPR curves. In the ROC curve, the TP rate was plotted against the FP rate at different thresholds in the statistical model. The TP rate is defined as the number of true positives identified by the model divided by the total number of positives. The FP rate is defined as the fraction of negatives misclassified as positives by the model, divided by the total number of negatives. When comparing the performance of two different classifiers in the ROC curve, the classifier with a higher TP rate at the same FP rate is considered to be a better classifier. The AUROC is a single measure for the accuracy of a model, as models with a higher AUROC are generally considered to be better models.

In the PR curve, the precision was plotted against recall at different thresholds in the statistical model. The recall is the same as the TP rate of the model (that is, the number of true positives identified by the model divided by the total number of positives). The precision is the fraction of positives predicted by the model that are correct (that is, the number of true positives identified by the model divided by the total number of positives predicted by the model). The AUPR is another measure of performance of a model. If the AUPR is high, the corresponding model has a low false-discovery rate and can better distinguish the positives from the negatives. PR curves are particularly useful to assess the performance of classifiers in skewed or imbalanced datasets in which one of the classes is observed much more frequently than the other class is[9]. For such skewed datasets, the AUROC values for two different models may be very similar even though they actually differ in performance with respect to their precision. Hence, the AUPR is a better reflection of the performance difference between two models with a similar AUROC in skewed datasets.

In Fig. 2, the positives are defined as the active peaks (intersecting with DHS or H3K27ac peaks) from a single STARR-seq experiment (single core promoter) or the union of active peaks from multiple STARR-seq experiments (multiple core promoters). The negatives are randomly chosen non-STARR-seq-peak regions in the genome that had the same lengths distribution as the enhancers from the STARR-seq. We required most of the regions to contain some H3K27ac signals, as negatives with no H3K27ac signal wouldn't provide enough information for training. We typically chose five to ten times the number of negatives than the number of positives in Figs. 2–4, as the number of enhancers and promoters in the genome (positives) is far less than the number of negatives; moreover, the AUPR is dependent on the ratio of negatives to positives during the tenfold cross-validation.

To evaluate the impact of the training-sample size on model performance, we did a saturation analysis in which we down-sampled the training data to different levels of fractions and evaluated the model performance on the remaining data. For each down-sampling fraction from 10% to 90%, with 10% as the step, we performed the 10-fold cross-validations. In each fold, the whole model, including the aggregation of signals, was based on the training dataset. The performance was tested on the remaining data and was independent of the training data. We found that the average AUPR increased with an increasing size of the training data. The AUPR of the SVM model started to saturate with 80–90% of the experiment data for training (Extended Data Fig. 4). The average AUROC remained comparable, although the variances decreased with increasing training-data size. This might suggest that a fivefold cross-validation would be sufficient.

Promoters and enhancers. In the STARR-seq experiment, each peak functions as an enhancer within the plasmid environment in the S2 cell line. However, to delineate the native role of the region, we classified them as promoters and enhancers on the basis of their distance to the transcription start sites (TSSs) in the genome. In Fig. 3, the active promoters were defined as active STARR-seq peaks (multiple core promoter) within 1 kb of a TSS (Ensembl release 78); enhancers were defined as active STARR-seq peaks more than 1 kb from any TSS in *Drosophila*. However, a few of the promoters may also regulate distal genes in addition to their promoter activity[50].

Validating enhancers in mammalian species. We downloaded tissue-specific epigenetics data from the ENCODE portal (https://www.encodeproject.org). The histone signals were converted to log-fold enrichment (with respect to control signal). We ran the integrated matched filter to get the enhancer and promoter predictions for six different mouse tissues (forebrain, midbrain, hindbrain, limb, heart and neural tube) at the e11.5 stage (genome-wide predictions are available through our website at https://goo.gl/E8fLNN). These tissues were selected as their epigenetic signals have been highly studied in mouse ENCODE, providing us with a rich source of raw data that could be utilized for making enhancer and promoter predictions. In addition, the VISTA database contains close to 100 validated enhancers that could be used to test predictions in each of these tissues. Using our model, we predicted 31,000 to 39,000 regulatory regions in individual tissues in mice, with each region ranging from 300 bp to 1,100 bp. Notably, a consistent proportion of two-thirds (66–70%) of these predicted regulatory regions were distal regulatory elements for all 6 tissues, with the other one-third (30–34%) being proximal regulators (Supplementary Table 10). These numbers agree with a previous enhancer evolution study[51], and suggest that the amounts of enhancers and promoters are likely comparable in different tissues.

Similarly, we performed a genome-wide prediction of regulatory regions in the ENCODE top-tier human cell lines, including H1-hESC, GM12878, K562, HepG2, A549 and MCF-7. Predicted active regions within 2 kb of any TSS were annotated as promoters, and regions that were more than 2 kb from any TSS were annotated as enhancers. The distribution of the expression of the closest gene (GENCODE v19 TSS[52]) from the ENCODE RNA-seq dataset for H1-hESCs was compared with the expression of all genes from H1-hESCs. The Wilcoxon test was used to measure the significance of changes in gene expression.

To assess the predictions, we ranked all the tested candidate elements by either the matched-filter scores of individual features, or the final prediction (probability of being an enhancer) from the integrated SVM model. We then took the labels of the candidate elements from the experiment readout to assess the predictions using ROC and PR curves.

Validation in mouse embryos. In Fig. 4, the enhancers were tested by transgenic mouse reporter assays[9,53]. Predicted enhancers were PCR amplified and cloned into a plasmid upstream of a minimal hsp68 promoter and a *lacZ* reporter gene. Resulting plasmids were linearized and injected into single-cell FVB/NCrl strain *Mus musculus* embryos. After reimplantation into surrogate mothers, resulting embryos were collected at e11.5, stained for β-galactosidase activity, and imaged. Elements were scored positive for enhancer activity if at least three resulting transgenic embryos had reporter-gene expression in the same tissue and pattern. Elements were scored negative if at least five transgenic embryos were recovered and no reproducible staining pattern was observed.

All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory (LBNL) Animal Welfare Committee. All mice used in this study were housed at the Animal Care Facility (ACF) at LBNL. Mice were monitored daily for food and water intake, and animals were inspected weekly by the Chair of the Animal Welfare and Research Committee and the head of the animal facility in consultation with the veterinary staff. The LBNL ACF is accredited by the American Association for the Accreditation of Laboratory Animal Care International.

Validation in human cell lines. We used a third-generation, self-inactivating (SIN) HIV-1-based vector system in which the enhanced GFP (eGFP) reporter was driven by the DNA element of interest to test putative enhancers after stable transduction of four cell lines, including H1-hESCs (Extended Data Fig. 11). The predicted enhancers were PCR amplified from human genomic DNA and separately inserted immediately upstream of a basal Oct-4 promoter of 142 bp within the self-inactivating (SIN) HIV vector. Each putative enhancer was tested in triplicate for both forward and reverse orientation in H1-hESCs. We used empty SIN HIV vector and FG12 as the negative and the positive control, respectively. Note that the empty vector had the basal Oct-4 promoter, along with the IRES-eGFP reporter cassette. We assessed putative enhancer activity by flow-cytometric readout of eGFP expression 48–72 h post-transduction, normalized to the negative control.

We selected a total of 23 predicted intergenic enhancers for validation. These predictions were chosen at random to ensure that they truly represented the whole spectrum of predicted enhancers and not just the top tier of predicted enhancers. Of these 23 putative enhancers, 20 were successfully PCR-amplified and cloned into the SIN HIV vector in both directions. To measure the distribution of gene expression in the absence of enhancer, we also amplified and cloned 20 non-repetitive elements with a similar length distribution that were predicted to be inactive into the same SIN HIV vector. All positive and negative DNA elements were transduced and tested for activity in both forward and reverse orientations, as enhancers are thought to function in an orientation-independent manner. Following the same procedures, we performed functional testing in duplicate in HOS, TZM-bl and A549 cell lines in addition to H1-hESCs.

Performance comparison with other computational methods. We compared the performance of the matched filter to the peak-based models of the different epigenetic marks (Fig. 2). We used the histone (or DHS) peaks that overlapped with at least 50% (10%) of the STARR-seq peak to rank that prediction. We used a

smaller threshold for DHS peaks as they are much shorter than histone peaks. We achieved similar results with thresholds of 25% for both histone and DHS peaks. The *P* value of the intersecting peak was used to rank the peak-based predictions. The modENCODE histone peaks and DHS peaks[44] were compared with the matched-filter scores in Fig. 2.

We compared with other published enhancer prediction tools, including ChromHMM, a multivariate hidden Markov model[54]; CSIANN, a neural network based approach[55]; DELTA, an ensemble model integrating different histone modifications[56]; RFECS, a random forest model based on histone modifications[39]; and REPTILE, a more recent published method that integrates histone modifications and whole-genome bisulfite sequencing data[57]. We used the mouse experimental data published in REPTILE for the comparison, and assessed the performance of our method compared with the four published methods mentioned above for all four mouse tissues with available experimental data, ChIP–seq data and DNase data.

Our integrated model outperformed ChromHMM in all four tissues, with an AUROC of 0.76 in hindbrain (versus ChromHMM 0.69), 0.81 in limb (versus ChromHMM 0.75) and so on (Extended Data Fig. 13a). For comparison with supervised algorithms such as CSIANN, DELTA and REPTILE, our method had the highest AUROC in three out of four tissues: hindbrain, limb and neural tube (Extended Data Fig. 13b). In midbrain, the AUROC for our prediction was slightly lower than REPTILE and RFECS, possibly because the DNase experiment performed in midbrain was very noisy; the DNase experiment of mouse e11.5 midbrain was marked as 'low SPOT score' in ENCODE, where SPOT stands for signal portion of tag. We found that, while 75–81% of the genome regions had DNase signals in the other three tissues, only 52% of the genome regions showed DNase signal in the experiment in midbrain. Overall, the comparison shows that our model trained using the *Drosophila* STARR-seq data had better performance than the other methods that were trained directly using mouse experimental data.

For humans, we did not have an extensive amount of validated enhancer data. For comparison, we first checked the overlap of our predicted enhancers with the enhancer predictions from ChromHMM[54] and Segway[58]. We observed that a majority of our predictions overlap with predictions from either of them (Supplementary Figs. 7–10). In addition, we compared our cell-type-specific enhancer predictions with the integrative annotation of ChromHMM and Segway using CAGE-defined enhancers from the FANTOM5 Atlas[59]. We found that the percentage of overlap for our predicted enhancers was more than three times higher than that of the combined ChromHMM and Segway enhancers in each of these cell lines. Despite the fact that our framework predicted a smaller number of enhancers, the number of overlaps was still higher for our predictions. We also compared the predicted promoters from our model with their promoter annotations using FANTOM5 promoter sets. Again, the promoters predicted in our model had a higher fraction of overlaps with the FANTOM promoters (Extended Data Fig. 14). In addition to the integrative ENCODE annotation, we again made comparisons using other supervised enhancer predictions, such as CSIANN[55], DEEP[60] and RFECS[39], using the FANTOM5 enhancer dataset. We found that our predicted K562 enhancers had a similar fraction of overlap with FANTOM5 enhancers compared with that of CSIANN, but the fraction was more than twice as high as that of DEEP and RFECS (Extended Data Fig. 14).

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

To check whether the STARR-seq-based enhancer predictions had different TF-binding patterns, we referred to the fraction of TF occupancy of predicted enhancer from other methods. The comparison demonstrated in Extended Data Fig. 15 shows that the TF-binding pattern of our prediction was very similar to that reported in previous literature[39].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A detailed description of the datasets used in each part of the study is in the corresponding section of Supplementary Methods. Specifically, the *Drosophila* epigenetics datasets used in this study were generated by the modENCODE consortium, available online (http://data.modencode.org). The mouse epigenetics datasets were generated by the ENCODE and Roadmap Epigenomics consortium, available online (https://www.encodeproject.org). We downloaded the *Drosophila* STARR-seq data[28] and the mouse FIREWACh data[32] from previous studies. Results from transgenic-mouse enhancer assays were generated by the Pennacchio lab at LBNL. Experimental results are summarized in Supplementary Tables 4–9, with the mouse images and additional details available on the VISTA Enhancer Browser (https://enhancer.lbl.gov). The human-cell-line enhancer reporter assay results were generated by the Sutton lab at Yale University. Experiment results are summarized in Supplementary Table 10. More detailed results for each cell line are available in Supplementary Data 1.

Code availability

We have implemented our methods in Python. The source code and the output annotations referenced in the paper are available at the website http://matchedfilter.gersteinlab.org. A dockerized image is also provided at this site.

References

43. Kumar, V. B. V. K., Mahalanobis, A. & Juday, R. D. *Correlation Pattern Recognition* (Cambridge University Press, 2005).
44. Mod, E. C. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).
45. Blanchard, G., Bousquet, O. & Massaer, P. Statistical performance of support vector machines. *Ann. Stat.* **36**, 489–531 (2008).
46. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
47. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
48. Stuart, R. & Norvig, P. *Artificial Intelligence: A Modern Approach* 2nd edn (Pearson, 2003).
49. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Diao, Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).
51. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
52. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
53. Kothary, R. et al. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105**, 707–714 (1989).
54. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
55. Firpi, H. A., Ucar, D. & Tan, K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* **26**, 1579–1586 (2010).
56. Lu, Y., Qu, W., Shan, G. & Zhang, C. DELTA: a distal enhancer locating tool based on adaboost algorithm and shape features of chromatin modifications. *PLoS One* **10**, e0130622 (2015).
57. He, Y. et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl Acad. Sci. USA* **114**, E1633–E1640 (2017).
58. Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
59. Arner, E. et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).
60. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* **43**, e6 (2015).

Acknowledgements

M. Gerstein was supported by NIH grant HG009446-01. A.V. and L.A.P. were supported by NHLBI grant R24HL123879 and NHGRI grants R01HG003988, U54HG006997 and UM1HG009421, where research was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. We thank A. Paccanaro, D. Galeano, M. Torres and Y. Wu for their insightful scientific discussions. We thank all ENCODE consortium members for their feedback on this work.

Author contributions

A.S. and M. Gu conceptualized and developed the matched-filter model under the supervision of M. Gerstein. L.C., K.-K.Y., J.R. and K.Y.Y. performed many initial explorations and analysis of ChIP–seq data. C.Y. did model-performance comparisons. E.G. and R.S. performed the transduction reporter assay in human cell lines. I.B., V.A., J.A.A, I.P.-F., C.S.N., M.K., T.H.G., Q.P., A.H., B.J.M., E.A.L., Y.F.-Y., A.V., D.E.D. and L.A.P. performed enhancer assays in transgenic mice. A.S., M. Gu and M. Gerstein designed the model, coordinated the experimental validation and wrote the manuscript with input from coauthors.

Competing interests

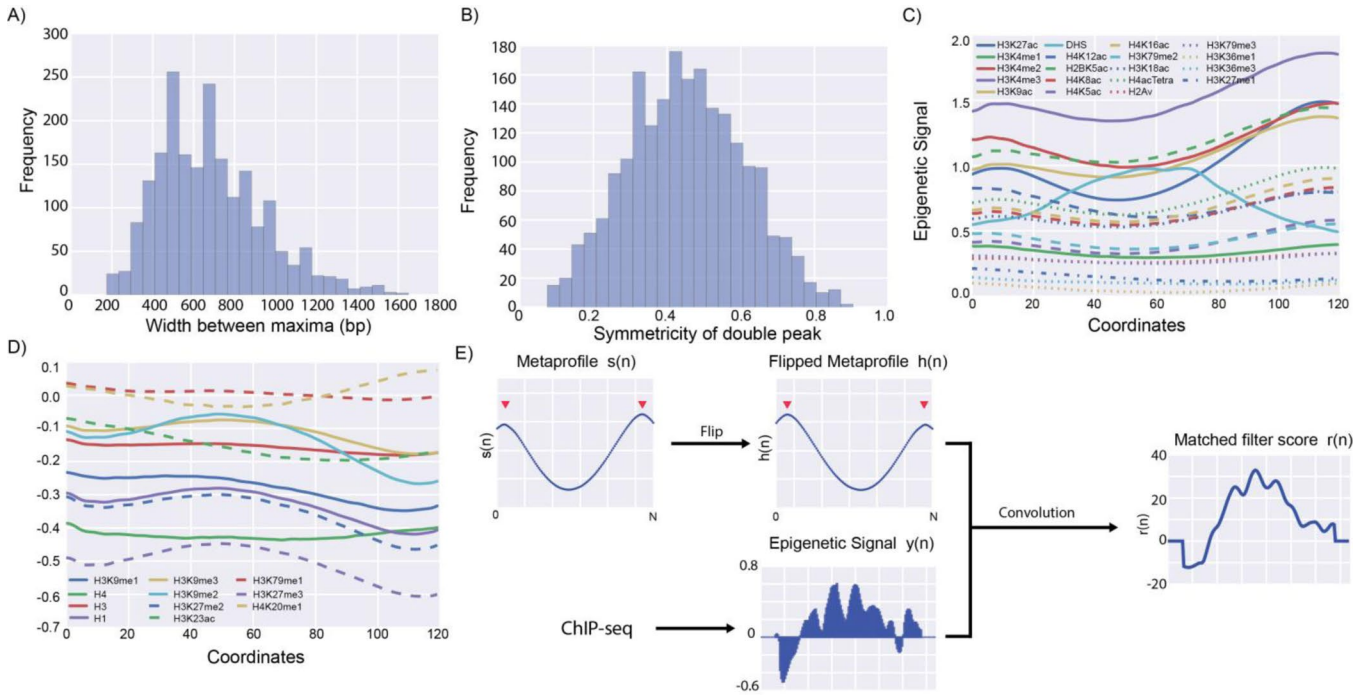
The authors declare no competing interests.

Additional information

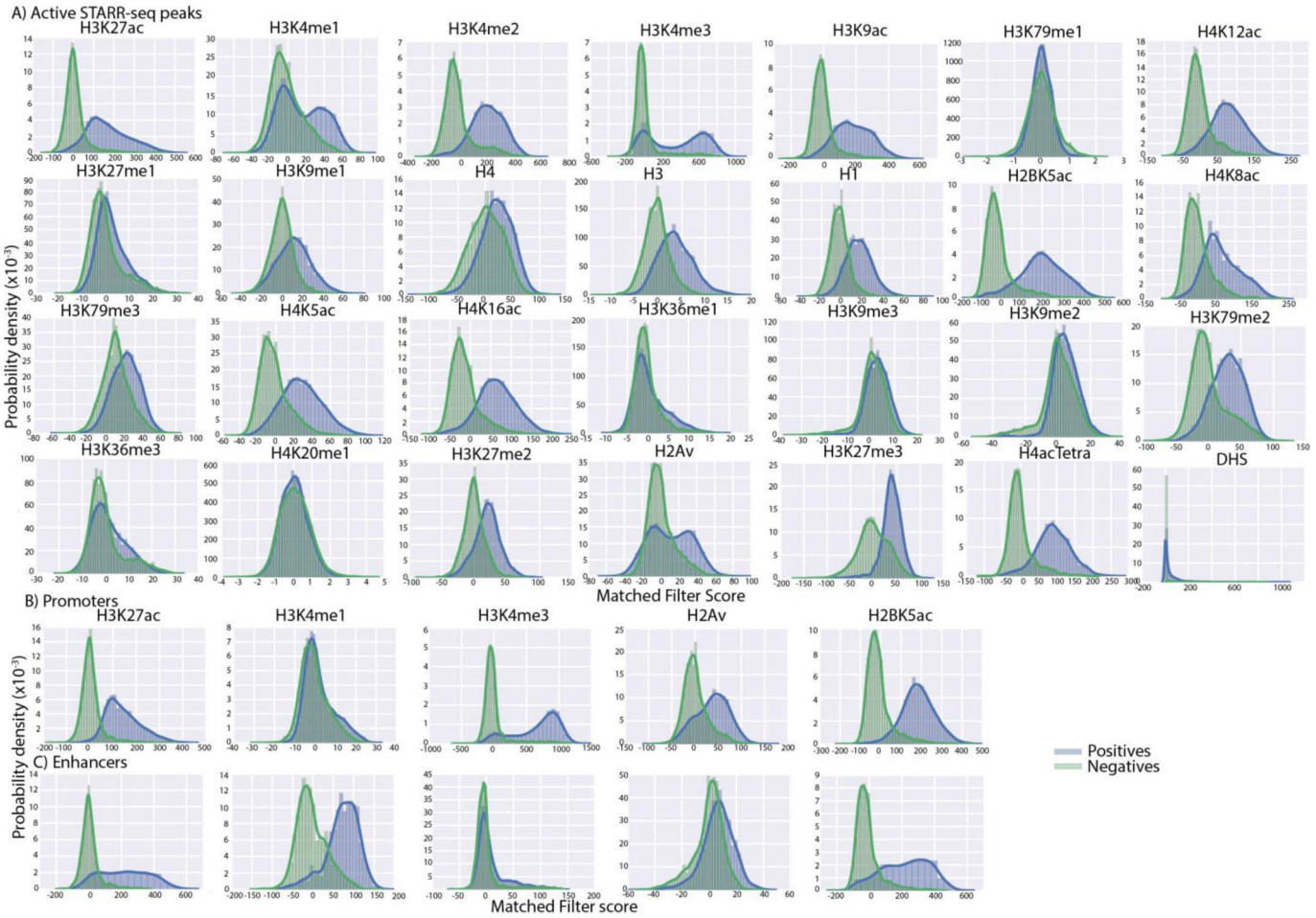
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0907-8.

Correspondence and requests for materials should be addressed to M.G.

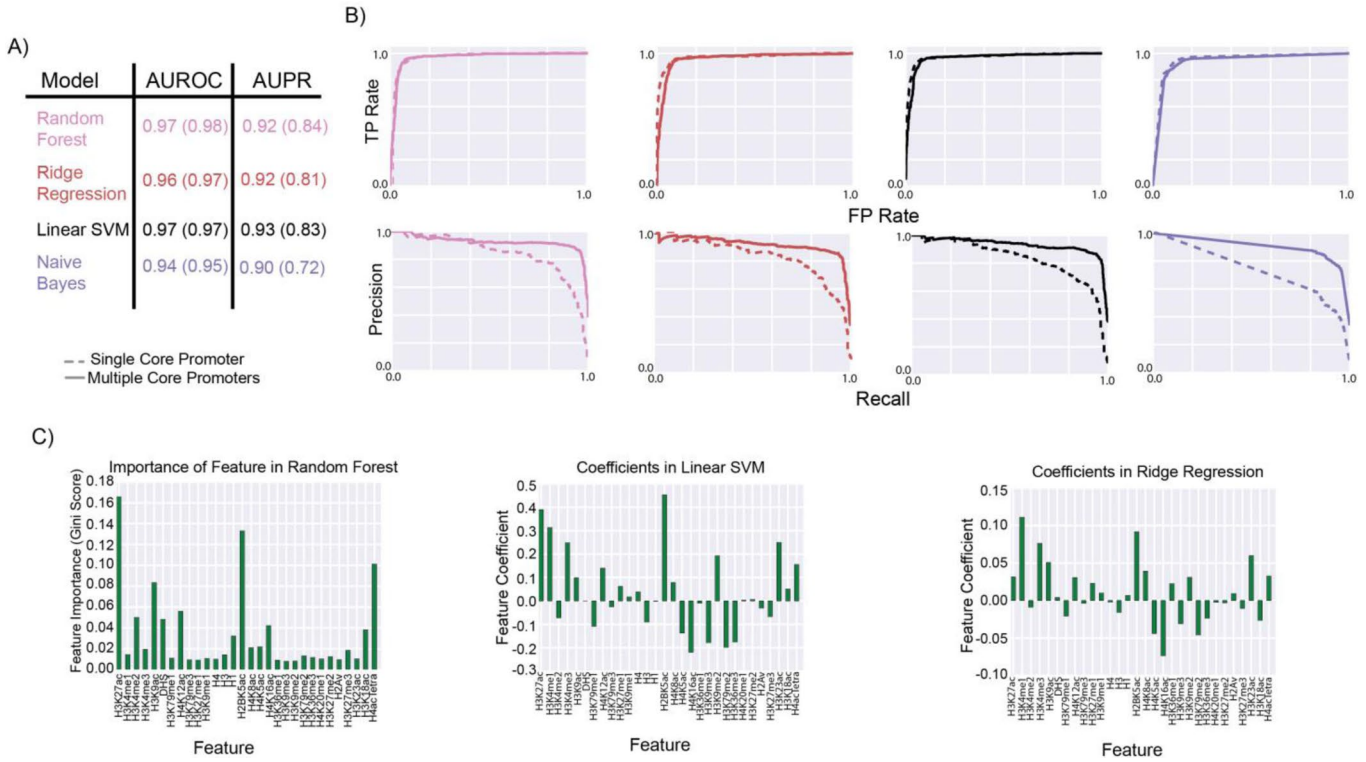
Peer review information Nicole Rusk and Lin Tang were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



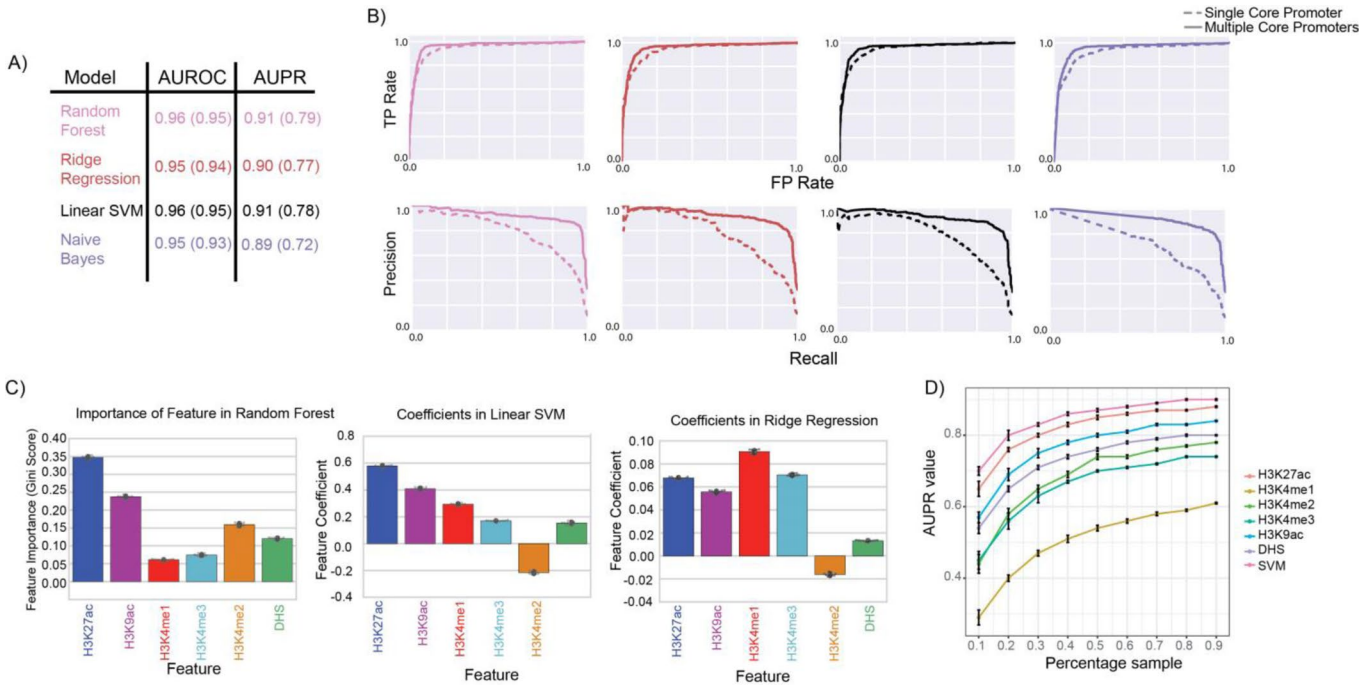
Extended Data Fig. 1 | Properties of double peak metaprofile. a) The frequency of distance between the two maxima in a double peak flanking active STARR-seq peaks is plotted. **b**) The symmetricity of the double peak pattern is plotted. The ratio of the distance between the two peaks to the ratio between one of the maxima and the minima is plotted. While there is large amount of variability in the distance between the two peaks (mostly between 300-1100 bp), the trough in the double peak tends to occur in the center of the two peaks. (**c**) The metaprofile around active STARR-seq peaks is plotted for different epigenetic marks. Histone marks that are enriched near STARR-seq peaks display the characteristic double peak pattern shown in **c**) due to the depletion of histone proteins at active regulatory regions. In addition, DHS displays a single peak at the center of these regulatory regions as shown in **c**). **d**) On the other hand, no such double peak pattern is observed on depleted histone marks at STARR-seq peaks. **e**) The matched filter score is calculated using the convolution of the flipped metaprofile and the epigenetic signal using a sliding window of variable length. The significant peaks in the final matched filter score are used to identify active regulatory regions.



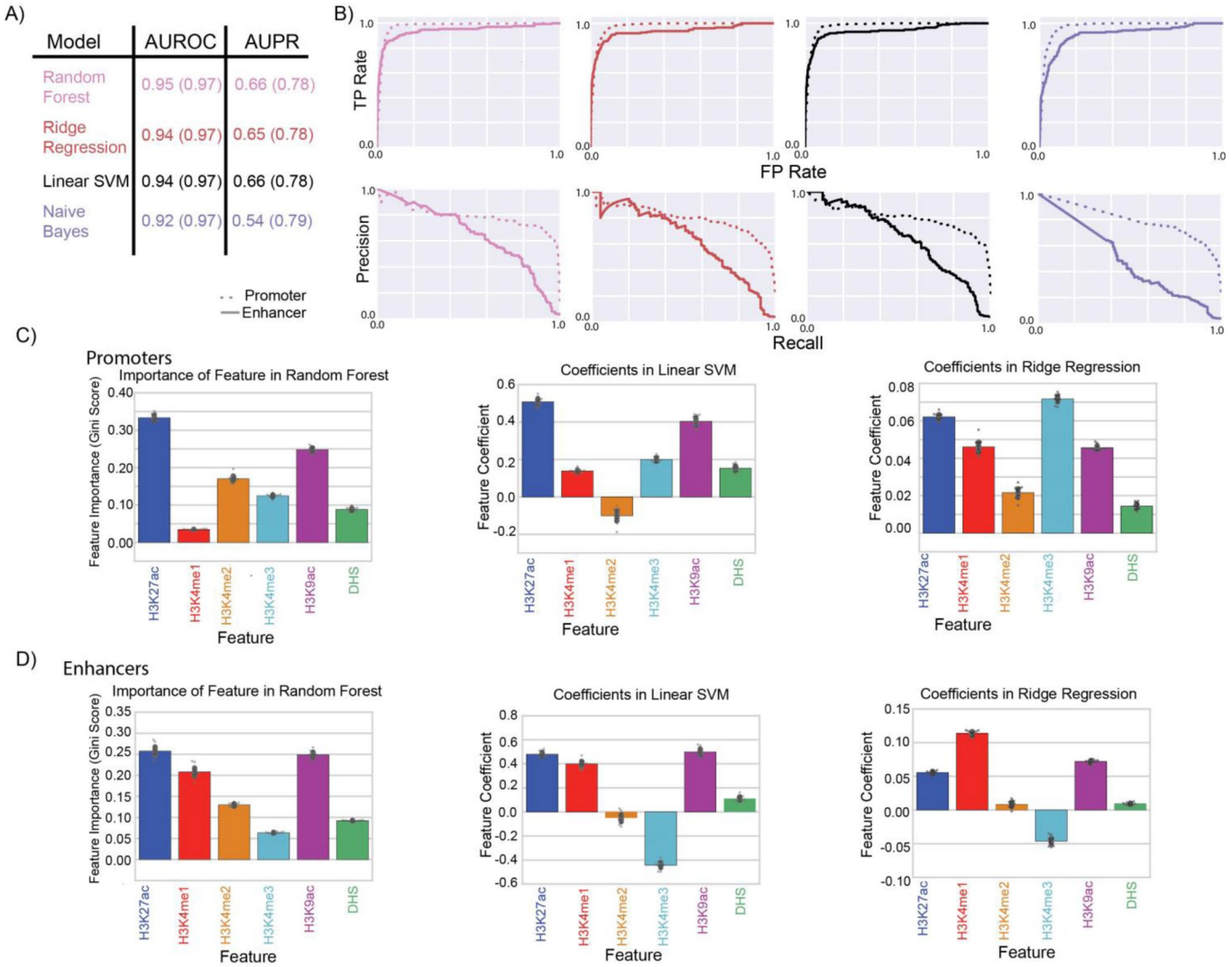
Extended Data Fig. 2 | Histogram of matched filter scores. a) The probability density of matched filter scores for different epigenetic marks for STARR-seq peaks (positives) and random regions of the genome (negatives) with H3K27ac signal. In most cases, the matched filter scores for positives and negatives are Gaussian curves. The amount of overlap between these two curves determines the accuracy of the matched filter for predicting STARR-seq peaks using the matched filters for the corresponding epigenetic feature. **b**) The histogram of matched filter scores for small set of epigenetic features on promoters is compared to random regions of the genome. **c**) The histogram of matched filter scores for small set of epigenetic features on enhancers is compared to random regions of the genome. The features chosen in **b**, **c** were chosen to display distinct features of epigenetic marks around promoters and enhancers.



Extended Data Fig. 3 | Comparison of different statistical models for predicting all STARR-seq peaks using a 30-feature model. The performance of the different statistical models to integrate the information from 30 epigenetic features is shown. **a**) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. **b**) The individual ROC and PR curves for each statistical model. **c**) The contribution of the matched filter score for each epigenetic feature to the different integrated models.



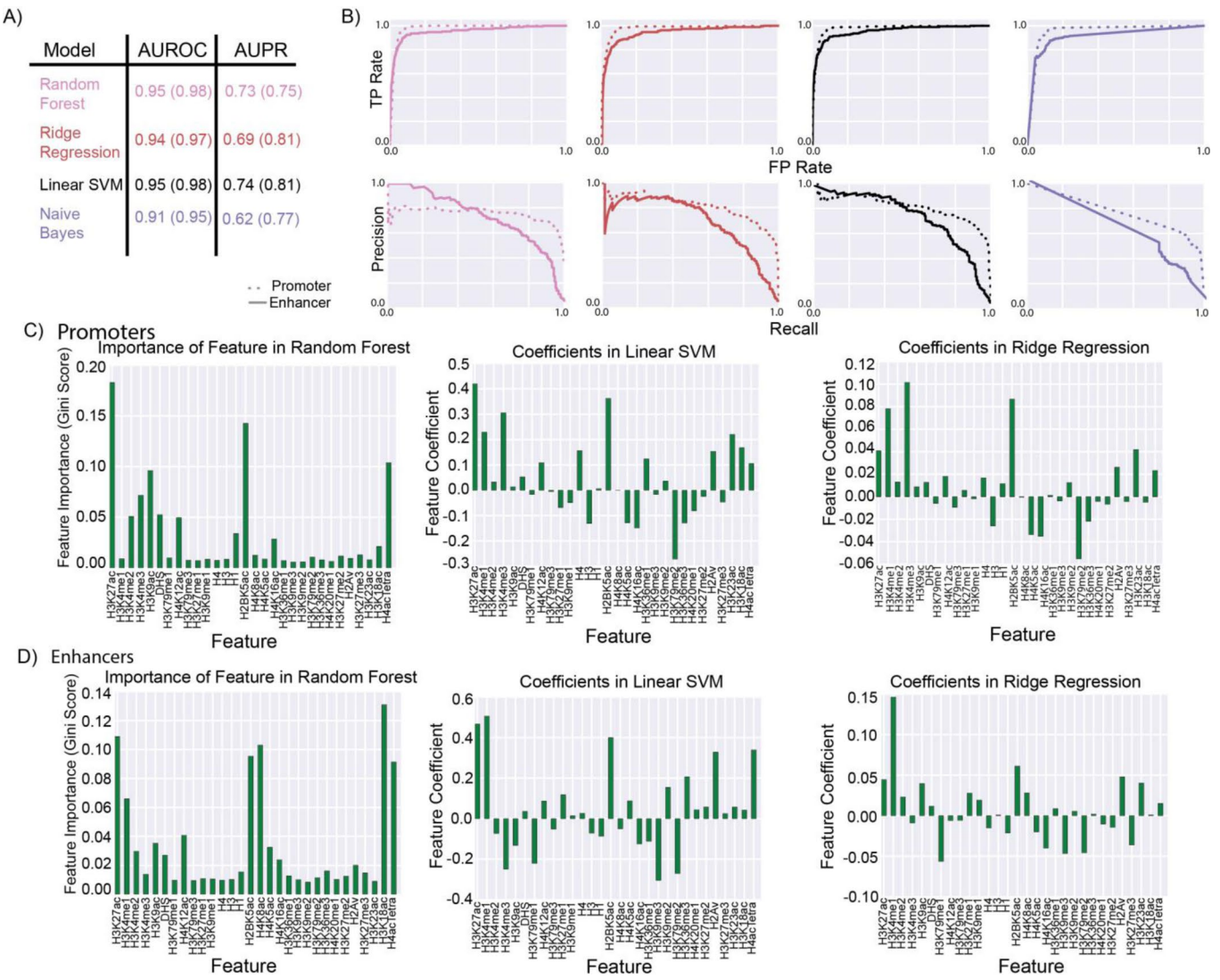
Extended Data Fig. 4 | Comparison of different statistical models for predicting all STARR-seq peaks using a 6-feature model. The performance of the different statistical models to integrate the information from six epigenetic features is shown. **a**) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. **b**) The individual ROC and PR curves for each statistical model. **c**) The contribution of the matched filter score for each epigenetic feature to the different integrated models. The mean value is displayed in the bar plot while the error bars show the standard deviation of feature weights measured by ten-fold cross validation. **d**) We evaluated the accuracy of the models using different amounts of training data. The AUPR of the model increases with increasing amount of training data until it starts to saturate around 70% of the data. The mean value is displayed in the bar plot while the error bars show the standard deviation of feature weights measured by ten-fold cross validation.



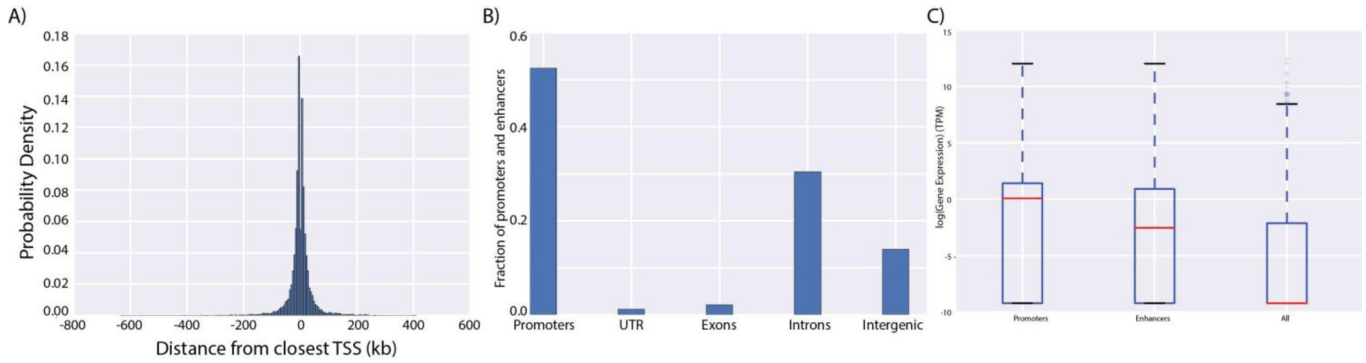
Extended Data Fig. 5 | Comparison of different statistical models for predicting enhancers and promoters using six features. The performance of the different statistical models to integrate the information from six epigenetic features for promoter and enhancer prediction is shown. **a**) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. **b**) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (**c**) and enhancer prediction (**d**) are shown. The mean value is displayed in the bar plot while the error bars show the standard deviation of feature weights measured by ten-fold cross validation.

A)

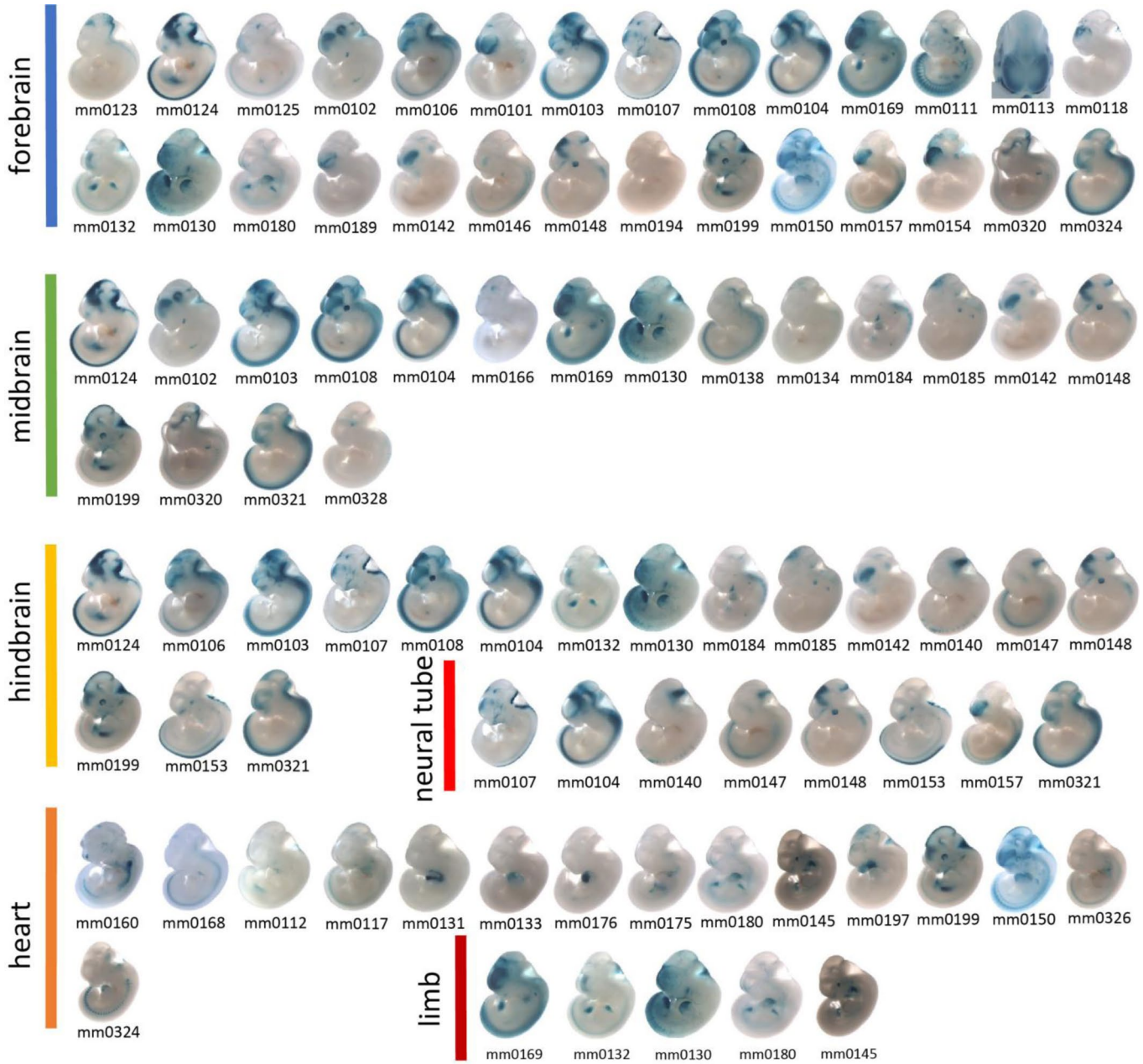
Model	AUROC	AUPR
Random Forest	0.95 (0.98)	0.73 (0.75)
Ridge Regression	0.94 (0.97)	0.69 (0.81)
Linear SVM	0.95 (0.98)	0.74 (0.81)
Naive Bayes	0.91 (0.95)	0.62 (0.77)



Extended Data Fig. 6 | Comparison of different statistical models for predicting enhancers and promoters using 30 features. The performance of the different statistical models to integrate the information from thirty epigenetic features for promoter and enhancer prediction is shown. **a**) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. **b**) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (**c**) and enhancer prediction (**d**) are shown.



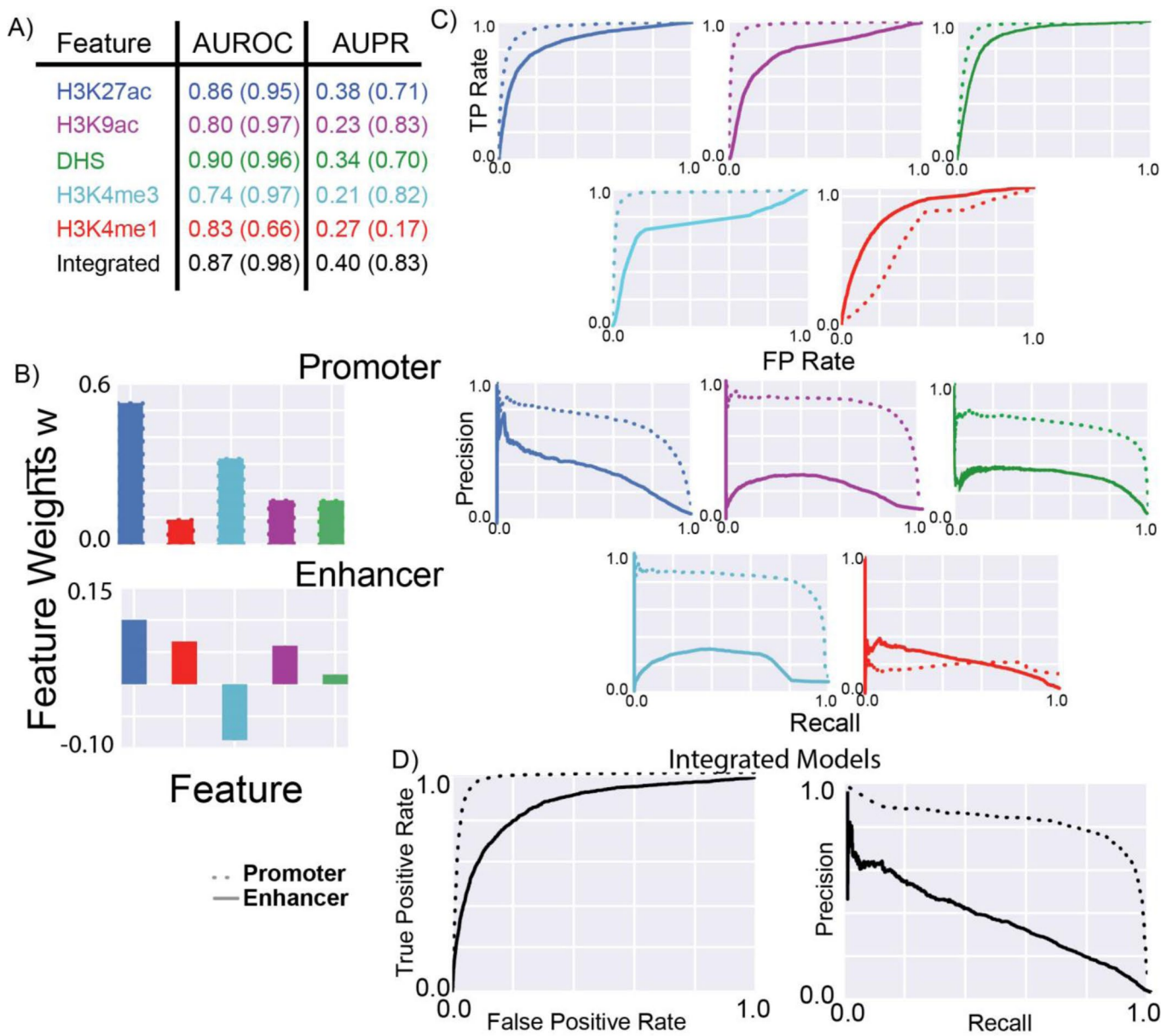
Extended Data Fig. 7 | Location of H1-hESC predictions. a) The probability density of the distance of the predicted promoter and enhancer from the closest TSS is shown. **b**) The location of the enhancers and promoters on genomic elements are shown. Promoters are defined as TSS +/- 2kb. All TSS, UTR, exons, introns, and intergenic elements are calculated based on GENCODE 19 definitions. A regulatory region is considered to overlap with the elements if more than 50% of the matched filter region overlaps with the corresponding element in **b**. **c**) The distribution of gene expression of gene closest to the enhancer/promoters are plotted and compared to the gene expression of all genes in H1-hESC. A two-sided Wilcoxon test shows that P-value for differences in gene expression of genes close to enhancers and promoters are significantly higher than expression of all genes in H1-hESC ($< 10^{-100}$ each). The center line in each category represents the median expression level for all genes close to corresponding category while the lower and upper boundaries of the box indicate the 25th and 75th percentile of the expression levels for genes within that category.



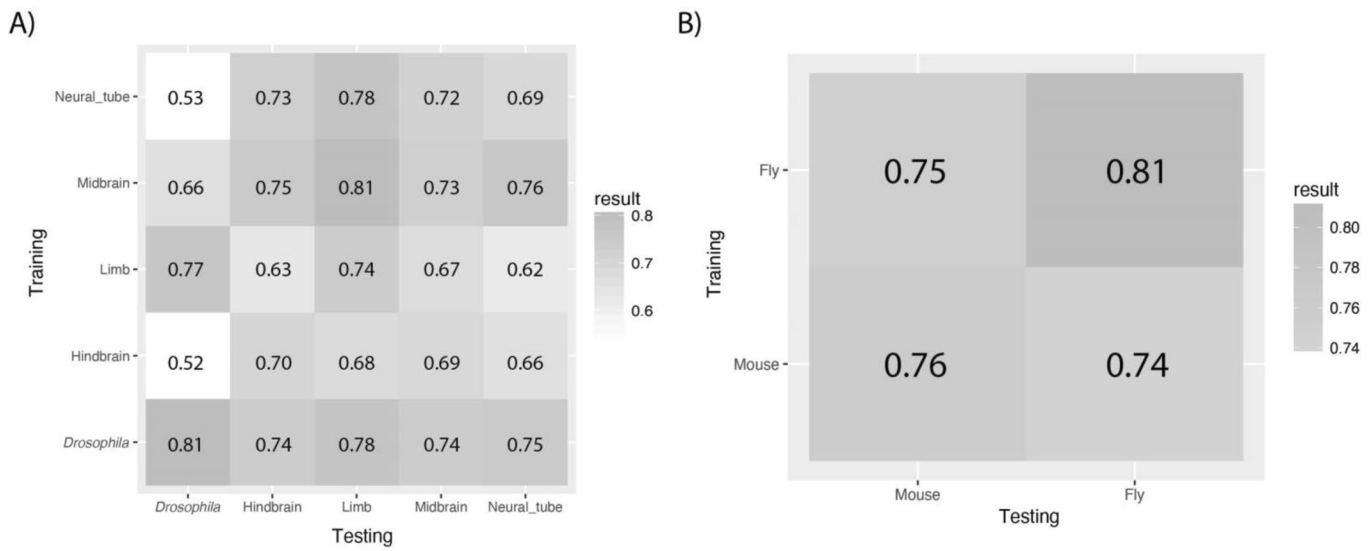
Extended Data Fig. 8 | Testing predicted enhancers using transgenic mouse enhancer assay. Representative embryo images are shown for transgenic mice at the e11.5 stage. Blue staining indicates enhancers displaying reproducible activity in expected tissues (forebrain, midbrain, hindbrain, heart, neural tube, or limb). The unique identifiers under each image (accession number starting with 'mm') correspond to the element numbers in Supplementary Tables 4-9. Details of each experiment can be found in the VISTA enhancer browser (https://enhancer.lbl.gov) under the corresponding accession number.

A)

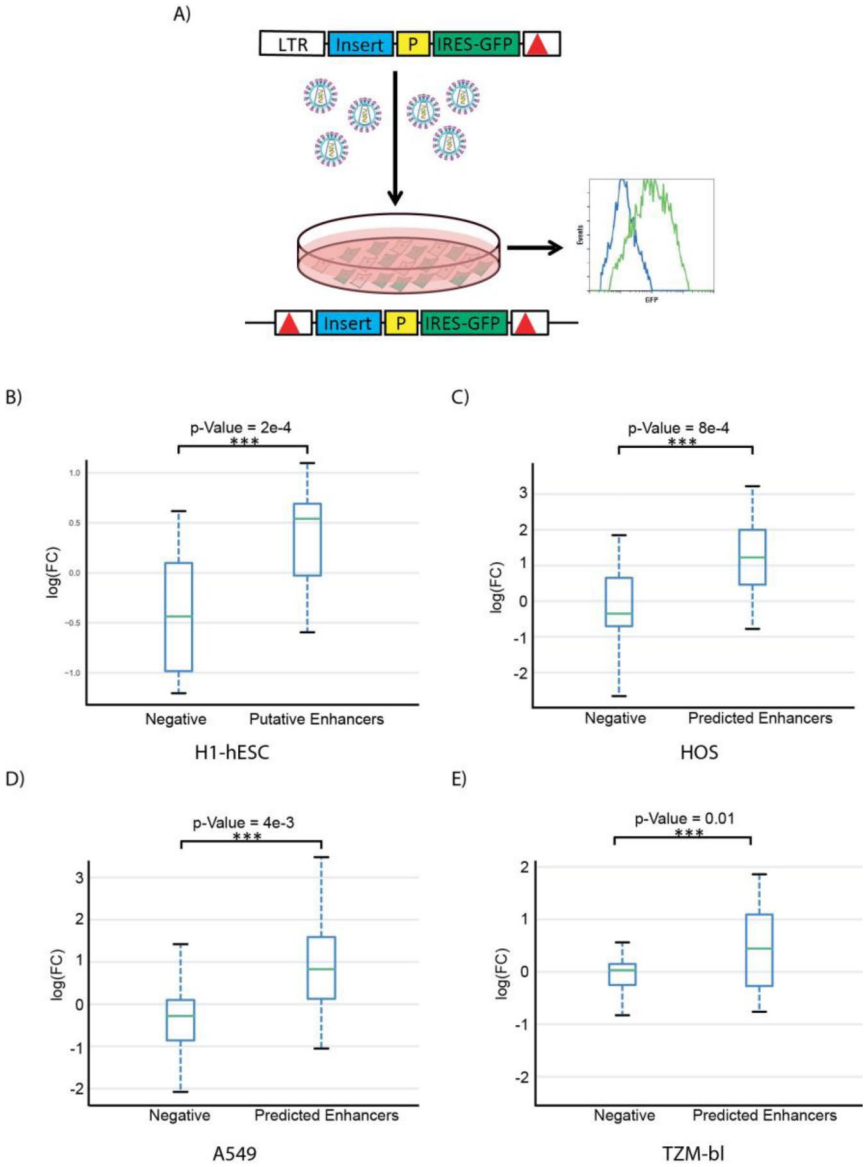
Feature	AUROC	AUPR
H3K27ac	0.86 (0.95)	0.38 (0.71)
H3K9ac	0.80 (0.97)	0.23 (0.83)
DHS	0.90 (0.96)	0.34 (0.70)
H3K4me3	0.74 (0.97)	0.21 (0.82)
H3K4me1	0.83 (0.66)	0.27 (0.17)
Integrated	0.87 (0.98)	0.40 (0.83)



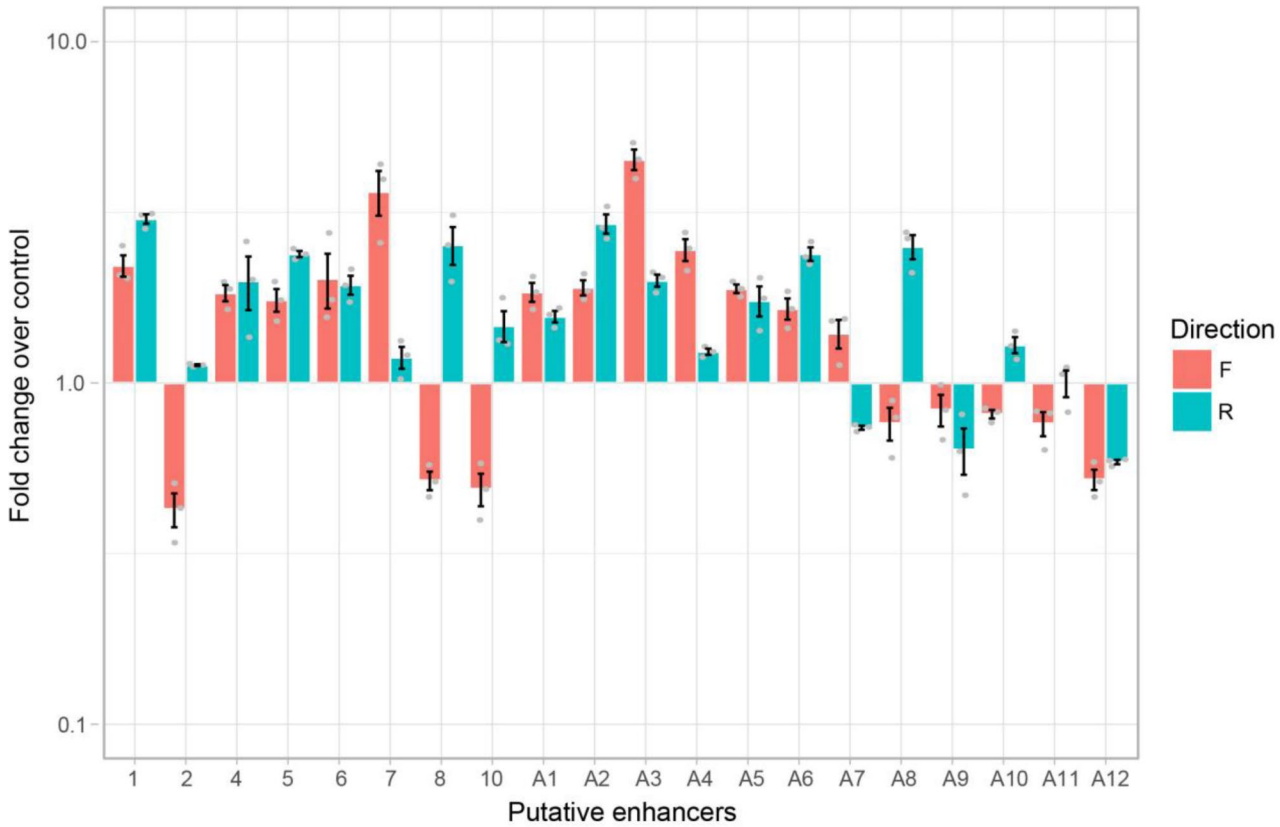
Extended Data Fig. 9 | Conservation of epigenetic features. The performance of the fly-based matched filters and the integrated model for predicting active promoters and enhancers in mouse embryonic stem cells identified using FIREWACh. **a** Similar to Fig. 3, the numbers within parentheses refer to 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 **d**) 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.



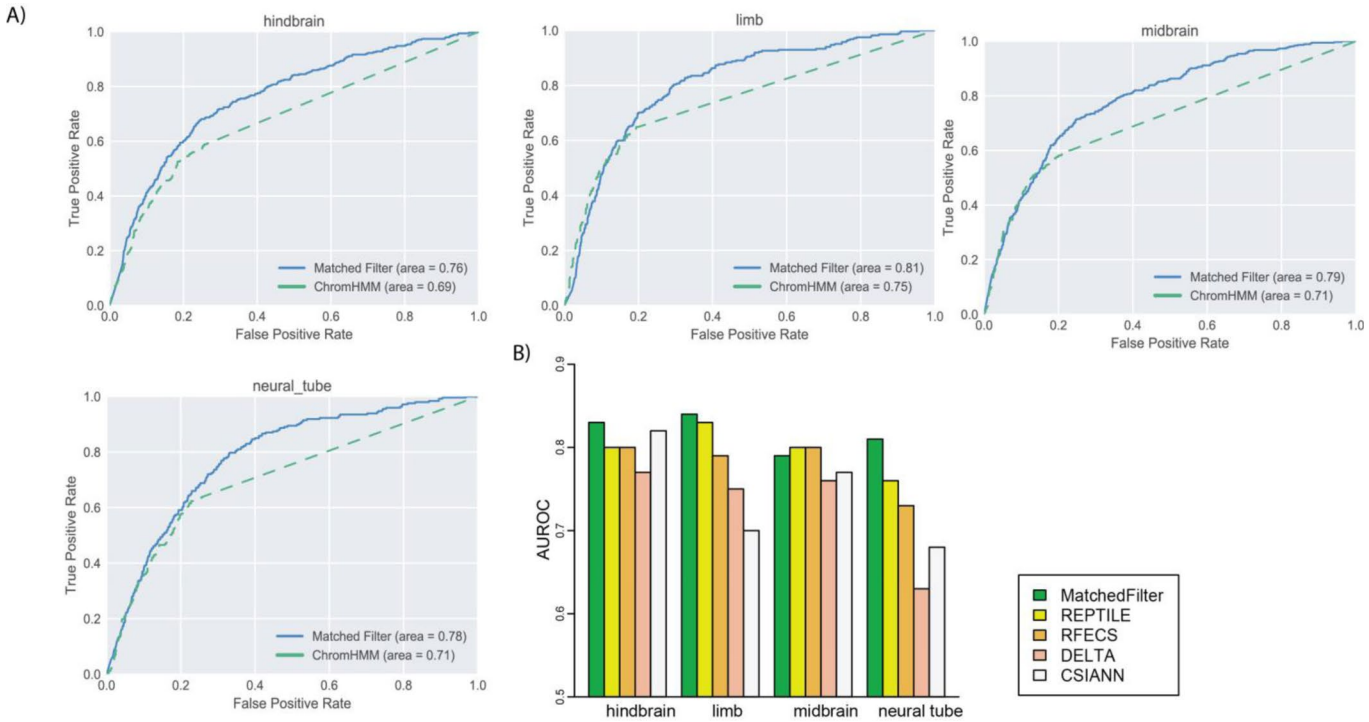
Extended Data Fig. 10 | Cross-comparison of integrated models for enhancer prediction. Cross test results of the integrated model on mouse and fly. **a**) Models were trained in a cell line- and tissue-specific fashion. Row names show the context where the model is trained. Column names show the cell line or tissue where the model is tested. **b**) Similar to **a**), assuming identical distribution of matched filter scores for active enhancer regions in each tissue in mouse, we combined the normalized matched filter scores to get a larger training set for the model.



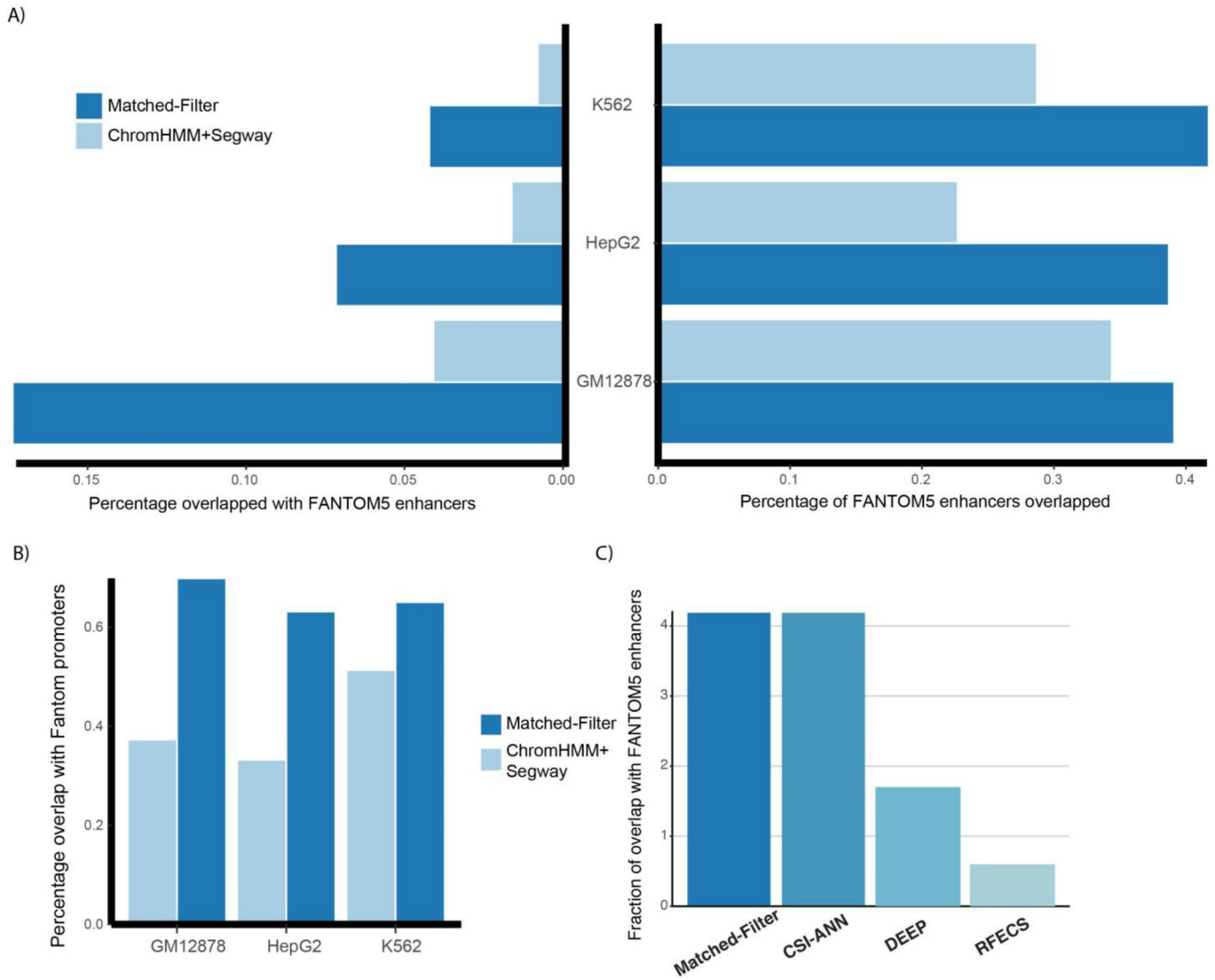
Extended Data Fig. 11 | Enhancer validation experiments in human cell lines. a) Schematic of the enhancer validation experiment flow. At top is the third-generation HIV-based self-inactivating vector (deletion in 3′ LTR indicated by red triangle), with PCR-amplified test DNA (blue, cloned in both orientations) inserted just 5′ of a basal Oct4 promoter (P) driving IRES-eGFP (green). Vector supernatant was prepared by plasmid co-transfection of 293T cells. Cells of interest were transduced and then analyzed by flow cytometry a few days later. Shown below is the expected post-transduction structure of the SIN HIV vector, with a duplication of the 3′ LTR deletion rendering both LTRs non-functional. **b**) Fold changes of gene expression of eGFP was compared between negative elements (n=20 biologically independent samples) and putative enhancers (n=20 biologically independent samples) chosen at random. Each sample in the plot is the average log fold change of the replicates for each element. **c**–**e**) Predicted enhancers increase gene expressions in A549, HOS, and TZM-bl cell lines. The enhancers were predicted in H1-hESCs. The activities of these enhancers (N=20 in each plot) were compared to control regions (N=20 in each plot) in three other cell lines: **c**) HOS, **d**) A549, and **e**) TZM-bl. The p-value were calculated by the two-sided t-test. The center value represented by the green line in the box plot shows the median log FC of each group. The 25th and 75th percentiles of the log fold changes in gene expressions for each group are represented by the upper and lower lines of the box, with whiskers connecting to the maximum and the minimum value.



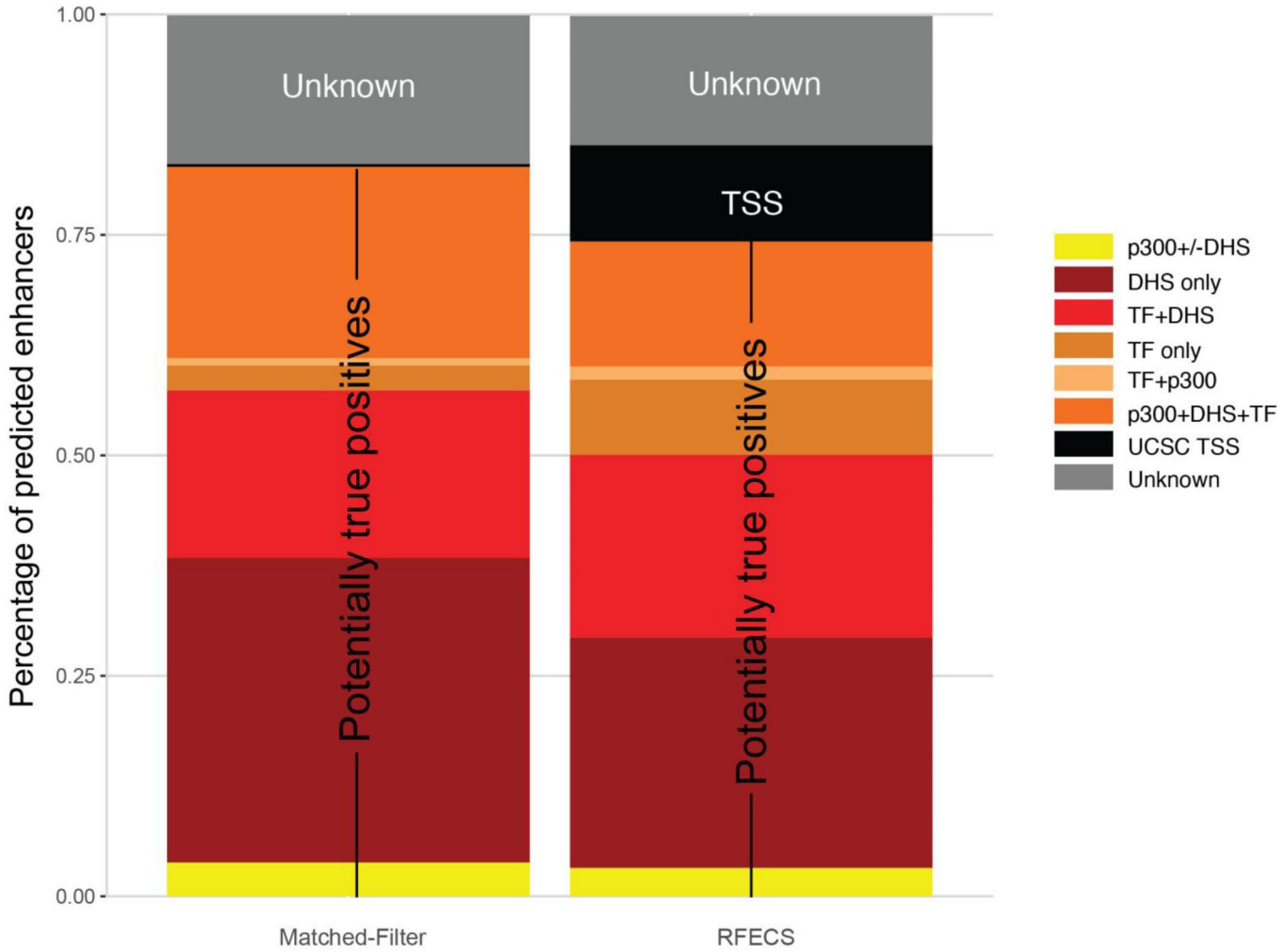
Extended Data Fig. 12 | Activity of putative enhancers tested in H1-hESCs. Each element was tested in triplicate (biologically independent experiments) by SIN HIV vector transduction of H1-hESCs. The bar plot shows the average of the activity measured in three replicates for each element by FACS analysis gating on eGFP+ cells, with error bars showing the standard deviations. F, forward orientation; R, reverse orientation.



Extended Data Fig. 13 | Performance comparison of the Matched-filter model in four mouse tissues. We compared the performance of the matched filter model to the other state-of-the-art predictive models on four mouse tissues where data is available. **a**) Comparing the performance of the matched filter model and ChromHMM with ROC curves using experimental results from transgenic mouse enhancer assays. The ROC curves for matched filter are plotted in blue solid lines, and the ROC curves for ChromHMM are plotted in green dashed lines. ROC curves are shown for all four tissues in embryonic mice at the e11.5 stage. **b**) Comparing the performance of the matched filter model with the reported performance of other published methods, including REPTILE, RFECS, DELTA, and CSI-ANN. Bar plots show the areas under the ROC curve (AUROC) of each methods in different tissues of embryonic mice at the e11.5 stage.



Extended Data Fig. 14 | Evaluating the Matched-filter prediction using FANTOM5 experimental data. We assessed the percentage overlap of the matched filter prediction with the FANTOM5 enhancers/promoters, and compared the percentages with other state-of-the-art methods. A) Comparison of the matched filter enhancer prediction in human cell lines with the integrated ChromHMM and Segway annotations using the FANTOM5 enhancer set. Bar plots on the left show the percentage of predicted enhancers overlapping with FANTOM enhancers; bar plots on the right show the percentage of FANTOM5 enhancers overlapping with predicted enhancers. B) Comparison of the overlap of matched filter promoter predictions with the FANTOM5 promoter set to that of the integrated ChromHMM and Segway annotations. The bar plots show the percentage of predicted promoters overlapping with FANTOM5 promoters, with dark blue denoting the matched filter model and light blue denoting the integrated ChromHMM and Segway annotations. C) Comparison of the overlap of K562 enhancers predicted by matched filter and other published methods with the FANTOM5 enhancer set. The bar plots show the percentage of predicted enhancers overlapping with FANTOM5 K562 enhancers for the matched filter model, CSI-ANN, DEEP, and RFECS.



Extended Data Fig. 15 | Comparison of the transcription factor binding pattern of matched filter and RFECS in H1-hESCs. Potentially positive enhancers were considered as regions with either DNase-I hypersensitive sites (DHS), or bound by transcription factors (TFs) such as NANOG, OCT4, SOX2, or p300. TSS were defined as within 2.5kb of any known GENCODE TSS. Predictions that fell out of the above categories were classified as unknown.

nature research

Corresponding author(s): Mark Gerstein

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   For validation experiments in human cell lines, we estimated the enhancer effects to do power analysis to determine the number of elements we need to test. For transgenic mouse enhancer tests, we tested 150 elements in addition to the validated enhancers in the VISTA database, which allow us more than sufficient samples to test our method.

2. Data exclusions

   Describe any data exclusions.

   No data were excluded from the analyses

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   For experiments in human cell lines, each tested elements has multiple biological replicates. The result is reported positive only if the enhancing effect is statistically significant from the biological replicates. For mouse transgenic assay, each element is tested in multiple transgenic embryos which can be viewed in VISTA enhancer browser. Elements were scored positive for enhancer activity if at least three resulting transgenic embryos had reporter gene expression in the same tissue and pattern. Elements were scored negative if at least five transgenic embryos were recovered and no reproducible staining patterns was observed.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   The predicted mouse enhancers are selected from three rank tiers for validation. The predicted human enhancers are selected randomly for validation.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Experiments are done with blinding to group allocations

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Our tool is implemented in Python2.7 with the following packages:
numpy 1.10.4
scipy 0.17.0
scikit-learn 0.16.1
matplotlib 1.5.1
seaborn 0.7.0
metaseq 0.5.5.4
pybedtools 0.7.1

A dockerized image has been provided on our website

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

There were no unique materials used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

H1 hESC was obtained from WiCell. HOS and A549 were obtained from ATCC and TZMbl from the AIDS Reagent Repository

b. Describe the method of cell line authentication used.

Cell lines are ATCC authenticated

c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines were tested negative for mycoplasma contamination

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly mis-identified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Enhancer transgene analysis was performed at embryonic day 11.5 in FVB strain male and female mice.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants.