Construction and analysis of active enhancer-target networks in 935 samples of human primary cells, tissues and cell lines

Qin Cao[1], Xihao Hu[1], Xiaodan Fan[2] and Kevin Y. Yip[1,3,4,5,*]

[1]Department of Computer Science and Engineering,
[2]Department of Statistics,
[3]Hong Kong Bioinformatics Centre,
[4]CUHK-BGI Innovation Institute of Trans-omics,
[5]Hong Kong Institute of Diabetes and Obesity,
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Abstract

Large-scale genomic studies have identified a large number of non-coding genetic variations associated with various diseases, many of which overlap transcriptional enhancers. Accurate evaluation of the functional impacts of these genetic variations and their links to the disease phenotypes require the identification of active enhancers, their target genes, and the quantitative relationships between enhancer features and gene expression levels in the relevant cell and tissue types. Here we construct and analyze active enhancer-target networks in 935 human samples that cover a large variety of primary cells, tissues and cell lines. Based on our networks, chromatin and expression features of multiple enhancers jointly infer the expression levels of their target genes with high accuracy. The enhancer-target connections are consistent with chromatin conformation contacts and topological domains. Similarity of networks from different samples follows closely their biological origins, and groups related samples into distinct clusters. Enhancers specifically active in a group of related samples tend to regulate genes specifically expressed in these samples. Enhancers commonly targeting a gene display simultaneous, independent or mutually exclusive mode of action across different samples, revealing different ways of co-regulation relevant to the functions of the target genes. In particular, defense-related genes are commonly regulated by multiple enhancers simultaneously in immune cells. Some enhancers also work together non-linearly in regulating their common targets. Overall, our enhancer-target networks enable the study of enhancer functions in normal and disease states in a context-specific manner.

Introduction

Enhancers are important DNA functional elements that participate in transcriptional regulation. They are bound by transcription factors and interact with the promoters of their target genes through DNA looping (Shlyueva et al., 2014; Visel et al., 2009). The target genes can be far away from the regulating enhancers, and can be either upstream or downstream of them (Shlyueva et al., 2014), making it non-trivial to determine enhancer targets. Enhancers are involved in the control of precise gene expression in development, cell differentiation, homeostasis, and response to external stimuli (Bulger and Groudine, 2010).

Enhancer activities depend highly on the particular context, such as cell and tissue types (Heintzman et al., 2009). Abnormal gain or loss of enhancer activities is associated with various diseases (Shlyueva et al., 2014; Williamson et al., 2011). Some of these aberrations are associated with epigenetic changes at the enhancers, such as altered levels of DNA methylation and histone modifications (Aran et al., 2013; Sakabe et al., 2012).

High-throughput experimental methods have made it possible to probe various types of molecular activities genome-wide across many human cell and tissue types (Andersson et al., 2014; The ENCODE Project Consortium, 2012; The GTEx Consortium, 2015; Roadmap Epigenomics Consortium et al., 2015). Using these and other data sets, context-specific human enhancers have been predicted based on transcription factor binding, histone modifications, chromatin accessibility, bi-directional transcripts and

*To whom correspondence should be addressed. Email: kevinyip@cse.cuhk.edu.hk

DNA methylation (Andersson et al., 2014; Ernst et al., 2011; Heintzman et al., 2007; Hoffman et al., 2013; Rajagopal et al., 2013; Thurman et al., 2012; Yip et al., 2012). Some of these predictions were further experimentally tested using either low- or high-throughput reporter assays (The ENCODE Project Consortium, 2012; Kwasnieski et al., 2014).

To identify the target genes of these enhancers, previous methods have used genomic distance, activity correlations and sequence co-conservation of enhancers and genes across multiple contexts (Andersson et al., 2014; Ernst et al., 2011; He et al., 2014; Thurman et al., 2012; Yip et al., 2012). The completeness and reliability of these enhancer networks were limited by the number of contexts with the required data. Information about DNA long-range interactions based on high-throughput extensions of Chromosome Conformation Capture (3C) (Dekker et al., 2002), such as Hi-C and ChIA-PET, can be used to partially evaluate the accuracy of the enhancer-target associations identified. Such data have been produced for only a small number of human cell types thus far (Dixon et al., 2012; Heidari et al., 2014; Kalhor et al., 2012; Li et al., 2012; Lieberman-Aiden et al., 2009; Rao et al., 2014; Tang et al., 2015).

To study the general properties of enhancer-based gene regulation, here we construct context-specific enhancer-target networks of 935 samples of human primary cells, tissues and cell lines using a large amount of high-throughput sequencing data collected from multiple sources. Our main methodology for identifying target genes of enhancers is constructing statistical models that can accurately infer expression levels of the genes based on activity features of the enhancers across many contexts. Our work differs from previous studies by considering the combined effect of multiple enhancers on the same target genes, where different enhancers could regulate the same gene in same or different contexts, and by using information from each sample to identify the active enhancer-target network specific to the sample.

We use the enhancer-target networks to study several questions about enhancers, namely 1) to what extent expression variability can be explained by enhancer features, either alone or together with promoter and gene body features, 2) the necessity of considering multiple enhancers jointly rather than individually, 3) the roles of context-specific enhancer activities in gene regulation, and 4) the relationships between different enhancers that share common targets.

Results

Enhancer features can quantitatively infer target gene expression

We first studied each enhancer-target pair independently to see to what extent target gene expression can be explained by enhancer features in this simple setting. We collected multiple sets of predicted active enhancers for two human cell lines with RNA polymerase II ChIA-PET data available (Table S1). For the same cell lines, we also collected gene expression levels and various features previously proposed to be indicative of active or repressed enhancers. Aggregation plots around the active enhancers display clear patterns of these features, including a sharp peak of DNase I hypersensitivity in the center, double peaks of H3K4me1, H3K4me2 and H3K27ac in the flanking regions, divergent transcription on the two strands, and a relatively low level of DNA methylation (Figure 1a). In contrast, the inactive enhancers exhibit much lower levels of chromatin accessibility, active histone marks and divergent transcription, and a relatively high level of DNA methylation (Figure 1b). The clarity of these signature patterns suggests the reliability of the predicted enhancers.

We used chromosome conformation data from ChIA-PET experiments to identify target promoters of these enhancers (Table S1), and focused on the enhancers with a single target gene. Hierarchical clustering of the enhancer-target pairs based on features at the enhancers, target gene promoters and gene bodies in the K562 cell line shows a clear separation of the active pairs in K562 (Figure 1c, upper half) from the pairs not active in K562 but active in some other cell lines (Figure 1c, lower half). H3K4me1, H3K4me2, H3K27ac, enhancer RNA (eRNA) and DNase I hypersensitivity levels at the enhancers all correlate positively with target gene expression, while the repressive mark H3K27me3, and to some extent DNA methylation, correlate negatively. These visually apparent correlations are noteworthy because the enhancer-target pairs were defined based on ChIA-PET data alone without referring to any of these features or target gene expression levels. Weak spatial patterns are observed for some of these

features, including the particularly low signals of eRNA and DNase I hypersensitivity at the center of inactive enhancers (bin 3) as compared to the flanking regions. The promoter and gene body features are also correlated with gene expression in ways largely consistent with current knowledge about these features (Zhou et al., 2011).

Given the strong correlations between enhancer features and target gene expression, we hypothesized that these features would be able to quantitatively infer expression levels of the target genes. To confirm this, we constructed statistical models of gene expression based on features of their targeting enhancers, as was previously performed using promoter and gene body features (Cheng et al., 2012; Dong et al., 2012; Lou et al., 2014). By using a cross-validation procedure that learns model parameters based on training subsets of the enhancer-target pairs and tests the resulting models on left-out testing subsets, high inference accuracy would indicate that the models capture general relationships between enhancer features and gene expression.

In K562 cells, when all enhancer features were considered together, the models were able to distinguish genes of different expression classes with an area under the receiver operator characteristic (AUC) of 0.74, which is much higher than the expected value of 0.5 for random predictions ($p<1e-6$, Figure S1a). We then constructed models using one type of enhancer features at a time (Figures S1a), and found that the features most successful in inferring the expression class of the target genes were H3K4me1, H3K4me2, H3K27ac and DNase I hypersensitivity, followed by eRNA with a small drop of accuracy. The repressive mark H3K27me3 and DNA methylation were not as successful, but they still performed much better than random predictions ($p=0.007$ for K3K27me3, $p=0.06$ for DNA methylation). None of these models involving single feature types was as successful as the one containing all enhancer features.

We also compared the relative accuracy of models involving promoter or gene body features, or combinations of them (Figures S1b). Models based on promoter and gene body features were able to infer the expression class with higher accuracies. Incorporating enhancer features into these models (e.g., comparing the models with both enhancer and promoter features (E+P) with the models with only promoter features (P)) did not improve the accuracy, probably due to the already very high predictive power of the promoter and gene body features. In fact, when we used one replicate of the RNA-seq data as the predictor to infer the expression classes of the transcripts in another replicate, the average AUC of the four classes was 0.87, showing that some of the promoter and gene body models already reached this "best-case" accuracy.

To check whether the above results depend on the choice of particular data sets or the way expression classes are defined, we repeated the whole process for the MCF-7 cell line and other enhancer sets for the two cell lines, as well as performing regression of the log expression levels instead of predicting expression classes. All these results (Figures S2-S5) reconfirm that enhancer features are able to quantitatively indicate target gene expression level.

Gene expression is better indicated by the joint effect of multiple targeting enhancers

We next investigated whether gene expression levels can be better indicated by considering the joint effects of multiple enhancers. For this part of study, we could not use high-resolution chromatin conformation data to define enhancer targets since such data were only available for a small number of cell lines. Instead, we inferred both enhancer targets and the quantitative relationships between enhancer features and target gene expression levels simultaneously using statistical models. We used transcription start site (TSS) as the basic expression unit, such that expression levels of different transcripts of a gene with different TSSs were modeled separately. To ensure the generality of our findings, we collected enhancer features and expression levels for a large number of human primary cell and tissue types and cell lines (which we call "samples" in general) from two sources. The first source was ENCODE and Roadmap Epigenomics (ENCODE+Roadmap), which provided ChIP-seq data of various histone modifications as enhancer features and RNA-seq data as TSS expression levels for 127 samples (Roadmap Epigenomics Consortium et al., 2015). We collected active enhancers in these 127 samples predicted by ChromHMM (Ernst et al., 2011; Roadmap Epigenomics Consortium et al., 2015). The second source was FANTOM5, which provided CAGE (Cap Analysis of Gene Expression) data as both enhancer features

(eRNAs) and TSS expression levels for 808 samples (Andersson et al., 2014). This data set came with potential active enhancers for each of the 808 samples. We used two different methods for constructing the models, both of which looked for a minimal set of enhancers that could accurately infer the expression level of each TSS across all the samples from each source.

For each TSS, we learned the model parameters using both enhancer features and expression data of some training samples, and then applied the models to the remaining samples without disclosing the expression levels of the TSSs in these testing samples. Modeling accuracy was then computed based on the difference between the predicted and actual expression levels of the TSS in the testing samples. To make sure that the models do not simply memorize the expression levels of the TSS in a training sample that is highly similar to a testing sample, we grouped biologically similar samples and assigned all samples in the same group to the training or testing set together. For the ENCODE+Roadmap samples, we defined four separate data sets based on the inclusion or exclusion of imputed data (Ernst and Kellis, 2015).

For all five data sets, considering different enhancers that could regulate a TSS jointly ("LASSO" and "Elastic Net") resulted in significantly better ($p<2.2e-16$ in all cases, Wilcoxon signed rank test) modeling accuracy than when each enhancer was considered separately ("independent") (Figure S6). Among the ENCODE+Roadmap data sets, the modeling accuracy was higher when imputed data were involved, especially when both enhancer features and expression levels involved imputed data. Although the AUC values based on the ENCODE+Roadmap data sets cannot be directly compared with those from the FANTOM5 data set due to the different enhancer features involved and ways to measure expression levels (RNA-seq vs. CAGE), the consistent relative performance of the three types of model in the five data sets suggests that TSS expression levels are better indicated by considering the joint effect of multiple enhancers in general.

Accurate construction of context-specific enhancer-target networks

The above statistical models identified for each TSS the set of enhancers that infer the expression level of the TSS most accurately across many samples. To further identify the enhancers that regulate a TSS in each specific sample, we developed a method that combines the global models with sample-specific TSS and enhancer activities and their genomic distance. Using this method, we inferred an active enhancer-target network for each of the 935 samples (the network files can be downloaded at `http://yiplab.cse.cuhk.edu.hk/enhancernetworks`).

To evaluate the accuracy of these networks, we first checked the overlaps between the inferred enhancer-target pairs and published ChIA-PET, Hi-C and expression quantitative trait loci (eQTL) data (collectively referred to as the "validation data") available for some of the samples (Table S2). While the DNA interactions in these data form neither a precise nor complete list of enhancer-target pairs in the respective samples, they provide independent structural and statistical information for evaluating whether our inferred enhancer-target pairs are likely correct. As negative controls, we considered four types of random enhancer-TSS pairs, taking into account the distance between enhancers and TSSs, background distribution of contact distances in chromatin conformation data, and the ratio between positive and negative pairs (Figure 2a). In all cases (Figures 2b-e, S7-S9), the enhancer-target pairs inferred by our models overlapped the enhancers and promoters connected by the validation data significantly more often than the negative controls, regardless of the sample involved, the way these negatives were drawn and the measure used to quantify the enrichment of overlap over the negative controls. The enrichment of inferred enhancer-target pairs overlapping ChIA-PET data was strongest among the three types of validation data, consistent with the notion that the ChIA-PET data we used describe direct contacts between DNA regions mediated by a transcription-related factor (POLR2A or H3K4me2), while eQTL and Hi-C data include also indirect interactions or DNA contacts unrelated to transcriptional regulation.

Among the different statistical models, the ones that considered the joint effect of multiple enhancers produced enhancer-target pairs that overlapped the validation data consistently more than the model that considered each enhancer individually. Consistent with previous studies, using genomic distances alone could also identify enhancer-target pairs with some overlaps with the validation data (He et al., 2014), but not as strongly as either the joint or independent models involving enhancer features.

We also compared our results with a recently published method for calling enhancer-target pairs, IM-PET (He et al., 2014), which combines multiple types of features in calling enhancer targets. Applying our modeling method on the same data set in He et al. (2014), our models obtained better overlap with the validation data (AUROC of 0.874 for CD4+ T cells and 0.862 for K562) than IM-PET (AUROC of 0.796 and 0.743, respectively), confirming the validity of our modeling approach.

Studies have shown that enhancer-target connections usually occur within topological domains in three-dimensional genome structures (Dixon et al., 2012; Heidari et al., 2014; Jin et al., 2013; Rao et al., 2014; Tang et al., 2015). In the second set of analysis, we therefore checked whether our inferred enhancer-target pairs are within topologically associating domains (TADs) defined by Hi-C data in IMR90 cells (Dixon et al., 2012) and within chromatin contact domains (CCDs) defined by CTCF ChIA-PET data in GM12878 cells (Tang et al., 2015). In both cases, the inferred enhancer-target pairs were significantly more within these domains than randomly shuffled domains of the same sizes for most cutoff thresholds of enhancer-target calls (Figure 2f). The statistical significance of the results was stronger for CCDs, which is in line with the recent report that most transcription activities occur within CTCF looped chromatin structures (Tang et al., 2015).

Given the importance of CTCF in binding the boundaries of CCDs and insulators, we expected a depletion of intervening CTCF binding between enhancers and their inferred target TSSs (Figure 2g). We found that this was the case for all 14 samples with CTCF ChIP-seq data available, with p-values $<$ 2.2e-16 for each of the 14 cases when compared to randomly drawn enhancer-TSS pairs with the same distance distribution. Among the intervening CTCFs in GM12878 from which the CCDs were defined, we also found that they were significantly less involved in defining CCD boundaries than would be expected by chance (Fisher exact test p$<$8.08e-9; Figure 2h).

In the third set of analysis, we investigated the well-known enhancer-target connections between the human beta-globin genes and its locus control region (LCR). In K562 cells, the interactions between the LCR and the beta-globin genes as well as flanking olfactory genes have been well-quantified by 3C (Dostie et al., 2006). We compared the 3C interaction frequencies with the predicted enhancer-TSS connection scores by our LASSO model, and found them highly consistent (Figure 2i), with a Pearson correlation of 0.80 for the 8 genes involved.

Collectively, these results show that the networks inferred by our methods reliably describe the active enhancer-target pairs in the samples.

Basic properties and context specificity of the enhancer networks

Since the FANTOM5 data set contained several times more samples than the ENCODE+Roadmap one, we focused on the FANTOM5 results for the remaining analyses. Overall, 20,693 TSSs were inferred to have a regulating enhancer and 29,359 enhancers were inferred to have a target TSS in at least one of the FANTOM5 samples based on the LASSO model, forming 128,725 enhancer-target pairs in total. On average, each TSS is regulated by 6.2 enhancers and each enhancer regulates 4.4 TSSs. In each specific sample, the average numbers of regulated TSSs, regulating enhancers and active enhancer-target pairs are 923, 523 and 1,136, respectively (Figure 3a). The average number of regulating enhancers per TSS in each sample ranges from 1 to 1.7 (Figure 3b), and the average number of target TSSs per enhancer in each sample ranges from 1.7 to 3.3 (Figure 3c). As an example, in K562 and CD4+ T cells, most TSSs in the networks have 1 regulating enhancer (Figure 3d), while the average numbers in these cell types are 1.1 and 1.3, respectively. The number of target TSSs per enhancer in these two cell types ranges from 1 to 13 and 1 to 12, respectively (Figure 3e), with averages of 2.4 and 2.0.

Previous studies have suggested that the number of regulated TSSs per enhancer ranges from 1.0 to 3.0 (He et al., 2014; Jin et al., 2013), which completely overlaps our range, and our larger range of values is likely due to the much larger number of samples considered. On the other hand, these studies suggested 2.0-2.6 regulating enhancers per TSS, which is larger than our values. Considering the number of enhancers predicted in FANTOM5 is much less than that predicted by the methods based on histone modifications (H3K4me1 and H3K27ac in Jin et al. (2013) and CSI-ANN in He et al. (2014)), we do expect that our inferred number of regulating enhancers per TSS would be smaller in FANTOM5. To verify this explanation, we also checked the ENCODE+Roadmap data in which enhancers were predicted by

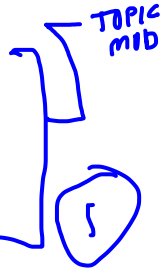
histone marks and the number of enhancers is similar to that of Jin et al. (2013) and He et al. (2014). We found that the mean number of regulating enhancers per TSS ranges from 1.4 to 3.3, which is consistent with the numbers suggested by the two previous studies.

Most enhancers regulate in no more than 50 samples, with a median of 5 samples (Figure 3g). However, there is a very long tail, with an enhancer regulating in 765 samples in the most extreme case.

The median distance between an enhancer and a target TSS it regulates in a sample ranges from 49kbp to 83kbp (Figure 3f), which is similar to the 58kbp-123kbp predicted in He et al. (2014) (where for 10 of the 12 samples, the range was 58kbp to 89kbp) and smaller than a previously reported value of 124kbp based on Hi-C data (Jin et al., 2013), possibly due to some inactive enhancer-TSS pairs in the Hi-C data. In K562 and CD4+ T cells, the enhancer-target distance peaks at 20-40kbp (Figure 3g), with median distances of 64kbp and 55kbp, respectively. In each sample, on average 43% of the active enhancers regulate the closest TSS in addition to any other TSSs (Figure 3h), which is close to the value of 40% reported by FANTOM5 (Andersson et al., 2014). Interestingly, among all the enhancer-target pairs in each sample, only 12%-30% of them involve the closest TSS of an enhancer (Figure 3h), indicating that many enhancers also regulate TSSs farther away.

Since a separate network of active enhancer-target pairs was inferred for each sample, we were able to study the context specificity of enhancer targeting. First, we used the network connections as the signature of each sample to perform a hierarchical clustering of the samples, and found that biologically related samples formed clear clusters (Figures 4a, S10-S11). For example, CD4+ T cells, CD14+ monocytes, endothelial cells, respiratory epithelia cells and renal epithelia cells all formed their own clusters. These results suggest that the inferred enhancer networks serve as distinctive signatures of the samples, with related samples having more similar networks.

Next, we identified subnetworks that were active in a single group of related samples. Visualizing the networks using a force directed layout and coloring nodes and edges active in only a single sample group (Figures 4b, S12), the enhancers, TSSs and enhancer-target interactions specifically active in the same groups were clustered together. This indicates that group-specific enhancers usually regulate TSSs that are only expressed in this group of samples, and conversely TSSs expressed only in a group of samples are usually regulated by enhancers that are active only in this group of samples. Indeed, among genes specifically expressed in a sample group, between 41.0% (muscle and bone) and 75.3% (cardiovascular) of them are regulated by an enhancer that regulates only in this group of samples. Likewise, among enhancers specifically regulate in a sample group, between 25.3% (immune) and 54.2% (neurosystem) of them regulate a TSS expressed only in this group of samples.

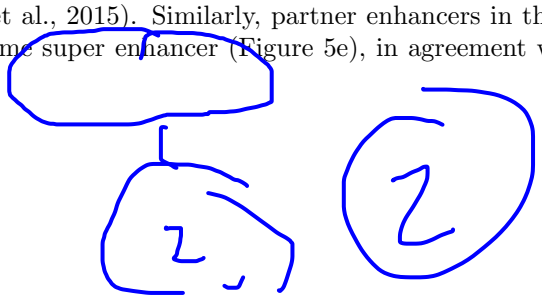


Finally, we compared the expressed TSSs regulated or not regulated by an enhancer active in the same sample, and found that the former TSSs had higher expression levels (Figure 4c) and were more specifically expressed (Figure 4d) in these samples. We also investigated the effect of having multiple enhancers regulating the same TSS in a sample, and found that TSSs with more regulating enhancers in a sample had higher expression level (Figure 4e) and strong expression specificity (Figure 4f).

Enhancers regulate common targets in several major modes

In the inferred enhancer-target networks, two enhancers can regulate a common TSS in same or different samples, forming a small regulatory module involving the two partner enhancers and their target. We defined three coregulation modes for these modules (Figure 5a), namely the Simultaneous, Independent and Mutually Exclusive modes, in which the two partner enhancers tend to regulate the target in the same samples, independent sets of samples, and different samples, respectively. A quick check of the modules revealed that the activity correlation of the partner enhancers across all samples were highest in the Simultaneous mode, followed by the Independent mode, and the lowest in the Mutually Exclusive mode as expected (Figure 5b).

In the Simultaneous mode, the partner enhancers were more frequently located within the same topological domain in the 3D structure of the genome (Figure 5c,d), which is consistent with the role of topological domains in facilitating enhancer-target connections (Dixon et al., 2012; Heidari et al., 2014; Rao et al., 2014; Tang et al., 2015). Similarly, partner enhancers in the Simultaneous mode were more likely to reside in the same super enhancer (Figure 5e), in agreement with the definition of super



enhancers as large regulatory units working in concert (Hnisz et al., 2013; Whyte et al., 2013). These two observations are likely partially due to a shorter genomic distance between the two enhancers in the Simultaneous modules (Figure 5f), which appears to be an intrinsic property of this mode.

Interestingly, we found that the partner enhancers in the Simultaneous mode shared fewer transcription factor (TF) binding motifs than the Independent and Mutually Exclusive modes (Figure 5g). We hypothesized that two enhancers could complement each other more effectively if they differ from each other by at least one attribute. In the Simultaneous mode in which the partners tend to regulate the common target in the same contexts, having different motifs increases the chance that at least one of them can function when only some of these TFs are expressed. On the other hand, if the partner enhancers contain the same TF binding motifs, they could complement each other by being active in different contexts as controlled by other regulatory mechanisms. Indeed, the accessibility correlation of the partner enhancers based on DNase I hypersensitivity was significantly higher for the Simultaneous mode than the Mutually Exclusive mode (Figure 5h).

We investigated whether there are any differences among the targets regulated by partner enhancers in the three modes. We found that the targets in the Simultaneous mode were less conserved (Figure 5i), suggesting that this mode of regulation is used by genes with more adaptive functions. Supporting this, we found that target genes regulated by the Simultaneous mode, but not those regulated the other two modes, were enriched in functional terms related to defensive mechanisms, including "defense response" (BH-corrected p=0.002), "inflammatory response" (BH-corrected p=0.01) and "immune response" (BH-corrected p=0.04). Correspondingly, we found that the enhancers in the Simultaneous mode regulate their targets significantly more frequently in immune cells than the other two modes (Figure 5j). We also observed that the targets regulated by the Simultaneous mode were slightly depleted of TF genes than those regulated by the Mutually Exclusive mode (Figure 5k).

For the targets regulated by the Simultaneous mode, we further identified a subset of which the partner enhancers had non-linear interaction effects. We found that these targets had higher mean expression levels (Figure 5l) and were slightly more specifically expressed (Figure 5m).

Discussion

In this manuscript, we have investigated the relationships between enhancers and their targets in a context-specific manner. We have shown that enhancer features can quantitatively infer the expression levels of their targets with high precision. Features commonly used as indicators of enhancers and enhancer activities, including H3K4me1, H3K4me2 (Rajagopal et al., 2013), H3K27ac (Rada-Iglesias et al., 2011) and DNase I hypersensitivity (Thurman et al., 2012), were among the most informative features. Enhancer RNA, a recently proposed feature of enhancer activities (Andersson et al., 2014), was also fairly informative in inferring target expression levels. The slightly lower accuracy of eRNA models could be due to the importance of its bidirectional signal shape (Kim et al., 2010), which is more difficult to use for quantifying the enhancer activity level than simple signal levels. Repressive marks such as H3K27me3 and DNA methylation were less informative, which is reasonable since lack of a repressive mark does not guarantee an enhancer to be active, leading to weaker (anti-)correlations with target expression levels. In general, our results have clearly shown that enhancer features not only can tell whether a target is expressed or not, but also some details of their quantitative expression levels. Such details would be critical in evaluating the consequence of aberrant enhancer activities on their targets.

Our results indicated that combining multiple enhancer features resulted in models more accurate than those based on individual features. This could be due to either complementary information contained in the different types of features, imperfect quantification of the feature values, or a combination of them. In many studies, due to resource and sample constraints it is impossible to obtain all these different types of features experimentally. The relative accuracy of the different types of features in our models could serve as a practical guide as to which features one may want to obtain first in such situations.

Using the large amount of sequencing data from ENCODE, Roadmap Epigenomics and FANTOM5, we were able to both infer enhancer targets in each sample and model the relationships between their activities across all samples. Our models were able to consider the joint effect of multiple enhancers on

targets that they commonly regulate in same or different samples. We have shown that modeling the effects of multiple enhancers jointly led to more accurate models in terms of both the inferred target expression levels in left-out test sets and enhancer-target pairs based on external validation data sets. Intuitively, if one gene can be regulated by multiple enhancers in different samples, the correlation of its expression level with the activity level of any of these enhancers across all samples may not be strong. If these enhancer-target pairs were to be identified by considering each enhancer separately, loose correlation thresholds would need to be used, which is problematic since many false positive enhancer-target pairs could be called. Modeling the joint effect of multiple enhancers also has implications to disease studies, since each enhancer that regulates a driver gene may only be disrupted in a subset of the samples, and analyzing differential features at all these enhancers together could give much stronger disease associations.

The constructed context-specific enhancer-target networks have allowed us to study their sample-specificity. We have shown that these networks can serve as signatures of the corresponding cell/tissue types for clustering related samples together. These results clearly demonstrate the dynamics of these networks and the importance of studying enhancer networks in a context-specific manner.

The networks have also enabled us to define three co-regulation modes of two enhancers regulating the same target. The Simultaneous mode, in which the enhancers tend to regulate their common target in the same samples, have less-conserved enhancers, more defense-related genes as the targets, and are more often active in immune cells. Interestingly, these enhancers share fewer TF binding motifs than the enhancer pairs in the two modes. The evolutionary driving force behind these observations is yet to be investigated, but we hypothesize that such settings can allow multiple transcription factors to regulate the expression of the target genes in the specific cell types, thereby increasing the flexibility of responding to multiple signaling pathways.

It should also be noted that since the sequencing data we used in our models were obtained from cell populations, the Simultaneous mode does not necessarily mean multiple enhancers are interacting with the target promoter in the same cell at the same time. While previous studies on chromosome long-range interactions have suggested the presence of anchor points at which multiple genes and enhancers interact with each other (Fullwood et al., 2009; Tang et al., 2015), such data were also obtained from cell populations and may reflect the average of many structures rather than a single structure. Whether some enhancers tend to regulate the same target in the same cell in order to drive a higher expression level of the target is still uncertain, although we did observe that a target having more enhancers regulating it in a sample tends to have a higher expression level.

In contrast, in the Mutually Exclusive mode, the partner enhancers tend to share common TF binding motifs but regulate their common target in different samples. We have hypothesized that in this setting the activity of the enhancers is more controlled by some other factors such as DNA accessibility. A possible alternative hypothesis is a competition of the enhancers in interacting with a common set of TFs. This hypothesis seems less likely, since a low abundance of the TFs is required for the competition to happen, and even if it does happen, in a population of cells in which each enhancer is active in a portion of the cells, both would appear to regulate the target in this sample and thus the enhancers would not appear to regulate the target in a mutually exclusive manner.

Materials and Methods

Collection and processing of data sets for modeling gene expression with one-to-one enhancer-target associations

We collected computationally predicted active enhancers in human cell lines from multiple sources (Table S1). To facilitate model construction, we resized each enhancer to 2,500bp by trimming or extending it while maintaining the original center of the enhancer. The purpose of using this relatively long span was to capture spatial signal patterns of features in the flanking regions of each enhancer, which have been shown to be important for indicating enhancer activities (Andersson et al., 2014; Core et al., 2014). We have also tried aligning enhancers based on their peak locations of DNase I hypersensitivity, and got

similar results.

For a given data source, we defined active and inactive enhancers of a cell line as follows. The active enhancers were taken directly from the ones predicted active in this cell line. The inactive ones were the predicted active enhancers in other cell lines but which did not overlap any active enhancers in this cell line. Including both active and inactive enhancers allowed the modeling procedure to capture the differences in enhancer features and target gene expression levels between the active and inactive cases.

The enhancers were paired with potential target genes using ChIA-PET chromosome conformation data. If an active enhancer was connected to the transcription start site (TSS) of a RefSeq (GRCh37) protein-coding gene based on the ChIA-PET data of the target cell line, the enhancer-TSS pair was defined as an active pair in this cell line. If 1) an inactive enhancer was connected to the TSS of a RefSeq protein-coding gene based on the ChIA-PET data of another cell line, 2) it was not involved in any connections based on the chromosome conformation data in the current cell line, and 3) the TSS was not involved in any active pairs, the enhancer-TSS pair was defined as an inactive pair in this cell line. Therefore in this procedure, no expression or activity features of the TSSs and their gene bodies were involved in defining the active and inactive pairs.

To study the simple setting of considering each enhancer-target associations independently, if a gene was involved in multiple enhancer-target pairs, we randomly retained only one of them and randomly selected an annotated transcript for the TSS. Finally, we randomly removed pairs from the active or inactive pairs until the two sets had the same size.

Each (active or inactive) enhancer was then sub-divided into five 500bp bins, and the average signals of H3K4me1, H3K4me2, H3K27ac, H3K27me3, DNase I hypersensitivity, bi-directional transcription and DNA methylation were computed for each bin. Bi-directional transcript levels defined by CAGE tags were obtained from FANTOM5. The raw levels x were transformed by the function $f(x) = \log_2(x + 1)$. The other features defined by ChIP-seq, DNase-seq and reduced representation bisulfite sequencing (RRBS) were obtained from ENCODE. H3K4me1 and H3K4me2 data were available only for K562 but not MCF-7.

Similarly, each promoter was composed of five 500bp bins that spanned from 2,000bp upstream of a TSS to 500bp downstream of a TSS. For each bin, the average values of the following features were computed: H3K4me3, H3K27me3, H3K9ac, DNase I hypersensitivity and DNA methylation. H3K9ac data were available only for K562 but not MCF-7.

For gene bodies, instead of defining bins of a fixed size, they were instead sub-divided into regions according to the gene structure as previously proposed (Li et al., 2010), namely first exon, internal exons, last exon, first intron, internal introns, and last intron. The features considered were H3K27me3, H3K36me3, DNase I hypersensitivity and DNA methylation. In order to compute feature values in these regions, for models that were compared with the ones involving gene body features, only genes with at least three introns were considered.

The expression level of each gene was defined as the number of reads mapped to the 1kbp region centered on the TSS, based on ENCODE poly-A enriched whole-cell RNA-seq data. Replicates were combined by taking their average. In the regression analyses, raw expression levels were log-transformed by the same function applied to eRNA levels.

Machine learning procedure for modeling gene expression with one-to-one enhancer-target associations

For the classification task, we divided all the included TSS into four equal-sized classes based on their expression levels. We randomly selected one-fifth of the pairs as a left-out testing set, and used the remaining four-fifth for training a Random Forest model (Breiman, 2001) for each of the four expression classes using the one-against-all setting. For example, when the highest expression class was defined as the positives, the other three classes were considered as negatives. The area under the receiver operator characteristic (AUC) was computed for each class, and the average AUC of the four classes was computed. This training-testing process was then repeated for a total of five disjoint testing sets. The average and standard deviation of the resulting AUC values from the five testing set results were computed.

To evaluate the best-case accuracy that a statistical model can achieve based on the data, we used the expression levels of the TSSs in one replicate as the only feature to predict the expression class of the

TSSs based on their expression levels in the other replicate. We took each of the two replicates at the predictor in turn and computed their average AUC.

For the regression task, a similar procedure was used. In this case, the target was the log-transformed expression values instead of the discrete expression classes. We constructed linear regression models implemented in Weka (Hall et al., 2009). The accuracy of a model was quantified by the percentage error. Suppose a gene in the testing set has an actual expression level of y_i and a predicted expression level of $\tilde{y}_i$, the percentage error is defined as $|y_i - \tilde{y}_i|/y_i$. The median of these error values among all genes was recorded. The whole procedure was repeated 10 times for 10 random testing sets. Error bars represent the standard deviation of these 10 values.

Collection and processing of data sets for modeling the joint effects of multiple enhancers on a target TSS

We collected ChIP-seq data of H3K4me1, H3K27ac and H3K27me3 and RNA-seq data for 127 human cell types, tissue types and cell lines (which we call "samples" in general) from ENCODE and Roadmap Epigenomics (ENCODE+Roadmap). Processed, replicates-combined signal values were downloaded from the Roadmap Epigenomics Web site in .bigwig format. Some of these ChIP-seq and RNA-seq data were imputed from other data files. The full list of samples can be found from the Roadmap Epigenomics metatdata page (`http://egg2.wustl.edu/roadmap/web_portal/meta.html`).

We also downloaded ChromHMM-predicted active enhancers (of states 6, 7 and 12) for each of the 127 samples from the same Web site. We took the union of the predicted enhancers from all samples, filtered those larger than 2500bp, merged the remaining enhancers that overlapped, and filtered the ones larger than 2500bp again after merging. Finally, we got a list of 489,581 enhancers.

For each of these enhancers, we computed the average H3K4me1, H3K27ac and H3K27me3 signal in each of the 127 samples based on the imputed data. For each TSS of each annotated protein-coding gene in Gencode version 19, we also computed the average RNA-seq signal in each of the 127 samples based on the imputed data. This whole set of data involving all 127 samples is denoted as "ENCODE+Roadmap 127 samples, imp-imp".

These 127 samples can be grouped into 19 categories (according to the "group" column in the spreadsheet linked from the Roadmap Epigenomics metadata page). We used these categories to define training and testing sets for the expression models, as described below.

Among the 127 samples, 48 of them contained non-imputed data for both RNA-seq and all three types of histone modification. Using these non-imputed data only, we defined another data set denoted as "ENCODE+Roadmap 48 samples, non-non". We also used the same 48 samples to construct two additional data sets, namely one involving only imputed data for both ChIP-seq and RNA-seq (denoted as "ENCODE+Roadmap 48 samples, imp-imp"), and a data set involving only imputed data for ChIP-seq but not for RNA-seq (denoted as "ENCODE+Roadmap 48 samples, imp-non").

In addition, we downloaded CAGE data from the FANTOM5 Web site for 808 samples (`http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.primary_cell.hCAGE/`, `http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.cell_line.hCAGE/` and `http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.tissue.hCAGE/`), and normalized the data as described in Andersson et al. (2014). The predicted active enhancers in each of these 808 samples and their processed CAGE signals were also downloaded from the FANTOM5 Web site. We took the union of these enhancers, and computed the log of the CAGE signal for each enhancer in each sample. We also computed the average CAGE signal at the promoter (from 500bp upstream of a TSS to 500bp downstream of it) of each RefSeq protein-coding gene, as in Andersson et al. (2014).

The primary cells can be grouped into 69 facets as suggested by the FANTOM5 consortium (Table 10 in the Supplement of Andersson et al. (2014)). When evaluating the accuracy of the expression models using the left-one-facet-out procedure (described below), we used only the 363 samples in these 69 facets.

Machine learning procedure for modeling the joint effects of multiple enhancers on a target TSS

To model the joint effects of multiple enhancers on a TSS, each time we took a target TSS and all enhancers within 1Mbp from it from the set of all processed enhancers collected from different samples. We considered only TSSs with non-zero expression in at least one sample. We used the histone modification data as enhancer features for the four ENCODE+Roadmap data sets, and the CAGE data as the only enhancer feature for the FANTOM5 data set. We used a left-one-category-out (in the case of the ENCODE+Roadmap) / left-one-facet-out (in the case of FANTOM5) procedure to evaluate the accuracy of expression models. Specifically, we constructed an expression model for this TSS using enhancer features and log expression values of all but one sample category/facet, and then applied the model to predict the log expression value of the TSS in the samples of the left-out category/facet.

We considered three different types of model, namely 1) simple linear regression involving one enhancer at a time (denoted as "independent"), 2) linear regression with LASSO (Tibshirani, 1996), and 3) linear regression with Elastic Net (Zou and Hastie, 2005). We used the R package glmnet for LASSO and Elastic Net. We have also tried some non-linear models such as Random Forest but found that the results were not better than the regularized linear regression models.

The accuracy of a model was quantified by the percentage error in the testing set. The median of the percentage error values among all TSSs and all samples was computed.

Construction and validation of sample-specific active enhancer-target networks

To determine the enhancer-target pairs active in a specific sample, we combined 1) the information from the above global models, 2) enhancer features and TSS expression levels in the sample, and 3) distance between enhancers and TSSs. The main idea is that the global models describe whether an enhancer regulates a TSS in any sample and the quantitative relationship between the features of the enhancer and the expression level of the TSS in general, but when focusing on a particular sample, whether the enhancer regulates the TSS in this sample depends on the activity of the enhancer and the expression level of the TSS in this sample, and how much the enhancer features can infer the expression level of the TSS in this sample. The distance between the enhancer and the TSS serves as a prior in addition to such sample-specific information.

Specifically, suppose the global linear regression model for a TSS is $y = a_0 + \sum_k a_k \cdot x_k$, where y is the expression level of the TSS, a_0 is the bias term, x_k is the features of enhancer k, a_k is the corresponding coefficients, and the summation is over all enhancers within 1Mbp from the TSS. To determine the regulating enhancers of the TSS in a particular sample j, we first computed the absolute error of the partial model of each enhancer k, $e_{kj} = |y_j - (a_0 + a_k \cdot x_{kj})|$, where y_j is the expression level of the TSS in sample j and x_{kj} is the features of enhancer k in sample j. This error term e_{kj} and the genomic distance between enhancer k and the TSS were then used as two features of this enhancer-TSS pair in learning a Random Forest classifier using a set of validation data (ChIA-PET, Hi-C or eQTL) to define the positive pairs and one of the four methods described below to define the negative pairs. An enhancer-TSS pair respectively overlapping a corresponding pair of genomic regions defined by the validation data was considered positive.

The first set of negatives was based on random DNA contacts (He et al., 2014). Specifically, given two genomic locations separated by a distance of $s$bp, the probability density for them to have a detected contact in a chromosome conformation capture experiment was $f(s) = ks^{-3/2}e^{-1400/s^2}$, where k reflects the efficiency of cross-linking (Dekker et al., 2002). We drew a positive enhancer-target pair randomly, used the above formula to sample a random contact point of the enhancer, and identified the TSS closest to this contact point. If the resulting enhancer-TSS pair was not supported by the validation data, we included it as a negative pair. We then repeated this procedure until getting the desired positive-to-negative ratio as described below.

The second set of negatives was based on random enhancers, which paired TSSs in the positive pairs with other random enhancers within 1Mbp. The third set was based on random targets, which paired

enhancers in the positive pairs with other random TSSs within 1Mbp. The fourth set was based on random pairs of enhancers and TSSs within 1Mbp, where both the enhancer and the TSS in each pair were randomly chosen.

In all four cases, we determined the number of negative pairs using the following method. For a given sample, let E, P and R be the number of active enhancers, the total number of enhancer-TSS pairs within 1Mbp, and the average number of active enhancers per TSS, respectively. E and P have fixed values for the sample, while N should take a value around 1-4 according to previous studies (He et al., 2014; Jin et al., 2013). The potential numbers of active and inactive enhancer-gene pairs in the sample are $E \times R$ and $P - E \times R$, respectively, and thus the positive-to-negative ratio should be $\frac{E \times N}{P - E\ timesN}$, and the number of negative enhancer-TSS pairs should be the number of positive enhancer-TSS pairs (based on ChIA-PET connections) multiplied by this ratio. The actual ratios used for ENCODE+Roadmap and FANTOM5 were 0.15 and 0.1, respectively. In the case of ENCODE+Roadmap, the ratio was achieved by sampling negative pairs using one of the four methods. In the case of FANTOM5, due to a smaller set of enhancers defined, sometimes the ratio could be achieved by considering all non-positive pairs to be negaive, and no sampling was necessary. In this case, we report the results as random pairs in Figure S9.

Taking the positive pairs with validation data support and one of these negative pairs, we performed 5-fold cross-validation tests to compute the area under the receiver operator characteristics (AUROC), area under the precision-recall curve (AUPR) and F-measure of the testing sets in Figures 2b-e and S7-S9. For computing the F-measure and in other analyses that required a set of binary enhancer-target pairs, we thresholded the ranked list of enhancer-TSS pairs using the Random Forest scores. By default, we used 0.5 as the threshold value.

To evaluate the performance of these models, we also constructed two baseline models. In the first baseline model, for each enhancer k within 1Mbp of a TSS, a linear regression model was formed using the features of the enhancer to infer the expression level of the TSS across all samples, $y = a_0 + a_k \cdot x_k$. When determining whether enhancer k regulates the TSS in a particular sample j, the error term $e_{kj} = |y_j - (a_0 + a_k \cdot x_{kj})|$ and the genomic distance between them were combined using Random Forest. In the second baseline model, the ranking of enhancer-TSS pairs was based on the genomic distance between an enhancer and a TSS alone, where pairs with a smaller distance were ranked higher.

To compare our method with IM-PET, we downloaded the enhancer-TSS pairs in K562 and CD4+ T cells predicted by IM-PET from `http://www.healthcare.uiowa.edu/labs/tan/EP_predictions.xlsx`. We extracted the enhancers from the 12 samples used in the IM-PET paper (He et al., 2014), and paired them with all TSSs within 1Mbp from them. We then ran IM-PET to rank these pairs by their chance of being active in K562 and in CD4+ T cells, and used the corresponding validation data to compute AUROC values. We also used our method for modeling each enhancer-target pair individually to rank the same pairs in K562, and computed a testing AUROC value using 5-fold cross-validation. We then applied the trained model in K562 to rank the pairs in CD4+ T cells to compute its AUROC value. We could not apply our method for modeling the joint effect of multiple enhancers, because it required the enhancers active in a large number of other samples as the inactive enhancers, which were not available for this set of data.

Since we only had validation data for several samples, to construct enhancer-target networks for all samples, we used the Random Forest model learned from the K562 cell line to combine the error terms computed from each individual sample and genomic distances for ranking the enhancer-target pairs. We chose the K562 cell line since it was one of the samples with the most reliable ChIA-PET data. We focused on the results of the LASSO models in the remaining analyses since the results for the LASSO and Elastic Net models were highly similar.

Analysis of chromosome domains and CTCF binding sites

We collected TADs in IMR90 and CCDs in GM12878 from Dixon et al. (2012) and Tang et al. (2015), respectively. Each TAD/CCD (referred to as "chromosome domains" in general) was represented by a contiguous genomic region. Different chromosome domains do not overlap, and there are genomic regions not covered by any chromosome domains. To evaluate whether our inferred enhancer-target pairs were more within a chromosome domain than expected by chance, we first computed the number of pairs

residing entirely within a chromosome domain. We then shuffled the chromosome domains by repeatedly swapping the location of two random domains. The genomic locations of all domains between them were also shifted accordingly based on the size difference of the two swapping domains. The number of enhancer-target pairs residing entirely within a chromosome domain was recorded, and the process was repeated to produce 1,000 shuffled lists of chromosome domains. The recorded numbers for these shuffled lists were then used to fit a Gaussian distribution, and a p-value of the number from the real chromosome domains was computed based on its z-score.

CTCF binding peaks were collected for 14 cell lines from ENCODE, namely GM12878, H1-hESC, HeLa-S3, HepG2, MHEC, HSMM, HSMMt, HUVEC, K562, NHA, NHDFAd, NHEK, NHLF and Osteobl. An enhancer-TSS pair was defined to have an intervening CTCF binding peak if the binding peak was anywhere between the enhancer and the TSS. We sampled random enhancer-TSS pairs matching both the total number and the distance distribution of the inferred enhancer-target pairs, and computed the number of pairs with intervening CTCF binding. These numbers from 1,000 sets of random pairs were used to fit a Gaussian distribution, and the p-value of the actual number of pairs with intervening CTCF binding was computed based on its z-score in the distribution.

A CTCF binding peak was considered to define a CCD boundary if it was within a certain distance from the boundary. We considered multiple distance values, including 1kbp, 2kbp, 3kbp, 4kbp and 5kbp. For each distance value, we then defined a 2×2 contingency table of whether a CTCF binding peak defined a CCD boundary, and whether it intervened an enhancer-target pair. A p-value was then computed using Fisher's exact test, making sure that the intervening CTCF binding peaks were significantly depleted rather than enriched in defining the CCD boundaries. Finally, among the 5 p-values obtained from the 5 distance values, the least significant one was reported.

Clustering of samples based on enhancer networks

For each sample, we represented its inferred enhancer-target network by a binary vector. The vector has an entry for every enhancer-TSS pair within 1Mbp from each other, which tool value 1 if this pair was inferred to be active in the sample, and value 0 otherwise. The distance between any two samples was defined as one minus the Pearson correlation of their vectors. The resulting distance matrix was then used to perform average-link hierarchical clustering using the ape package in R.

Visualizing group-specific parts of the enhancer target network

To visualize parts of the global enhancer-target network that are specifically active in a subset of related samples, we grouped all samples into a small number of groups (Tables S3-S5). A global network was first formed using all enhancers and TSSs involved in an enhancer-target pair in any sample. TSSs regulated only in a group, enhancers regulating only in a group, and enhancer-target pairs active only in a group were colored. A virtual node was then created for each sample group, and it was connected to all enhancer and TSS nodes active only in this sample group by virtual edges. The virtual nodes and virtual edges were not shown in the figures to avoid creating visual artifacts. The placement of network nodes and edges were then determined by the force directed layout option of Cytoscape (Shannon et al., 2003). The virtual nodes and virtual edges encouraged nodes with the same color to cluster together, but the actual network layout depended on the overall connections.

Definitions of expression specificity

We defined expression specificity scores based on the ideas of Schug et al. (2005). For a TSS i with expression level y_{ij} in sample j, the overall expression specificity score was defined as

$$S(i) = \log_2 m - \sum_{j=1}^{m} \left[\frac{y_{ij}}{\sum_{l=1}^{m} y_{il}} \log_2 \frac{y_{ij}}{\sum_{l=1}^{m} y_{il}} \right],$$

where m is the total number of samples. A TSS with uniform expression across the m samples would receive the minimum specificity score of 0, while a TSS specifically expressed in only one sample would have a maximum specificity score of $\log_2 m$.

We also defined a similar score to quantify whether a TSS i is specifically expressed in a particular sample j:

$$S(i,j) = S(i) + \log_2 \frac{y_{ij}}{\sum_{l=1}^{m} y_{il}}$$

In this function, $S(i,j)$ is bounded above by the overall specificity score $S(i)$, with a reduction of $\log_2 \frac{y_{ij}}{\sum_{l=1}^{m} y_{il}}$ depending on how specific TSS i is expressed in sample j.

In the calculation of $S(i,j)$, in order to avoid taking the logarithm of 0, a small positive constant is added to all expression values.

Definition and anlayses of enhancer coregulation modes

For any regulation module involving two enhancers and a TSS that they individually regulate in at least one sample, we represented the samples in which each enhancer regulated the TSS as a binary vector. We then computed both the significance of dependence of the two vectors by Fisher's exact test and their degree of overlap by Jaccard Index. The module was defined to be in Simultaneous coregulation mode if the one-sided Fisher's exact test (right tail) p-value was less than 0.1 and the Jaccard Index was larger than 0.45. The module was defined to be in Mutually Exclusive coregulation mode if the one-sided Fisher's exact test (left tail) p-value was less than 0.1 and the Jaccard Index was smaller than 0.05. A module not satisfying either definition was considered to be in Independent coregulation mode.

In the analysis of coregulation modules, super enhancers were downloaded form the dbSUPER database (Khan and Zhang, 2016). For the motif analysis, we first produced a FASTA file of the enhancer regions using bedtools (Quinlan and Hall, 2010) and downloaded the human motif data from hocomoco (Kulakovskiy et al., 2013) (`http://autosome.ru/HOCOMOCO/`). We then scanned the enhancer regions for the occurrences of the motifs using fimo (Grant et al., 2011). For each enhancer region, all unique motifs with a p-value less than 1e-4 were recorded, and the number of unique motifs in each region was counted. Accessibility correlations were based on the DNase I sensitivity of partner enhancers across 97 ENCODE+Roadmap samples, downloaded from `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase/`. Sequence conservation was based on PhastCons scores (Siepel et al., 2005) among 100 species obtained from UCSC Genome Browser. In the TF analysis, the list of TF genes was downloaded from The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2004). To classify regulatory modules in the Simultaneous mode into the Linear and Non-Linear sub-modes, we added a cross-term to the linear model involving the two partner enhancers as predictors for the expression level of the target. If the cross-term was significant ($p<0.01$), the two partner enhancers were classified as having non-linear interaction when regulating the target; Otherwise, they were considered to operate in a linear manner.

Acknowledgments

We would like to thank Yijun Ruan and Zhonghui Tang for providing the list of Chromatin Contact Domains in GM12878, and Wai-Lun Chan, Jie Chen and Xin Ma for helpful discussions. This project is supported by HKSAR RGC TBRS T12-401/13-R, T12-402/13-N and T12C-714/14-R, CRF C4017-14G, and a CUHK VC discretionary fund.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.*, 2014. An atlas of active enhancers across human cell types and tissues. *Nature*, **507**(7493):455–461.

Aran, D., Sabato, S., and Hellman, A., 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology*, **14**:R21.

Breiman, L., 2001. Rnadom forests. *Machine Learning*, **45**:5–32.

Bulger, M. and Groudine, M., 2010. Enhancers: The abundance and function of regulatory sequences beyond promoters. *Developmental Biology*, **339**:250–257.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-k., Dong, X., Djebali, S., Ruan, Y., *et al.*, 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, **22**(9):1658–1667.

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K., 2012. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research*, **22**:490–503.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T., 2014. Analysis of nascent rna identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, **46**(12):1311–1320.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N., 2002. Capturing chromosome conformation. *Science*, **295**(5558):1306–1311.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398):376–380.

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigo, R., Birney, E., *et al.*, 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, **13**:R53.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., *et al.*, 2006. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**:1299–1309.

Ernst, J. and Kellis, M., 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, **33**(4):364–376.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., 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**(7345):43–49.

Firpi, H. A., Ucar, D., and Tan, K., 2014. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**:1579–1586.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.*, 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**(7269):58–64.

Grant, C. E., Bailey, T. L., and Noble, W. S., 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**:1017–1018.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, **11**:10–18.

He, B., Chen, C., Teng, L., and Tan, K., 2014. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, **111**:E2191–E2199.

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbanian, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P., 2014. Genome-wide map of regulatory interactions in the human genome. *Genome Research*, **24**:1905–1917.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., *et al.*, 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243):108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., *et al.*, 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, **39**(3):311–318.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., Hoke, H. A., and Young, R. A., 2013. Super-enhancers in the control of cell identity and disease. *Cell*, **155**:934–947.

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., *et al.*, 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, **41**:827–841.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A., and Ren, B., *et al.*, 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475):290–294.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L., 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, **30**(1):90–98.

Khan, A. and Zhang, X., 2016. dbSUPER: A database of super-enhancers in mouse and human genome. *Nucleic Acids Research*, **44**:164–171.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., *et al.*, 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**(7295):182V–187.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J., 2013. HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research*, **41**:D195–D202.

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., and Cohen, B. A., 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, **24**:1595–1602.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., *et al.*, 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**:84–98.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., *et al.*, 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLOS Biology*, **8**:e1000533.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.*, 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950):289–293.

Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., Chan, L. L., Lam, V. K. L., So, W.-Y., Wang, Y., *et al.*, 2014. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, **15**:408.

Quinlan, A. R. and Hall, I. M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**:841–842.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**(7333):279V–283.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B., 2013. RFECS: A random-forest based algorithm for enhancer identification from chromatin state. *PLOS Computational Biology*, **9**:e1002968.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., *et al.*, 2014. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**:1665–1680.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.*, 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539):317–330.

Sakabe, N. J., Savic, D., and Nobrega, M. A., 2012. Transcriptional enhancers in development and disease. *Genome Biology*, **13**:238.

Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert Jr, C. J., 2005. Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biology*, **6**:R33.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T., 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**:2498–2504.

Shlyueva, D., Stampfel, G., and Stark, A., 2014. Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, **15**:272–286.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**:1034–1050.

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., *et al.*, 2015. CTCF-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, **163**:1611–1627.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., and Gersbach, C. A., 2015. Highly specific epigenome editing by CRISPR-cas9 repressors for silencing of distal regulatory elements. *Nature Methods*, **12**(12):1143V–1149.

The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2004. A promoter-level mammalian expression atlas. *Nature*, **507**(7493):462–470.

The GTEx Consortium, 2015. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235):648–660.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., *et al.*, 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**(7414):75–82.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**:267–288.

Visel, A., Rubin, E. M., and Pennacchio, L. A., 2009. Genomic views of distant-acting enhancers. *Nature*, **461**(7261):199–205.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**:307–319.

Williamson, I., Hill, R. E., and Bickmore, W. A., 2011. Enhancers: From developmental genetics to the genetics of common human disease. *Developmental Cell*, **21**:17–19.

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., *et al.*, 2012. Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biology*, **13**:R48.

Zhou, V. W., Goren, A., and Bernstein, B. E., 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, **12**(1):7–18.

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodological)*, **67**:301–320.

Figures

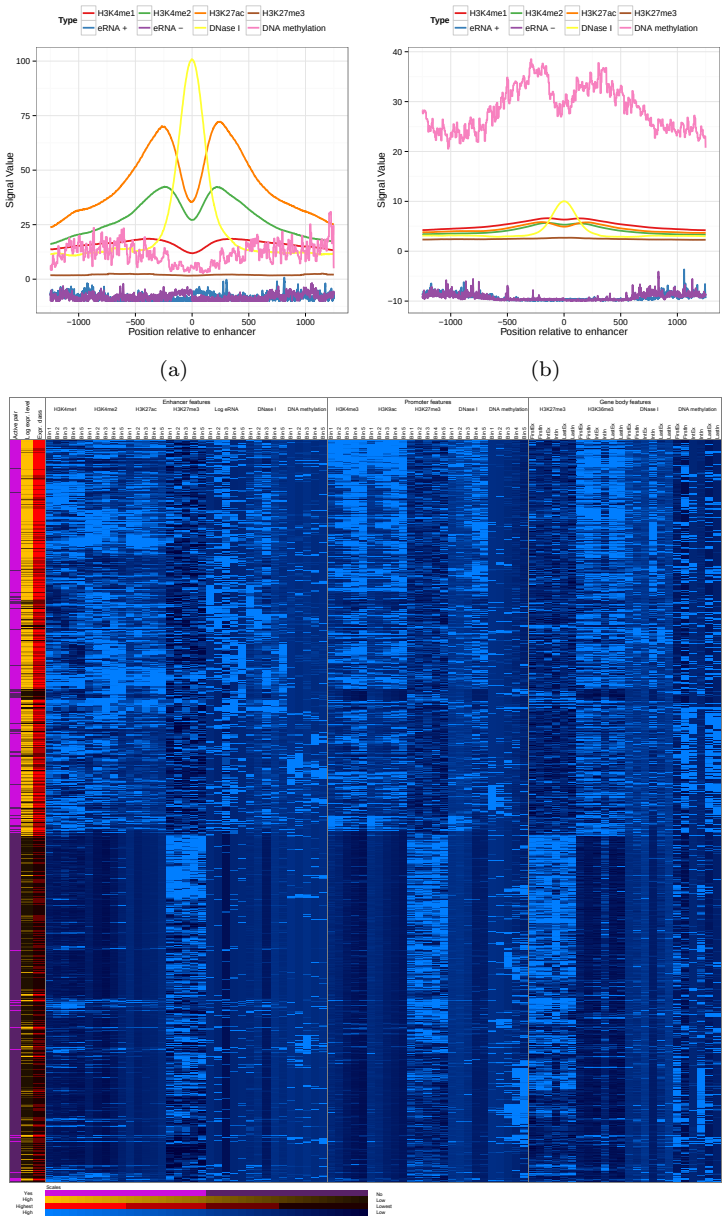


Figure 1: Patterns of enhancer and gene features. Aggregation plot of enhancer features around (a) active and (b) inactive enhancers as defined by FANTOM5 in K562. For features other than DNA methylation, each point is the average signal of all active/inactive enhancers at the corresponding base pair. For DNA methylation, smoothing was performed by averaging the signals within the 6bp local window since most signals come from CpG sites only. For eRNA, the log-transformed signals on the two strands are plotted separately. In the y-axis, the unit for eRNA is log-transformed signal per bp, while the unit for all other features is signal per bp. (c) A heatmap showing the correlations of enhancer, promoter and gene body features with gene expression. Each row in the heatmap corresponds to an enhancer-target pair that is either predicted to be active or inactive in the K562 cell line, based on FANTOM5 enhancers. The first three columns show the predicted activity of the pair in K562, the log-transformed expression level of the transcription start site (TSS) of the target gene, and the corresponding expression class. The remaining columns show the selected features at the enhancers, promoters, and gene bodies, respectively. For each enhancer and promoter feature, the signal values in five consecutive genomic bins are shown. For each gene body feature, the signal values in six regions, namely first exon (FirstEx), first intron (FirstIn), internal exons (IntEx), internal introns (IntIn), last exon (LastEx) and last intron (LastIn) are shown. Each column is standardized to have zero mean and unit variance. The rows are ordered by an average-link agglomerative hierarchical clustering using one minus Pearson correlation as the distance.

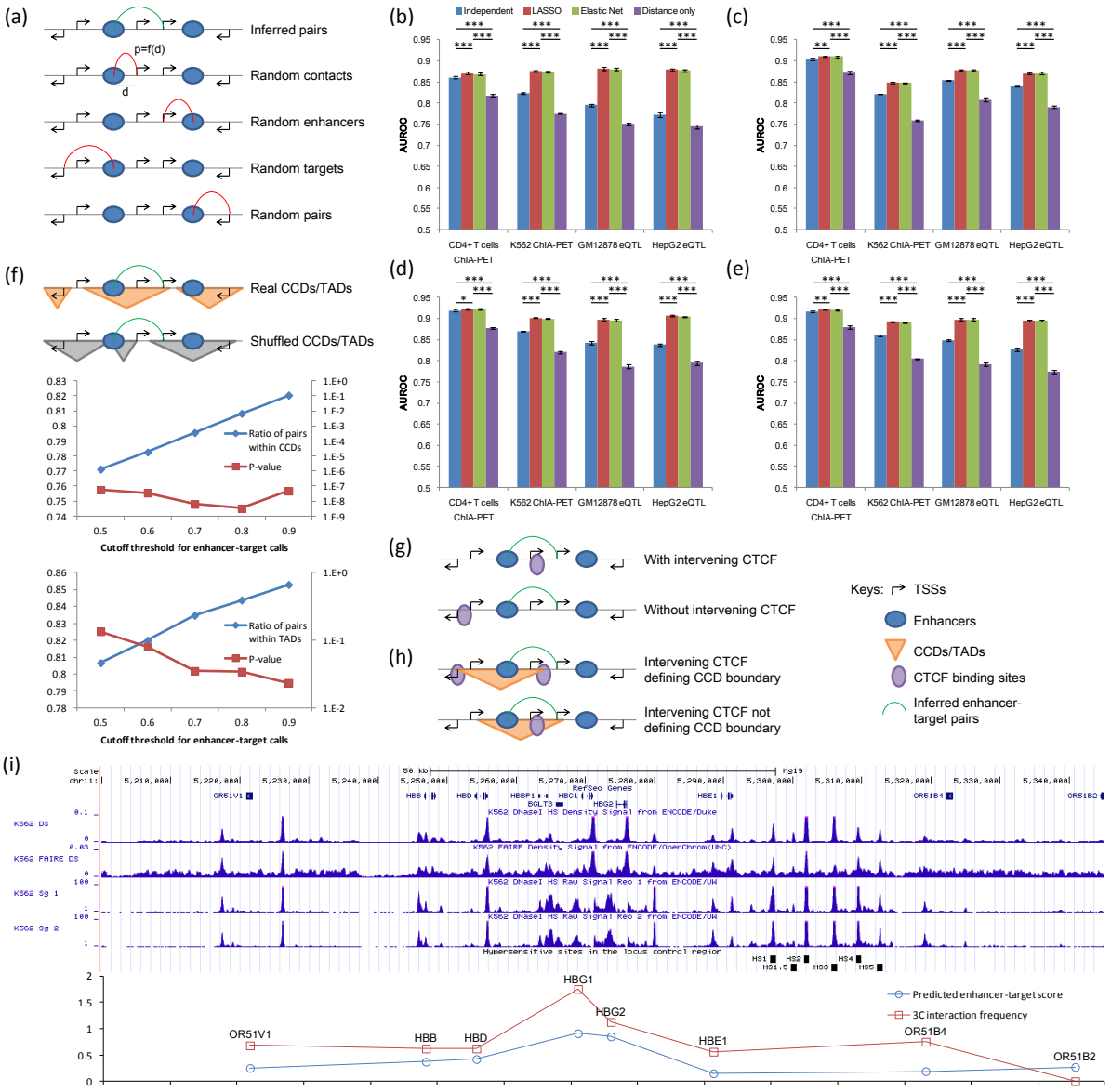


Figure 2: Reliability of the inferred enhancer networks. (a) The four types of random enhancer-TSS pairs for comparing with the inferred enhancer-target pairs based on ChIA-PET, eQTL and Hi-C data. (b-e) Consistency of the enhancer-target pairs inferred by non-imputed data from 48 ENCODE+Roadmap samples, using the three types of validation data as reference. The four panels correspond to four different ways to draw negative pairs, namely (b) random contacts, (c) random enhancers, (d) random targets and (e) random pairs. "Independent" shows the results of the models that consider each enhancer separately, "LASSO" and "Elastic Net" show the results of the two types of model that consider multiple enhancers jointly, and "Distance only" shows the results when only distance information was used to infer enhancer-target pairs. Error bars show the standard deviations of 10 repeated runs based on different random seeds. P-values were based on two-sided T-test. *:p<0.05; **:P<0.001; ***:p<0.0001. (f) Comparing the fractions of inferred enhancer-target pairs within GM12878 CCDs and IMR90 TADs with the fractions within shuffled domains. (g) Enhancer-target pair with or without intervening CTCF binding. (h) Intervening CTCF defining CCD boundary or not. (i) Locus control region (LCR) around the human beta-globin locus and flanking olfactory genes in K562. The upper panel shows the locations of the genes, open chromatin signals and locations of hypersensative sites (HS1-HS5) in the LCR. The locations of the hypersensitive sites were taken from Thakore et al. (2015). Lower panel shows the predicted enhancer-target scores between HS4 and various genes, and the corresponding chromatin conformation capture (3C) signals for HS5 from Dostie et al. (2006).

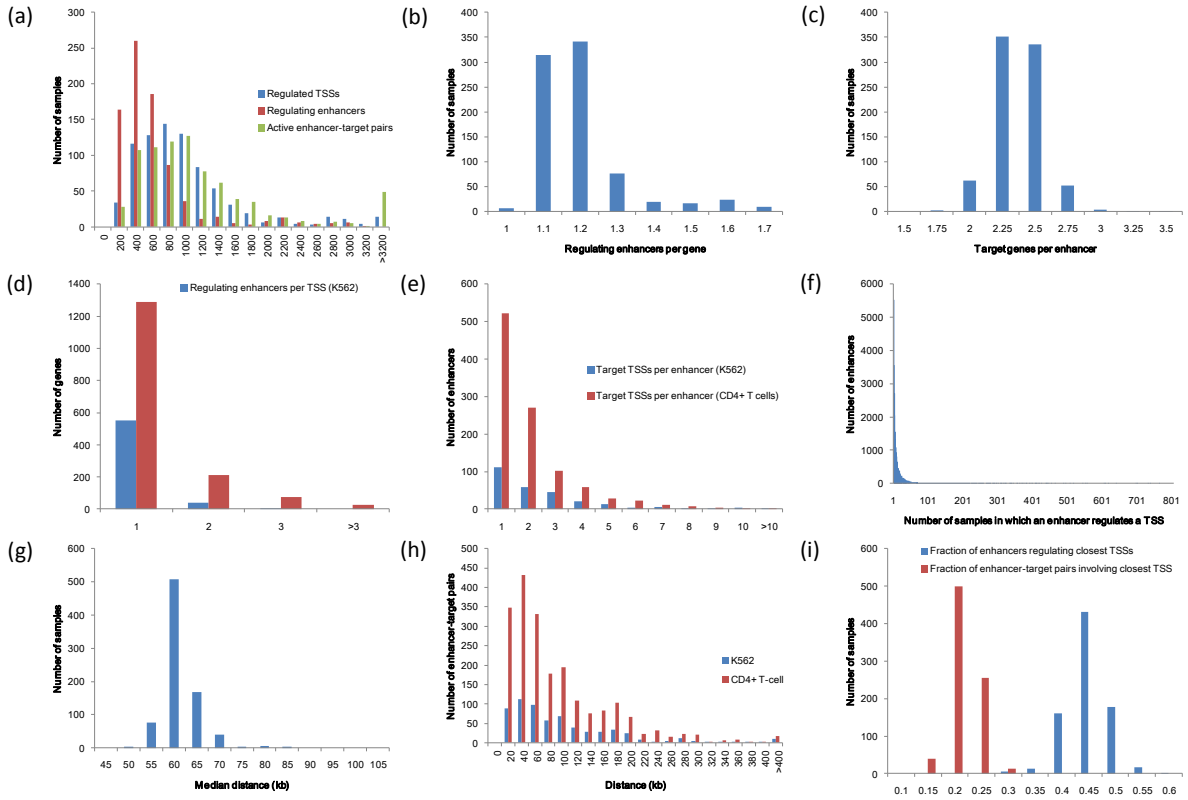
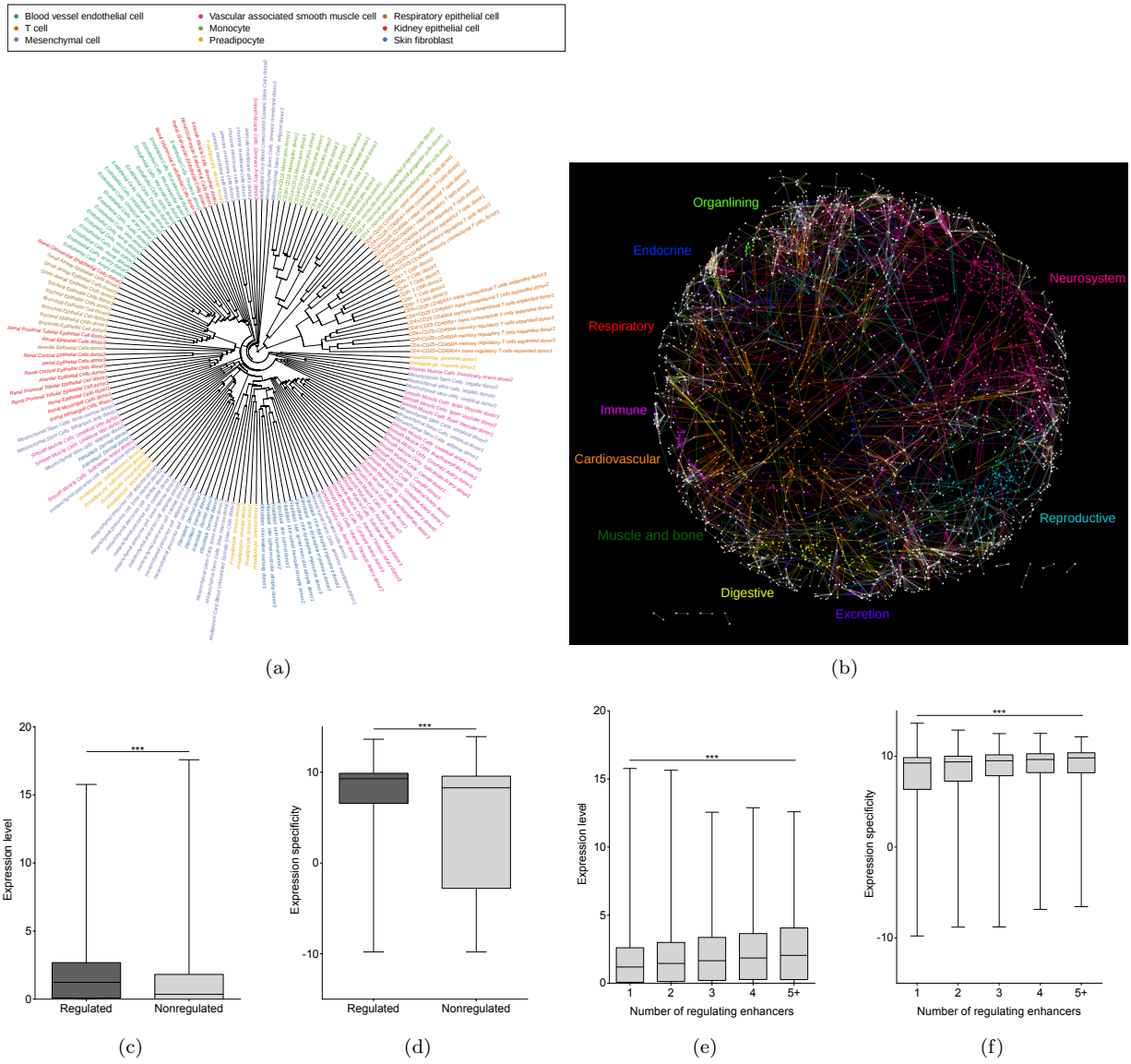


Figure 3: Basic properties of enhancer-target networks. (a) Number of TSSs regulated by at least one enhancer, number of enhancers regulating at least one target TSS, and number of enhancer-target pairs in each sample. (b) Average number of regulating enhancers per TSS and (c) average number of target TSSs per enhancer in each sample. (d) Number of regulating enhancers of each TSS and (e) number of target TSSs per enhancer in K562 and CD4+ T cells. (f) Number of samples in which each enhancer regulates a TSS. (g) Median distance between an enhancer and a target TSS that it regulates in each sample. (h) Distance between an enhancer and a TSS in each enhancer-target pair in K562 and CD4+ T cells. (i) Fraction of enhancers regulating the closest TSS and fraction of enhancer-target pairs involving an enhancer regulating the closest TSS in each sample.



(a) (b) (c) (d) (e) (f)

Figure 4: Context specificity of the inferred enhancer networks. (a) Clustering of samples using the inferred enhancer-target network of each sample as its signature for computing similarity values among FANTOM5 cell samples of the largest facets. Samples are colored based on the sample facets defined by FANTOM5. (b) Subnetworks active in only a single group of FANTOM5 tissues for the enhancers ans TSSs in chromosome 1. Enhancers and TSSs are respectively presented as circle and rhombus nodes, while enhancer-target connections are represented as edges. Each color corresponds to enhancers, TSSs and enhancer-target interactions that are respectively regulating, regulated and active in a single group of samples obtained by grouping the FANTOM5 facets. (c) Expression level and (d) expression specificity of expressed TSSs with or without a regulating enhancer in the expressed FANTOM5 sample. (e) Expression level and (f) expression specificity of expressed TSSs with various numbers of regulating enhancers in the expressed FANTOM5 sample. P-values in (c) and (d) were based on two-sided Wilcoxon rank-sum test, and those in (e) and (f) were based on Pearson correlation values. *:p<0.05; **:P<0.001; ***:p<0.0001

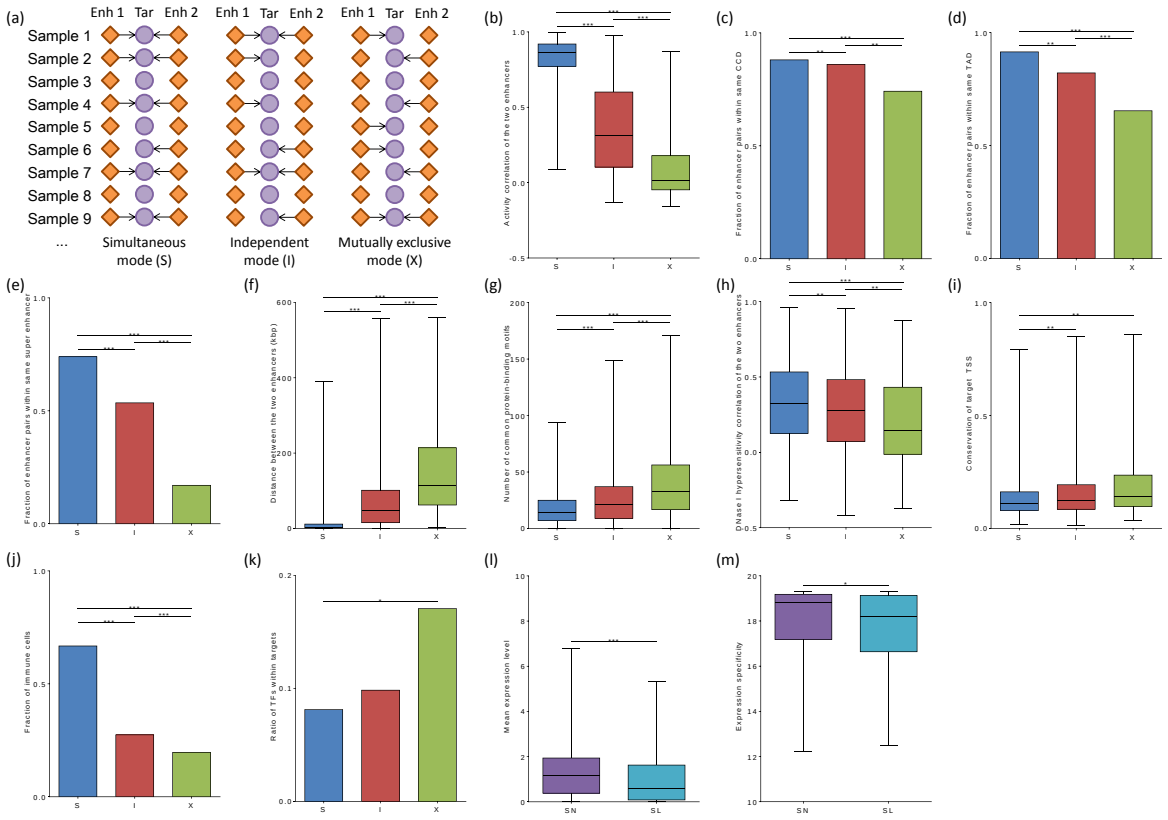


Figure 5: Enhancer coregulation modes. (a) Definition of the three coregulation modes, namely Simultaneous (S), Independent (I) and Mutually Exclusive (X). (b) Correlation between the activities of the two partner enhancers in each module based on their eRNA levels across all samples. (c)-(e) Fraction of modules in which the two enhancers reside in the same CCD (c), TAD (d) or super enhancer (e). (f) Genomic disntace between the two enhancers in each module. (g) Number of common transcription factor binding motifs found in partner enhancers. (h) Accessibility correlation of partner enhancers based on DNase I hypersensitivity. (i) Conservation of targets based on PhastCons scores 100 species. (j) Among the samples in which an enhancer in a module regulates the target, the fraction of them being immune cells. (k) Fraction of TF genes within targets. (l)-(m) Mean expression levels (l) and expression specificity (m) of targets regulated by Simultaneous mode with linear (SL) or non-linear (SN) relationship among the partner enhancers. In all box plots and bar plots, p-values were based on two-sided Wilcoxon rank-sum test and two-sided Fisher's exact test, respectively. *:p<0.05; **:p<0.001; ***:p<0.0001

Supplementary information

Supplementary tables

Table S1: Summary of datasets used for studying each enhancer-target pair independently.

Cell line	Enhancers (features used in prediction)	Enhancer-target associations
K562	ChromHMM (Ernst et al., 2011) (histone marks, protein binding)	POLR2A ChIA-PET (Li et al., 2012)
K562	FANTOM5 (Andersson et al., 2014) (eRNA)	POLR2A ChIA-PET (Li et al., 2012)
MCF-7	CSI-ANN (Firpi et al., 2014) (histone marks)	POLR2A ChIA-PET (Li et al., 2012)
MCF-7	FANTOM5 (Andersson et al., 2014) (eRNA)	POLR2A ChIA-PET (Li et al., 2012)

Table S2: Summary of datasets used for evaluating the reliability of the inferred enhancer-target pairs.

Data type	Sample	Data used in inferring enhancer targets
ChIA-PET (H3K4me2) (Chepelev et al., 2012)	CD4+ T cells	ENCODE+Roadmap, FANTOM5
ChIA-PET (POLR2A) (Li et al., 2012)	K562	ENCODE+Roadmap, FANTOM5
ChIA-PET (POLR2A) (Li et al., 2012)	MCF-7	FANTOM5
eQTL (The GTEx Consortium, 2015)	GM12878	ENCODE+Roadmap
eQTL (The GTEx Consortium, 2015)	HepG2	ENCODE+Roadmap
Hi-C (Jin et al., 2013)	IMR90	ENCODE+Roadmap

Table S3: Sub-grouping of the ENCODE+Roadmap samples.

Group	Samples
Cardiovascular	E033, E034, E037, E038, E039, E040, E041, E042, E043, E044, E045, E047, E048, E062, E065, E083, E095, E104, E105
Digestive	E075, E077, E079, E084, E085, E092, E094, E101, E102, E106, E109, E110
Immune	E029, E030, E031, E032, E035, E036, E046, E050, E051, E093, E112
Muscle cells	E052, E076, E078, E089, E090, E100, E103, E107, E108, E111
Neurosystem	E053, E054, E067, E068, E069, E070, E071, E072, E073, E074, E081, E082
Protective	E027, E028, E055, E056, E057, E058, E059, E061, E063
Respiratory	E017
Stem cells	E001, E002, E003, E004, E005, E006, E007, E008, E009, E010, E011, E012, E013, E014, E015, E016, E018, E019, E020, E021, E022, E023, E024, E025, E026, E049
Others	E066, E080, E086, E087, E088, E091, E096, E097, E098, E099, E113, E114, E115, E116, E117, E118, E119, E120, E121, E122, E123, E124, E125, E126, E127, E128, E129

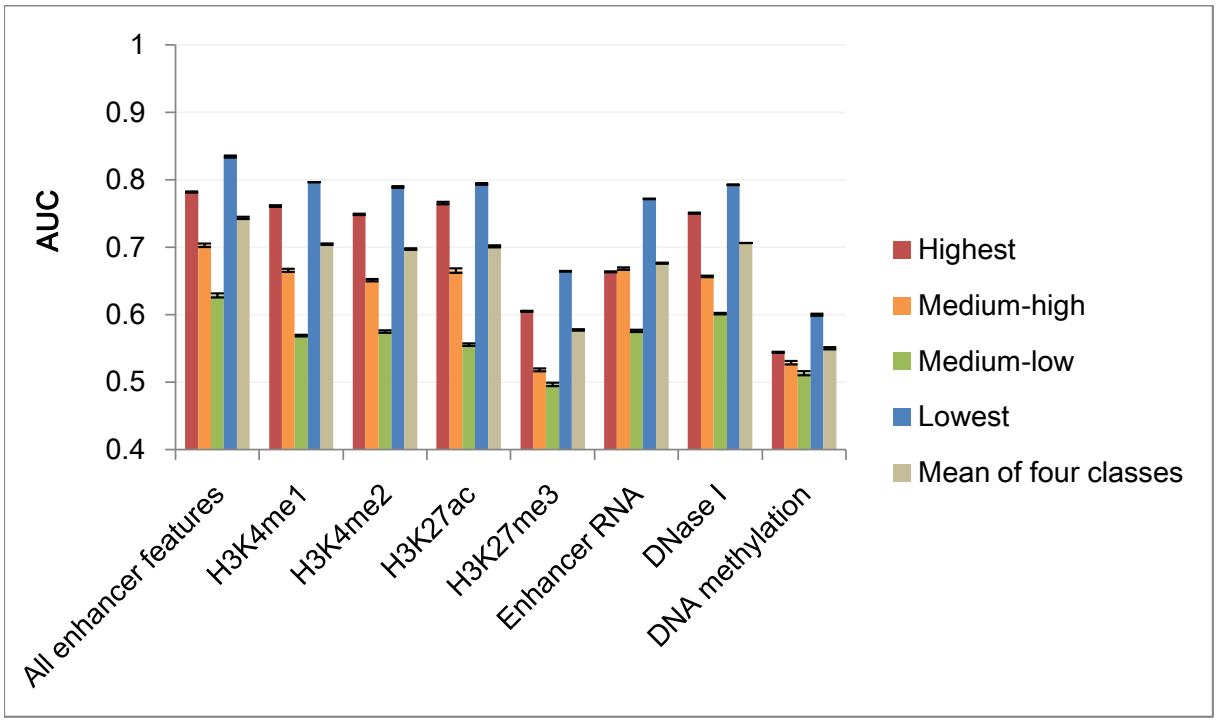
Table S4: Sub-grouping of the FANTOM5 primary cells.

Group	Samples
Cardiovascular	CNhs12075, CNhs12092, CNhs12341, CNhs12350, CNhs12571, CNhs13552, CNhs13553
Digestive	CNhs10847, CNhs11051, CNhs11054, CNhs11335, CNhs11371, CNhs11951, CNhs12017, CNhs12067, CNhs12068, CNhs12069, CNhs12093, CNhs12340, CNhs12349, CNhs12494, CNhs12626, CNhs12810, CNhs12811, CNhs12812
Fibroblast	CNhs10848, CNhs10866, CNhs10867, CNhs10874, CNhs10878, CNhs11052, CNhs11065, CNhs11074, CNhs11082, CNhs11319, CNhs11322, CNhs11337, CNhs11339, CNhs11351, CNhs11352, CNhs11353, CNhs11354, CNhs11378, CNhs11379, CNhs11902, CNhs11907, CNhs11909, CNhs11911, CNhs11912, CNhs11913, CNhs11914, CNhs11952, CNhs11953, CNhs11961, CNhs11962, CNhs11971, CNhs11981, CNhs11982, CNhs11996, CNhs12006, CNhs12011, CNhs12013, CNhs12027, CNhs12028, CNhs12038, CNhs12039, CNhs12052, CNhs12055, CNhs12057, CNhs12059, CNhs12061, CNhs12065, CNhs12095, CNhs12118, CNhs12344, CNhs12493, CNhs12498, CNhs12499
Immune	CNhs10843, CNhs10852, CNhs10853, CNhs10854, CNhs10855, CNhs10857, CNhs10858, CNhs10859, CNhs10861, CNhs10862, CNhs10865, CNhs11062, CNhs11073, CNhs11334, CNhs11897, CNhs11899, CNhs11901, CNhs11904, CNhs11905, CNhs11906, CNhs11954, CNhs11955, CNhs11956, CNhs11957, CNhs11959, CNhs11997, CNhs11998, CNhs11999, CNhs12000, CNhs12001, CNhs12003, CNhs12091, CNhs12122, CNhs12343, CNhs12352, CNhs12354, CNhs12566, CNhs12575, CNhs12592, CNhs12593, CNhs12594, CNhs13195, CNhs13202, CNhs13203, CNhs13204, CNhs13205, CNhs13206, CNhs13207, CNhs13208, CNhs13215, CNhs13216, CNhs13223, CNhs13224, CNhs13468, CNhs13480, CNhs13484, CNhs13491, CNhs13512, CNhs13513, CNhs13535, CNhs13536, CNhs13537, CNhs13538, CNhs13539, CNhs13540, CNhs13541, CNhs13547, CNhs13548, CNhs13549, CNhs13811, CNhs13812, CNhs13813, CNhs13814
Muscle and bone	CNhs10838, CNhs10839, CNhs10863, CNhs10868, CNhs10869, CNhs10870, CNhs10877, CNhs10883, CNhs11083, CNhs11084, CNhs11085, CNhs11086, CNhs11087, CNhs11088, CNhs11090, CNhs11091, CNhs11305, CNhs11309, CNhs11311, CNhs11320, CNhs11324, CNhs11329, CNhs11372, CNhs11373, CNhs11385, CNhs11900, CNhs11908, CNhs11920, CNhs11923, CNhs11927, CNhs11963, CNhs11964, CNhs11965, CNhs11973, CNhs11976, CNhs11980, CNhs11987, CNhs11988, CNhs11989, CNhs11990, CNhs11991, CNhs11992, CNhs12004, CNhs12007, CNhs12008, CNhs12015, CNhs12020, CNhs12021, CNhs12035, CNhs12036, CNhs12043, CNhs12044, CNhs12045, CNhs12046, CNhs12048, CNhs12049, CNhs12050, CNhs12053, CNhs12056, CNhs12060, CNhs12080, CNhs12569, CNhs12597, CNhs12639, CNhs12640, CNhs12641, CNhs12731, CNhs12894
Neurosystem	CNhs10864, CNhs11321, CNhs11960, CNhs12005, CNhs12081, CNhs12117, CNhs12338, CNhs12726, CNhs13815
Organlining	CNhs10837, CNhs10842, CNhs10850, CNhs10851, CNhs10871, CNhs10872, CNhs10875, CNhs10882, CNhs10884, CNhs11061, CNhs11077, CNhs11079, CNhs11092, CNhs11303, CNhs11317, CNhs11323, CNhs11325, CNhs11330, CNhs11331, CNhs11332, CNhs11333, CNhs11336, CNhs11338, CNhs11340, CNhs11375, CNhs11376, CNhs11377, CNhs11382, CNhs11383, CNhs11386, CNhs11896, CNhs11903, CNhs11925, CNhs11926, CNhs11966, CNhs11967, CNhs11972, CNhs11975, CNhs11977, CNhs11978, CNhs11993, CNhs12009, CNhs12010, CNhs12012, CNhs12014, CNhs12016, CNhs12022, CNhs12023, CNhs12024, CNhs12026, CNhs12032, CNhs12033, CNhs12037, CNhs12051, CNhs12054, CNhs12058, CNhs12062, CNhs12074, CNhs12079, CNhs12084, CNhs12086, CNhs12087, CNhs12088, CNhs12120, CNhs12121, CNhs12123, CNhs12124, CNhs12342, CNhs12347, CNhs12348, CNhs12495, CNhs12496, CNhs12497, CNhs12568, CNhs12570, CNhs12572, CNhs12574, CNhs12589, CNhs12596, CNhs12624, CNhs12728, CNhs12732, CNhs12733, CNhs13080, CNhs13550, CNhs13551, CNhs13816, CNhs13817, CNhs13818, CNhs13819
Outerlayer	CNhs10879, CNhs11064, CNhs11381, CNhs11979, CNhs12030, CNhs12031, CNhs12339, CNhs12501
Respiratory	CNhs11328
Stem cells	CNhs10844, CNhs10845, CNhs11057, CNhs11063, CNhs11341, CNhs11344, CNhs11345, CNhs11347, CNhs11349, CNhs11350, CNhs11384, CNhs12100, CNhs12104, CNhs12105, CNhs12125, CNhs12126, CNhs12127, CNhs12363, CNhs12364, CNhs12365, CNhs12366, CNhs12367, CNhs12368, CNhs12369, CNhs12370, CNhs12371, CNhs12379, CNhs12380, CNhs12492, CNhs12502, CNhs12503, CNhs12504, CNhs12506, CNhs12730, CNhs12922, CNhs13098

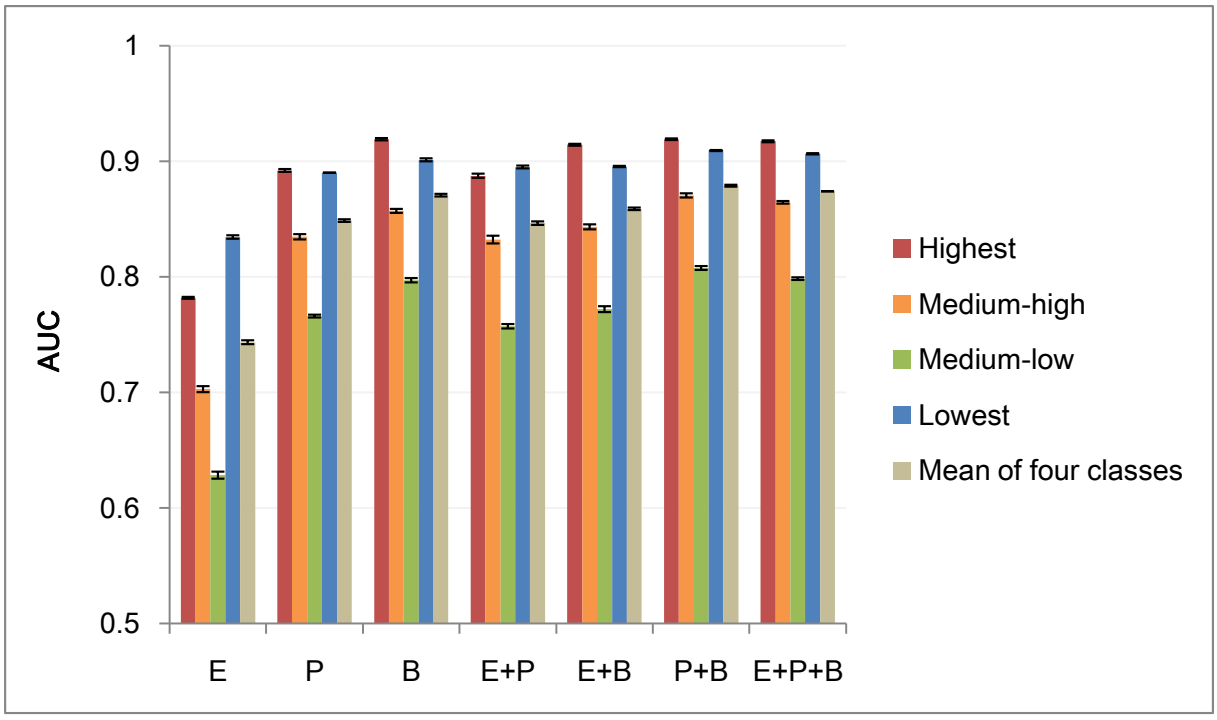
Table S5: Sub-grouping of the FANTOM5 tissues.

Group	Samples
Cardiovascular	CNhs10621, CNhs10653, CNhs11075, CNhs11076, CNhs11671, CNhs11672, CNhs11673, CNhs11757, CNhs11758, CNhs11760, CNhs11761, CNhs11789, CNhs11790, CNhs11948, CNhs11949, CNhs12844, CNhs12855, CNhs12856, CNhs12857
Digestive	CNhs10619, CNhs10620, CNhs10624, CNhs10630, CNhs11677, CNhs11768, CNhs11771, CNhs11773, CNhs11777, CNhs11780, CNhs11794, CNhs11798, CNhs12842, CNhs12848, CNhs12849, CNhs12852, CNhs12853
Endocrine	CNhs10633, CNhs10634, CNhs10650, CNhs11756, CNhs11769
Excretion	CNhs10616, CNhs10622, CNhs10652
Immune	CNhs10631, CNhs10651, CNhs10654, CNhs11788
Muscle and bone	CNhs10629, CNhs11755, CNhs11776, CNhs11779, CNhs13454
Neurosystem	CNhs10617, CNhs10636, CNhs10637, CNhs10638, CNhs10640, CNhs10641, CNhs10642, CNhs10643, CNhs10644, CNhs10645, CNhs10646, CNhs10647, CNhs10649, CNhs11762, CNhs11764, CNhs11772, CNhs11781, CNhs11782, CNhs11784, CNhs11787, CNhs11795, CNhs11796, CNhs11797, CNhs12227, CNhs12228, CNhs12229, CNhs12310, CNhs12311, CNhs12312, CNhs12314, CNhs12315, CNhs12316, CNhs12317, CNhs12318, CNhs12319, CNhs12320, CNhs12321, CNhs12322, CNhs12323, CNhs12324, CNhs12610, CNhs12611, CNhs12996, CNhs12997, CNhs13449, CNhs13793, CNhs13794, CNhs13795, CNhs13796, CNhs13797, CNhs13798, CNhs13799, CNhs13800, CNhs13801, CNhs13802, CNhs13804, CNhs13805, CNhs13807, CNhs13808, CNhs13809
Organlining	CNhs10615, CNhs10648, CNhs11774, CNhs12840
Reproductive	CNhs10618, CNhs10626, CNhs10627, CNhs10628, CNhs10632, CNhs11676, CNhs11763, CNhs11765, CNhs12846, CNhs12847, CNhs12850, CNhs12851, CNhs12854, CNhs12998
Respiratory	CNhs10625, CNhs10635, CNhs11680, CNhs11766, CNhs11770, CNhs11786, CNhs12858

Supplementary figures

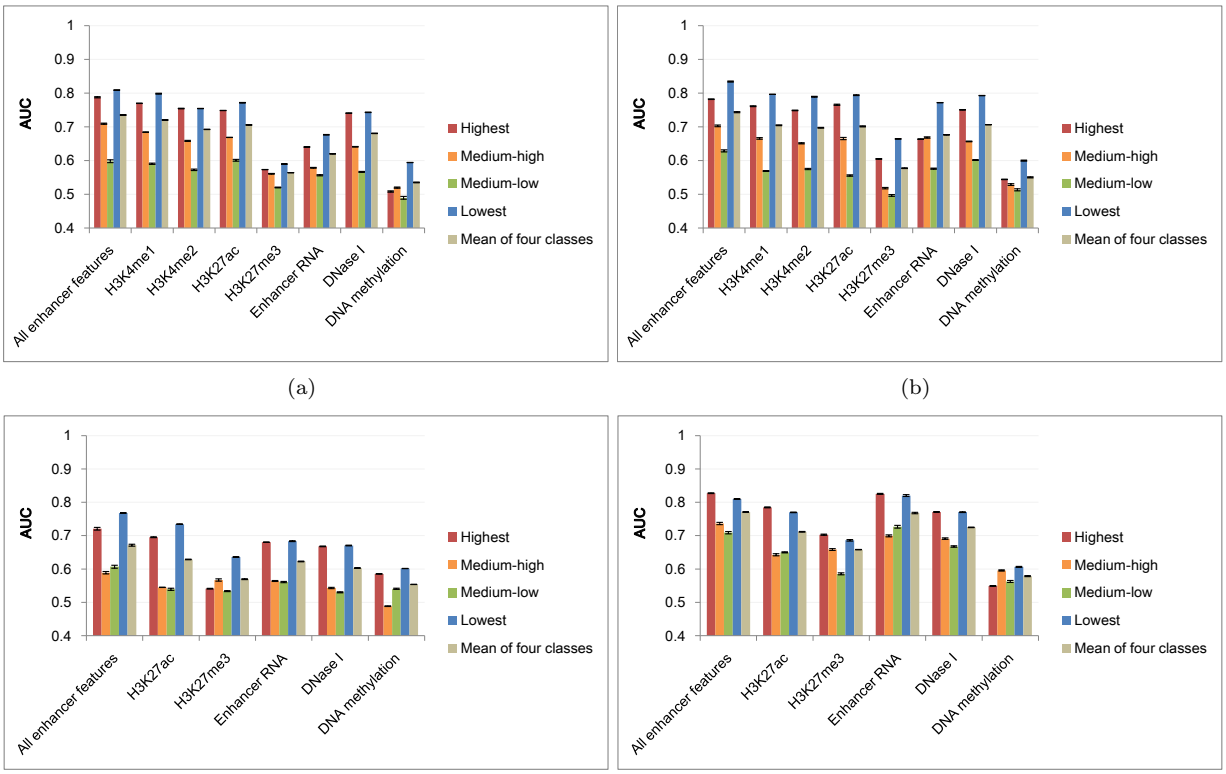


(a)



(b)

Figure S1: Classification accuracy of the statistical models when each enhancer-target pair is considered separately, based on FANTOM5 enhancers and ChIA-PET-defined active and inactive enhancer-target pairs in K562. (a) Accuracy of the models involving all enhancer features together, or one type of features at a time. (b) Accuracy of the models based on enhancer features (E), promoter features (P), gene body features (B), or their combinations. In both panels, error bars show the standard deviations of five repeated runs of the cross-validation procedure, each with a different random partition of all enhancer-target pairs into five subsets for defining training and testing sets.



(a) (b)

(c) (d)

Figure S2: Classification accuracy of the statistical models when each enhancer-target pair is considered separately, involving all enhancer features together, or one type of features at a time. The active enhancers were predicted (a) by ChromHMM in K562, (b) by FANTOM5 in K562, (c) by CSI-ANN in MCF-7 and (d) by FANTOM5 in MCF-7. Error bars show the standard deviations of five repeated runs of the cross-validation procedure, each with a different random partition of all enhancer-target pairs into five subsets for defining training and testing sets.

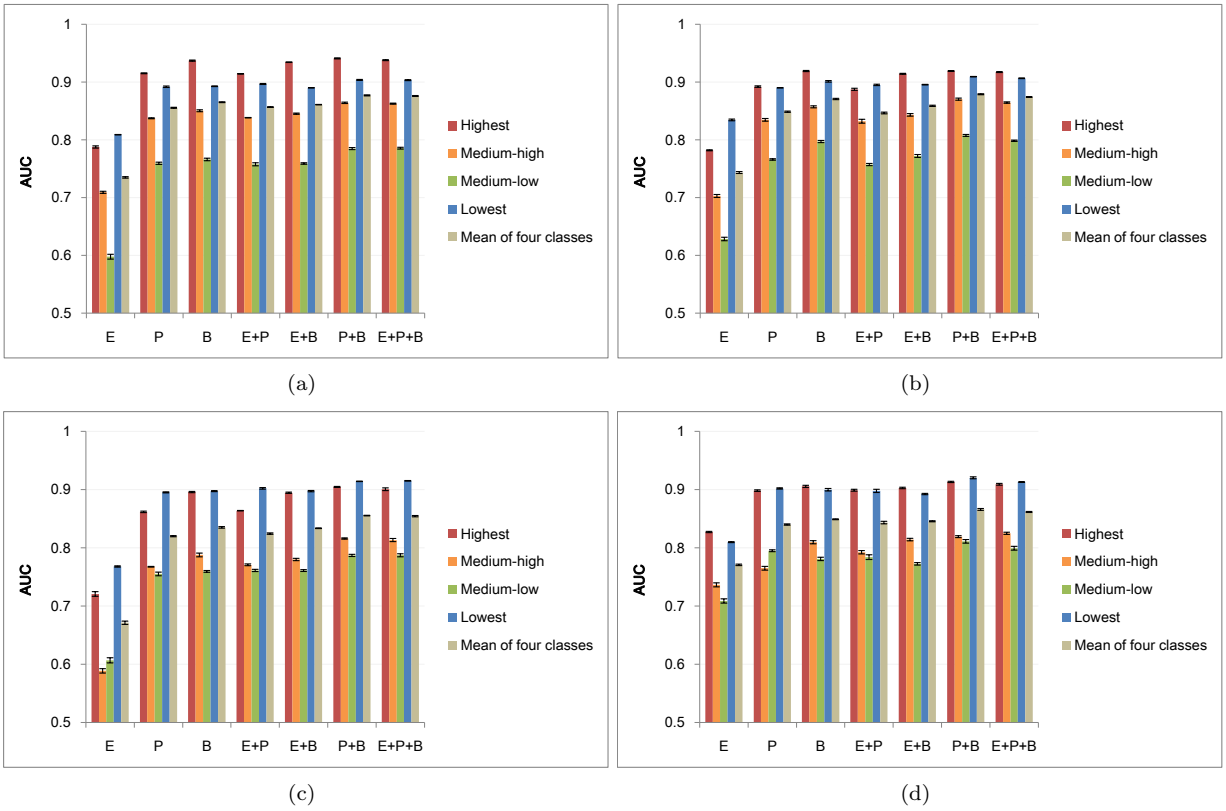


Figure S3: Classification accuracy of the statistical models when each enhancer-target pair is considered separately, involving enhancer features (E), promoter features (P), gene body features (B), or their combinations. The active enhancers were predicted (a) by ChromHMM in K562, (b) by FANTOM5 in K562, (c) by CSI-ANN in MCF-7 and (d) by FANTOM5 in MCF-7. Error bars show the standard deviations of five repeated runs of the cross-validation procedure, each with a different random partition of all enhancer-target pairs into five subsets for defining training and testing sets.

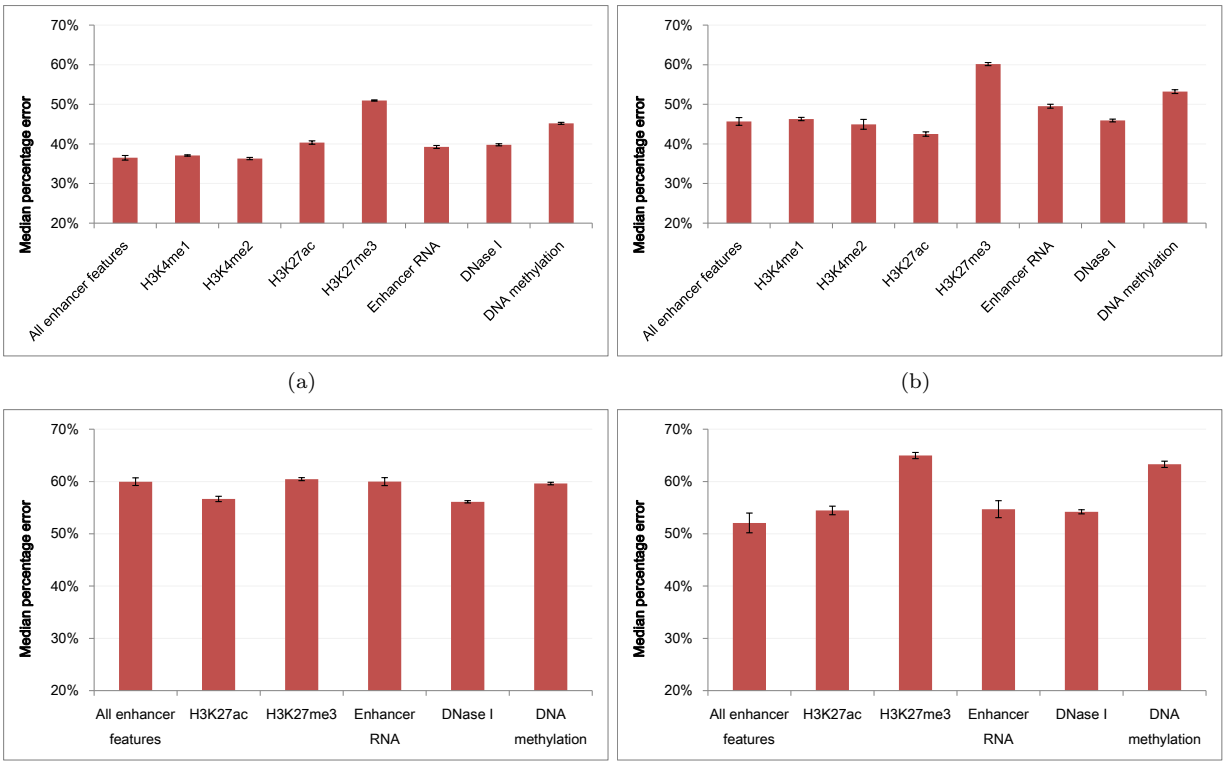
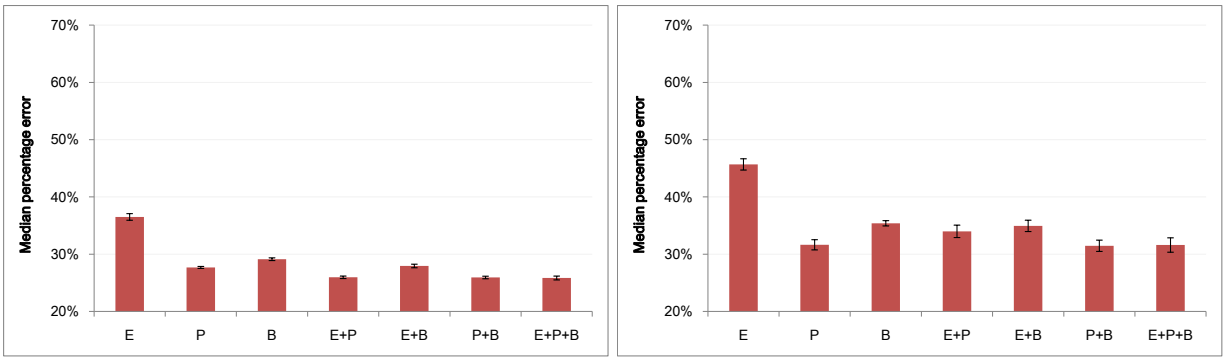
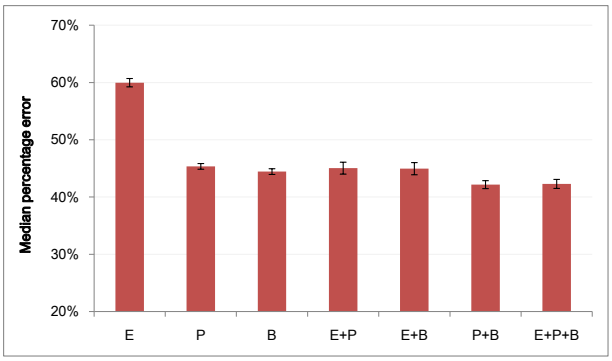


Figure S4: Regression errors of the statistical models when each enhancer-target pair is considered separately, involving all enhancer features together, or one type of features at a time. The active enhancers were predicted (a) by ChromHMM in K562, (b) by FANTOM5 in K562, (c) by CSI-ANN in MCF-7 and (d) by FANTOM5 in MCF-7.

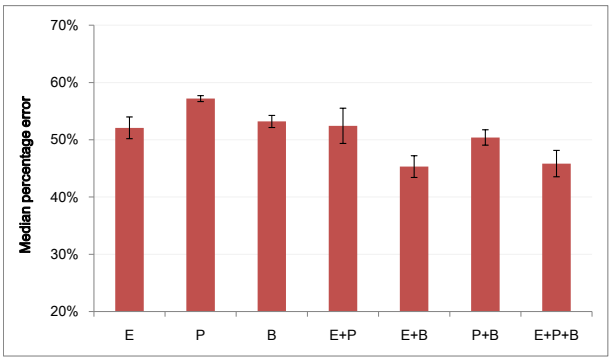


(a)

(b)



(c)



(d)

Figure S5: Regression errors of the statistical models when each enhancer-target pair is considered separately, involving enhancer features (E), promoter features (P), gene body features (B), or their combinations. The active enhancers were predicted (a) by ChromHMM in K562, (b) by FANTOM5 in K562, (c) by CSI-ANN in MCF-7 and (d) by FANTOM5 in MCF-7.

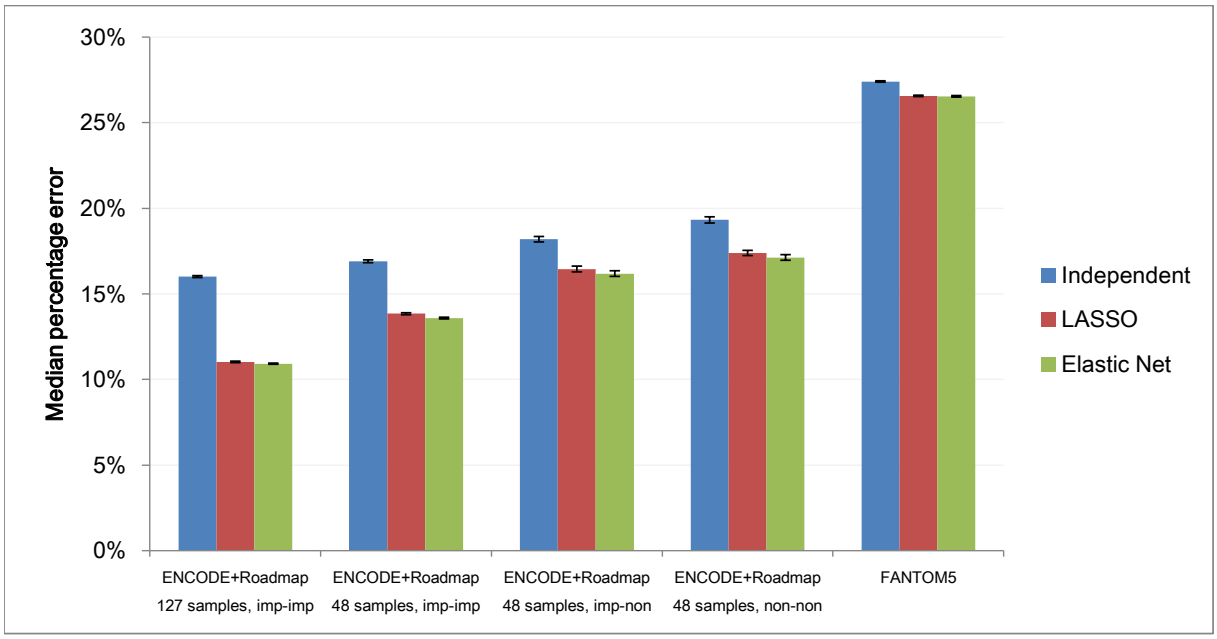


Figure S6: Regression errors of the statistical models when multiple enhancers potentially regulating a TSS are considered at the same time. Each bar group corresponds to the results based on one data set. Each data set from ENCODE+Roadmap is marked by the number of samples involved, whether the enhancer features involve (imp) or not involve (non) data imputation, and whether the expression data involve or not involve data imputation. For example, "Roadmap 48 samples, imp-non" is the data set involving 48 samples, where the enhancer features involve some imputed data while the expression data do not involve imputation. Within each bar group, "Independent" shows the results of the models that consider each enhancer separately, while "LASSO" and "Elastic Net" show the results of the two types of model that consider multiple enhancers jointly. Error bars show the standard deviations of the median percentage errors in ten random sub-samples.

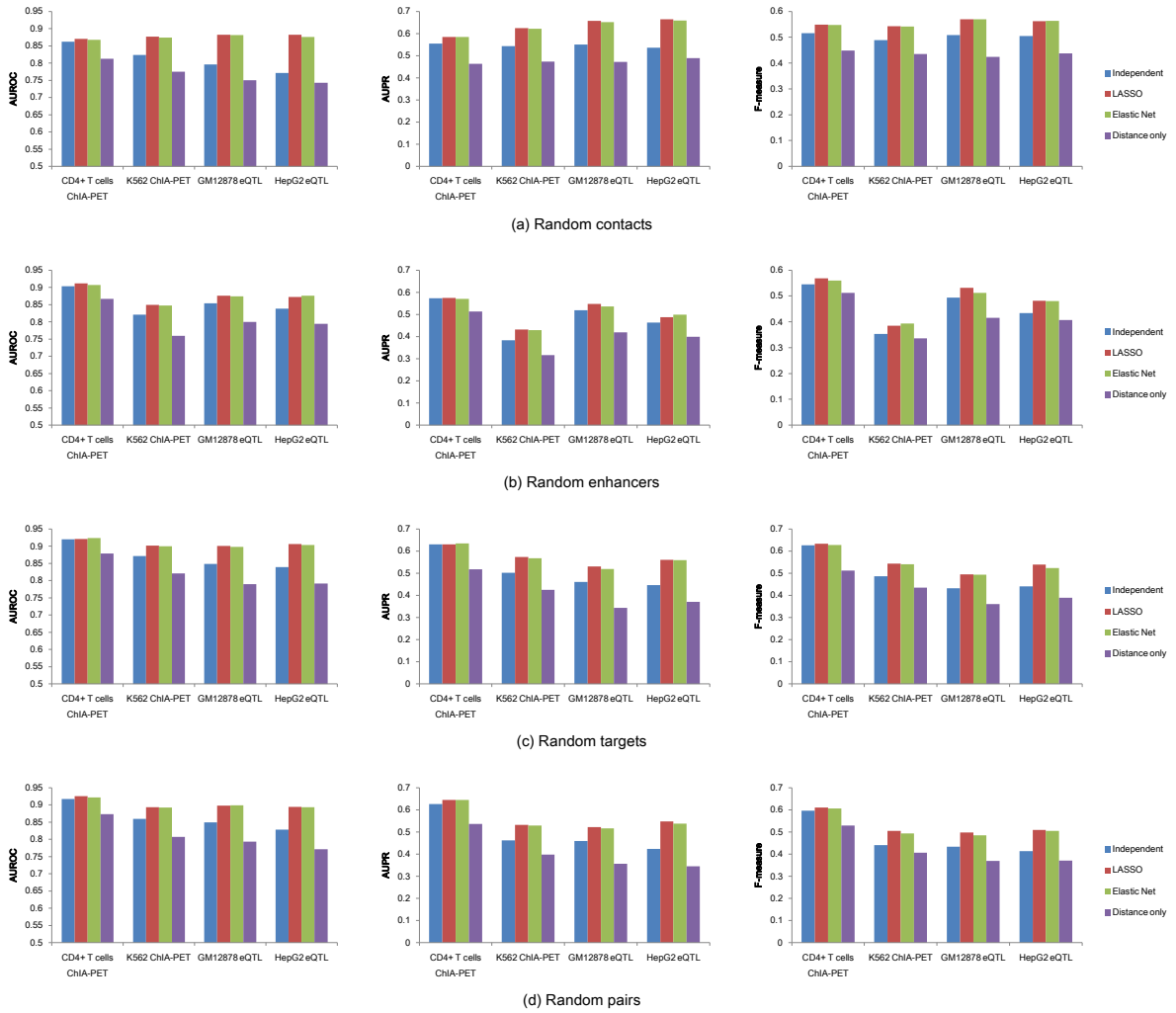


Figure S7: Consistency of the enhancer-target pairs inferred by non-imputed data from 48 ENCODE+Roadmap samples, using ChIA-PET, eQTL and Hi-C data as reference. The four panels correspond to four different ways to draw negative pairs, namely (a) random contacts, (b) random enhancers, (c) random targets and (d) random pairs. Each panel consists of three sub-figures that correspond to three different consistency measures. "Independent" shows the results of the models that consider each enhancer separately, "LASSO" and "Elastic Net" show the results of the two types of model that consider multiple enhancers jointly, and "Distance only" shows the results when only distance information was used to infer enhancer-target pairs.

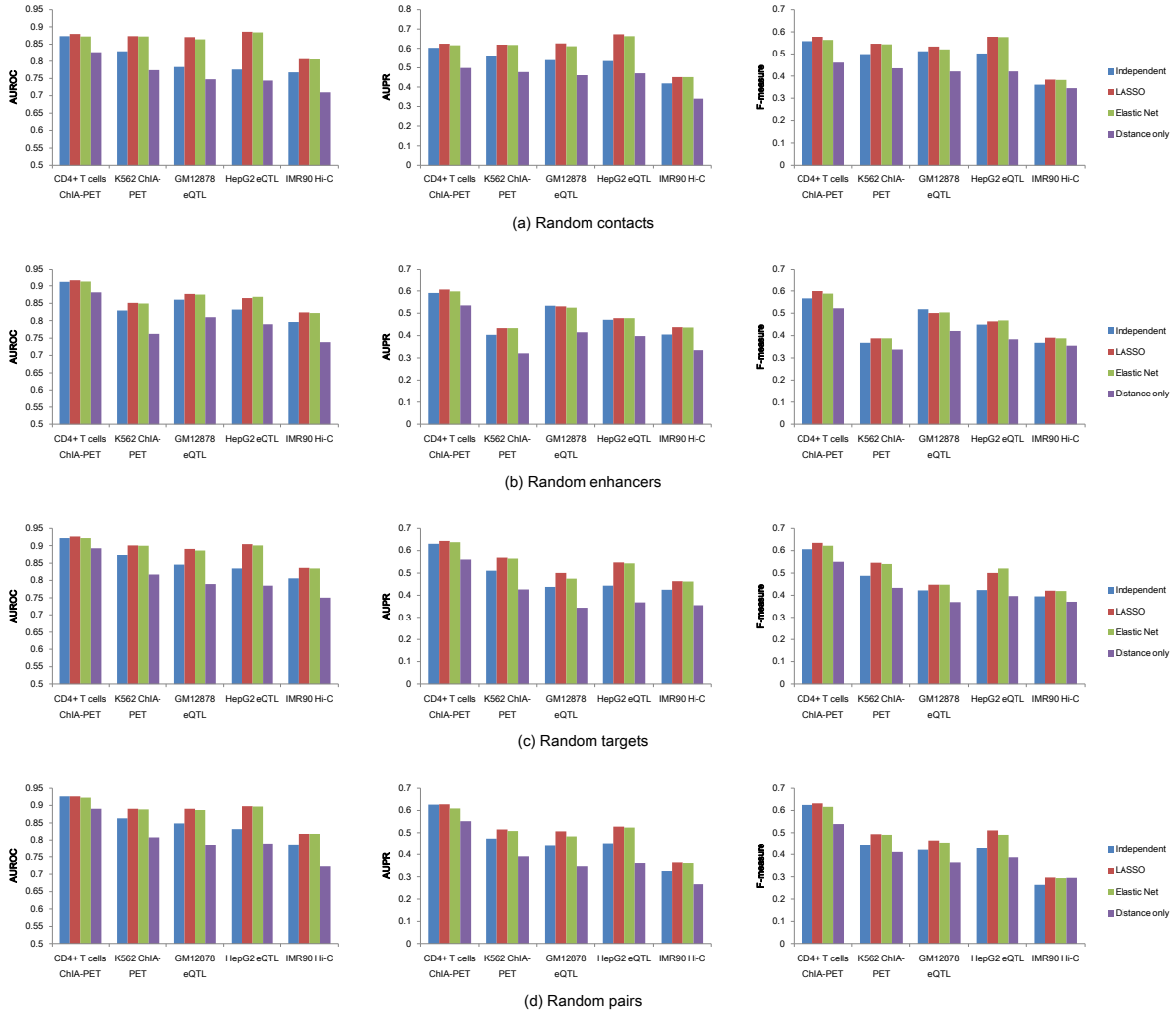


Figure S8: Consistency of the enhancer-target pairs inferred by imputed and non-imputed data from 127 ENCODE+Roadmap samples, using ChIA-PET, eQTL and Hi-C data as reference. The four panels correspond to four different ways to draw negative pairs, namely (a) random contacts, (b) random enhancers, (c) random targets and (d) random pairs. Each panel consists of three sub-figures that correspond to three different consistency measures. "Independent" shows the results of the models that consider each enhancer separately, "LASSO" and "Elastic Net" show the results of the two types of model that consider multiple enhancers jointly, and "Distance only" shows the results when only distance information was used to infer enhancer-target pairs.

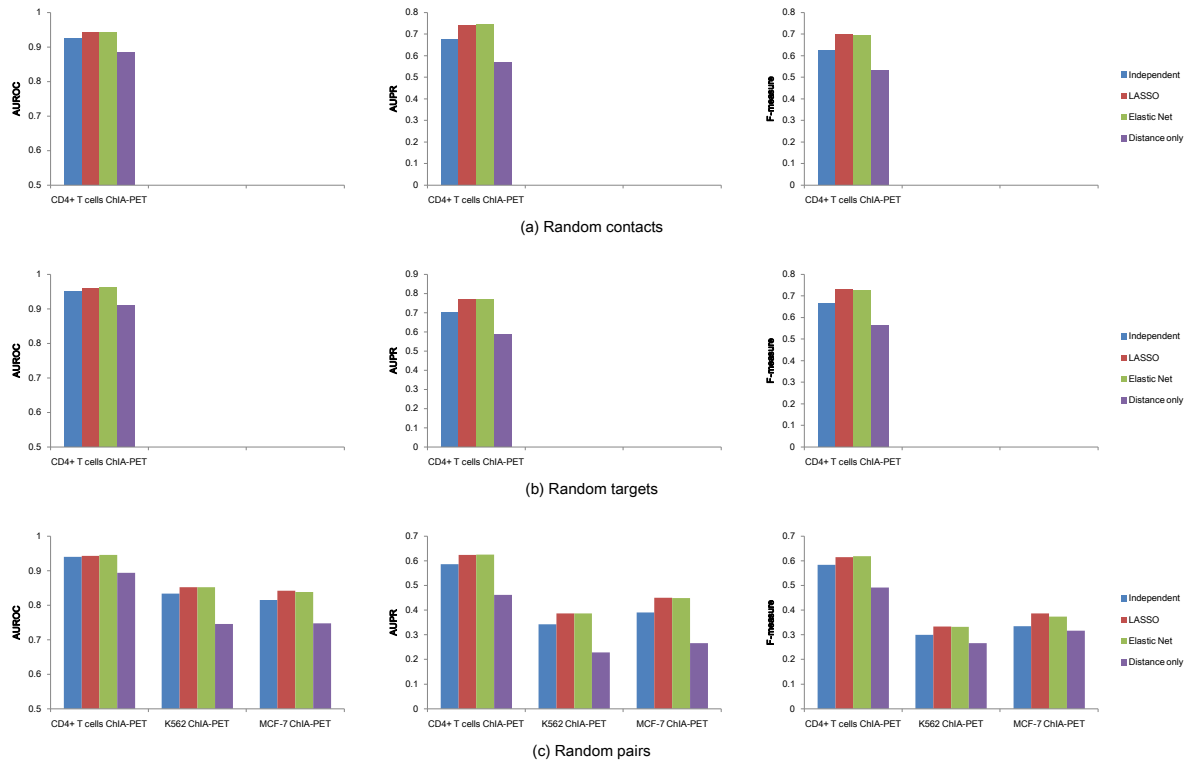


Figure S9: Consistency of the enhancer-target pairs inferred by data from FANTOM5 samples, using ChIA-PET data as reference. The four panels correspond to four different ways to draw negative pairs, namely (a) random contacts, (b) random targets and (c) random pairs. As explained in Materials and Methods, due to the desired positive-to-negative ratios, some of the negative sets could not be formed. Each panel consists of three sub-figures that correspond to three different consistency measures. "Independent" shows the results of the models that consider each enhancer separately, "LASSO" and "Elastic Net" show the results of the two types of model that consider multiple enhancers jointly, and "Distance only" shows the results when only distance information was used to infer enhancer-target pairs.

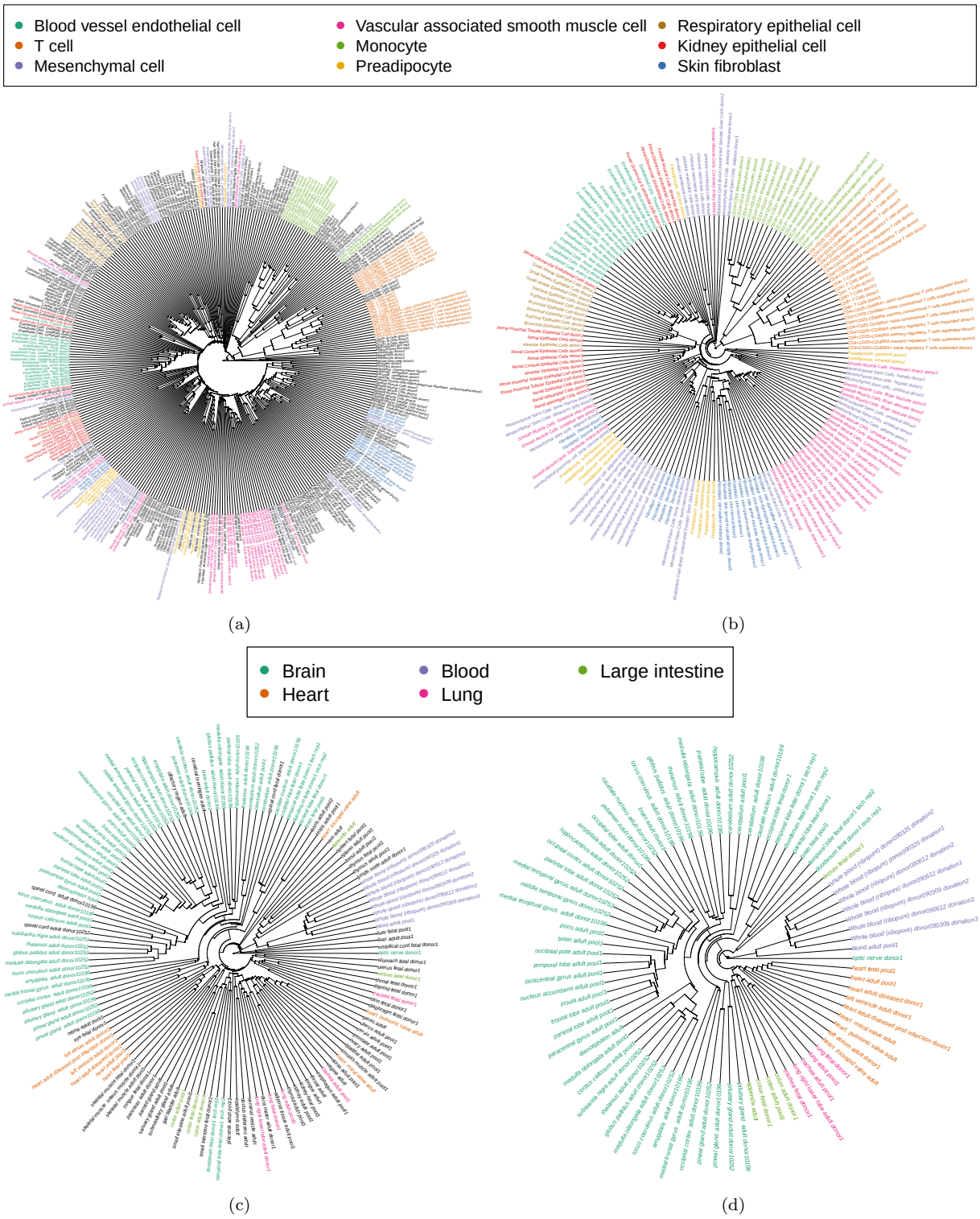
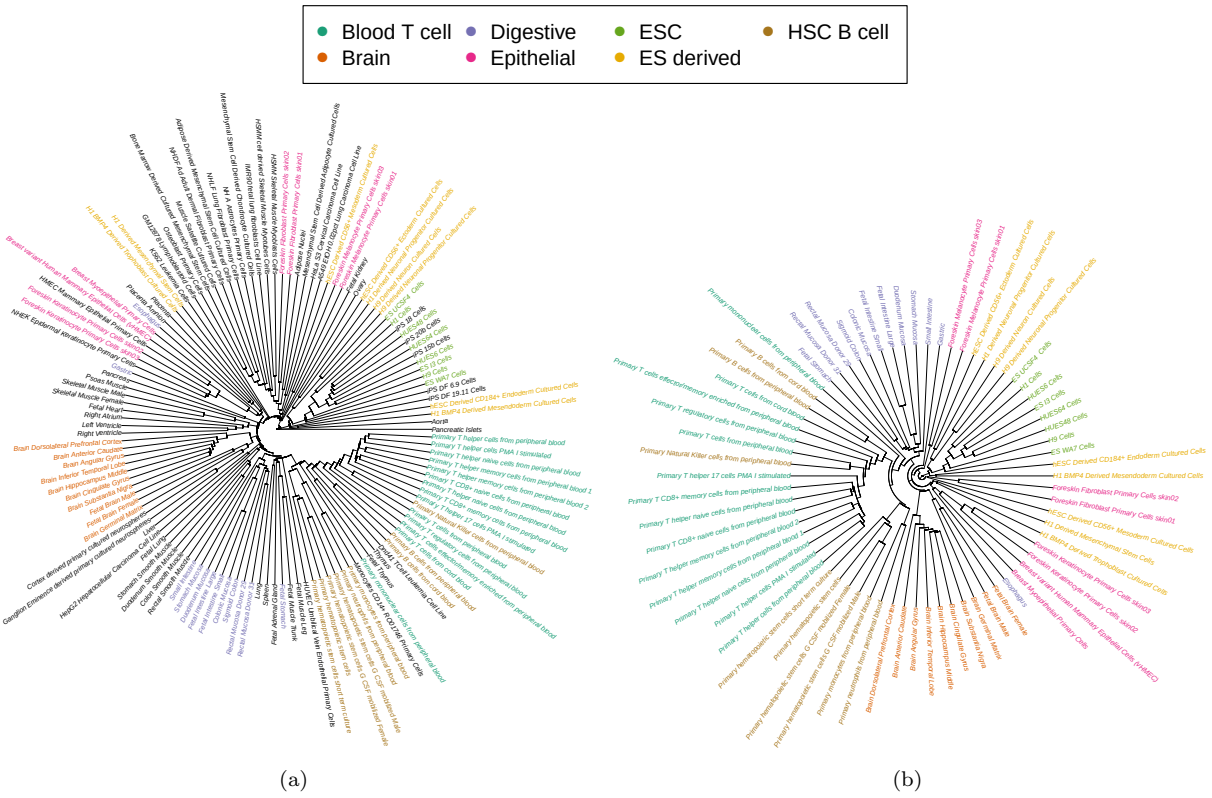
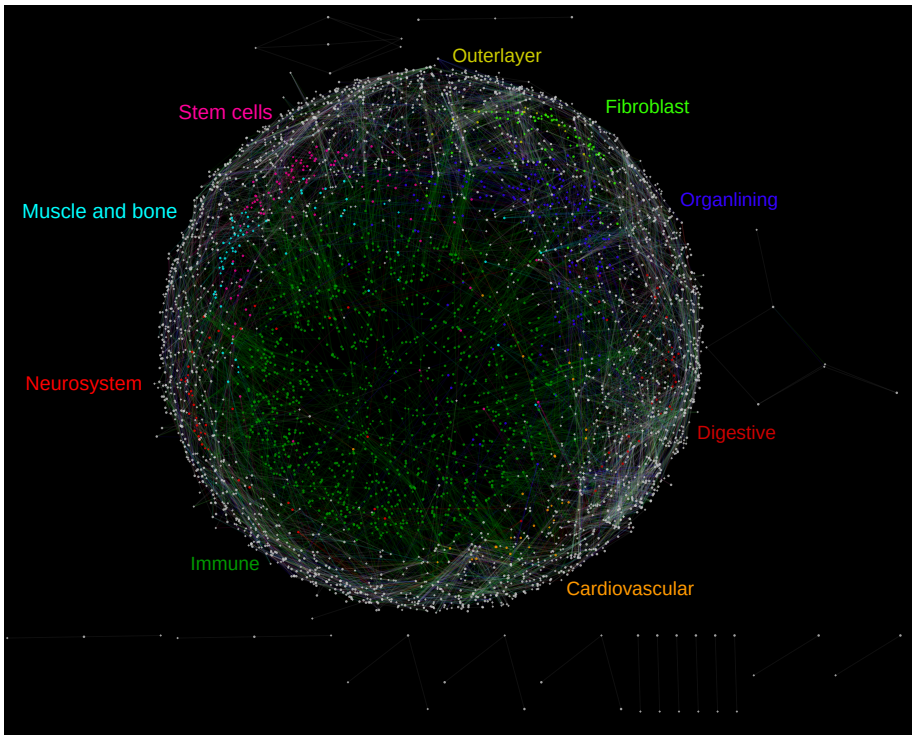


Figure S10: Clustering of samples using the inferred enhancer-target network of each sample as its signature for computing similarity values among samples. The clustering involves (a) All FANTOM5 cell samples; (b) FANTOM5 cell samples of the 9 largest facets; (c) All FANTOM5 tissue samples; and (d) FANTOM5 tissue samples of the 5 largest facets. Samples are colored based on the sample facets provided by FANTOM5.

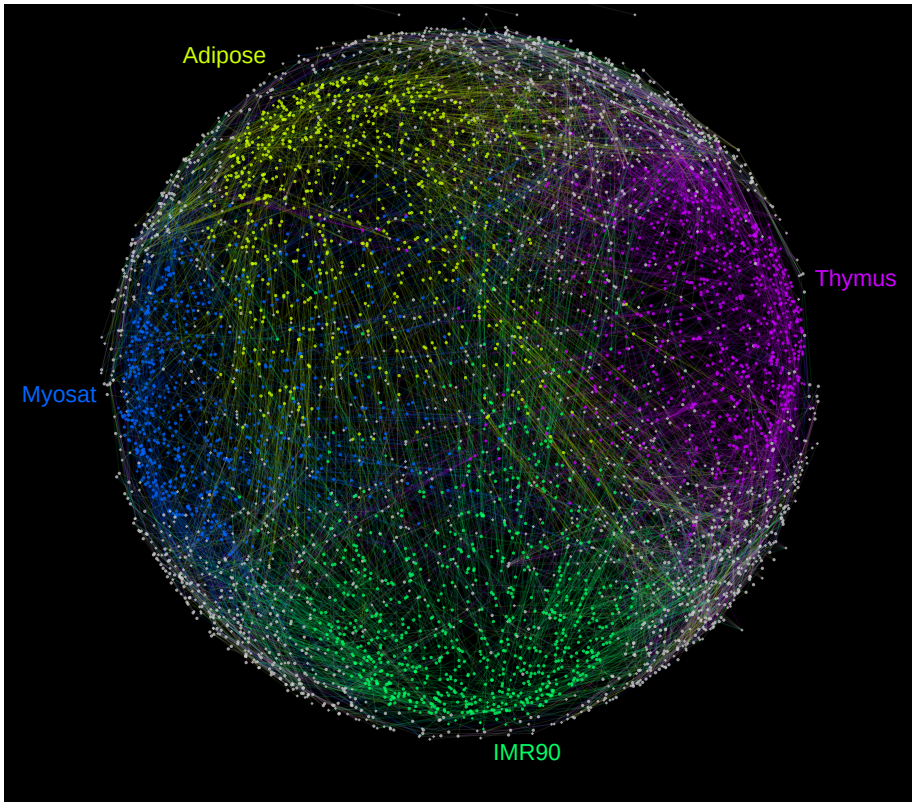


(a) (b)

Figure S11: Clustering of samples using the inferred enhancer-target network of each sample as its signature for computing similarity values among samples. The clustering involves (a) All ENCODE+Roadmap samples; and (b) ENCODE+Roadmap samples of the 7 largest groups. Samples are colored based on the sample groups provided by Roadmap Epigenomics.



(a)



(b)

Figure S12: Subnetworks active in only a single group of (a) FANTOM5 primary cells and (b) ENCODE+Roadmap samples. Enhancers and TSSs are respectively presented as circle and rhombus nodes, while enhancer-target connections are represented as edges. Each color corresponds to enhancers, TSSs and enhancer-target interactions that are active in a single group of samples obtained by grouping the FANTOM5 facets or Roadmap Epigenomics sample groups.