nature genetics

Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines

Qin Cao[1], Christine Anyansi[1,2], Xihao Hu[1], Liangliang Xu[3], Lei Xiong[4], Wenshu Tang[3], Myth T S Mok[3], Chao Cheng[5], Xiaodan Fan[6], Mark Gerstein[7–9], Alfred S L Cheng[3] & Kevin Y Yip[1,10–12]

We propose a new method for determining the target genes of transcriptional enhancers in specific cells and tissues. It combines global trends across many samples and sample-specific information, and considers the joint effect of multiple enhancers. Our method outperforms existing methods when predicting the target genes of enhancers in unseen samples, as evaluated by independent experimental data. Requiring few types of input data, we are able to apply our method to reconstruct the enhancer–target networks in 935 samples of human primary cells, tissues and cell lines, which constitute by far the largest set of enhancer–target networks. The similarity of these networks from different samples closely follows their cell and tissue lineages. We discover three major co-regulation modes of enhancers and find defense-related genes often simultaneously regulated by multiple enhancers bound by different transcription factors. We also identify differentially methylated enhancers in hepatocellular carcinoma (HCC) and experimentally confirm their altered regulation of HCC-related genes.

Enhancers are bound by transcription factors and interact with the promoters of their target genes through DNA looping[1]. Because enhancers can be far away from the genes they regulate and can be either upstream or downstream of them[1,2], determining enhancer targets is difficult. Direct DNA contacts obtained from Hi-C[3], chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[4] or similar experiments would provide valuable information, but thus far these data are only available for a few human cell types[3,5–11]. As a result, previous studies have attempted to computationally infer enhancer targets[12–18]. Limited by the availability of epigenomic and other data required by these methods, they have only been applied to a small number of human cell types.

To study general properties of enhancer-mediated gene regulation, here we reconstruct and analyze active enhancer–target networks in 935 samples, covering a major fraction of human cell and tissue types. The method we employ only requires several types of input data, making it easily applicable to many samples. It combines both global trends across all samples and sample-specific information, and considers the joint effect of multiple enhancers on the same target genes. Our method is more accurate than existing methods in predicting enhancer targets in unseen samples. We are also able to experimentally confirm some of our predictions in HCC samples.

RESULTS

Gene expression is more accurately inferred by considering the joint effect of multiple targeting enhancers

To identify enhancer targets, some previous studies have assumed that the activity of an enhancer can partially explain the expression levels of its target genes[13,19]. Indeed, using data from two human cell lines (**Supplementary Table 1**) with predicted active (**Fig. 1a**) and inactive (**Fig. 1b**) enhancers, as well as ChIA-PET-defined active and inactive enhancer–target interactions (**Fig. 1c**), we found that enhancer features were highly correlated with target gene expression (**Fig. 1d**). These visually apparent correlations are remarkable because the enhancer–target pairs were defined by ChIA-PET alone, without referring to any of these features. Using statistical modeling[20–22], we confirmed that, for these pairs, enhancer activities alone could partially determine the expression level of target genes in both cell lines (**Supplementary Figs. 1–5** and **Supplementary Note**).

On the other hand, because a gene can be regulated by multiple enhancers, considering each enhancer independently could lead to important enhancer–target interactions being missed. For example, if a gene is regulated by several enhancers in disjoint cell types, the activity of each enhancer may only correlate weakly with the expression of the gene across the cell types. To investigate the generality of this

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [2]Department of Computer Science, Vrije Universiteit, Amsterdam, the Netherlands. [3]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [4]Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [5]Department of Biomedical Data Sciences, Dartmouth College, Hanover, New Hampshire, USA. [6]Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [7]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [8]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA. [9]Department of Computer Science, Yale University, New Haven, Connecticut, USA. [10]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [11]CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [12]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. Correspondence should be addressed to K.Y.Y. (kevinyip@cse.cuhk.edu.hk).

Received 19 December 2016; accepted 14 August 2017; published online 04 September 2017; doi:10.1038/ng.3950

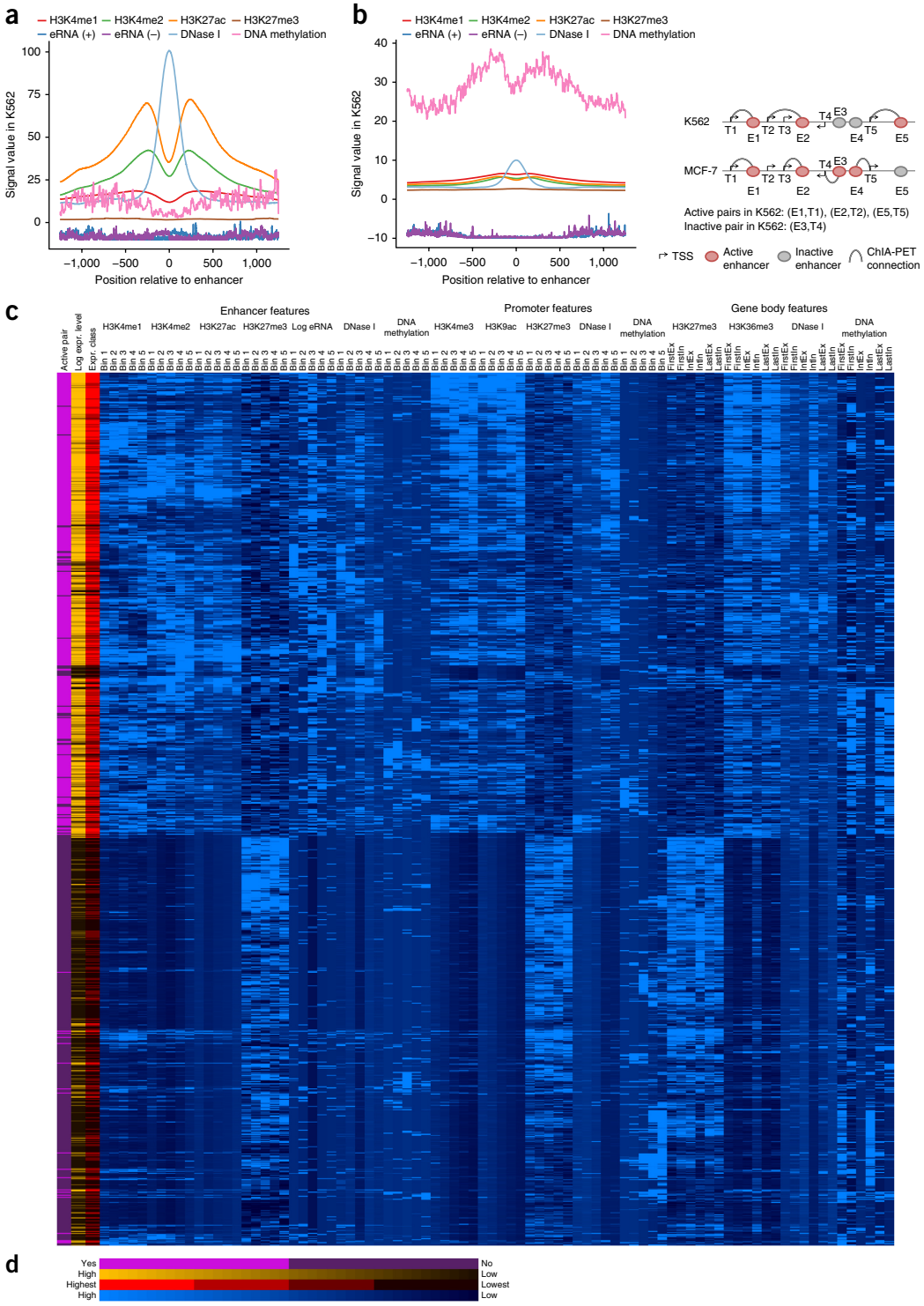

Figure 1 Quantitative relationships between FANTOM5 enhancer and target activities. (**a**,**b**) Aggregation plots of K562 features around enhancers active in K562 cells (**a**) or inactive in K562 cells but active in other cell lines (**b**). Each point along the *x* axis is a 6-bp window for DNA methylation or a single base pair for other features. The *y* axis shows the log-transformed signal on one strand for eRNAs or the average signal for other features. (**c**) Definitions of active and inactive enhancer–target pairs in K562 cells. An inactive pair must have both the enhancer and target not involved in any active pairs. (**d**) Heat map showing correlations between gene expression and enhancer, promoter and gene body features. Each row is an enhancer–target pair. Column 1 shows the activity of the pair in K562 cells. Columns 2 and 3 show the log-transformed transcript level around the TSS of the gene and its expression class, respectively. The remaining columns show enhancer, promoter and gene body features. For each enhancer and promoter feature, signal values in five consecutive 500-bp bins are shown. For each gene body feature, signal values in the first exon (FirstEx), the first intron (FirstIn), the internal exons (IntEx), the internal introns (IntIn), the last exon (LastEx) and the last intron (LastIn) are shown. Each column is standardized to zero mean and unit variance. Rows are ordered by average-link hierarchical clustering using one minus the Pearson correlation as the distance, with the activity of each pair not involved in the clustering.

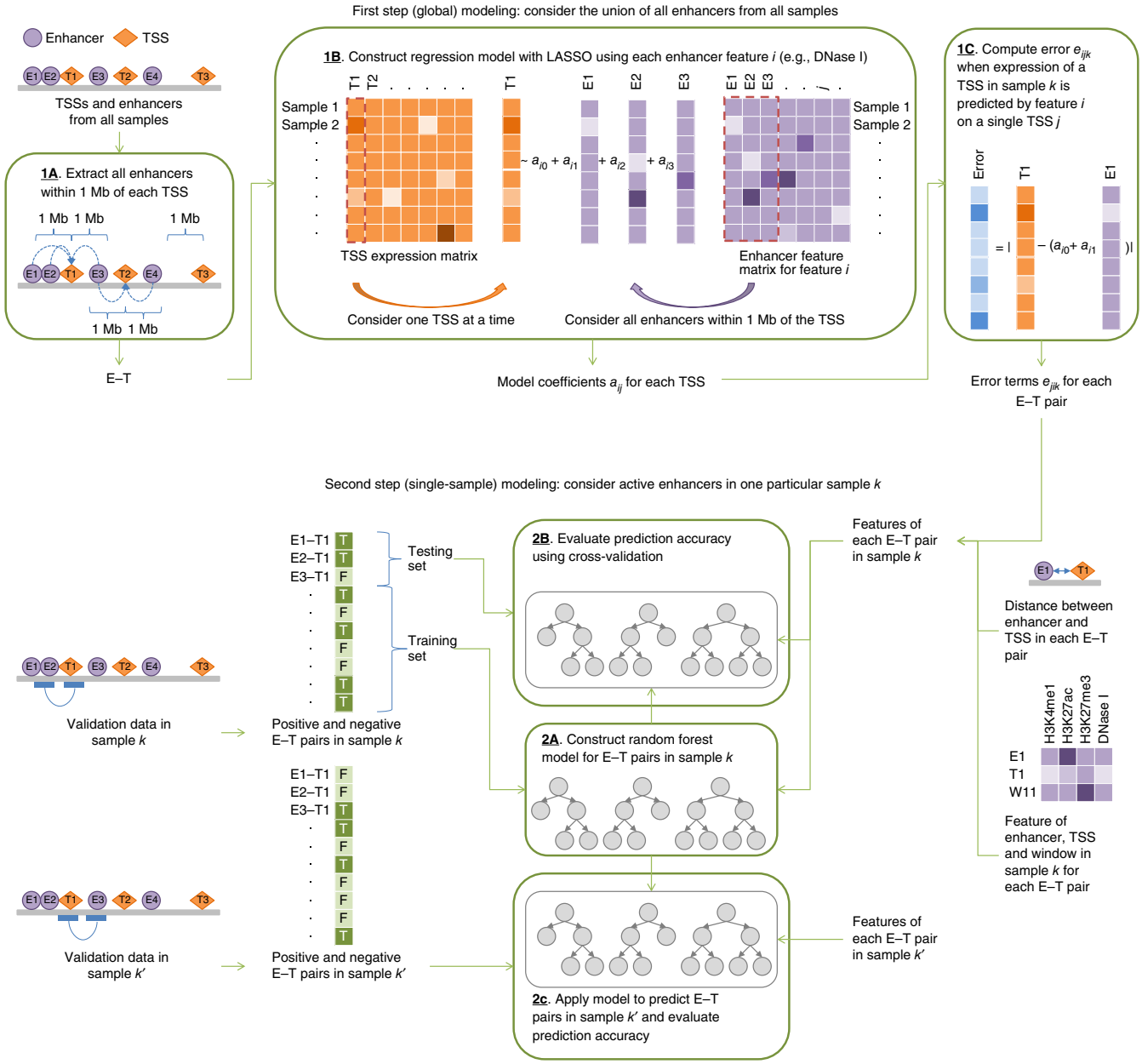

Figure 2 The JEME method. This schematic diagram illustrates the JEME method for determining active enhancer–target (E–T) pairs in each specific sample. Detailed explanations are given in the Online Methods.

issue, we collected enhancer features and gene expression levels for 935 human primary cell types, tissue types and cell lines (which we call 'samples' in general hereafter) from ENCODE + Roadmap Epigenomics (127 samples)[23] and FANTOM5 (808 samples)[24], in which active enhancers were defined by histone modifications and enhancer RNAs (eRNAs), respectively. Because no Hi-C or ChIA-PET data were available for most samples, we considered all enhancers within 1 Mb of each transcription start site (TSS) as potential regulating enhancers and allowed the statistical model to select the most promising ones.

On the basis of left-out TSSs not involved in model training, the models that considered the joint effect of multiple enhancers performed significantly better ($P < 2.2 \times 10^{-16}$ in all cases, Wilcoxon signed-rank test) (**Supplementary Fig. 6**), on both ENCODE + Roadmap and FANTOM5 data sets and irrespective of whether imputed data[25] were included. These results suggest that the joint effect of multiple enhancers should be considered when identifying enhancer targets.

Accurate reconstruction of enhancer–target networks in 935 human cell types, tissue types and cell lines

On the basis of the above findings, we designed JEME (joint effect of multiple enhancers), a new method for determining the target genes of enhancers in specific samples. JEME involves two main steps (**Fig. 2**, Online Methods and **Supplementary Table 2**). In the first step, it identifies enhancers that potentially regulate each TSS on the basis of multiple regression of all enhancers in the genomic neighborhood of a TSS across all samples, without requiring any known

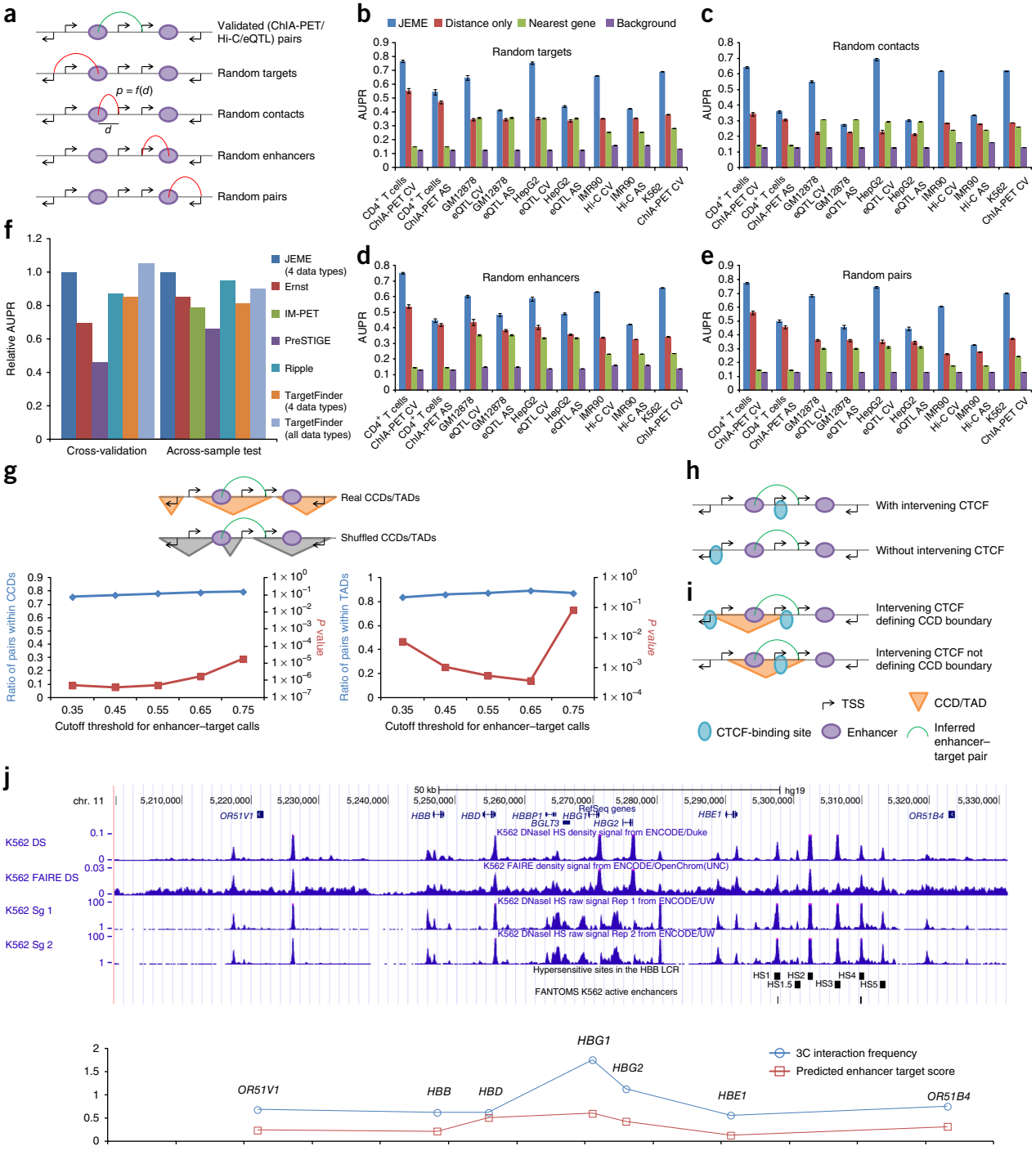

Figure 3 Reliability of the inferred enhancer networks. (**a**) Four types of background enhancer–TSS pairs. *d*, distance; *p*, probability of interaction, which is a function *f* of *d*. (**b**–**e**) Validation of enhancer–target pairs inferred by JEME with different background pairs (random targets (**b**), random contacts (**c**), random enhancers (**d**) and random pairs (**e**), based on 127 ENCODE + Roadmap samples involving imputed data. "Distance only," "Nearest gene" and "Background" correspond to the results, respectively, when the genomic distance between the enhancer and TSS was used as the only feature to train a random forest model, when each enhancer was predicted to regulate the nearest gene only and when predictions were made randomly. AS and CV correspond, respectively, to across-sample prediction (trained in K562 cells and tested on the specified sample) and cross-validation. Error bars show the s.d. of ten random training and testing sets. AUPR, area under the precision-recall curve. (**f**) Relative AUPRs of the enhancer–target prediction methods based on cross-validation results in GM12878 cells and across-sample results with training in K562 cells and testing in GM12878 cells. For JEME and TargetFinder, the number of data types used in predicting enhancer targets (not counting data for predicting the enhancers) is shown. (**g**) Comparison of the fractions of inferred enhancer–target pairs within GM12878 CCDs and IMR90 TADs with the fractions within shuffled domains. (**h**) An enhancer–target pair with and without intervening CTCF binding. (**i**) Intervening CTCF binding defining the CCD boundary or not. (**j**) LCR around the human β-globin locus in K562 cells. Top, locations of the genes, open chromatin signals, HS1–HS5 (ref. 50) and K562 active enhancers defined in FANTOM5 in the LCR. Bottom, predicted enhancer–target scores from JEME based on FANTOM5 data and corresponding 3C signals for the LCR[26].

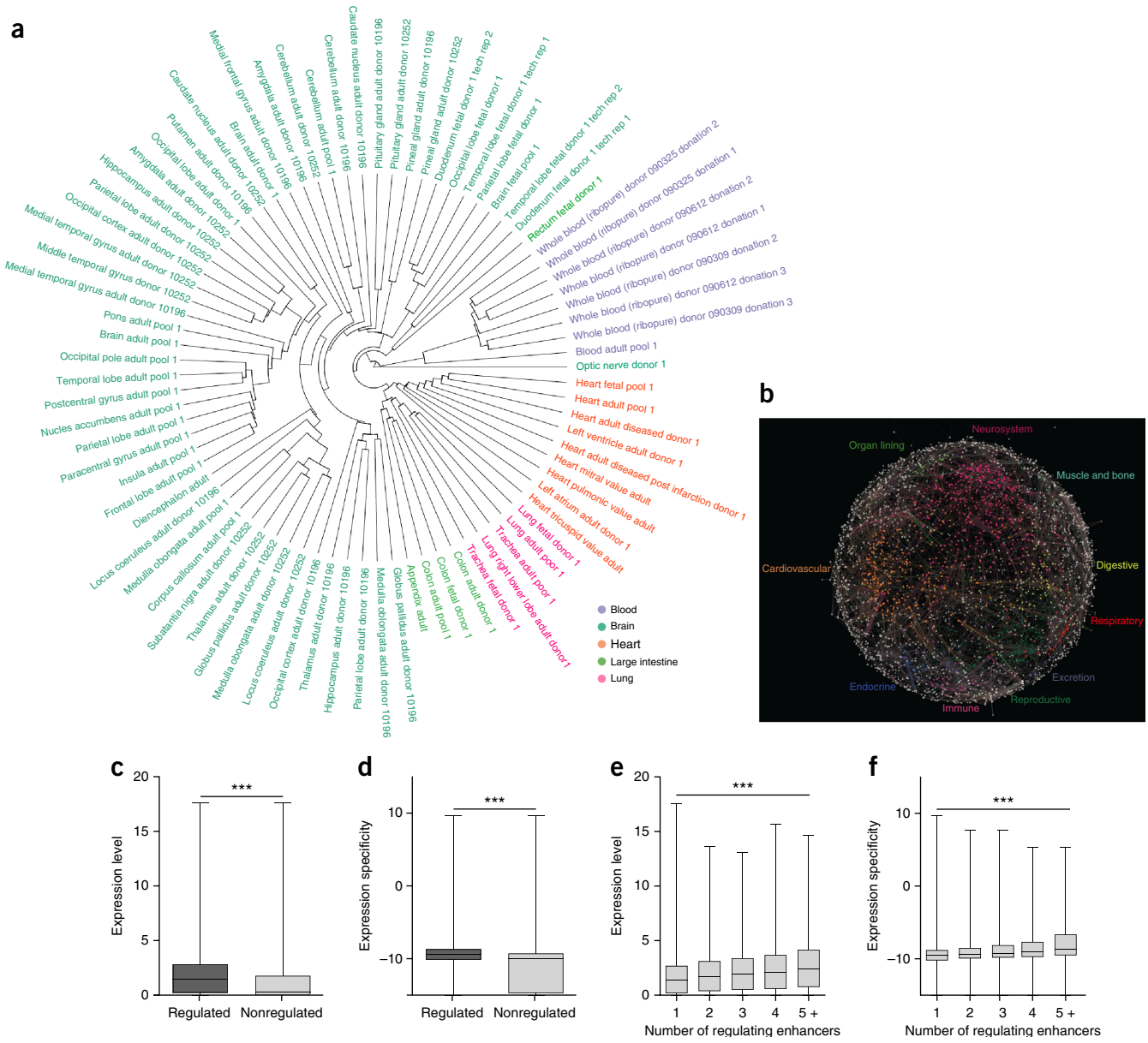

Figure 4 Context specificity of the inferred enhancer networks. (**a**) Clustering of FANTOM5 samples from the largest facets, based on similarities of their enhancer–target networks. (**b**) Subnetworks active in only a single group of FANTOM5 tissues for the enhancers and TSSs on chromosome 1. Enhancers and TSSs are presented, respectively, 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 on the basis of FANTOM5 facets. (**c**,**d**) Expression level (**c**) and sample-specific expression specificity (**d**) of TSSs with or without a regulating enhancer in each FANTOM5 sample. (**e**,**f**) Expression level (**e**) and sample-specific expression specificity (**f**) of TSSs with various numbers of regulating enhancers in each FANTOM5 sample. *P* values in **c** and **d** are based on two-sided Wilcoxon rank-sum tests, and those in **e** and **f** are based on *t* tests of Pearson correlation values: ***$P < 0.0001$. In box plots, the top and bottom bars mark the maximum and minimum; the upper and lower edges of the box mark the 75th and 25th percentiles; and the line within the box marks the median.

enhancer–target pairs as examples. In the second step, sample-specific information is incorporated to identify enhancers that regulate the TSS in each particular sample. To the best of our knowledge, JEME seems to be the first method that simultaneously considers the joint effect of multiple enhancers and incorporates both global and sample-specific information.

We applied JEME to identify enhancer targets in each of the 935 samples and conducted a series of analyses to evaluate the accuracy of JEME. First, we checked the overlaps between the inferred enhancer–target pairs and published ChIA-PET, Hi-C and expression quantitative trait locus (eQTL) data (collectively referred to as the 'validation data') in all samples for which such data were available (**Supplementary Table 3**). We used a rigorous evaluation procedure, applying both cross-validation tests within the same sample and across-sample tests, in which the second step of JEME involved training with ChIA-PET data from K562 cells and the resulting model was applied to a different sample and evaluated using validation data in that sample. In both cases, the validation data used to evaluate the

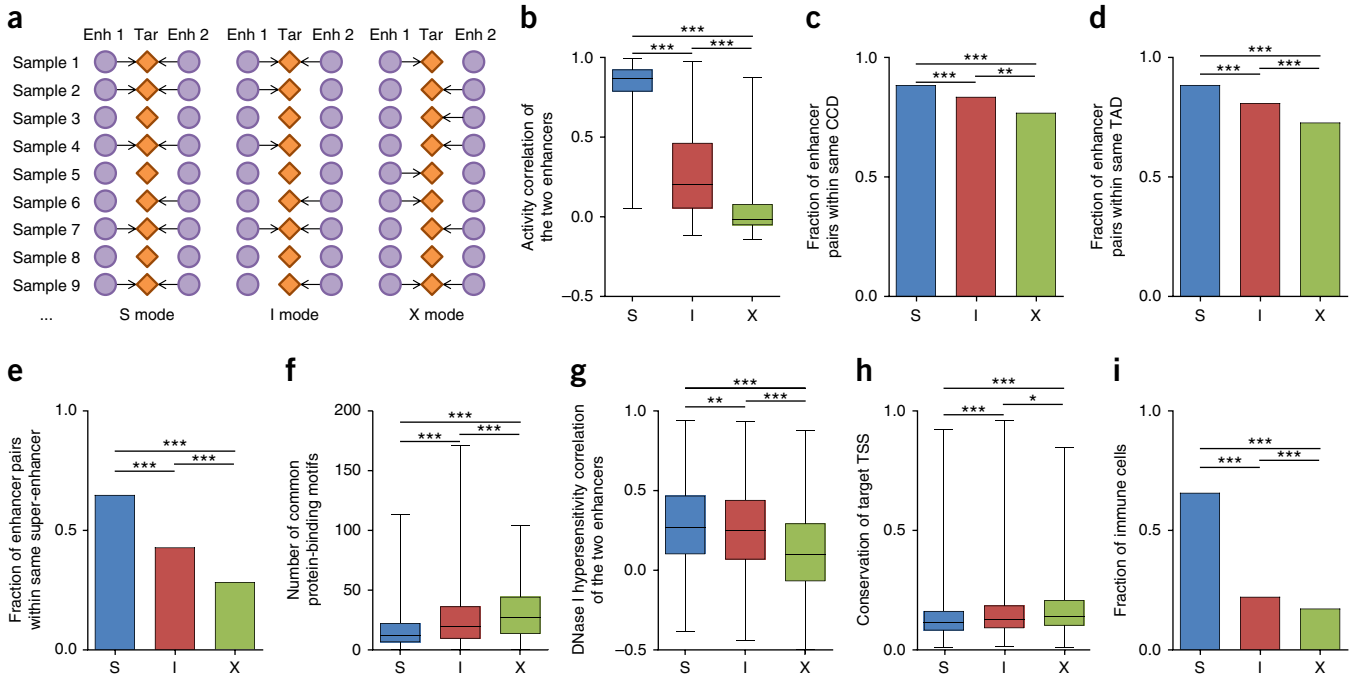

Figure 5 Enhancer co-regulation modes. (**a**) Definitions of the simultaneous (S), independent (I) and mutually exclusive (X) co-regulation modes. (**b**) Correlation between the activities of the two enhancers in each module based on eRNA levels across all samples. (**c**–**e**) Fraction of modules in which the two enhancers reside in the same CCD (**c**), TAD (**d**) or super-enhancer (**e**). (**f**) Number of common transcription factor binding motifs found in partner enhancers. (**g**) Accessibility correlation of partner enhancers based on DNase I hypersensitivity. (**h**) Conservation of target genes based on 100-species phastCons scores. (**i**) Of the samples in which an enhancer in a module regulates the target, the fraction that are immune cells. $*P < 0.05$; $**P < 0.001$; $***P < 0.0001$, two-sided Wilcoxon rank-sum test.

accuracy of JEME were not involved in training the corresponding models. We considered four different ways to sample background enhancer–TSS pairs, taking into account the distance between enhancers and TSSs, the background distribution of contact distances in chromatin conformation data, and the ratio between positive and background pairs (**Fig. 3a** and Online Methods).

In all cases (**Fig. 3b–e** and **Supplementary Figs. 7–9**), the enhancer–target pairs inferred by JEME were supported by the validation data significantly more often than the background pairs, regardless of (i) the evaluation procedure (cross-validation or across-sample tests), (ii) the data set (ENCODE + Roadmap or FANTOM5, with or without imputed data), (iii) the sample involved, (iv) how the background pairs were drawn and (v) the measure used to quantify the enrichment of overlap over the background. We also used genomic distance and nearest gene as baseline methods for predicting enhancer targets and found that JEME was much more accurate than these baselines (**Fig. 3b–e** and **Supplementary Figs. 7–9**).

Second, we compared JEME with other state-of-the-art methods for predicting enhancer targets, including Ernst[13], IM-PET[14], PreSTIGE[12], Ripple[15] and TargetFinder[17]. On the basis of exactly the same data sets, JEME was the most accurate of all the methods in across-sample tests (**Fig. 3f** and **Supplementary Fig. 10**). In cross-validation tests, JEME was slightly less accurate than TargetFinder when TargetFinder was allowed to use all (70–100) input data types, but not when it used only the top 4–16 data types. Because JEME requires only four types of input data for enhancer–target predictions when an input set of enhancers is given, it can be easily applied to many samples.

Third, studies have shown that enhancer–target connections usually occur within topological domains in 3D genome structures[5–7,10,11]. We found that, indeed, a large fraction of the enhancer–target pairs inferred by JEME were contained within topologically associating domains (TADs)[5] and chromatin contact domains (CCDs)[11] (**Fig. 3g**).

Fourth, given the importance of CTCF in binding CCD boundaries and insulators, we expected a depletion of intervening CTCF binding between enhancers and their target TSSs (**Fig. 3h**). We found that this was the case for all 14 samples for which CTCF ChIP–seq data were available, with significantly less intervening CTCF binding between JEME-inferred enhancers and targets than random enhancer–TSS pairs with the same distance distribution ($P < 2.2 \times 10^{-16}$ for all 14 cases, Z test; **Supplementary Table 4**). For the intervening CTCF sites, we also found that these were significantly less involved in defining CCD boundaries ($P < 1.0 \times 10^{-14}$, Fisher's exact test; **Fig. 3i** and **Supplementary Table 5**).

Finally, we examined the well-studied enhancer–target connections between human β-globin genes and their locus control region (LCR). In K562 cells, the interactions between the LCR and both the β-globin genes and the flanking olfactory receptor genes have been quantified by chromosome conformation capture (3C)[26]. We compared the 3C interaction frequencies with the predicted enhancer–TSS interaction scores from JEME and found that they were significantly correlated (Pearson $r = 0.75$; $P = 0.027$, t test) (**Fig. 3j**). Interestingly, JEME predicted the LCR to interact with *HBG1* with the highest score, which is consistent with the 3C results even though *HBG1* is not the gene closest to the LCR.

Collectively, these results indicate that JEME reliably inferred the active enhancer–target networks in the specific samples.

Basic properties and context specificity of the enhancer networks

Basic statistics for the networks were largely consistent with previous studies (**Supplementary Fig. 11** and **Supplementary Note**), with the

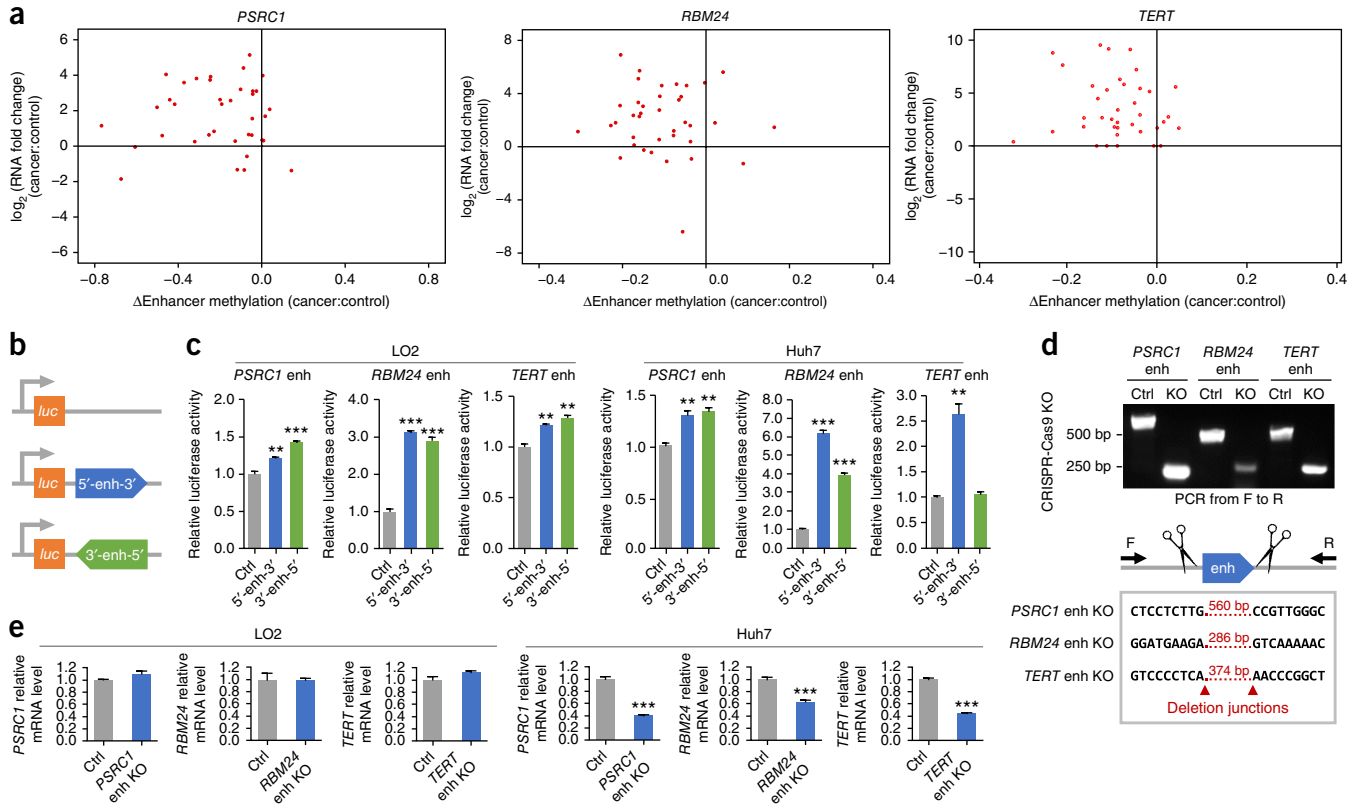

Figure 6 HCC cancer genes identified by the liver-related enhancer–target networks. (**a**) Differential enhancer methylation and differential target gene expression of the target genes *PSRC1*, *RBM24* and *TERT*. A positive value corresponds to a stronger methylation or expression level in the tumor sample. (**b**) Schematic diagram showing the pGL3-Promoter vector with or without insertion of a gene enhancer in the forward or reverse orientation. (**c**) Luciferase activity in the presence or absence of gene enhancers was examined in Huh7 and LO2 cells. Ctrl, control without enhancer. (**d**) Representative genotyping results showing CRISPR–Cas9-mediated knockout (KO) of gene enhancers and the flanking sequences at deletion loci. (**e**) The effects of enhancer deletions on the transcription of associated genes were examined in Huh7 and LO2 cells. In **c** and **e**, data represent the mean ± s.d. of at least three independent experiments, each performed in triplicate, and are presented relative to control. $^{**}P < 0.01$; $^{***}P < 0.001$, two-sided Wilcoxon rank-sum test.

small discrepancies likely due to the much larger number of samples involved in our study and the relatively small number of enhancers called by FANTOM5 per sample.

The hundreds of enhancer–target networks inferred by JEME enabled us to study the context specificity of enhancer targeting in general. First, we clustered samples on the basis of the similarity of their networks and found that biologically related samples formed clear clusters (**Fig. 4a** and **Supplementary Figs. 12** and **13**). For example, all the blood samples formed a cluster completely separated from the other samples—as did the heart, lung, and all but two large intestine and brain samples—suggesting that enhancer–target networks serve as distinctive signatures of cell and tissue lineages.

Next, we identified enhancers and TSSs specifically active in a single group of related samples (**Supplementary Tables 6–8**) and found that their regulatory relationships were highly specific to these samples (**Fig. 4b**). For FANTOM5 samples, of the genes specifically expressed in a sample group, between 46.8% (respiratory) and 80.9% (reproductive) were regulated by an enhancer that had regulatory activity only in this group of samples ($P < 2.2 \times 10^{-16}$ for all facets, Fisher's exact test). Likewise, of the enhancers that had regulatory activity specifically in a sample group, between 16.9% (respiratory) and 45.8% (neurosystem) regulated a TSS that was regulated only in this group of samples ($P < 2.2 \times 10^{-16}$ for all facets, Fisher's exact test).

As expected, we found that, when a TSS was regulated by an enhancer in a sample, it was both more highly expressed (**Fig. 4c**) and more specifically expressed (**Fig. 4d**) in this sample than other TSSs. Intriguingly, we also found that TSSs regulated by more enhancers in a sample generally had higher expression levels (**Fig. 4e**) and stronger expression specificity (**Fig. 4f**) in this sample.

Multiple enhancers co-regulate common targets in three major modes

In the inferred networks, two enhancers could regulate a common TSS in the same or different samples, forming a regulatory module involving the two enhancers and the TSS. We defined three co-regulation modes for these modules (**Fig. 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. As expected, the activity correlations of the partner enhancers across all samples were highest in the simultaneous mode and lowest in the mutually exclusive mode (**Fig. 5b**).

In the simultaneous mode, the partner enhancers were more frequently located within the same topological domain (**Fig. 5c**,**d**), consistent with the compartmentalization role of these domains[5,6,10,11]. Partner enhancers in the simultaneous mode were also more likely to reside in the same super-enhancer (**Fig. 5e**), in agreement with

the definition of super-enhancers as large regulatory units working in concert[27,28].

Interestingly, we found that partner enhancers in the simultaneous mode shared fewer transcription factor binding motifs (**Fig. 5f**). We hypothesize that two enhancers could complement each other more effectively if they differ from one another by at least one attribute. In the simultaneous mode, in which partner enhancers tend to regulate a common target in the same context, having different motifs increases the chance that at least one of the enhancers can function when only some of the transcription factors are expressed. In contrast, if the partner enhancers contain the same transcription factor binding motifs, they could complement each other by being active in different contexts. This is in agreement with the observation that partner enhancers in the mutually exclusive mode had significantly lower correlations of DNA accessibility than those in the simultaneous mode (**Fig. 5g**).

We also found that the target genes in the simultaneous mode were the least evolutionarily conserved (**Fig. 5h**), suggesting that this co-regulation mode is used by more adaptive functions. Supporting this hypothesis, only genes regulated by the simultaneous mode were enriched in functional terms related to defense mechanisms, including 'inflammatory response' (Benjamini–Hochberg (BH)-corrected $P = 3.5 \times 10^{-5}$, EASE score), 'cytokine–cytokine receptor interaction' (BH-corrected $P = 3.5 \times 10^{-5}$) and 'immunity' (BH-corrected $P = 1.0 \times 10^{-4}$). Correspondingly, we found that enhancers in the simultaneous mode regulated their targets significantly more frequently in immune cells than enhancers in the other two modes (**Fig. 5i**).

Identification of cancer-related genes using the inferred networks

As an application, we used our inferred enhancer–target networks to identify differentially expressed genes in HCC potentially caused by aberrant enhancer activities (**Supplementary Fig. 14**). Some previous studies have suggested that differential methylation at enhancers could be associated with inverse differential expression of their target genes[29–31]. We collected methylome and transcriptome data from 38 pairs of matched HCC tumor–control samples from The Cancer Genome Atlas (TCGA) (**Supplementary Table 9**). Among the active enhancers in FANTOM5 liver-related normal and cancer samples (**Supplementary Table 10**), we identified 172 with significant differential methylation between the tumor and control samples. We then looked up the target genes of these enhancers in our inferred liver-related networks and found seven of them to exhibit inverse differential expression (**Supplementary Table 11**). The genomic distance between the enhancer and the TSS ranged from 7.5 kb to 187 kb.

We selected three genes for further experimental validations, namely *TERT*, *PSRC1* and *RBM24* (**Fig. 6a**). *TERT* is well known for its promoter mutations in cancer[32–34], which are also early events in the tumor progression of some human HCCs[35]. Similarly to previous studies[36], we found that *TERT* expression was substantially higher in tumor samples. *PSRC1* is a downstream target of TP53 (ref. 37) involved in activation of the β-catenin pathway in mouse[38]. It is involved in TP53 signaling and is considered to be a potential oncogene in HCC[39]. *RBM24* encodes an RNA-binding protein and is also a downstream target of TP53 (ref. 40).

We first confirmed the activities of the three enhancers in both immortalized, non-tumorigenic hepatocyte (LO2) and HCC (Huh7) cell lines. We performed luciferase reporter assays by inserting the enhancers next to a *luc* (luciferase) gene and measuring the change in luciferase activity (**Fig. 6b**). All three enhancers stimulated *luc* transcription in both cell lines (**Fig. 6c**), with the *RBM24* enhancer exerting the strongest induction in both orientations. Next, we examined the functional roles of these enhancers *in vivo*. Using CRISPR–Cas9 genome editing, we successfully deleted the three enhancers in the two cell lines (**Fig. 6d**). The transcriptional levels of all three genes were consistently diminished after deletion of their respective enhancers in Huh7 cells but not in LO2 cells (**Fig. 6e**), consistent with the overexpression of these genes in TCGA HCC samples. To investigate whether these enhancers are under epigenetic control, we treated seven liver cell lines with the DNA demethylation agent 5-aza-2′-deoxycytidine (5-aza-dC). All three genes were reactivated by 5-aza-dC in at least two cell lines (**Supplementary Fig. 15a**). Using bisulfite sequencing, we further found that reactivation of *PSRC1* and *RBM24* was associated with demethylation of their enhancers but not their promoters, whereas *TERT* reactivation could result from demethylation of both the promoter and enhancer (**Supplementary Fig. 15b**). These results further support the notion that these three genes are regulated by their respective enhancers and that regulation could be modulated epigenetically.

DISCUSSION

The encouraging performance of JEME in predicting enhancer–target interactions was attributed to its novel combination of both global trends and sample-specific information, and consideration of the joint effect of multiple enhancers. The increased accuracy was most apparent in across-sample tests, which constitute the more important setting, as most cell types do not currently have ChIA-PET or Hi-C data available. In addition, each ChIA-PET data set investigates only DNA contacts related to one particular factor owing to the use of immunoprecipitation, while Hi-C data contain a lot of DNA contacts unrelated to enhancer–target interactions. Even with modified experimental protocols for specifically probing enhancer–target interactions[41,42] and the increasing popularity of using CRISPR–Cas9 to identify functional elements and their targets[43,44], the low requirements for input data (eRNAs or several types of epigenomic signals) still make JEME an inexpensive method for predicting enhancer targets in a large number of cell and tissue types and for cross-checking the interactions identified from these experimental methods by integrating additional information from other types of data and data from other samples.

The 935 enhancer–target networks inferred (provided on our supplementary website; see URLs) can be used to systematically study gene regulation in both normal and disease states on a large scale. In particular, they can identify genes potentially affected by genetically or epigenetically perturbed enhancers from disease studies, including enhancers that become either more active or less active in disease-related samples. If the cell or tissue types most relevant to the disease are not covered by the ENCODE + Roadmap and FANTOM5 samples or a more comprehensive set of human enhancers becomes available, the prediction models of JEME can be easily retrained and applied to predict enhancer targets in the study samples using the software we provide on our supplementary website, if the input data (eRNAs or four types of epigenomic data) are available for these cell and tissue types.

Along this line, we have identified dysregulated genes in HCC likely caused by epigenetic changes at their enhancers. Currently, most cataloged cancer-related genes are identified on the basis of recurrent coding-region mutations or differential expression in cancer. Few of them have well-documented aberrant regulatory regions. *TERT* was one notable exception, with extensive study of its recurrent promoter mutation in many types of cancer[32–35,45–49]. Our results suggest that enhancer methylation could be another mechanism regulating *TERT* expression. Further experimental validations are needed, and the role of *TERT* enhancer methylation in HCC pathogenesis is to be investigated.

URLs. Supplementary website, http://yiplab.cse.cuhk.edu.hk/jeme/; source code for JEME in GitHub, https://github.com/yiplabcuhk/JEME/; Roadmap Epigenomics metadata, http://egg2.wustl.edu/roadmap/web_portal/meta.html; CAGE data from FANTOM5, 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/, http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.tissue.hCAGE/; HOCOMOCO, http://autosome.ru/HOCOMOCO/; DNase I sensitivity data across 97 ENCODE + Roadmap samples, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase/; CCTop: CRISPR–Cas9 target online predictor, http://crispr.cos.uni-heidelberg.de/; TCGA Research Network, http://cancergenome.nih.gov/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank Y. Ruan and Z. Tang for providing the list of CCDs in the GM12878 cell line and W.-L. Chan, J. Chen, M. Gu, S. Hu, X. Hu, X. Ma and B. Zou for helpful discussions. The data for patients with HCC were generated by the TCGA Research Network (see URLs). This project is supported by HKSAR RGC TRS T12-401/13-R, T12-402/13-N and T12C-714/14-R, CRF C4017-14G, GRF 14145916, and grants 3132964 and 3132821 from the Research Committee of CUHK.

AUTHOR CONTRIBUTIONS

K.Y.Y. conceived the study. Q.C. and K.Y.Y. developed the JEME method. Q.C., C.A., X.H., M.T.S.M., C.C., X.F., M.G., A.S.L.C. and K.Y.Y. analyzed the data. L. Xu, L. Xiong, W.T. and M.T.S.M. performed the molecular experiments. Q.C., C.A., M.T.S.M. and K.Y.Y. prepared the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
2. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
3. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
4. Fullwood, M.J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
5. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
6. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).
7. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
8. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).
9. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
10. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
11. Tang, Z. *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
12. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).
13. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
14. He, B., Chen, C., Teng, L. & Tan, K. HeB. Global view of enhancer–promoter interactome in human cells. *Proc. Natl. Acad. Sci. USA* **111**, E2191–E2199 (2014).
15. Roy, S. *et al.* A predictive modeling approach for cell line–specific long-range regulatory interactions. *Nucleic Acids Res.* **43**, 8694–8712 (2015).
16. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
17. Whalen, S., Truty, R.M. & Pollard, K.S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
18. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).
19. Yip, K.Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
20. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22**, 1658–1667 (2012).
21. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
22. Lou, S. *et al.* Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol.* **15**, 408 (2014).
23. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
24. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
25. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).
26. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
27. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
28. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
29. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
30. Heyn, H. *et al.* Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* **17**, 11 (2016).
31. Yao, L., Shen, H., Laird, P.W., Farnham, P.J. & Berman, B.P. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 105 (2015).
32. Bojesen, S.E. *et al.* Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371–384 (2013).
33. Horn, S. *et al. TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
34. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
35. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
36. Shay, J.W., Zou, Y., Hiyama, E. & Wright, W.E. Telomerase and cancer. *Hum. Mol. Genet.* **10**, 677–685 (2001).
37. Lo, P.K. *et al.* Identification of a novel mouse p53 target gene DDA3. *Oncogene* **18**, 7765–7774 (1999).
38. Hsieh, P.-C. *et al.* p53 downstream target DDA3 is a novel microtubule-associated protein that interacts with end-binding protein EB3 and activates β-catenin pathway. *Oncogene* **26**, 4928–4940 (2007).
39. Yang, J.D. *et al.* Genes associated with recurrence of hepatocellular carcinoma: integrated analysis by gene expression and methylation profiling. *J. Korean Med. Sci.* **26**, 1428–1438 (2011).
40. Jiang, Y. *et al.* Rbm24, an RNA-binding protein and a target of p53, regulates p21 expression via mRNA stability. *J. Biol. Chem.* **289**, 3164–3175 (2014).
41. Javierre, B.M. *et al.* Lineage-specific genome architecture links enhancers and non-coding dsease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
42. Pellacani, D. *et al.* Analysis of normal human mammary epigenomes reveals cell-specific active enhancer states and associated transcription factor networks. *Cell Rep.* **17**, 2060–2074 (2016).
43. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR–Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
44. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
45. Killela, P.J. *et al. TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. USA* **110**, 6021–6026 (2013).
46. Landa, I. *et al.* Frequent somatic *TERT* promoter mutations in thyroid cancer: higher prevalence in advanced forms of the disease. *J. Clin. Endocrinol. Metab.* **98**, E1562–E1566 (2013).
47. Liu, X. *et al.* Highly prevalent *TERT* promoter mutations in aggressive thyroid cancers. *Endocr. Relat. Cancer* **20**, 603–610 (2013).
48. Rachakonda, P.S. *et al. TERT* promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc. Natl. Acad. Sci. USA* **110**, 17426–17431 (2013).
49. Vinagre, J. *et al.* Frequency of *TERT* promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).
50. Thakore, P.I. *et al.* Highly specific epigenome editing by CRISPR–Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

ONLINE METHODS

Human reference genome. All genomic locations were based on the human reference genome hg19. TAD locations were originally provided in hg18 and were converted to hg19 using the liftover tool of the UCSC Genome Browser[51].

Collection and processing of data sets for studying the joint effect of multiple enhancers on a target TSS. We collected ChIP–seq data for H3K4me1, H3K27ac and H3K27me3, DNase–seq data 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, replicate-combined signal values were downloaded from the Roadmap Epigenomics website in .bigwig format. Some of these ChIP–seq, DNase–seq and RNA-seq data were imputed from other data files[25]. The full list of samples can be found on the Roadmap Epigenomics metadata page (see URLs). Non-imputed data for 48 samples were available for the three histone modifications and RNA-seq.

We also downloaded ChromHMM-predicted active enhancers (from states 6 (genic enhancers), 7 (enhancers) and 12 (bivalent enhancers) in the core 15-state model) for each of the 127 samples from the same site. We took the union of the predicted enhancers from all samples, removed those larger than 2,500 bp, merged the remaining enhancers that overlapped and removed the ones larger than 2,500 bp again after merging. Finally, we obtained a list of 489,581 enhancers.

For each of these enhancers, we computed the average H3K4me1, H3K27ac, H3K27me3 and DNase–seq signal in each of the 127 samples on the basis of the imputed data. For each TSS of each annotated protein-coding gene in GENCODE version 19 (we used GENCODE with ENCODE + Roadmap data because it was the default gene annotation set), we also computed the average RNA-seq signal in each of the 127 samples on the basis of the imputed data. This whole set of data involving all 127 samples is denoted as 'ENCODE + Roadmap 127, 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 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, non-non'. We also used the same 48 samples to construct two additional data sets, namely one involving imputed data for both ChIP–seq and RNA-seq (denoted as 'ENCODE + Roadmap 48, imp-imp') and a data set involving only imputed data for ChIP–seq but not for RNA-seq (denoted as 'ENCODE + Roadmap 48, imp-non').

In addition, we downloaded CAGE data from the FANTOM5 website for 808 samples (see URLs) and normalized the data as described in Andersson *et al.*[24]. The predicted active enhancers in each of these 808 samples and their processed CAGE signals were also downloaded from the FANTOM5 website. 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 flanking regions of the TSS (from 500 bp upstream to 500 bp downstream of it) of each RefSeq protein-coding gene, as in Andersson *et al.*[24].

The 363 samples of primary cells can be grouped into 69 facets as suggested by the FANTOM5 Consortium (Table 10 in the supplement of Andersson *et al.*[24]). When evaluating the accuracy of the expression models using the leave-one-facet-out procedure (described below), we used only the 363 samples in these 69 facets.

Among the FANTOM5 samples, some are listed as misidentified cell lines by the International Cell Line Authentication Committee, including AZ521, HEp-2, SK-N-MC, SKW-3 and TCO-1. We did not experimentally test whether these FANTOM5 cell lines were contaminated, but, because we only used them in our modeling and analyses without studying their individual phenotypic properties (for instance, we did not include Hep-2 cells in the HCC analysis; **Supplementary Table 10**), the inclusion of them does not affect our findings and conclusions.

Statistical tests. In all the sections below involving statistical tests, except when the testing procedure is stated explicitly, whenever two distributions of continuous values were compared, the non-parametric two-sided Wilcoxon rank-sum test was used. Student's *t* test was used for evaluating the statistical significance of Pearson correlations. Fisher's exact test was used for evaluating the significance of the dependence of two binary variables based on a 2 × 2 contingency table.

The JEME method for identifying enhancer targets in specific samples. JEME involves two main steps, namely a first step for finding potential enhancers of each TSS based on data from all samples and a second step for determining the enhancers that actively regulate each TSS in a specific sample. The main idea of the first step is that enhancers that regulate a TSS in different samples should together be able to explain a good portion of the expression level of the TSS based on their activity signals. The second step further integrates these modeling results with sample-specific information to identify the enhancers most likely regulating a TSS in a specific sample. The high-level procedures involved are as follows (**Fig. 2**):

Step 1A. Identify all enhancers within 1 Mb of each TSS as its candidate regulating enhancers.

Step 1B. For each TSS, form a regression model that predicts its expression by one type of activity signal of its candidate regulating enhancers.

Step 1C. On the basis of the regression models, determine the error of each simplified model that involves only one type of activity signal of one candidate enhancer in predicting the expression of a TSS in each sample.

Step 2A. To predict the regulating enhancers of a TSS in a particular sample, construct a random forest model based on the corresponding error terms and sample-specific features in this sample, and the distance between the TSS and its candidate regulating enhancers. We used random forest to capture any nonlinear relationships between the features. We also tried gradient-boosting trees, but the performance in across-sample tests was not very good (data not shown).

Step 2B. Perform cross-validation to evaluate the accuracy of the predictions for a sample if validation data are available for this sample.

Step 2C. Apply the random forest model constructed based on data for a sample to predict enhancer targets in another sample and evaluate these predictions using validation data from that other sample if available.

Specifically, in step 1B, for each enhancer feature i, the expression level y of a TSS is modeled as

$$y = a_{i0} + \sum_j a_{ij} x_{ij},$$

where the summation is over all enhancers j within 1 Mb of the TSS and x_{ij} is the value of feature i of enhancer j. The coefficients a_{ij} of the enhancers are learned by LASSO, which minimizes the regression error over all samples while selecting a small number of enhancers to have nonzero coefficients. The features considered include H3K4me1, H3K27ac, H3K27me3 and DNase I hypersensitivity for ENCODE + Roadmap and CAGE signal for FANTOM5.

On the basis of these models, in step 1C, for each sample k an error term is computed to check how much the expression of the TSS in this sample, y_k, can be explained by considering each feature i of each enhancer j alone, that is,

$$e_{ijk} = \left| y_k - \left(a_{i0} + a_{ij} x_{ijk} \right) \right|,$$

where x_{ijk} is the value of feature i of enhancer j in sample k. These error terms are computed for all enhancer features i.

In step 2A, these error terms, as well as (i) the values of the features used in the first step of JEME of enhancer j, the TSS and the genomic region between them in sample k and (ii) the genomic distance between enhancer j and the TSS, form a set of predictors for predicting whether enhancer j regulates the TSS in sample k. The actual prediction is performed by learning a random forest model, which is trained on multiple enhancer–TSS pairs based on 'gold-standard' answers defined by a set of validation data (ChIA-PET, Hi-C or eQTL) from sample k. A pair is considered positive if and only if it is supported by the validation data.

The learned random forest model can then be used to predict enhancer–target interactions either on other pairs in sample k in a cross-validation

setting (step 2B) or in another sample (step 2C). More details about the evaluation of prediction accuracy are provided below.

In this study, JEME was applied to predict the target TSSs of active enhancers in the 127 ENCODE + Roadmap samples and separately in the 808 FANTOM5 samples. Methods for turning the numeric predictions into binary networks are described in the **Supplementary Note** and **Supplementary Figure 16**.

JEME can also be used to predict enhancer targets using other data sets. If the ENCODE + Roadmap or FANTOM5 enhancers are considered to be a suitable set of human enhancers but the enhancer targets in some new samples are to be predicted, step 1B onward can be carried out with the TSS expression and enhancer activity signals of these new samples incorporated. On the other hand, if a new set of enhancers is to be used instead, all the steps, including step 1A, need to be carried out with data from the enhancers incorporated.

Checking the consistency between JEME's predictions and validation data. We compared the predictions of JEME with the validation data using both cross-validation and across-sample tests. Because the validation data are expected to miss some active enhancer–target interactions, assuming all pairs not supported by the validation data to be negative would create an unrealistic positive-to-negative ratio. We therefore sampled background pairs as negatives using four different methods, which collectively provide a better evaluation of the accuracy of JEME. The actual ratios used are shown in **Supplementary Tables 12–14**. Details are provided in the **Supplementary Note**.

Comparisons with other enhancer–target prediction methods. We compared JEME with five other enhancer–target prediction methods. To ensure the fairness of the comparisons, for each method compared we used the same data for it and for JEME. Because GM12878 and K562 were used by all five methods, we performed the comparisons using them. By default, we used the enhancers defined by ENCODE + Roadmap in these two cell lines. Positive pairs were those supported by the validation data, and negative pairs were drawn by the random target method. In comparisons with some methods, additional filtering criteria were applied, as detailed below. In the case of Ripple, the enhancer–TSS pairs from the supplementary website of Roy *et al.*[15] were used directly.

Details of running these methods are provided in the **Supplementary Note**.

Definition and analyses of enhancer co-regulation modes. We defined a regulation module as any two enhancers and a TSS where the two enhancers individually regulate the TSS in at least one sample. For each regulation module, we represented the samples in which each enhancer actively regulated the TSS as a binary vector. We then computed the significance of the 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 co-regulation mode if the one-sided Fisher's exact test (right-tailed) *P* value was less than 0.1 and the Jaccard index was larger than 0.5. The module was defined to be in mutually exclusive co-regulation mode if the one-sided Fisher's exact test (left-tailed) *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 co-regulation mode.

In analysis of the co-regulation modules, super-enhancers were downloaded from the dbSUPER database[52]. For motif analysis, we first produced a FASTA file of the enhancer regions using BEDtools[53] and downloaded the human motif data from HOCOMOCO[54] (see URLs). We then scanned the enhancer regions for occurrences of the motifs using FIMO[55]. For each enhancer region, all unique motifs with a *P* value less than 1.0×10^{-4} were recorded as occurring, and the number of unique motif types occurring in each region was counted. Accessibility correlations were based on the DNase I sensitivity of partner enhancers across 97 ENCODE + Roadmap samples (see URLs). Sequence conservation was based on phastCons scores[56] among 100 species obtained from the UCSC Genome Browser.

Identification of differentially expressed genes in HCC with inverse differential methylation at the regulating enhancer. Preprocessed DNA methylation data for patients with liver hepatocellular carcinoma (LIHC) produced using Illumina Infinium Human Methylation 450 arrays were downloaded from TCGA. Likewise, raw read counts of RNA-seq V2 data were downloaded from the same source. Thirty-eight patients matched the inclusion criteria for consideration in our study, which required samples to have matched tumor and control data from both data platforms (**Supplementary Table 9**). All analysis was conducted within the R computing environment.

For each active enhancer in FANTOM5 liver-related samples (**Supplementary Table 10**), we computed the average β value of all the CpG sites within the enhancer for each of the tumor and control samples. An enhancer region was labeled as differentially methylated if it achieved a Bonferroni-adjusted *P* value of less than 0.01 in a paired Student's *t* test based on these average methylation levels for the 38 pairs of tumor and control samples.

For RNA-seq data, raw read counts were first used to filter out genes with fewer than two counts in total in all tumor and control samples. The read counts of the remaining genes were then input into DESeq2 (ref. 57) to determine differentially expressed genes. A gene was considered to be differentially expressed if it had an absolute $\log_2$-transformed fold change greater than 1.5 and an adjusted *P* value of less than 0.01 between the tumor and control samples based on the paired test of DESeq2.

We selected genes with differential expression and inverse differential methylation at an enhancer predicted by JEME to regulate it in at least one FANTOM5 liver-related (normal or HCC) sample.

Human cell culture and demethylation treatment. Human immortalized non-tumorigenic liver cell lines and HCC cell lines were obtained from the American Type Culture Collection (ATCC). All cell lines were obtained from the original source where short tandem repeat (STR) profiling was used to confirm cell identities. All cell lines were free of mycoplasma contamination as verified by PlasmoTest (Invivogen). All cells were cultured in high-glucose DMEM (Gibco) supplemented with 10% FBS (HyClone) at 37 °C in a humidified incubator with 5% CO_2. For demethylation treatment, cell lines were treated with 10 μM 5-aza-2′-deoxycytidine (Sigma-Aldrich) for 3 d.

Luciferase reporter assays. The enhancer fragments of *PSRC1*, *RBM24* and *TERT* were amplified from human genomic DNA and cloned into the BamHI site of the pGL3-Promoter vector (Promega). The cloning primers are listed in **Supplementary Table 15**. The constructs were transfected into LO2 and Huh7 cells using jetPRIME (Polyplustransfection), with a *Renilla* luciferase reporter construct as an internal control and pGL3-Promoter plasmid as the normalization control. Transfected cells were incubated for 24 h, and luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) and the luminometer microplate (GLOMAX) instrument.

CRISPR–Cas9 knockout cell line construction. Guide RNAs (gRNAs) targeting different gene enhancers were designed using CCTop: CRISPR–Cas9 target online predictor (see URLs) and were cloned into the gRNA expression vector MLM3636 (Addgene, 43860); constructs were confirmed by sequencing. To generate the knockout cell lines, parental cells were transfected with multiple gRNA plasmids and hCas9 expression vector (Addgene, 41815) using Lipofectamine 3000 (Invitrogen), and the resulting transfected clones were subjected to G418 selection for 2–4 weeks. Successful knockout of gene enhancers was confirmed by extraction of genomic DNA from the cells, followed by PCR amplification and sequencing of the enhancer-flanking regions. The CRISPR primers are listed in **Supplementary Table 16**.

RNA extraction, reverse transcription and quantitative real-time PCR. Total RNA was extracted from samples using RNAiso (Takara), treated with DNase I (Invitrogen) and used to synthesize cDNA with PrimeScript reverse transcriptase (Takara). Quantitative real-time PCR was performed using SYBR Premix Ex Taq (Takara) in QuantStudio 7 (Applied Biosystems). The primers are listed in **Supplementary Table 17**.

Bisulfite sequencing. Genomic DNA was extracted using the QIAamp DNA Mini kit (Qiagen) and bisulfite modified using the EZ DNA Methylation kit (Zymo Research) according to the manufacturer's protocols. Bisulfite genomic sequencing (BGS) PCR was performed using Taq-Gold polymerase (Applied Biosystems). PCR products were used for further Sanger sequencing to calculate relative methylation levels at CpGs. The BGS primers are

doi:10.1038/ng.3950

listed in **Supplementary Table 18**. A full-length gel photo is provided in **Supplementary Figure 17**.

Code availability. Source code for all computer programs used in the analyses can be downloaded from GitHub (https://github.com/yiplabcuhk/JEME/).

Data availability. The data that support the findings of this study are available from our supplementary website (http://yiplab.cse.cuhk.edu.hk/jeme/). A **Life Sciences Reporting Summary** is available.

51. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
52. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164–D171 (2016).
53. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
54. Kulakovskiy, I.V. *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* **41**, D195–D202 (2013).
55. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
56. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
57. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

nature research

Corresponding Author: Kevin Yip

Date: 25 July 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

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

Experimental design

1. Sample size

Describe how sample size was determined.

The study did not involve the determination of sample size.

2. Data exclusions

Describe any data exclusions.

No data were excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

For the molecular experiments involved in the HCC part, the reported results are based on multiple replicates.

4. Randomization

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

The study did not involve group allocation.

5. Blinding

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

The study did not involve group allocation.

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

6. Statistical parameters

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

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

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

Software

Policy information about availability of computer code

7. Software

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

We implemented some programs for the data analyses performed. The

Nature Genetics: doi:10.1038/ng.3950

source code is available at our GitHub page (https://github.com/yiplabcuhk/JEME) and supplementary web site (http://yiplab.cse.cuhk.edu.hk/jeme/).

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	This study did not involve the use of unique materials.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	This study did not involve the use of antibodies.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Human immortalized non-tumorigenic liver cell lines and HCC cell lines were obtained from the American Type Culture Collection (ATCC).
b. Describe the method of cell line authentication used.	All cell lines were obtained from the original source where Short Tandem Repeat (STR) profiling was used to confirm the cell identities.
c. Report whether the cell lines were tested for mycoplasma contamination.	All cell lines were free of mycoplasma contamination as verified by PlasmoTest (Invivogen).
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Among the FANTOM5 samples, some are listed as misidentified cell lines by the International Cell Line Authentication Committee, including AZ521, HEp-2, SK-N-MC, SKW-3 and TCO-1. We did not experimentally test whether these FANTOM5 cell lines were contaminated, but since we only used them in our modeling and analyses without studying their individual phenotypic properties (for instance, we did not include the Hep-2 cells in the HCC analysis, see Table S10), the inclusion of them does not affect our findings and conclusions.

Animals and human research participants

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

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	This study did not involve the use of research animals.

Policy information about studies involving human research participants

12. Description of human research participants

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

Nature Genetics: doi:10.1038/ng.3950