Accurate prediction of functional states of *cis*-regulatory modules reveals the universal epigenetic code in mammals

Pengyu Ni[1],

Zhengchang Su[1*]

1.Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, Charlotte, NC 28223, USA

* To whom correspondence should be addressed.

Zhengchang Su

Tel: +01-704-687-7996

Fax: +01-704-687-8667

Email: zcsu@uncc.edu

Running title: Accurate prediction of functional states of CRMs

Key words: *cis*-regulatory modules, epigenetic marks, functional states, machine learning, cell types

ABSTRACT

Predicting *cis*-regulatory modules(CRMs) in a genome and predicting their functional states in various cell/tissue types of the organism are two related challenging computational tasks. Most current methods attempt to achieve both simultaneously using epigenetic data. Though conceptually attractive, they suffer high false discovery rates and limited applications. To fill the gaps, we proposed a two-step strategy to first predict a map of CRMs in the genome, and then predict functional states of the CRMs in various cell/tissue types of the organism. We have recently developed an algorithm for accurately predicting CRMs in a genome by integrating numerous transcription factor ChIP-seq datasets. Here, we showed that only three or four epigenetic marks data in a cell/tissue type were sufficient for a machine-learning model to accurately predict functional states of all CRMs. Our predictions are substantially more accurate than the best achieved so far. Interestingly, a model trained on different cell/tissue types in a mammal can accurately predict functional states of CRMs in different cell/tissue types of the mammal as well as in various cell/tissue types of a different mammal. Therefore, epigenetic code that defines functional states of CRMs in various cell/tissue types is universal at least in mammals. Moreover, we found that from tens to hundreds of thousands of CRMs were active in a human and mouse cell/tissue type, and up to 99.98% of them were reutilized in different cell/tissue types, while as small as 0.02% of them were unique to a cell/tissue type that might define the cell/tissue type.

INTRODUCTION

Since the completion of the Human Genome Project 19 years ago(Lander et al. 2001; Venter et al. 2001), we have largely categorized coding sequences(CDSs) in the genome(Temple et al. 2009) and gained a good understanding of their functions in various human cell/tissue types. In contrast, although *cis*-regulatory sequences are as important as CDSs in specifying human traits(Maston et al. 2006; Narlikar and Ovcharenko 2009; Alexander et al. 2010), our understanding of them falls largely behind due to more technical difficulties in categorizing them in the genome and characterizing their functional states in various human cell/tissue

types(Kleftogiannis et al. 2016; Lim et al. 2018). *cis*-regulatory sequences are also known as *cis*-regulatory modules(CRMs), as they often function independently of their locations and orientations relative to their target genes. A CRM is generally made of clusters of transcription factors(TFs) binding sites(TFBSs), to which cognate TFs can bind(Davidson 2006). A CRM regulates gene transcription through interacting with the promoter of its target gene in a cell/tissue type, where the CRM is accessible and cognate TFs of the constituent TFBSs are available to be bund(Davidson 2006; Hardison 2012).

As active CRMs are nucleosome free(Armstrong and Emerson 1996; Tirosh and Barkai 2008; He et al. 2010) and often modified by certain histone marks in a cell/tissue type specific manner(Heintzman et al. 2007; Morse 2010; Rada-Iglesias et al. 2011; Zentner and Scacheri 2012; Trynka et al. 2013; Huang et al. 2015), many machine-learning methods have been developed to simultaneously predict CRMs and their functional states based on sequences' chromatin accessibility(CA) as measured by DNase-seq(Boyle et al. 2008) or assay for transposase accessible chromatin by sequencing(ATAC-seq(Buenrostro et al. 2013) and histone modifications as measured by ChIP-seq(Johnson et al. 2007). Earlier methods include hidden Markov models(ChromHMM)(Ernst and Kellis 2010; Ernst et al. 2011), dynamic Bayesian networks(segway)(Hoffman et al. 2012; Hoffman et al. 2013), neural networks(CSI-ANN)(Firpi et al. 2010), random forest(RFECS and REPTILE)(Rajagopal et al. 2013; Ghandi et al. 2014; Kleftogiannis et al. 2015; He et al. 2017), support vector machines(SVMs)(Fernandez and Miranda-Saavedra 2012; Kleftogiannis et al. 2015) and others(Gao et al. 2016; Fishilevich et al. 2017; Gao and Qian 2020). Although conceptually attractive and a great deal of insights into CRMs have been gained from the studies, these methods suffer quite high false discovery rates(FDRs)(Kheradpour et al. 2013; Kwasnieski et al. 2014; Dogan et al. 2015; Kleftogiannis et al. 2015; Catarino and Stark 2018; Arbel et al. 2019; Ni and Su 2021) duo probably to two reasons, in our opinion. First, a sequence segment that has CA(Arnold et al. 2013) and/or histone marks such as H3K3me1(Dorighi et al. 2017; Rickels et al. 2017;

Rada-Iglesias 2018), H3H4me3(Howe et al. 2017) and H3K27ac(Zhang et al. 2020) are not necessarily CRMs(Dogan et al. 2015; Young et al. 2017; Catarino and Stark 2018; Arbel et al. 2019; Ni and Su 2021). Second, in many supervised machine-learning methods, a mark such as H3K27ac was used both as a feature and as the label of training sets(Liu et al. 2016; Sethi et al. 2020). However, since an active CRM may not necessarily have the mark(Catarino and Stark 2018; Zhang et al. 2020), these models were actually trained to differentiate sequences with and without the label, instead of the functional states (active and non-active) of CRMs. At the same time, various forms of massively parallel reporter assays(MPRAs) have been developed to characterize enhancers(Santiago-Algarra et al. 2017). In particular, self-transcribing assay of regulatory regions sequencing(STARR-seq) clones randomly sheared *D. melanogaster* genomic DNA between a minimal-promoter-driven GFP open reading frame and a downstream polyA sequence(Arnold et al. 2013). If a sequence is an active enhancer, this results in transcription of the enhancer sequence, allowing to assess more and relatively longer candidate sequences than earlier MPRAs that used short (~200bp) synthetic sequences(Klein et al. 2020). Variants of STARR-seq have been developed to accommodate to large mammalian genomes, such as whole genome STARR-seq(WHG-STARR-seq)(Liu et al. 2017) and ATAC enrichment coupled with STARR-seq(ATAC-STARR-seq)(Wang et al. 2018). However, all forms of STARR-seq methods are based on episomal expression vectors, the results may not reflect the native chromosomal contexts(Inoue et al. 2017; Klein et al. 2020). Moreover, sequences that can be assessed by STARR-seq are still much shorter(~500bp) than a typical mammalian CRM (~2,000bp), thus they suffer high FDRs and probably high false negative rates(Arnold et al. 2013; Inoue et al. 2017; Liu et al. 2017; Barakat et al. 2018; Muerdter et al. 2018; Wang et al. 2018; Klein et al. 2020; Peng et al. 2020).

Based on the fact that TF binding is a more reliable predictor of CRMs than CA and histone marks(Dogan et al. 2015), we recently developed a new CRM predictor dePCRM2(Ni and Su 2021) that identifies closely located clusters of TFBSs in a genome by

integrating thousands of TF ChIP-seq data in the organism. Using dePCRM2, we are able to predict more accurate and more complete maps of CRMs and constituent TFBSs in mammalian genomes(Ni and Su 2021). Although dePCRM2 can also predict functional states (TF-binding or non-binding) of CRMs that overlap binding peaks of ChIP-ed TFs in a cell/tissue type, the fraction of the CRMs whose functional states can be predicted depends on the availability of ChIP-seq data for active TFs in the cell type(Ni and Su 2021). As in most cell/tissue types, only few or even no TF ChIP-seq datasets are available, this fraction of CRMs is generally very low or even zero(Ni and Su 2021) (also see RESULTS). Clearly, to predict functional states of all CRMs in a cell type in this way, one might need ChIP-seq data for all active TFs in the cell/tissue type. This can be too costly or currently unfeasible. As a result, CRMs and TFBSs predicted by dePCRM2 are largely cell type agnostic(Ni and Su 2021). Fortunately, it has been shown that when the locus of a CRM is accurately anchored by the binding of key TFs, epigenetic marks can be an accurate predictor of the functional state of the CRM(Podsiadlo et al. 2013; Kwasnieski et al. 2014; Dogan et al. 2015; Kleftogiannis et al. 2015; Arbel et al. 2019). It appears that a pattern of epigenetic marks on a CRM is sufficient to define its functional state, however, a sequence with such a pattern is not necessarily even a CRM(Dogan et al. 2015; Young et al. 2017; Arbel et al. 2019; Zhang et al. 2020; Ni and Su 2021).

Although dozens of epigenetic marks have been found to modify CRMs in different cellular contexts(Zhao and Garcia 2015; Sethi et al. 2020), which and how many of them are required to define functional states of CRMs remains elusive(Dogan et al. 2015; He et al. 2017; Howe et al. 2017; Rada-Iglesias 2018; Zhang et al. 2020). Recently, it was found that machine-learning models trained using as few as six marks(H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac and CA) performed equally well as the same models trained using as many as 30 marks in differentiating STARR peaks and negative control sequences, with H3K27ac being the most important feature(Sethi et al. 2020). This study also showed that patterns of the six marks learned on STARR peaks in *D. drosophila* S2 cells could accurately

predict functional states of candidate enhancers and promoters in mouse and human cell/tissue types(Sethi et al. 2020). As STARR peaks in *D. drosophila* S2 cells contained ~30% false positives(Arnold et al. 2013), the study used STARR peaks that overlapped H3K27ac peaks as positive sets to reduce false positives. In our opinion, however, this practice created the aforementioned dilemma that H3K27ac was used both as a feature and as the label of the STARR peaks and candidate CRMs(Sethi et al. 2020), and thus, could be problematic. Since our program dePCRM2 predicts CRMs in a genome and functional states of some CRMs in a cell/tissue type without using any histone marks and CA data(Ni and Su 2021), we can avoid the dilemma by using the predicted active CRMs as a positive training set. Therefore, it is interesting to see how well the functional states of our predicted cell type agonistic CRMs in a cell/tissue type can be predicted using a minimal set of epigenetic marks as features by machine-learning classifiers trained on the smaller subsets of the CRMs whose functional states can be predicted by dePCRM2. It is also interesting to investigate how universal are epigenetic code that define functional states of CRMs across different cell/tissue types in a species as well as in different species. We show that only four common epigenetic marks (CA, H3K4me1, H3K4me3 and H3K27ac) are sufficient to very accurately predict functional states of CRMs in a cell/tissue type. We find that machine-learning models trained on different cell/tissue types in a species can accurately predict functional states of CRMs in different cell/tissue types in the same species as well as in various cell/tissue types in different species, and CA is the most important feature in all the scenarios. Therefore, epigenetic code that defines functional states of CRMs are universal across different cell/tissue types of the same and different species. Our predictions are substantially more accurate than the best achieved so far, suggesting that our training sets better capture the distinct epigenetic patterns that differentiate active and non-active CRMs.

RESULTS

Machine-learning models trained on our predicted active CRMs and non-CRMCs outperform those trained on positive and negative sets defined by other methods

Due to the lack of a large gold standard positive (active) and negative (non-active) CRMs in any human and mouse cell/tissue types, different methods have been used to define positive and negative sets of CRMs in a cell/tissue type for training machine-learning models(Firpi et al. 2010; Rajagopal et al. 2013; Lu et al. 2015; Liu et al. 2016; He et al. 2017; Sethi et al. 2020). To find a better way of defining positive and negative sets in a cell/tissue type, we evaluated seven methods as detailed in METHODS in the six human cell lines (A549, HCT116, HepG2, K562, MCH-7 and SH-HY5Y) where WHG-STARR-seq data were available. We first trained a logistic regression (LR) classifier model on each pair of the positive and the negative sets defined by each method(Table 1) in each cell line, using four more widely available epigenetic marks(CA, H3K4me1, H3K27ac, H3K4me3) as features(Figure 1A), and validated the performance by 10-fold cross-validation(TFCV). To our surprise, the models trained on the Bin+S^+/Bin+S^- sets only achieved a mediocre median AUROC of 0.621(Figure 1B), indicating that the models had little capability to discriminate sequence bins that overlapped STARR peaks(Bin+S^+) and those that did not (Bin+S^-). This result, however, is consistent with the facts that the two sets had little differences in their patterns of the four epigenetic marks(Figure S2A). To further investigate the underlying reasons, we plotted the distribution of the GERP conservation scores of non-coding nucleotide positions of the positive Bin+S^+ sets and the negative Bin+S^- sets, pooled from the six cell lines. As shown in Figure S3A, although positions in the Bin+S^+ sets are under slightly more evolutionary constraints than those in the Bin+S^- sets, both sets evolve much like all non-coding positions of the 85% regions of the human genome covered by TF binding peaks(Ni and Su 2021), suggesting that like the negative Bin+S^- sets, a large portion of the positive Bin+S^+ sets were not even CRMs. Consistently, genes closest to the Bin+S^+ sequences had quite low expression levels, which were slightly, yet significantly higher($p<4.64\times10^{-58}$) than those closest to the Bin+S^- sequences, notwithstanding (Figure 1C). Such low expression levels of the genes are understandable, as it has been shown that episomal expression vectors used to define WHG-STARR-seq peaks do not mimic native chromosomal contexts of assessed sequences, resulting in up to 87.3% FDRs(Liu et al.

2017). When trained on the Bin+ac$^+$/Bin+ac$^-$ sets or the Bin+S$^+$&ac$^+$/Bin+S$^-$&ac$^-$ sets, the models could very accurately discriminate the positive sets and the negative sets with a very highly mean AUROC of 0.996 and 0.973, respectively(Figure 1B), consistent with the fact that the positive sets and the negative sets defined by the two methods had distinguishable patterns of the four epigenetic marks (Figures S2B, S2C). However, we suspected that the superior performance might be artifacts due to the aforementioned reason that H3K27ac was used both as one of the features and as the label of the training sets in both methods(Table 1). Indeed, when the models were trained on the Bin+ac$^+$/Bin+ac$^-$ sets or the Bin+S$^+$&ac$^+$/Bin+S$^-$&ac$^-$ sets using only the other three marks(CA, H3K4me1 and H3K4me3) as features, their performance reduced 17.7% and 8.0% with an intermediate median AUROC of 0.826 and 0.895, respectively(Figure 1B). Although non-coding positions of positive Bin+ac$^+$ and Bin+S$^+$&ac$^+$ sets are under moderately more evolutionary constraints than those in the respective negative Bin+ac$^-$ and Bin+S$^-$&ac$^-$ sets, their GERP score distributions differ only slight from those of the 85% regions of the genome(Figures S3B, S3C), suggesting that a considerable portion of the positive sets defined by both methods might not be even CRMs, although they were heavily marked by H3K27ac (Figures S2B, S2C) and genes closest to the Bin+ac$^+$ and Bin+S$^+$&ac$^+$ sequences had higher expression levels ($p<2.23x10^{-302}$) than those closest to the Bin+ac$^-$ and Bin+S$^-$&ac$^-$ sequences, respectively (Figure 1C). Since sequences marked by H3K27ac are not necessarily CRMs(Dogan et al. 2015; Arbel et al. 2019), and H3K27ac is not essential for active CRMs(Zhang et al. 2020), it appears that the models were trained to differentiate sequence bins that were marked by H3K27ac or not, rather than active and non-active CRMs. When trained on the CRM+TF$^+$/CRM+S$^-$ sets or the CRM+TF$^+$/non-CRM sets, the models performed equally very well with a median AUROC of 0.976 and 0.975, respectively (Figure 1B), consistent with the facts that the positive CRM+TF$^+$ sets (our predicted active CRMs) had distinct patterns(Figures S2D, S2E) of the four marks than both negative sets, i.e., non-CRM (our predicted non-CRMs) and CRM+S$^-$ (our predicted CRMs that did not overlap STARR beaks)(Table 1). Interestingly, the two negative sets(CRM+S$^-$ and non-CRM) had

indistinguishable patterns of the four marks(Figure S2F), suggesting that our predicted non-active CRMs and non-CRMs in a cell/tissue types might have highly similar patterns of the four marks. Non-coding positions of the positive CRM+TF$^+$ sets are subject to slightly more evolutionary constraints than those of the negative CRM+S$^-$ sets, but both are substantially more evolutionarily constrained than those of the 85% regions of the genome(Figure S3D), and thus might be true CRMs as we argued earlier(Ni and Su 2021). Nonetheless, the CRM+TF$^+$ sets and the CRM+S$^-$ sets differed in their epigenetic marks (Figure S2D), and thus, also might differ in their functional states. In stark contrast, non-coding positions of the negative non-CRM sets are more likely evolutionarily neutral than those in the 85% regions of the genome(Figure S3E), and are unlikely CRMs as we argued earlier(Ni and Su 2021). Consistently, genes closest to the CRM+TF$^+$ sequences had significantly higher expression levels($p<2.23x10^{-302}$) than those closest to the CRM+S$^-$ or non-CRM sequences, and the latter two groups of genes had very low expression levels(Figure 1C). When trained on the CRM+S$^+$/CRM+S$^-$ sets and the CRM+S$^+$/non-CRM sets, the models also performed well with a median AUROC of 0.943 and 0.972, respectively, but slightly worse than when trained on the CRM+TF$^+$/CRM+S$^-$ sets (median AUROC=0.976) and the CRM+TF$^+$/non-CRM sets(median AUROC=0.975)(Figure 1B). As expected, non-coding positions of the CRM+S$^+$/CRM+S$^-$ sets(Figure S3F) and of the CRM+S$^+$/non-CRM sets (Figure S3G) evolve quite similarly to those of the CRM+TF$^+$/CRM+S$^-$ sets (Figure S3D) and those of the CRM+TF$^+$/non-CRM sets (Figure S3E), respectively. Consistently, genes closest to the positive CRM+S$^+$ sequences had significantly higher expression levels ($p<2.23x10^{-302}$) than those closest to the negative CRM+S$^-$ or non-CRM sequences(Figure 1C). Unlike cases of the Bin+ac$^+$/Bin+ac$^-$ and Bin+S$^+$&ac$^+$/Bin+S$^-$&ac$^-$ sets that used H3K27ac as the label, omitting H3K27ac as one of features had little effects on the performance of models trained on all other five types of datasets that did not use H3K27ac as the label(Figure 1B, Table 1). Notably, when H3K27ac was omitted as a feature, models trained on the CRM+TF$^+$/CRM+S$^-$ sets and the CRM+TF$^+$/non-CRM sets performed significantly better ($p<0.001$) than those trained on all other datasets(Figure 1B). Consistently, H3K27ac had the highest weights

($p<0.001$) in the LR models trained on the Bin+ac$^+$/Bin+ac$^-$ and Bin+S$^+$&ac$^+$/Bin+S$^-$&ac$^-$ sets, while CA had the highest weights ($p<0.001$) in the models trained on the other datasets except for the CRM+S$^+$/CRM+S$^-$ sets where H3K4me3 had the highest weights ($p<0.001$)(Figure 1D). These results suggest that a mark should not be used both as a feature and as the label in training datasets to avoid artifacts.

To see why models trained on the CRM+TF$^+$/CRM+S$^-$ sets and the CRM+TF$^+$/non-CRM sets perform better than those trained on the CRM+S$^+$/CRM+S$^-$ sets and the CRM+S$^+$/non-CRM sets (Figure 1B), we plotted heat maps of signals of the four epigenetic marks around each positive set and its two matched negative sets. As shown in Figures 2 and S4, in all the cell lines (data in the SH-HY5Y cell line were not available to us), the two negative sets for either positive set indeed had virtually indistinguishable patterns of the four marks as indicated earlier(S2F), while the positive CRM+TF$^+$ set had stronger CA and H3K4me3 signals than the positive CRM+S$^+$ set, and the reverse was true for the H3K27ac and H3K4me1 signals. The stronger CA signals of the CRM+TF$^+$ set might largely account for the better performance of models trained on it than those trained on the CRM+S$^+$ set, given that CA was the most important feature in the models(Figure 1D). In summary, LR models trained on the positive and negative sets defined by the seven methods(Table 1), have from mediocre to very accurate performance for differentiating the defined active and non-active CRMs in a cell line, and models trained on the CRM+TF$^+$/non-CRM sets or the CRM+TF$^+$/CRM+S$^-$ sets achieve the highest accuracy.

We next asked whether other machine-leaning methods (AB, DT, linear SVM, NB, NN and RF) could also accurately discriminate the CRM+TF$^+$/non-CRM sets as well as the CRM+TF$^+$/CRM+S$^-$ sets in the six cell lines using the four marks as features. As shown in Figure 1E, like LR all the other six classifiers also achieved a very high mean AUROC (>0.970) in differentiating the positive sets and the negative sets defined by the two methods, indicating that our predicted active CRMs (CRM+TF$^+$) can be very accurately and

robustly differentiated from our predicted non-CRMs or non-active CRMs (CRM+S$^-$) using the four epigenetic marks. Notably, RF, DT and AB slightly outperformed the other four classifiers. However, we chose LR for further analysis, as the weights (coefficients) in the models are consistent with those of the linear SVN models (data not shown), and more explainable. Notably, the classifiers trained on the CRM+TF$^+$/CRM+S$^-$ sets slightly outperformed those trained on the CRM+TF$^+$/non-CRM sets. Nonetheless, since STARR-seq data were available in only few human cell lines, and since the two negative sets had virtually indistinguishable patterns of the four marks(Figures 2, S2F,S6~S8), we used the CRM+TF$^+$/non-CRM sets as the training sets in the remaining predictions and analysis.

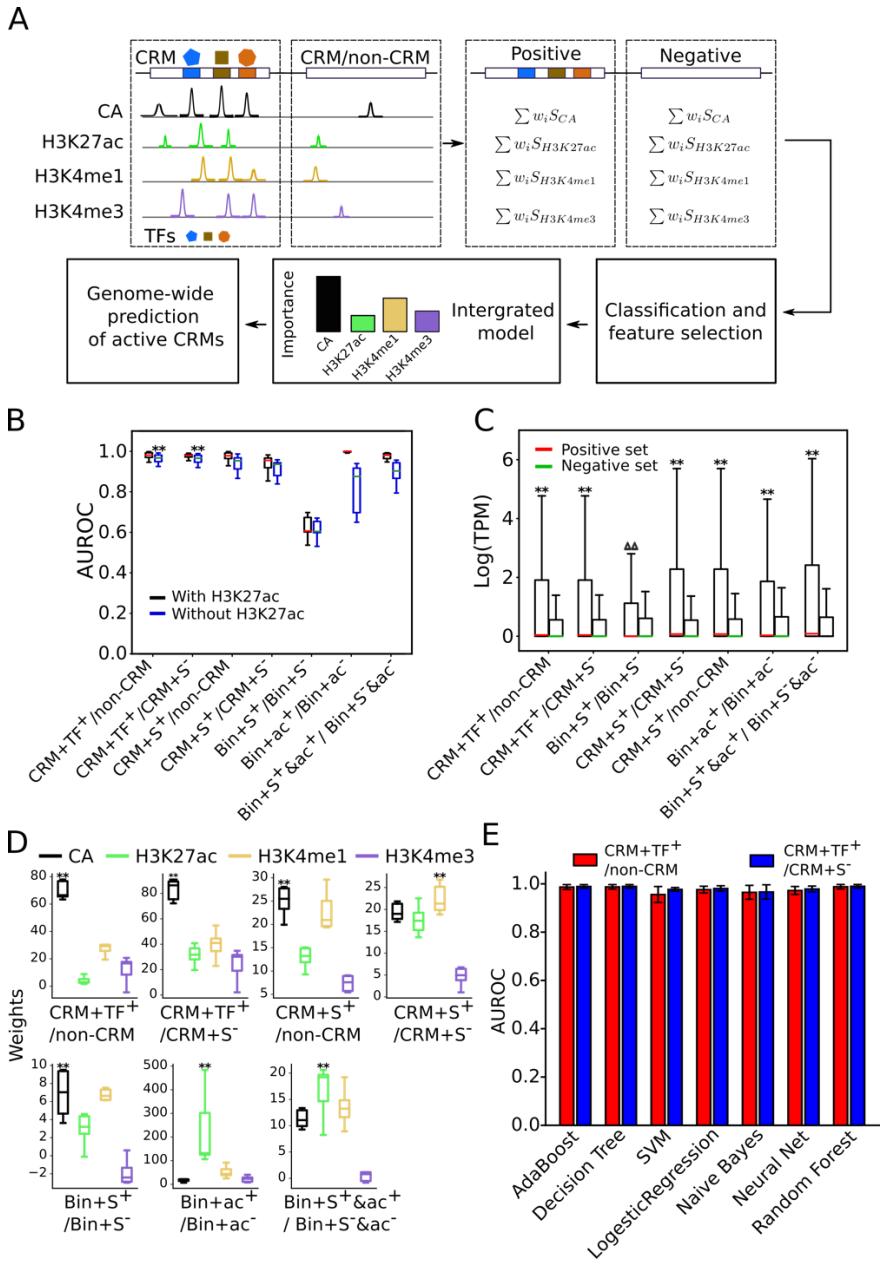

Figure 1. Evaluation of the classifiers and methods for defining active and non-active CRMs. A. Flowchart of our method. **B.** Performance of LR models trained on the positive and the negative sets defined by the seven methods (Table 1) in the six cell lines using the four (CA, H3K4me1, H3K4me3 and H3K27ac) or three epigenetic marks (omitting H3K27ac) as features. Models trained on the CRM+TF$^+$/non-CRM sets or the CRM+TF$^+$/CRM+S$^-$ sets outperformed significantly those trained the other datasets when three epigenetic marks (omitting H3K27ac) were used as features; **p<0.001, Wilcoxon signed-rank test. **C.** Boxplot of expression levels of genes closest to the positive and negative sets defined by each method, pooled from the six cell lines with expression data available. $^{\Delta\Delta}p<4.64x10^{-56}$, **$p<2.2x10^{-302}$, , Mann Whitney U test . **D.** Boxplot of weights (coefficients) of the four marks in the LR models for discriminating the positive and the negative sets defined by different methods in the six cell lines. The weight of the mark is significantly higher than those on the other three marks; **p<0.001 Wilcoxon signed-rank test. **E.** Performance of the seven classifiers trained on the CRM+TF$^+$/non-CRM sets or the CRM+TF$^+$/CRM+S$^-$ sets in the six cell lines using the four marks as features.

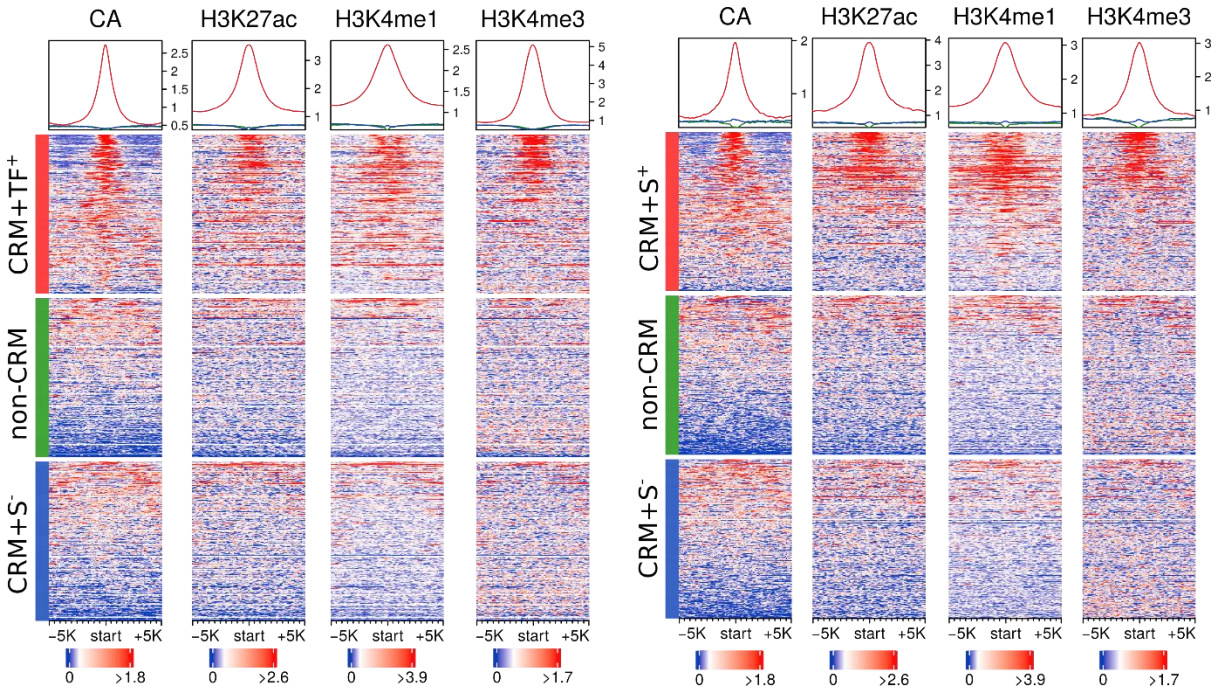

Figure 2. Heat maps of the four epigenetic marks signals around the positive CRM+TF$^+$ and CRM+S$^+$ sets, and the negative non-CRM and CRM+S$^-$ sets in the HCT116 cells. We plotted signal of CA, H3K4me1, H3K3me1, and H3K27ac around the sequences in each set. To make the density plots, we first extended the centers of sequences to 10 kb, and then for each 100-bp tiling window in the extended regions, and calculated the signal using EnrichedHeatmap(w0 mode)(Gu et al. 2018). Note that the

scales of vertical axes of the density plots are different in the right and left panels. The sequences in each set are sorted in the descending order of the CA signals for all the marks.

A minimum of four epigenetic marks is sufficient to accurately predict functional states of CRMs

It was recently reported that machine-learning models trained using six epigenetic marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac and CA) had almost the same power as models trained using 30 marks in differentiating STARR peaks and negative control sequences(Sethi et al. 2020). To find an even smaller set of epigenetic marks that can accurately predict the functional states of cell type agnostic CRMs predicted by dePCRM2(Ni and Su 2021), we first considered the 15 possible combinations of four arguably most important and widely availability marks, CA, H3K4me1, H3K27ac and H3K4me3. Using each combination of the four marks as features, we trained a LR classifier on the CRM+TF^{+}/non-CRMC sets in each of the 67 human and the 64 mouse cell/tissue types. In the case of the human datasets, H3K27ac peaks on average covered the largest number of genome positions, followed by H3K4me1, H3K4me3 and CA peaks(Figure 3A). Although there were extensive overlaps among the four marks in a cell/tissue type, only a small number of genome positions were covered by peaks of all the four marks(Figures 3A). Each mark alone had varying capability of predicting active and non-active CRMs, with CA(model 3) performing best, followed by H3K4me1(model 1), H3K27ac(model 2), and H3K4me3(model 4), with a median AUROC of 0.953, 0.818, 0.803, and 0.755, respectively (Figure 3A). This result is in excellent agreement with the earlier finding that CA alone can be a good predictor of activities of CRMs when combined with TF binding data(Kwasnieski et al. 2014). However, we noted that the models perform poorly in some cell/tissue types for each mark(the first row in Figure S5), due probably to the low quality of the relevant datasets. Among the six models using a combination of two marks as features, model 10 using CA+H3K4me1 achieved the highest median AUROC of 0.974(Figure 3A). Of the four models using a combination of three marks as features, model 13 using CA+H3K4me1+H3K27ac

and model 14 using CA+H3K4me1+H3K4me3 performed almost equally better with a median AUROC of 0.981 and 0.980, respectively, than the other two models($p<0.001$) (Figure 3A). Model 15 using all the four marks (CA+H3K4me1+H3K27ac+ H3K4me3) had the highest median AUROC of 0.986 among all the 15 models ($p<0.001$)(Figures 3A, 3B). CA had the highest contribution among the four marks to predicting functional states of CRMs in model 15(Figure 3C), in agreement with our earlier result based on the six cell lines(Figure 1D). This therefore is in stark contrast with the earlier result that H3K27ac was the most important feature for predicting H3K27ac labelled sequences(Sethi et al. 2020), due probably to the aforementioned reason that the mark was used both as a feature and as the label(Figures 1B, 1D). Moreover, with the increase in the number of marks used as features, the variation of performance of the models in different cell/tissue types decreased(Figure S5), suggesting that the effects of a low quality dataset for a mark can be compensated by using datasets of other marks in the cell/tissue type. Clearly, adding an additional mark to the feature list of a model always led to a new model that outperformed the original one(Figure 3A). However, the infinitesimal increments of 0.005(0.5%) and 0.006(0.6%) in the median AUROC of model 15 over models 13(CA+H3K4me1+H3K27ac) (0.981) and 14 (CA+H3K4me1+ H3K4me3)(0.980)(Figure 3A), respectively, suggests that the improvement of accuracy is already in the later phase of saturation. To verify this, we next trained LR models using five (adding H3K4me2 or H3K9ac to the four marks) or six (adding both H3K4me2 and H3K9ac to the four marks) marks as features on 22 human cell/tissue types in which all the six marks datasets were available. As shown in Figure S6A, the models trained on the four marks achieved a median AUROC of 0.9798 in the 22 human cell/tissue types, while adding H3K9ac or H3K4me2 to the four marks improved the median AUROC(0.9801 or 0.9807) by only 0.0003(0.03%) or 0.0009(0.09%), respectively, and adding both marks improved the median AUROC(0.9812) by only 0.0014(0.14%). Thus, improvement of accuracy by using more than four marks is indeed very limited. Therefore, four marks CA+H3K4me1+ H3K4me3+H3K27ac, or even three marks (CA+H3K4me1+H3K27ac, or CA+H3K4me1+H3K4me3) are sufficient to very accurately

predict functional states of our predicted CRMs in a cell/tissue type(Figure 3A). In the model using the six marks as features, CA again has the highest weights, followed by H3K4me1, H3K4me2, H3K4me3, H3K27ac and H3K9ac(Figure S6B).The rapid saturation of AUROC values with the increase in the number of marks used as features suggests redundancy of information in data of different marks. To reveal this, we computed Pearson's correlation coefficients(PCC) between each pair of the six marks on the CRMs in the positive set defined in each of the 22 human cell/tissue types. Indeed, all pairs of the six marks have varying levels of positive correlations(Figure S6C). Specifically H3K4me1 and H3K9ac(PCC=0.12) as well as H3K4me1 and H3K4me3(PCC=0.14) have low correlations; H3K4me3 and H3K9ac(PCC=0.86) as well as H3K27ac and H3K9ac(PCC=0.81) have high correlations; and all the other pairs have intermediate(0.27~0.73) of correlations. The correlations can also be seen from heat maps of signals of the six marks around the CRMs in the positive set(CRM+TF+) in each cell/tissue type as shown in Figure S6D for the GM12878 cells as an example. The same conclusions are drawn from the results obtained using the mouse datasets(Figures S7~S9). However, the models generally performed better in the mouse datasets(Figures S7~S9) than in the human datasets(Figures 3, S5, S6), due probably to the better quality of mouse datasets, as evidenced by the less variation of the performance of the models on the datasets using single marks (the first row in Figure S8 vs the first row in Figure S5). Therefore, we used the four marks (CA+H3K4me1+H3K4me3+ H3K27ac) as features in the remaining predictions and analysis, considering their wider availability, albeit H3K4me2 had slightly higher weights in the models than H3K4me3, H3K27ac and H3K9ac(Figures S6B, S9B).

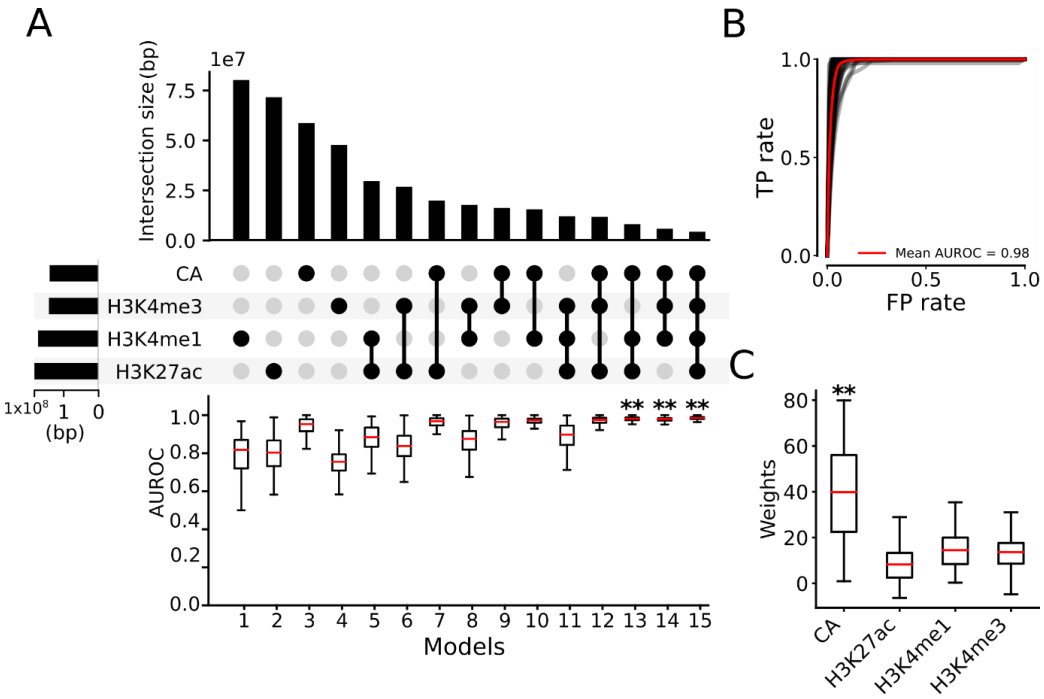

Figure 3. Importance of different epigenetic marks in predicting functional states of CRMs in human cell/tissue types. A. Upset plot showing intersections of genome positions covered by the peaks of the four epigenetic marks (upper and middle panels), and boxplot of AUROCs of the 15 LR models using the 15 different combinations of the four marks in 67 human cell/tissue types. Models 13, 14 and 15 have similar AUROCs, each is significantly higher than those of the other 12 models; **p<0.001, Wilcoxon signed-rank test. **B.** ROCs of model 15 using the four marks in the human cell/tissue types. Each gray curve is the ROC for a cell/tissue type, and red one is the average ROC of the 67 cell/tissue types. **C.** Boxplot of weights (coefficients) of the four marks in model 15. The weights of CA are significantly higher than those on the other three marks; **p<0.001, Wilcoxon signed-rank test.

Epigenetic code defining functional states of CRMs is universal in different cell types of a species and even in different species

To see whether distinct epigenetic patterns on active and non-active CRMs (Figures 2, S4) learned in a cell/tissue type in a species can be transferred to different cell types in the same species, we trained LR models on pooled positive and negative sets from n-1 human cell/tissue types (n=67), and tested it on the left-out cell/tissue type (leave one out cross validation, LOOCV, see METHODS). As shown in Figure 4A, the models trained on different cell/tissue types achieved quite a high median AUROC of 0.985 in left-out ones, which was

not significantly different from that (0.986) achieved by the models that were trained and tested on the same cell/tissue types using TFCV (Figure 4A, $p<0.327$). Similar results (median AUROC= 0.990 vs. 0.990) were obtained using the datasets from 64 mouse cell/tissue types (Figure 4B, $p<0.353$). These results strongly suggest that epigenetic patterns of the four marks that define active and non-active CRMs in different cell/tissue types in the same species are the same. To see whether epigenetic patterns of CRMs learned in multiple cell/tissue types in a species can be transferred to various cell types in a different species, we trained the models on the 64 mouse cell/tissue types, and tested it on each of the 67 human cell/tissue types, and vice versa. As shown in Figures 4A and 4B, very high median AUROCs of 0.969 and 0.954 were achieved in human and mouse cell types, which were only slightly but significantly lower ($p<2.01\times10^{-12}$ and $p<3.70\times10^{-12}$) than those achieved by the models that were trained and tested in the same species (median AUROC=0.985 and 0.990), respectively. Notably, the model trained on mouse datasets applying to human cell/tissue types performed better than that trained on human datasets applying to mouse cell/tissue types, due probably to the better quality of mouse datasets (Figures S5, S8). These results strongly suggest that epigenetic patterns that define active and non-active CRMs are highly conserved in human and mouse. As in the earlier case where the models were trained and tested in the same cell/tissue types (Figures 3C, S10C), CA was the most important feature for predicting functional states of CRMs in these two latter cases(Figures 4C~4F). Therefore, it appears that epigenetic code defining functional states of CRMs is universal in different cell types in same species and even in at least different mammalian species.

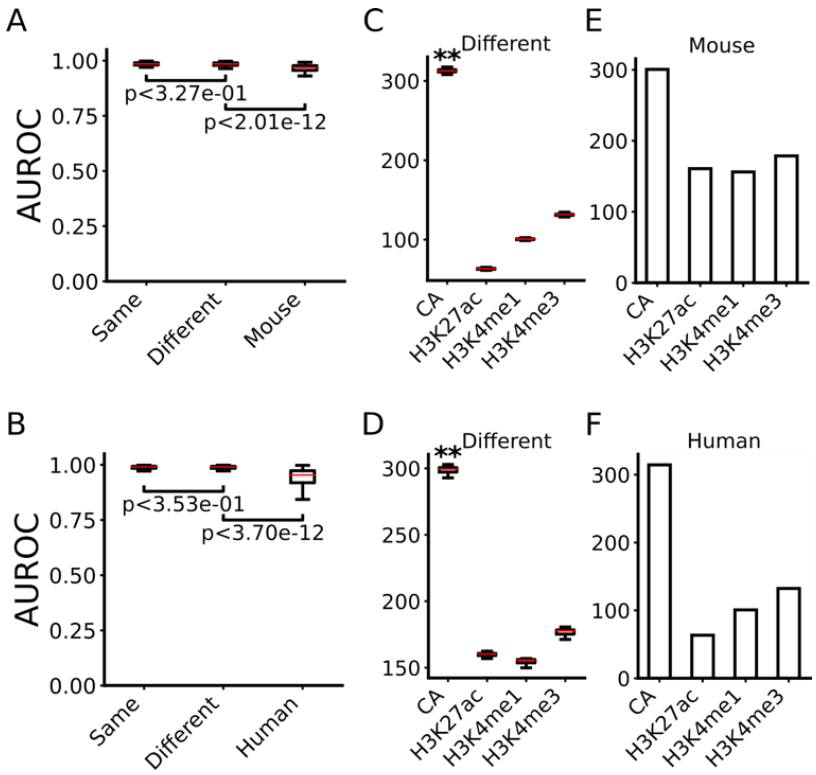

Figure 4. Universal epigenetic code defining functional states of CRMs in various cell types in different species. A, B. Boxplots of AUROCs of the models trained and tested in the same cell/tissue type in human (A) or mouse (B), of the models trained in multiple cell/tissue types in human (A) or mouse (B) and tested in different cell/tissue types in the same species, and of the model trained in multiple cell/tissue types in mouse (A) or human (B) and tested in various cell/tissue types in human or in mouse, respectively. The statistical tests were done using Wilcoxon signed-rank test. **C, D.** Boxplot of coefficients of the four marks in the LR models trained in multiple cell/tissue types in human (C or mouse (D) and tested in different cell/tissue types in the same species. The weights of CA are significantly higher than those on the other three marks, **p<0.001, Wilcoxon signed-rank test. **E, F.** Coefficients of the four marks in the model trained in multiple cell/tissue types in mouse (E) or in human (F) and tested in various cell/tissue types of the human or in Human.

The models have similar performance in predicting proximal and distal CRMs

CRMs can be largely classified into proximal and distal ones based on their distances to their nearest transcription start sites (TSSs). The former category often overlaps TSSs and functions as core or proximal promoters, while the latter category often functions as enhancers or silencers, though such classification is not clear-cut, as some promoters can also function as distal enhancers of remote genes(Andersson et al. 2015). To see how well

the LR models predict functional states of these two categories of CRMs, we split both a positive set and a negative set into a proximal set (within 1,000bp of nearest TSSs) and a distal set (more than 1,000bp away from nearest TSSs), and separately trained and tested the models on the proximal and distal positive and negative sets. When trained and tested in the same cell/tissue types, the models performed well in predicting active proximal(Figures 5A, 5B) and distal CRMs (Figures 6A, 6B) in both human and mouse cell/tissue types using TFCV, with a median AUROC of 0.989 and 0.983, and 0.992 and 0.988, respectively. The models trained on n-1 cell/tissue types in human and mouse also performed well in left-out cell types in the same species in predicting active proximal (Figures 5A, 5B) and active distal CRMs(Figures 6A, 6B), with a median AUROC of 0.988 and 0.984, and 0.993 and 0.988, respectively, which were not significantly different from or even better than those of models trained and tested in the same cell/tissue types(Figures 5B), due probably to the larger training stets (n-1 vs 1). The models trained on multiple mouse or human cell/tissue types also performed well in various cell/tissue types in the other species in predicting active proximal (Figures 5A, 5B) and active distal CRMs (Figures 6A, 6B), with a median AUROC of 0.962 and 0.964, and 0.966 and 0.942, respectively. However, the performance of the models reduced slightly but significantly ($p<1.12x10^{-12}$ and $p<3.53x10^{-12}$) compared with that of the models trained and tested in the same species. Interestingly, in all the models of the three scenarios in both human and mouse, CA and H3K4me3 contributed more ($p<0.001$) than H3K4me1 and H3K27ac for predicting active proximal CRMs (Figures 5C~5H), while CA and H3K4me1 contributed more ($p<0.001$) than H3K4me3 and H3K27ac for predicting active distal CRMs(Figures 6C~6H). These results are consistent with the findings that both active promoters and enhancers are nucleosome-free(Armstrong and Emerson 1996; Tirosh and Barkai 2008; He et al. 2010), but the former tend to have a higher level of H3K4me3 and a lower level of H3K4me1 in flanking regions, and the opposite is true for the latter(Santos-Rosa et al. 2002; F and RM 2005; Heintzman et al. 2009; Spicuglia and Vanhille 2012; Howe et al. 2017; Local et al. 2018; Rada-Iglesias 2018).

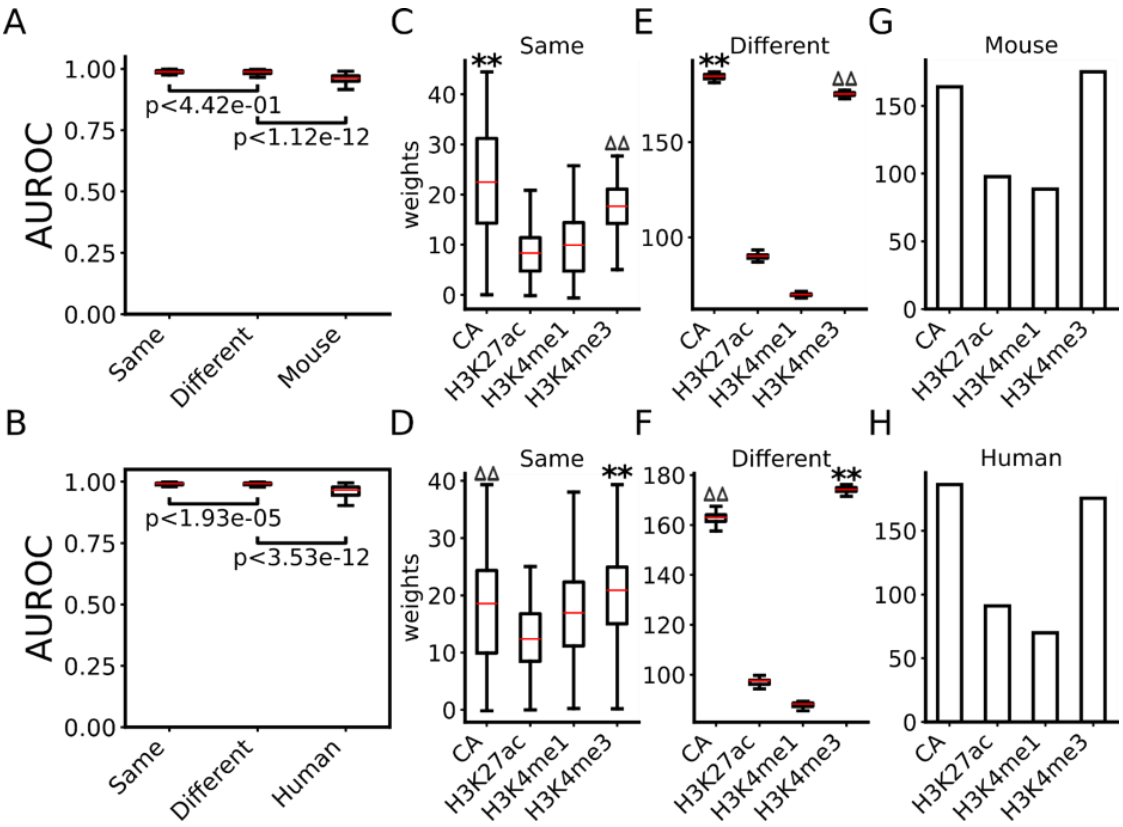

Figure 5. Performance of the model for prediction proximal CRMs. A, B. Boxplots of AUROCs of the model trained and tested in a same cell/tissue type in human (A) or mouse (B), of the model trained in multiple cell/tissue types in human (A) or mouse (B) and tested in different cell/tissue types in the same species, and of the model trained in multiple cell/tissue types in mouse (A) or human (B) and tested in various cell/tissue types in human or mouse, respectively. The statistical tests were done using Wilcoxon signed-rank test. **C, D.** Boxplot of Coefficients of the four marks in the model trained and tested in the same cell/tissue type in human (C) or mouse (D). **E, F.** Boxplot of coefficients of the four marks in the models trained in multiple cell/tissue types in human (E) or in mouse (F) and tested in different cell/tissue types in the same species. The weights of the mark are significantly higher than those on the other three marks, **p<0.001; the weight of the mark are significantly higher than those on the other two marks, $^{\Delta\Delta}$p<0.001; Wilcoxon signed-rank test. **G, H.** Coefficients of the four marks in the model trained in multiple cell/tissue types in mouse (G) or human (H) and tested in various cell/tissue type in human or mouse, respectively.

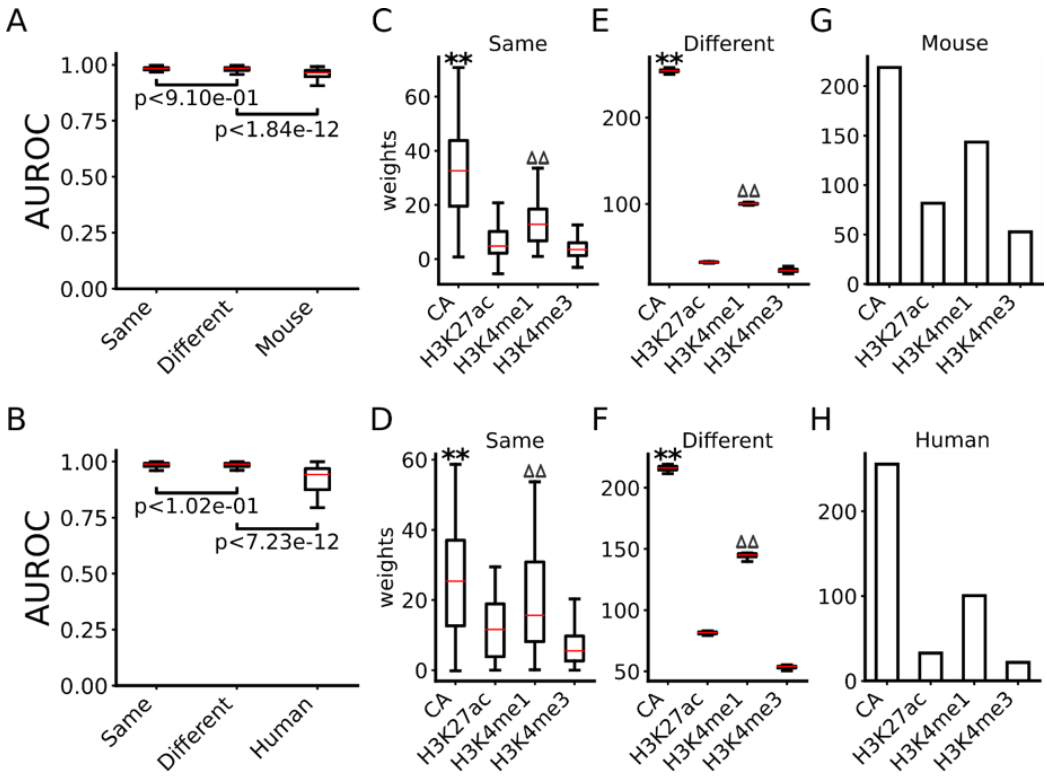

Figure 6. Performance of the model for prediction distal CRMs. A, B. Boxplots of AUROCs of the model trained and tested in a same cell/tissue type in human (A) or mouse (B), of the model trained in multiple cell/tissue types in human (A) or mouse (B) and tested in different cell/tissue types in the same species, and of the model trained in multiple cell/tissue types in mouse (A) or human (B) and tested in various cell/tissue types in human or mouse, respectively. The statistical tests were done using Wilcoxon signed-rank test. **C, D.** Boxplot of Coefficients of the four marks in the model trained and tested in the same cell/tissue type in human (C) or mouse (D). **E, F.** Boxplot of coefficients of the four marks in the models trained in multiple cell/tissue types in human (E) or in mouse (F) and tested in different cell/tissue types in the same species. The weights of the mark are significantly higher than those on the other three marks, **p<0.001; the weights of the mark are significantly higher than those on the other two marks, $^{\Delta\Delta}$p<0.001; Wilcoxon signed-rank test. **G, H.** Coefficients of the four marks in the model trained in multiple cell/tissue types in mouse (G) or human (H) and tested in various cell/tissue type in human or mouse, respectively.

Active proximal and distal CRMs cannot be accurately differentiated based on their epigenetic marks

We next asked whether the two categories of active CRMs can be discriminated based on their four epigenetic marks by a classifier trained on the active proximal CRMs and active

distal CRMs as the positive sets and the negative sets, respectively. When trained and tested on the datasets from even the same cell/tissue types, the models performed only moderately well in differentiating active proximal CRMs and active distal CRMs in both human and mouse with a median AUROC of 0.798 and 0.815, respectively(Figures 7A, 7B). The models trained on n-1 cell/tissue types in the human and mouse also performed only moderately well in left-out cell/tissue types in the same species with a median AUROC of 0.787 and 0.788, respectively, which were significantly ($p<1.41x10^{-10}$, $p<6.42x10^{-11}$) lower than those of the models trained and tested in the same cell/tissue types(Figures 7A, 7B). The models trained on mouse or human cell/tissue types also performed only moderately well human and mouse cell/tissue types with a median AUROC of 0.739 and 0.753, respectively, which were significantly lower ($p<1.12x10^{-12}$, $p<4.49x10^{-9}$) than those of the models trained and tested in cell/tissue types in the same species (Figures 7A, 7B). In almost all the models in the three scenarios, H3K4me3 had the highest positive weights ($p<5.49x10^{-9}$) and H3K4me1 the lowest negative weights ($p<0.001$), while CA and H3K27ac near zero weights (Figures 7C~7H). These results suggest that the two categories of active CRMs have largely opposite patterns of H3K4me3 and H3K4me1 modifications, but largely extinguishable CA and H3K27ac modifications, consistent with the current understanding of histone modifications on promoters and enhancers(Spicuglia and Vanhille 2012; Howe et al. 2017; Local et al. 2018; Rada-Iglesias 2018). Taken together, active proximal CRMs and active distal CRMs might have conserved different patterns of H3K4me1 and H3K4me3 modifications across various cell/tissue types in the same and even in different species, however, they cannot be accurately differentiated simply based on such differences, due probably to their sometimes overlapping functions(Andersson et al. 2015; Dao and Spicuglia 2018).

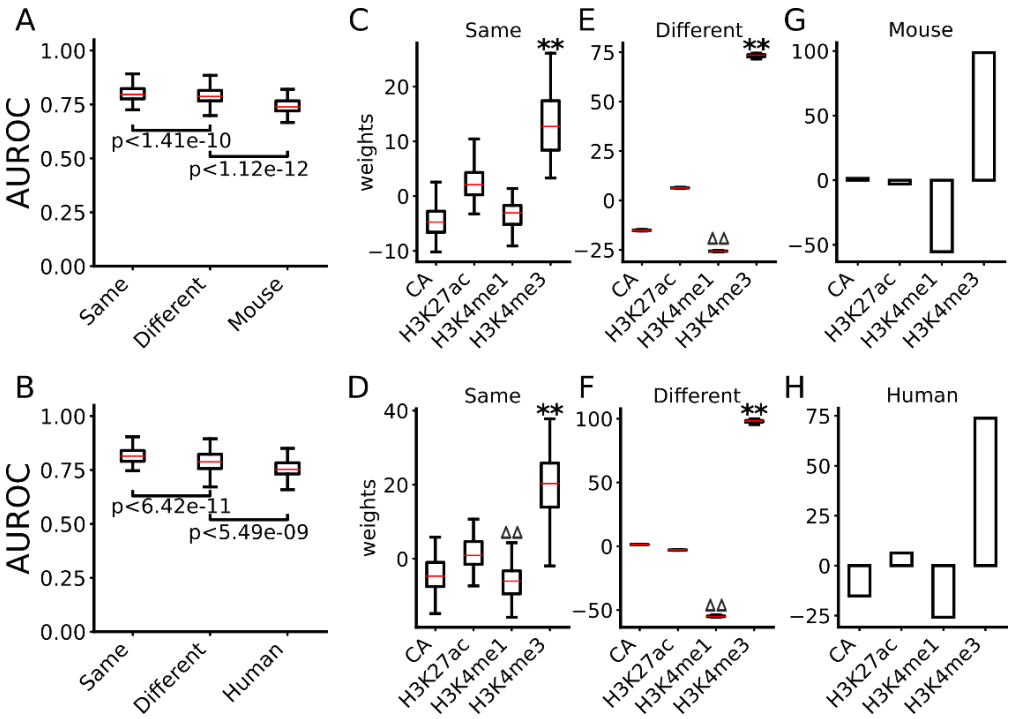

Figure 7. Performance of the model for differentiating active proximal CRMs and active distal CRMs. A, B. Boxplots of AUROCs of the model trained and tested in a same cell/tissue type in human (A) or mouse (B), of the model trained in multiple cell/tissue types in human (A) or mouse (B) and tested in different cell/tissue types in the same species, and of the model trained in multiple cell/tissue types in mouse (A) or human (B) and tested in various cell/tissue types in human or mouse, respectively. The statistical tests were done using Wilcoxon signed-rank test. **C, D.** Boxplot of coefficients of the four marks in the model trained and tested in the same cell/tissue type in human (C) or mouse (D). **E, F.** Boxplot of coefficients of the four marks in the models trained in multiple cell/tissue types in human (E) or in mouse (F) and tested in different cell/tissue types in the same species. The weights of the mark are significantly higher than those on the other three marks, **p<0.001; the weights of the mark are significantly lower than those on the other three marks, $^{\Delta\Delta}$p<0.001; Wilcoxon signed-rank test. **G, H.** Coefficients of the four marks in the model trained in multiple cell/tissue types in mouse (G) or human (H) and tested in various cell/tissue type in human, or mouse, respectively.

Most CRMs are re-utilized in different cell/tissue types

The human genome was estimated to encode 44K genes (20K protein genes plus 24K RNA genes)(Salzberg 2018). We recently estimated that the human genome might encode 1.47M CRMs(Ni and Su 2021). Although a cell type was estimated to express up to 15K

genes(Ramskold et al. 2009; García-Ortega and Martínez 2015), little is known about the number of active CRMs in a cell type. Thus, it is interesting to predict active CRMs in each of the 67 human and 64 mouse cell/tissue types, particularly, active CRMs that are missed by dePCRM2 due to the lack of sufficient TF binding data(Ni and Su 2021). To this end, we trained a LR model on pooled positive CRM+TF^{+} sets and negative non-CRM sets from all the cell/tissue types in human or mouse(METHODS), and used the model to predict functional states in each cell/tissue type of all the 1,225,115 and 798,258 predicted CRMs in the 85% and 80% regions of the human and mouse genomes, respectively(METHODS). We found that the number of predicted active CRMs in a cell/tissue type varied widely ranging from 37,792 to 313,389 and from 37,899 to 180,827 with a mean of 133,250 and 89,751 in human and mouse cell/tissue types, respectively. Thus, the number of active CRMs in a cell/tissue type is much larger than that predicted by dePCRM2 based on available TF binding data (Figures S1, 8A, S10A), indicating that the LR models predict active CRMs missed by dePCRM2. In both human(Figure 8A) and mouse(Figure S10A), embryonic stem cells and complex tissues(such as brain) generally had more active CRMs than terminally differentiated cell types(such as T or B lymphocytes). Genes closest to predicted active CRMs had significantly higher ($p<2.23X10^{-302}$) expression levels than those closest to predicted non-active CRMs in all the 19 human(Figure 8B) and 14 mouse(Figure S15B) cell/tissue types where gene expression data were available, suggesting that CRMs with the learned patterns of epigenetic modifications (Figures 2, S4) and bund by TFs might be in active states, or in non-active states, otherwise. Therefore, at least most of our predicted active CRMs and non-active CRMs in the cell/tissue types might be authentic.

Of the 1,225,115 and 798,258 predicted CRMs in the 85% and 80% regions of the human and the mouse genomes, 68.28%(836,527) and 63.26%(505,016) were predicted to be active in at least one of the 67 and 64 human and mouse cell/tissue types, respectively (Figures 8C, S10C). The remaining 31.72% and 36.14% of the predicted CRMs were predicted to be non-active in the human and mouse cell/tissue types analyzed, respectively.

It is likely that these non-active CRMs are active in other cell/tissue types that were not analyzed in this study due to the lack of datasets for the four marks. Interestingly, of all the predicted active CRMs in the human (n=836,527) and mouse (n=505,016) cell/tissue types, only 22.44% (187,688) and 20.61% (104,074) were used in only a single cell/tissue type, while the remaining 77.56% and 79.39% were reused in at least two cell/tissue types analyzed(Figure 8C, S10C), indicating that most CRMs were reused in different cell/tissue types. The number of uniquely active CRMs in a cell/tissue type ranged from 27 to 43,333 and from 32 to 6,914 with a mean of 2,801 and 1,626, comprising from 0.02% to 13.83% and from 0.08% to 5.62% of predicted active CRMs in a cell/tissue type in human (Figures S11A, S11B) and mouse (Figures S11C, S11D), respectively. Gene ontology (GO) term analysis (Ashburner et al. 2000; Mi et al. 2019; Consortium 2021) indicates that genes closest to uniquely active CRMs are involved in cell type specific functions. For example, uniquely active CRMs in human embryonic stem cells (H1) are closest to genes enriched for 236 GO terms for developments, such as tongue development (GO:0043586), establishment of epithelial cell polarity (GO:0090162), and negative regulation of axon extension (GO:0030517), to name a few (Table S3), while uniquely active CRMs in human brain tissues are closest to genes enriched for 177 GO terms for neuronal functions, such as inhibitory synapse assembly (GO:1904862), neuron cell-cell adhesion (GO:0007158), and ionotropic glutamate receptor signaling pathway (GO:0035235), etc, (Table S4). The similar results are seen for uniquely active CRMs in mouse embryonic stem cells(Table S5) and mouse brain cells (Table S6). These results suggest that a cell/tissue type might be determined by a set of the uniquely active CRMs in the cell/tissue type. On the other hand, there were a total of 39,942 and 29,857 genes (including non-protein genes) closest to the uniquely active CRMs in the human and mouse cell/tissue types, but only a total of 12,389(31.0%) and 11,741(39.2%) of them were unique to a human (Figure 8D) and a mouse(Figure S10D) cell/tissue type, respectively. The remaining majority of genes were shared by at least two cell/tissue types, but no gene was shared by all the 67 human or 64 mouse cell/tissue types(Figures 8D, S10D). This result is in agreement with the notion that a

terminal cell type is determined by a unique combination of otherwise more widely expressed genes(Hobert 2011; Doitsidou et al. 2013).

Interestingly, in both human and mouse cell/tissue types, the number of shared active CRMs decreased largely monotonously with the increase in the number of sharing cell/tissue types, with the exception that the number of active CRMs shared by all the cell/tissue types in human (9,537(1.14%)) and in mouse(8,869(1.11%)) analyzed was larger than that shared by fewer number of cell/tissue types(Figures 8C, S10C). The 9,537 and 8,869 active CRMs shared by all the human and mouse cell/tissue types comprised from 3.04% to 25.25% and from 4.91% to 23.4% of active CRMs in a human and mouse cell/tissue type, respectively. GO term analysis (Ashburner et al. 2000; Mi et al. 2019; Consortium 2021) indicates that genes closest to these all-shared active CRMs are enriched for 929 and 859 and GO terms for house-keeping functions in human (Table S7) and mouse(Table S8), respectively, such as amino acid activation (GO:0043038), cell death (GO:0008219), and ribosomal large subunit biogenesis (GO:0042273), to name a few, suggesting that the functions of all-shared CRMs are largely conserved. Interestingly, the all-shared active CRMs tended to be more evolutionarily conserved than the uniquely active CRMs ($p<2.23X10^{-302}$)(Figured 8D and S10D). As expected, like the uniquely active CRMs, the predicted non-active CRMs were less conserved than the all-shared active CRMs($p<2.23X10^{-302}$)(Figured 8D and S10D). It is likely that these non-active CRMs might be uniquely active in or shared by other cell/tissue types yet to be analyzed. To see whether the extent to which cell/tissue types share active CRMs reflects their known lineage relations, we hierarchically clustered the cell/tissue types based on the Jaccard index of predicted active CRMs of each pair of cell/tissue types. As shown in Figures 8F and S10F, cell/tissue types with related lineages indeed formed clusters. Taken together, these results support the notion that cell/tissue types are produced in a step-wise manner through cell differentiation, so that the closer cell types in a lineage, the more gene regulatory programs they share(Briggs et al. 2018; Farrell et al. 2018; Wagner et al. 2018).

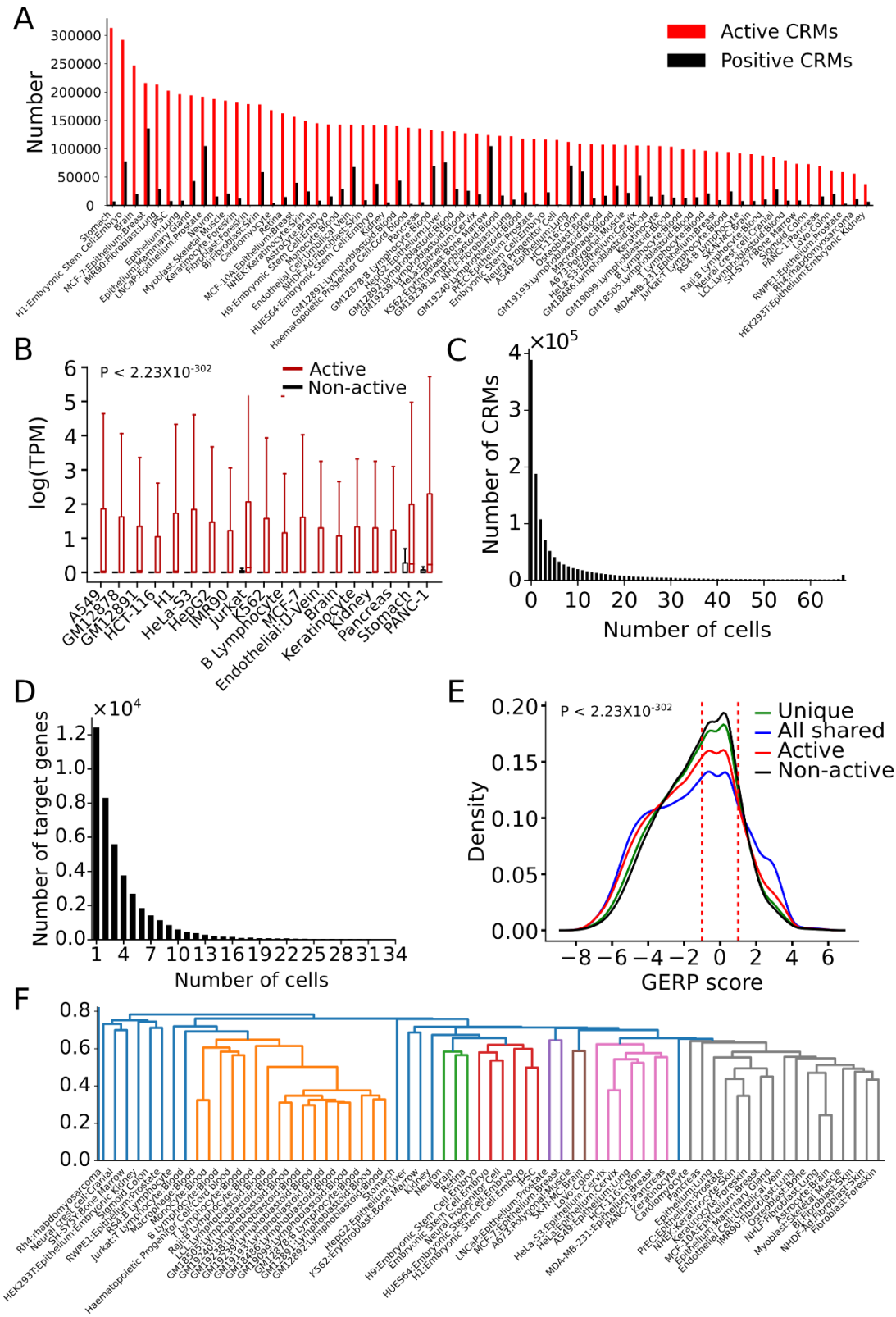

Figure 8. Reutilization of CRMs in different human cell/tissue types. A. Number of active CRMs predicted by the LR models (Active CRMs) and by dePCRM2 (Positive CRMs) in the human cell/tissue types. **B.** Boxplot showing that expression levels of genes closest to predicted active CRMs were significantly higher than those of genes closet to predicted non-

active CRMs in the human cell/tissue types with gene expression data available, $p<2.23X10^{-302}$, Mann Whitney U test. **C.** Number of predicted active CRMs shared by different numbers of cell/tissue types. **D.** Number of closest genes to the uniquely active CRMs, shared by different numbers of cell/tissue types. **E.** Distributions of GERP scores of all-shared active CRMs, uniquely active CRMs, all active CRMs and all non-active CRMs in the human cell/tissue types. All distributions are significantly different from one another, $p<2.23X10^{-302}$, K-S test. **F.** The levels of shared active CRMs reflect lineage relationships of the cell/tissue types. Cell/tissue types were clustered based on the Jaccard index of predicted active CRMs of each pair of the cell/tissue types.

Our predictions are substantially more accurate than earlier results

CRMs and their functional states have been predicted in several mouse tissues by various machine learning classifiers trained on positive and negative sets defined by different methods based on overlaps between epigenetic marks(Firpi et al. 2010; Rajagopal et al. 2013; Lu et al. 2015; He et al. 2017), and more recently between epigenetic marks and STARR peaks(Sethi et al. 2020). We compared our results with those of five state-of-the-art methods in four mouse tissues including hindbrain, limb, midbrain and neural tube, where abundant data including peaks for the four epigenetic marks are available. As shown in Figure 9, our LR models trained on either mouse (LOOCV) or human cell/tissue types substantially outperformed all the five earlier methods, namely, Matched Filter (Sethi et al. 2020), REPTILE (He et al. 2017), FRECS(Rajagopal et al. 2013), DELTA(Lu et al. 2015), CSI-ANN(Firpi et al. 2010) in predicating active CRMs in the four mouse tissue measured by AUROC. This result is remarkable as we only used four epigenetic marks while all these earlier methods used more marks in addition to the four we used.

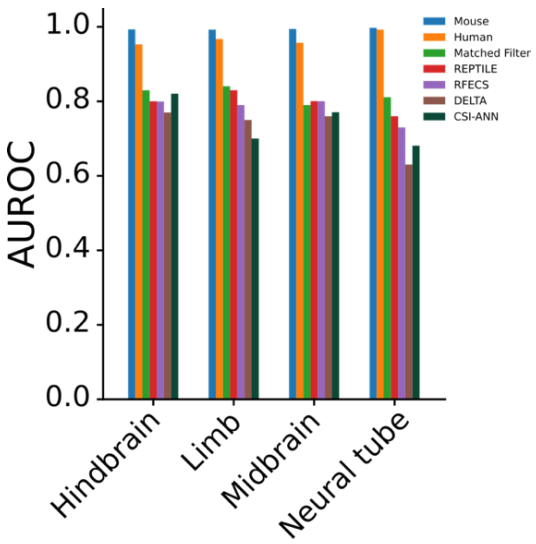

Figure 9. Comparison of the performance of our methods with earlier state-of-the-art methods.

AUROCs of our LR models trained on mouse or human positive and negative sets, and of the five earlier methods (Matched Filter (Sethi et al. 2020), REPTILE (He et al. 2017), FRECS(Rajagopal et al. 2013), DELTA(Lu et al. 2015), CSI-ANN(Firpi et al. 2010)), for predicting CRMs and their functional states in the four mouse tissues. The AUROC values of the five earlier methods were taken from relevant publications.

DISCUSSION

Annotation of CRMs in a genome has three tasks. The first is to identify all CRMs and constituent TFBSs in the genome; the second is to characterize the functional state (active or non-active) of each CRM in each cell/tissue type of the organism; and the third is to determine the target genes of each active CRM in each cell/tissue type. The first and the second tasks are clearly two facets of the same coin, solving one would facilitate solving the other, and thus, ideally should be solved simultaneously. Indeed, most existing methods attempt to simultaneously predict CRMs and their functional states in a cell/tissue type by integrating histone marks and CA data from the cell/tissue type using various machine-learning methods(Ernst and Kellis 2010; Firpi et al. 2010; Ernst et al. 2011; Hoffman et al. 2012; Hoffman et al. 2013; Rajagopal et al. 2013; Ghandi et al. 2014; Kleftogiannis et al. 2015; Gao et al. 2016; Fishilevich et al. 2017; Gao and Qian 2020). Although conceptually attractive, these methods in practice suffer from quite a high FDR(Kheradpour et al. 2013;

Kwasnieski et al. 2014; Dogan et al. 2015; Kleftogiannis et al. 2015; Catarino and Stark 2018; Arbel et al. 2019; Ni and Su 2021) due probably to the fact that a sequence segment with these epigenetic marks such as CA, H3K3me1, H3H4me3 and H3K27ac are not necessarily CRMs(Dogan et al. 2015; Young et al. 2017; Arbel et al. 2019; Zhang et al. 2020; Ni and Su 2021), although active CRMs bear a certain pattern of them(Heintzman et al. 2007; Morse 2010; Rada-Iglesias et al. 2011; Zentner and Scacheri 2012; Trynka et al. 2013; Huang et al. 2015). Moreover, it is difficult if not impossible, to de novo predict TFBSs in CRMs using histone marks and CA data alone. As a result, CRMs predicted by these methods are of low resolution and lack information of constituent TFBSs, although some methods scan predicted CRMs for TFBSs of known motifs(Fishilevich et al. 2017). One way to circumvent the limitations of these methods might be to integrate epigenetic marks data with TF ChIP-seq data in a cell/tissue type, since it has been shown that an active CRM can be more accurately predicted using information of both chromatin modifications and binding of key TFs (Podsiadlo et al. 2013; Kwasnieski et al. 2014; Dogan et al. 2015; Kleftogiannis et al. 2015; Arbel et al. 2019). However, the application of this approach is limited because sufficient TF ChIP-seq data are available in very few well-studied cell lines(Ni and Su 2021). Thus, the fraction CRMs whose functional states in a cell/tissue type can be predicted by this approach could be very small or even zero for most cell/tissue types. Another way to circumvent the limitations of existing methods might be to take a two-step approach as we proposed earlier(Niu et al. 2018; Ni and Su 2021). Septically, in the first step we predict a more accurate and more complete map of CRMs and constituent TFBSs in the genome at single-nucleotide resolution by integrating all available tens of thousands of ChIP-seq datasets for different TFs in various cell/tissues of the organism, because ChIP-seq data are more accurate than histone marks and CA data to locate CRMs (Podsiadlo et al. 2013; Dogan et al. 2015; Kleftogiannis et al. 2015; Arbel et al. 2019). Moreover, as CRMs and TFs are often reused in different cell/tissue types(Davidson 2006), CRM prediction can be enhanced by integrating TF ChIP-seq data from different cell/tissue types. We recently developed a new tool dePCRM2(Ni and Su 2021) for this step. In addition to predicting a

map of CRMs and constituent TFBSs in the genome regions covered by appropriately extended peaks of ChIP-seq datasets, dePCRM2 also can predict a varying fraction of active CRMs in a cell/tissue type, dependent on the number of TF ChIP-seq datasets available in the cell/tissue type(Figures 8A, S1, S10A). However, the functional states of most predicted CRMs in most cell/tissue types cannot be predicted by dePCRM2 due to the lack of sufficient TF ChIP-seq data in most cell/tissue types(Ni and Su 2021), and thus need to be predicted separately in a second step. In this study, we present a method for the second step to predict functional states of all predicted CRMs in any cell/tissue type using a minimal set of only three or four epigenomic marks data in the cell/tissue type, based on the observation that once the locus of a CRM is accurately located by key binding TFs, epigenetic marks can be an accurate predictor of the functional state of the CRM(Podsiadlo et al. 2013; Dogan et al. 2015; Kleftogiannis et al. 2015; Arbel et al. 2019).

We showed that this two-step strategy achieved substantially more accurate prediction of CRMs and their functional states than the existing methods (Figures 1B, 3A, S7A and 9). We attribute the outstanding performance of our strategy to two reasons. First, dePCRM2 provides a high quality positive set (CRM+TF$^+$) as well as a high quality negative set (CRM+S$^-$ or non-CRM) in a cell/tissue type, though the size of the positive set depends on the number of available TF ChIP-seq data in the cell/tissue type(Figure S1). The two negative sets(CRM+S$^-$ or non-CRM) have virtually indistinguishable patterns of light modifications of active epigenetic marks analyzed(Figures 2, S4). In stark contrast, the positive set (CRM+TF$^+$) shows correlated patterns of heavy modifications of active epigenetic marks analyzed (Figures 2, S4). Therefore, it is not surprising that machine-learning models trained on such distinct positive and negative sets performed substantially better than the same models trained on positive and negative sets defined by other methods(Figure 1B). It is worth noting that according to the definitions of the positive and negative sets(Table 1), a machine-learning classifier trained even on the CRM+TF$^+$/CRM+S$^-$ sets differentiated predicted CRMs with TF binding from predicted CRMs without STARR-

seq signals. In other words, the classifier was trained to predict whether or not a CRM was bound by TFs in a cell/tissue type given the epigenetic profile of the CRM. TF binding to a silencer might reduce gene expression, while TF binding to an enhancer may not necessarily enhance gene expression. Nevertheless, we found that genes closest to the positive sets tended to have significantly higher expression levels than those closest to the negative sets, which had very low expression levels(Figure 1C), suggesting that at least most CRMs bound by TFs might enhance gene expression, and thus were likely enhancers or promoters. Therefore, at least most sequences in the positive sets (CRM+TF^{+}) might be active enhancers, while at least most sequences in the negative sets are non-active enhancers (CRM+S^{-}) or even are not CRMs (non-CRM). Second, dePCRM2 provides an more accurate and more complete map of CRMs in up to 85% and 80% regions of the human and mouse genomes, making prediction of functional states of the CRMs less challenging, because as we indicated earlier, it has been shown that when the locus of a CRM is anchored by the binding of key TFs, epigenetic marks become an accurate predictor of the functional state of the CRM(Podsiadlo et al. 2013; Dogan et al. 2015; Kleftogiannis et al. 2015; Arbel et al. 2019). Indeed, using our predicted CRMs in the mouse genome as candidates, machine-learning models trained on either mouse (LOOCV) or human cell/tissue types substantially outperformed similar models using CRM candidate defined by the earlier state-of-the-art methods for predicting active enhancers in four mouse tissues(Figure 9), although different training sets used the earlier studies might also contribute to the performance discrepancies.

Moreover, we show that functional states of all predicted CRMs in a cell/tissue type can be very accurately predicted using peaks of a minimal of three (CA+H3K4me1+H3K4me3, or CA+H3K4me1+H3K27ac) or four(CA+H3K4me1+H3K4me3+ H3K27ac) widely available epigenetic marks in the cell/tissue type(Figures 3A, S7A). Using additional histone marks data could only infinitesimally increases the accuracy duo to the redundant information in the data as indicated by the positive correlations between peak signals of the marks analyzed(Figures S6, S9). Therefore, once a map of CRMs in a

genome is more accurately and more completely predicted, our two-step strategy can be highly cost-effective for predicting functional states of all the CRMs in all cell/tissue types in the organism by generating data of few (1~4) marks, although the more the better. The rank of the weights of the marks in the LR models might suggest a prioritization of data generation(Figures S6A, S9B). Furthermore, we show that machine-learning models trained on pooled positive and negative sets from multiple cell/tissue types in human or mouse can accurately predict functional states of CRMs in other cell/tissue types in the same species as well as in various cell/tissue types in a different species. These results confirm that the epigenetic code for defining functional states of CRMs is universal in various cell/tissue types of at least different mammals. However, we found that the most critical epigenetic mark for defining functional state of CRMs is CA, rather than H3K27ac as suggested earlier(Sethi et al. 2020). On contrary, our results show that H3K27ac is among the three less important marks among the six marks analyzed(Figure S6B, S9B), consistent with a recent report that the mark is dispensable in mouse embryonic stem cells(Zhang et al. 2020). It is highly likely that the earlier conclusion(Sethi et al. 2020) was erroneously drawn since H3K27ac was used both as a feature and as the label in training datasets as we replicated in this study(Figures 1B, 1D).

The high accuracy of our predicted active CRMs positioned us to address two interesting related questions: 1) how many of the 1.2M and 0.8M CRMs predicted by dePCRM2 in the human and mouse genome regions (METHODs), respectively, are active in a cell/tissue type of the organisms; and 2) how many active CRMs are needed to define a cell/tissue type? We found that different cell/tissue types of human and mouse have widely varying numbers of active CRMs, ranging from 37,792 to 313,389 and from 37,899 to 180,827, respectively. Of these active CRMs, from only 27 to 43,333 and from only 32 to 6,914 are unique to a human and mouse cell/tissue type, respectively. We show that genes closest to the uniquely active CRMs are enriched for GO terms related to the functions of the cell/tissue. Thus, it appears that uniquely active CRMs in a cell/tissue type largely specify

the cell/tissue type. In contrast, only about a third of genes closest to uniquely active CRMS are unique to a cell type(Figures 8D, S10D), supporting that the notion that a terminal cell type is determined a unique combination of otherwise more widely expressed genes(Hobert 2011; Doitsidou et al. 2013). In this regarding, we note that the human genome encodes 3 times CRMs (n=1.47M)(Ni and Su 2021) as genes(n=44K), making it possible to use a set of uniquely active CRMs to regulate a specific combination of otherwise more widely expressed genes that determine the cell type. On the other hand, the vast majority of active CRMs in a cell/tissue type is reutilized in at least one of other cell/tissue type analyzed. However, since many complex tissues in our analysis, such as brain, testis and stomach, to name a few, might contain multiple cell types, and since the numbers of cell type/tissue types we analyzed are still small, it is likely that we might have overestimated the upper bounders of uniquely active CRMs in both the human and mouse cell types. Interestingly, cell/tissue types with related lineages form clusters based on the extent to which they share active CRMs (Jaccard index), consistent with the notion that cell types are produced in a step wise manner during cell differentiation, and thus, cell types differentiate more recently from the last common ancestral cell type share more active CRMs(Briggs et al. 2018; Farrell et al. 2018; Wagner et al. 2018).

In summary, we present a two-step strategy that substantially outperforms existing state-of-the-art methods to predict functional states of CRMs in any cell type using as few as three or four widely available epigenetic marks. The next step would be to develop a method to more accurately predict target genes of active CRMs in a cell/tissue type beyond the “closest gene principle” as used in this study and many others(Gasperini et al. 2020).

METHODS

The datasets

We downloaded from CISTROME(Zheng et al. 2019) (12/20/2020) 11,348 and 9,060 TF ChIP-seq binding peaks files for 1,360 and 701 TFs in 722 and 569 human and mouse

cell/tissue types, respectively. Using dePCRM2(Ni and Su 2021), we predicted 1,225,115 and 798,258 CRMs as well as 1,755,876 and 1,270,937 non-CRMs in 85% and 80% regions of the human and mouse genomes that were covered by the TF binding peaks, respectively. Of these human and mouse cell/tissue types, 67 and 64, respectively, have datasets available for the four epigenetic marks (CA (measured by DNase-seq or ATAC-seq), H3K4me1, H3K27ac, H3K4me3). Of these 67 and 64 human and mouse cell/tissue types, 22 and 29 also have both H3K4me2 and H3K9ac datasets available, and 19 and 14 have RNA-seq datasets available, respectively. We downloaded from CISTROME peak files of these marks in the cell/tissue types. The peaks were uniformly produced by the CISTROME team using the peak-calling tool MACS(Zhang et al. 2008). Of the 67 human cell/tissue types, six cell lines(A549, HCT116, HepG2, K562, MCH-7 and SH-HY5Y) have WHG-STARR-seq data available(Liu et al. 2017). We downloaded gene expression data and WHG-SRARR peaks from the ENCODE data portal. These human and mouse datasets except the TF ChIP-seq data are summarized in Tables S1 and S2, respectively.

Construction of the positive and the negative training datasets

dePCRM2 partitions genome regions covered by TF binding peaks in two exclusive sets, the CRM candidates(CRMCs) set and the non-CRMCs set, and predicts a CRMC to be active(TF-binding) in a cell/tissue type if at least one of constituent TFBSs in the CRMC overlap the summit of a ChIP-seq peak of a ChIP-ed TF in the cell/tissue type(Ni and Su 2021). In each cell/tissue type, we defined the predicted active CRMCs as the positive set in the cell/tissue type, denoted as a CRM+TF$^+$ set (Table 1). We also constructed the negative set in the cell/tissue type by randomly selecting from the non-CRMCs with matched number of the positive sets, denoted as a non-CRM set (Table 1). Since the number of predicted active CRMs (CRM+TF$^+$) depends on the number of available ChIP-seq datasets for different TFs in the cell/tissue type (Figures S1A and S1B), and since in most cell types only few TF ChIP-seq datasets were available(Ni and Su 2021), the CRM+TF$^+$ set is usually only a small portion of all active CRMs in a cell/tissue type (see RESULTS). For comparison, we

binned the genome into 700bp segments, and defined additional six positive and negative sets in each of the six cell lines with WHG-STARR-seq data available using six earlier methods(Firpi et al. 2010; Rajagopal et al. 2013; Lu et al. 2015; Liu et al. 2016; He et al. 2017; Sethi et al. 2020). 1) Bin+S^+/Bin+S^-, where the positive set was formed by 700bp genomic sequence bins that overlapped STARR peaks in the cell line, and the negative set was formed by 700bp genomic bins that did not overlap any STARR peaks in the cell line. 2) Bin+ac^+/Bin+ac^-, where the positive set was formed by 700bp bins that overlapped H3K27ac peaks in the cell line, and the negative set was formed by 700bp bins that did not overlap any H3K27ac peaks in the cell line. 3) Bin+S^+&ac^+/Bin+S^-&ac^-, where the positive set was formed by 700bp bins that overlapped both H3K27ac and STARR peaks in the cell line, and the negative set was formed by 700bp bins that overlapped neither H3K27ac nor STARR peaks in the cell line. 4) CRM+S^+/CRM+S^-, where the positive set was formed by our predicted CRMs that overlapped STARR peaks in the cell line, and the negative set was formed by our predicted CRMs that did not overlap any STARR peaks in the cell line. 5) CRM+TF^+/CRM+S^-, where the positive set was formed by our predicted active CRMs in the cell line, and negative set was formed by our predicted CRMs that did not overlap any STARR peaks. 6) CRM+S^+/non-CRM, where the positive set was our predicted CRMs that overlapped STARR peaks in the cell line, and the negative set was randomly selected from our predicted non-CRMs with matched number and lengths of the positive set. The number of sequences in a negative set is the same as in its cognate positive set, but the number varies greatly for different methods and cell types as summarized in Table 1. If a sequence t (a CRM or a bin of genonic sequence) overlapped at least one peak of an epigenetic mark m by at least 50% of the length of the shorter one, we computed a score S_{signal} of t for m, defined as,

$$S_{signal}\,(t,k) = \sum_{i=1}^{n_k} w_{i,k} s_{i,k} \qquad (1)$$

where $w_{i,k}$ is the ratio of the length of t over the length of the i-th peak of k, $s_{i,k}$ is the MACS signal score(Zhang et al. 2008) of the i-th peak of k. Clearly, if t did not overlap any peak of k, $S_{signal}(t,k) = 0$.

Table 1. Summary of the definitions of training datasets defined by the seven methods

Methods	Labels	Size (sequences)	Positive set	Negative set
CRM+S+/non-CRM	STARR peaks	22,610~ 71,176	CRMCs overlapping STARR peaks	Randomly selected non-CRMCs
CRM+S+/CRM+S-	STARR peaks	22,610~ 71,176	CRMCs overlapping STARR peaks	CRMCs not overlapping STARR peaks
CRM+TF+/CRM+S-	TF binding	17,558~ 272,128	CRMCs overlapping ChIP-seq peaks	CRMCs not overlapping STARR peaks
CRM+TF+/non-CRM	TF binding	17,558~272,128	CRMCs overlapping ChIP-seq peaks	Randomly selected non-CRMCs
Bin+S+/Bin+S-	STARR peaks	60,668~109,118	700bp bin overlapping STARR peaks	700bp bin not overlapping STARR peaks
Bin+S+&ac+/Bin+S-&ac-	STARR & H3K27ac peaks	7,220~49,462	700bp bin overlapping STARR&H3K27ac peaks	700bp bin not overlapping STARR&H3K27ac peaks
Bin+ac+/Bin+ac-	H3K27ac peaks	175,530~1,688,868	700bp bin overlap with H3K27ac,	700bp bin not overlapping H3K27ac peaks

Model training, testing and evaluation

We evaluated seven machine-learning classifiers, including linear regression(LR), adaboost(AB), support vector machine(SVM), neural network(NN), naïve Bayes (NB), decision tree(DT), random forest(RF). In a cell/tissue type, we trained a classifier model on the positive set and the negative set defined in the cell type using epigenetic marks as features. We conducted 10-fold cross- validation(TFCV) of the model in the cell/tissue type. In addition, in each species (human or mouse), we constructed a positive set and a negative set by pooling positive sets and negative sets in all cell/tissue types in the species, respectively. We trained a classifier model on the pooled positive and negative sets in a species using epigenetic marks as features. We conducted leave-one-out cross validation (LOOCV) of a model in a species by training the model on datasets in $n-1$ cell/tissue types, and testing the model on the left-out cell/tissue type. We assessed the performance of a model using area under the ROC (receiver operator characteristic) curve (AUCROC). We also evaluated the importance of epigenetic marks using their coefficients in LR and SVM models. All of the classifier models were implemented using sci-kit learn v.0.24.2.

Genome-wide prediction of active CRMs and non-active CRMs

We predicted whether a predicted CRMs was active or not in a cell type using a comprehensive model trained on pooled positive sets and negative sets from all cell/tissue types in the same species using epigenetic marks as features.

Comparison with the existing methods

We compared the performance of our methods to the five earlier state-of-the-art models, including Matched Filter(Sethi et al. 2020), REPTILE (He et al. 2017), FRECS(Rajagopal et al. 2013), DELTA(Lu et al. 2015), CSI-ANN(Firpi et al. 2010)) on four mouse tissues, i.e., hindbrain, midbrain, neural tube and limb, where all these methods have been applied. We extracted the AUROC values from original publications.

AVAILABILITY

Not applicable.

SUPPLEMENTARY DATA

Supplementary Data are available at Genome Research online.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

ACKNOWLEDGEMENT

We would like to thank all lab members for their discussion.

AUTHORS' CONTRIBUTIONS

ZS conceived the project. ZS and PN developed the algorithms and PN carried out all computational experiments and analysis. PN and ZS wrote the manuscripts. All authors read and approved the final manuscript.

FUNDING

The work was supported by US National Science Foundation (DBI- 1661332). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

REFERENCES

Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. 2010. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**: 559-571.

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet* **31**: 426-433.

Arbel H, Basu S, Fisher WW, Hammonds AS, Wan KH, Park S, Weiszmann R, Booth BW, Keranen SV, Henriquez C et al. 2019. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proc Natl Acad Sci U S A* **116**: 900-908.

Armstrong JA, Emerson BM. 1996. NF-E2 disrupts chromatin structure at human beta-globin locus control region hypersensitive site 2 in vitro. *Mol Cell Biol* **16**: 5634-5644.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.

Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**: 276-288.e278.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311-322.

Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM. 2018. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**: 1213-1218.

Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32**: 202-223.

Consortium GO. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**: D325-D334.

Dao LTM, Spicuglia S. 2018. Transcriptional regulation by promoters with enhancer function. *Transcription* **9**: 307-314.

Davidson EH. 2006. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press.

Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, Keller CA, Cheng Y, Jain D, Visel A et al. 2015. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**: 16.

Doitsidou M, Flames N, Topalidou I, Abe N, Felton T, Remesal L, Popovitchenko T, Mann R, Chalfie M, Hobert O. 2013. A combinatorial regulatory signature controls terminal differentiation of the dopaminergic nervous system in C. elegans. *Genes & development* **27**: 1391-1405.

Dorighi KM, Swigut T, Henriques T, Bhanu NV, Scruggs BS, Nady N, Still CD, Garcia BA, Adelman K, Wysocka J. 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**: 568-576.e564.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43-49.

F Z, RM S. 2005. - RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Developmental biology* **283**: 40-57.

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**.

Fernandez M, Miranda-Saavedra D. 2012. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res* **40**: e77.

Firpi HA, Ucar D, Tan K. 2010. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* **26**: 1579-1586.

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**.

Gao T, He B, Liu S, Zhu H, Tan K, Qian J. 2016. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**: 3543-3551.

Gao T, Qian J. 2020. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* **48**: D58-d64.

García-Ortega LF, Martínez O. 2015. How Many Genes Are Expressed in a Transcriptome? Estimation and Results for RNA-Seq. *PloS one* **10**: e0130262.

Gasperini M, Tome JM, Shendure J. 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**: 292-310.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**: e1003711.

Gu Z, Eils R, Schlesner M, Ishaque N. 2018. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC genomics* **19**: 1-7.

Hardison RC. 2012. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J Biol Chem* **287**: 30932-30940.

He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M et al. 2010. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42**: 343-347.

He Y, Gorkin DU, Dickel DE, Nery JR, Castanon RG, Lee AY, Shen Y, Visel A, Pennacchio LA, Ren B et al. 2017. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc Natl Acad Sci U S A* **114**: E1633-e1640.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311-318.

Hobert O. 2011. Regulation of terminal differentiation programs in the nervous system. *Annual review of cell and developmental biology* **27**: 681-696.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473-476.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827-841.

Howe FS, Fischl H, Murray SC, Mellor J. 2017. Is H3K4me3 instructive for transcription activation? *BioEssays : news and reviews in molecular, cellular and developmental biology* **39**: 1-12.

Huang J, Marco E, Pinello L, Yuan GC. 2015. Predicting chromatin organization using histone marks. *Genome biology* **16**: 162.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38-52.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497-1502.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**: 800-811.

Kleftogiannis D, Kalnis P, Bajic VB. 2015. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res* **43**: e6.

Kleftogiannis D, Kalnis P, Bajic VB. 2016. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* **17**: 967-979.

Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**: 1083-1091.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**: 1595-1602.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lim LWK, Chung HH, Chong YL, Lee NK. 2018. A survey of recently emerged genome-wide computational enhancer predictor tools. *Comput Biol Chem* **74**: 132-141.

Liu F, Li H, Ren C, Bo X, Shu W. 2016. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci Rep* **6**: 28517.

Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. 2017. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol* **18**: 219.

Local A, Huang H, Albuquerque CP, Singh N, Lee AY, Wang W, Wang C, Hsia JE, Shiau AK, Ge K et al. 2018. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* **50**: 73-82.

Lu Y, Qu W, Shan G, Zhang C. 2015. DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications. *PloS one* **10**: e0130622.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29-59.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**: D419-D426.

Morse RH. 2010. Epigenetic marks identify functional elements. *Nat Genet* **42**: 282-284.

Muerdter F, Boryn LM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**: 141-149.

Narlikar L, Ovcharenko I. 2009. Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic* **8**: 215-230.

Ni P, Su Z. 2021. Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genom Bioinform* **3**: lqab052.
Niu M, Tabari E, Ni P, Su Z. 2018. Towards a map of cis-regulatory sequences in the human genome. *Nucleic Acids Res* **46**: 5395-5409.
Peng T, Zhai Y, Atlasi Y, Ter Huurne M, Marks H, Stunnenberg HG, Megchelenbrink W. 2020. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol* **21**: 243.
Podsiadlo A, Wrzesien M, Paja W, Rudnicki W, Wilczynski B. 2013. Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data. *BMC Syst Biol* **7 Suppl 6**: S16.
Rada-Iglesias A. 2018. Is H3K4me1 at enhancers correlative or causative? *Nat Genet* **50**: 4-5.
Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**: 279-283.
Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. 2013. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* **9**: e1002968.
Ramskold D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5**: e1000598.
Rickels R, Herz HM, Sze CC, Cao K, Morgan MA, Collings CK, Gause M, Takahashi YH, Wang L, Rendleman EJ et al. 2017. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* **49**: 1647-1653.
Salzberg SL. 2018. Open questions: How many genes do we have? *BMC biology* **16**: 94.
Santiago-Algarra D, Dao LTM, Pradel L, Espana A, Spicuglia S. 2017. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res* **6**: 939.
Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**: 407-411.
Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, Barozzi I, Afzal V, Akiyama JA, Plajzer-Frick I et al. 2020. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nature methods* **17**: 807-814.
Spicuglia S, Vanhille L. 2012. Chromatin signatures of active enhancers. *Nucleus* **3**: 126-131.
Temple G Gerhard DS Rasooly R Feingold EA Good PJ Robinson C Mandich A Derge JG Lewis J Shoaf D et al. 2009. The completion of the Mammalian Gene Collection (MGC). *Genome Res* **19**: 2324-2333.
Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Res* **18**: 1084-1091.
Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *NatGenet* **45**: 124-130. doi: 110.1038/ng.2504. Epub 2012 Dec 1023.
Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**: 1304-1351.
Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**: 981-987.
Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, Kellis M. 2018. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**: 5380.
Young RS, Kumar Y, Bickmore WA, Taylor MS. 2017. Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol* **18**: 242.
Zentner GE, Scacheri PC. 2012. The chromatin fingerprint of gene enhancer elements. *J Biol Chem* **287**: 30888-30896.

Zhang T, Zhang Z, Dong Q, Xiong J, Zhu B. 2020. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol* **21**: 45.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**: R137.
Zhao Y, Garcia BA. 2015. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol* **7**: a025064.
Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen CH, Brown M, Zhang X, Meyer CA et al. 2019. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* **47**: D729-d735.