Improved regulatory element prediction based on tissue-specific local epigenomic signatures

Yupeng He[a,b], David U. Gorkin[c], Diane E. Dickel[d], Joseph R. Nery[a], Rosa G. Castanon[a], Ah Young Lee[c], Yin Shen[e,f], Axel Visel[d,g,h], Len A. Pennacchio[d,g], Bing Ren[c,i], and Joseph R. Ecker[a,j,1]

[a]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; [b]Bioinformatics Program, University of California, San Diego, La Jolla, CA 92093; [c]Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA 92093; [d]Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [e]Institute for Human Genetics, University of California, San Francisco, CA 94143; [f]Department of Neurology, University of California, San Francisco, CA 94143; [g]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [h]School of Natural Sciences, University of California, Merced, CA 95343; [i]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093; and [j]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

Contributed by Joseph R. Ecker, January 9, 2017 (sent for review November 7, 2016; reviewed by Peter A. Jones and Uwe Ohler)

Accurate enhancer identification is critical for understanding the spatiotemporal transcriptional regulation during development as well as the functional impact of disease-related noncoding genetic variants. Computational methods have been developed to predict the genomic locations of active enhancers based on histone modifications, but the accuracy and resolution of these methods remain limited. Here, we present an algorithm, regulatory element prediction based on tissue-specific local epigenetic marks (REPTILE), which integrates histone modification and whole-genome cytosine DNA methylation profiles to identify the precise location of enhancers. We tested the ability of REPTILE to identify enhancers previously validated in reporter assays. Compared with existing methods, REPTILE shows consistently superior performance across diverse cell and tissue types, and the enhancer locations are significantly more refined. We show that, by incorporating base-resolution methylation data, REPTILE greatly improves upon current methods for annotation of enhancers across a variety of cell and tissue types. REPTILE is available at https://github.com/yupenghe/REPTILE/.

enhancer prediction | DNA methylation | bioinformatics | gene regulation | epigenetics

In mammals, genes are transcribed in a temporally and spatially specific manner during development. The precise regulation of gene expression is primarily driven by the activity of distal regulatory sequences, known as enhancers. Disruption of enhancers can cause developmental abnormalities and diseases (1–6). Moreover, the vast majority of genetic variants associated with human diseases by genome-wide association studies (GWASs) lie in noncoding regions, which potentially affect gene transcription and contribute to diseases through disrupting enhancer activity (7, 8). To identify causal noncoding variants and understand their functional consequences, methods for accurate enhancer annotation are essential.

Enhancers are bound by transcription factors (TFs), which in turn recruit cofactors such as the histone acetyltransferase EP300 to achieve transcription activation of target genes from a distance (9). Active enhancers are generally located in accessible chromatin and marked by enrichment of histone H3 lysine 4 monomethylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac) (10–12). Enrichment of histone modifications in the genome can be determined by chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq).

Computational approaches have been developed to predict active enhancers from the combinations of these genome-wide profiles [see review (13) for a list of representative methods]. They generally use machine-learning algorithms to learn the histone modification profiles of putative enhancers active in a given cell/tissue type and then predict enhancers in additional cell/tissue types. Although they have proven to be useful, these methods have several important limitations. First, the centers and boundaries of enhancer predictions are not well defined because of the broad enrichment of histone modifications in regions around enhancers. Second, existing methods often perform worse when tested on cells and tissues other than the cell/tissue types used for training of the algorithm. Third, existing methods consider only one cell/tissue type at a time, and thus neglect potentially useful information about the variation between cell/tissue types.

To address these limitations, we developed regulatory element prediction based on tissue-specific local epigenetic marks (REPTILE), an algorithm to predict enhancers by integrating whole-genome, base-resolution cell/tissue-specific DNA methylation data along with histone modification data. Cytosine DNA methylation (mC) is a type of chemical modification that plays critical roles in gene regulation, transposon repression, and the determination of cell identity (14–17). In mammalian genomes, it occurs in both CG and non-CG contexts (18–22) and can be quantified at nucleotide resolution using whole-genome bisulfite sequencing (WGBS) (18). In this study, we consider only the most prevalent form of cytosine methylation (mCG). Transcription factor binding sites (TFBSs) are generally depleted of mCG (18, 23). Whether mCG affects binding affinity is unclear for the majority of

Significance

In mammals, when and where a gene is transcribed are primarily regulated by the activity of regulatory DNA elements, or enhancers. Genetic mutation disrupting enhancer function is emerging as one of the major causes of human diseases. However, our knowledge remains limited about the location and activity of enhancers in the numerous and distinct cell types and tissues. Here, we develop a computational approach, regulatory element prediction based on tissue-specific local epigenetic marks (REPTILE), to precisely locate enhancers based on genome-wide DNA methylation and histone modification profiling. We systematically tested REPTILE on a variety of human and mouse cell types and tissues. Compared with existing methods, we found that enhancer predictions from REPTILE are more likely to be active in vivo and the predicted locations are more accurate.

Author contributions: Y.H. designed research; Y.H., D.U.G., D.E.D., J.R.N., R.G.C., A.Y.L., Y.S., A.V., L.A.P., and B.R. performed research; D.E.D., A.V., and L.A.P. collected the tissues from E11.5 mouse embryo, which were later profiled for epigenetic marks; D.U.G., A.Y.L., Y.S., and B.R. generated the histone modification data for the E11.5 tissues; D.E.D., A.V., and L.A.P. conducted transgenic mouse assay and generated the newly validated VISTA enhancers; J.R.N. and R.G.C. generated the whole-genome bisulfite sequencing data for the E11.5 tissues; Y.H. analyzed data; Y.H., D.U.G., B.R., and J.R.E. wrote the paper; and J.R.E. supervised the project.

Reviewers: P.A.J., Van Andel Institute; and U.O., Max Delbrueck Center for Molecular Medicine.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: ecker@salk.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618353114/-/DCSupplemental.

TFs, although recent studies suggest that there can be significant alteration of binding affinity (24–26). The anticorrelation of mCG and TF binding is predictive in inferring TFBS (27) and enhancers (23, 28). These observations led us to take advantage of mCG depletion as a high-resolution (~1 bp depending on density of CG sites) enhancer signature that is complementary to the lower-resolution histone modification data derived from ChIP-seq experiments (with fragment size ranging from 200 to 600 bp after sonication) (29). Our results indicate that, by incorporating mCG data, REPTILE achieves higher prediction accuracy and produces higher-resolution enhancer predictions than existing methods that rely solely on histone modification profiles.

Results

The REPTILE Algorithm. We designed REPTILE based on three observations: (*i*) active enhancers, which are bound by TFs in certain cells and tissues, show cell/tissue-specific hypomethylated, and such anticorrelation is an informative feature in predicting enhancers. It has been shown that regions that are differentially methylated across diverse cell and tissue types [also known as differentially methylated regions (DMRs)] strongly overlap with enhancers (19, 20, 30). (*ii*) With base-resolution mCG data, the centers and boundaries of DMRs can be accurately defined, which may be informative in identifying the precise location of enhancers. (*iii*) The known enhancers (31, 32) (~2 kb) are generally much larger than TFBSs (~10–20 bp) and likely include sequences that contribute little to enhancer activity. We used the term "query region" to describe such large regions where a small fraction of the sequences may have a regulatory role. Query regions also refer to negative regions (that showed no observable enhancer activity) and the genomic windows used by enhancer prediction methods. Because a large portion of an active query region may have little contribution to its enhancer activity, the epigenomic signature of the whole active query region may not be an ideal approximation to the epigenomic state of the bona fide regulatory sequences within it. To address this issue, we used DMRs (~500 bp) to pinpoint the possible regulatory subregions within the query regions and to capture informative local epigenomic signatures in both enhancer model training and prediction generation processes (Fig. 1 *A* and *B*).

Specifically, the REPTILE algorithm involves four major steps (Fig. 1*C*). First, DMRs are identified by comparing the mCG profiles of the target sample (in which enhancers will be predicted) and several different cell/tissue types (which serve as reference) (*Methods*). Next, REPTILE integrates epigenomic data and represents each DMR or query region as a feature vector, where each element is the value of either the intensity or the intensity deviation of an epigenetic mark (Fig. 1*D*). The intensity deviation feature captures the epigenomic variation between cell/tissue types and is a unique aspect of REPTILE, whereas existing methods rely on data of a single cell/tissue type (Fig. S1*A* and *Methods*). In the third step, REPTILE learns a model of enhancer epigenomic signatures from the feature values of (putative) known enhancers and negative regions as well as the DMRs within them. This model contains two random forest (33) classifiers, which predict enhancer activities of query regions and DMRs based on their own epigenomic signature (*Methods*). In the last step, REPTILE uses the two random forest classifiers to calculate enhancer confidence scores for DMRs and query regions, based on which the final predictions are generated (*Methods*).

Training Computational Models for Human and Mouse Enhancers. To evaluate the prediction accuracy of REPTILE, we systematically compared REPTILE with four widely used enhancer prediction methods, PEDLA (34), RFECS (35), DELTA (36), and CSIANN (37), using data from a wide variety of human and mouse cells and tissues (Fig. S1 *B–D* and *Methods*). These methods all use machine-learning techniques to predict active enhancers based on histone modification profiles, whereas PEDLA also considers evolutionary conservation (*SI Methods*). Unless specifically stated, six histone modifications were used in these analyses, including H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac, and H3K9ac (*Methods*). Notably, REPTILE uses mCG information in addition to histone marks.

For each method, we trained a model (a set of parameters) for human enhancers using epigenomic data from H1 human embryonic

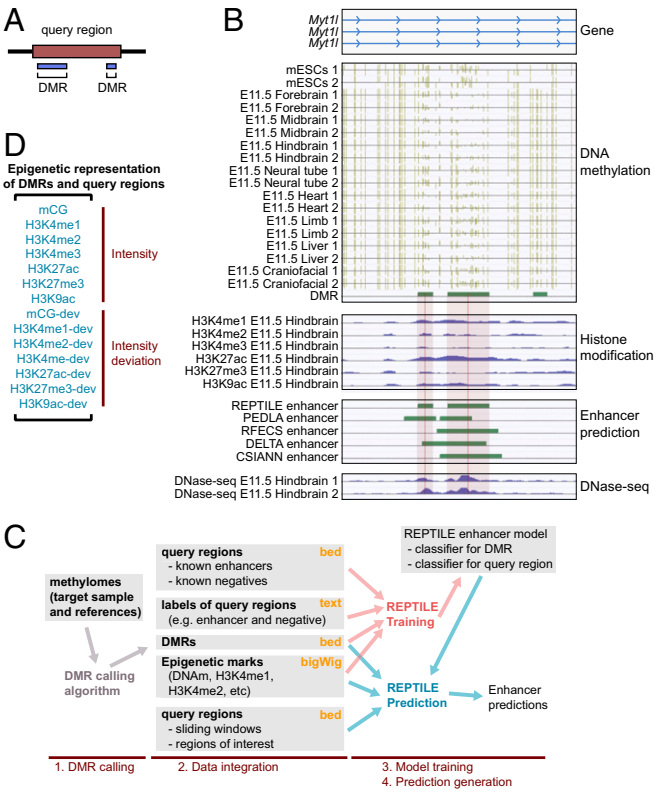


Fig. 1. REPTILE improves enhancer identification by incorporating tissue-specific DNA methylation data. (*A*) Differentially methylated regions (DMRs), typically smaller than query regions, serve as high-resolution enhancer candidates in overlapped query regions. (*B*) Example of a region (chr12: 29,660,800–29,668,600) where REPTILE uses base-resolution DNA methylation data to improve the resolution of enhancer prediction. Diagram of the gene model (GENCODE M2) in this region is shown at the top ("Gene"). "DNA methylation" displays mCG data of mESCs and eight E11.5 mouse tissues, where ticks represent methylated CG sites and their heights indicate the methylation level. Ticks on the forward strand are projected upward, and ticks on the reverse strand are projected downward. Last track shows DMRs across these samples. "Histone modification" shows the logtwofold change of histone modification ChIP-seq data relative to input. Predictions from four computational methods are visualized in "Enhancer prediction." Predictions from REPTILE best recapitulate the open chromatin data shown in "DNase-seq." Light red rectangles mark the REPTILE putative enhancers, whereas the genomic locations of the midpoints (i.e., centers) are highlighted in red. (*C*) Workflow of REPTILE, including four major steps. (1) DMRs are identified by comparing the CG methylation profiles of target sample and the reference samples. (2) REPTILE integrates data in input files and represents query regions and DMRs as feature vectors (*D*). Yellow text on the top right corner shows the format for each input data type. (3) REPTILE trains an enhancer model based on the epigenomic signatures of known enhancers and negative sequences as well as the DMRs within them (red arrows). (4) Predictions are generated based on the enhancer model, DMR, query regions, and epigenomic data (blue arrows). (*D*) Representation of one DMR or query region as a feature vector of intensity or intensity deviation of epigenetic marks. The 14 features used by REPTILE for the benchmark in this paper are shown. The "-dev" features in the vector are the intensity deviation features.

stem cells and a model for mouse enhancers using data from mouse embryonic stem cells (mESCs). During the training process, EP300 binding sites were used as putative active enhancers (positive instances), whereas promoters and randomly chosen genomic regions were used as negative instances (*SI Methods*). When the REPTILE human enhancer model was trained, data of four H1-derived cell types were also included as the reference and DMRs were called for the methylomes of H1 and these cell types. During training of the REPTILE mouse enhancer model, data for eight mouse tissues from embryonic day 11.5 (E11.5) embryo was used as the reference and DMRs were called across the methylomes of mESCs and all of these tissues. In the prediction step, all samples except the target sample were used as the reference. For example, when we applied REPTILE to generate enhancer predictions for E11.5 forebrain, mESCs and the remaining E11.5 tissues were used as the reference.

Unless explicitly stated, all putative enhancers in human cell types and tissues were generated for each method using the human enhancer model, trained using H1 data as described above. Similarly, all enhancer predictions in mouse cell types and tissues were based on the mouse enhancer model, trained using data from mESCs.

REPTILE Shows Superior Prediction Accuracy Compared with Existing Methods. We first used cross-validation to evaluate the learned human enhancer models and mouse enhancer models in H1 and mESCs, where the models were trained. In both cell types, REPTILE showed the best performance among all of the tested methods (Fig. S2 *A* and *B*). In addition, we found that, in H1 cells, putative enhancers from REPTILE and RFECS had the greatest overlaps with distal TFBSs and/or distal open chromatin regions [DNase hypersensitivity sites (DHSs)], whereas REPTILE outperformed all other methods in mESCs (Fig. 2 *A* and *B*, and *SI Methods*). Also, REPTILE showed one of the highest validation rates (fraction of predictions that are within 1 kb to distal DHSs but not in promoters) and one of the lowest misclassification rates (fraction of predictions that are within promoters; Fig. S3 *A–D*). We then tested REPTILE on the 211 experimentally validated regions in mESCs from Yue et al. (32), and it showed superior performance compared with all other methods (Fig. 2*C* and *SI Methods*). Furthermore, we found that REPTILE predictions recaptured the most distal regulatory DNA elements that were identified by multiplexed editing regulatory assay (MERA), a high-throughput genome mutation screening approach (38) (Fig. S2*C* and *SI Methods*).

Because training datasets (e.g., EP300 data) are often not available for the cells or tissues of interest (target samples), it is extremely desirable that the enhancer model learned on one cell/tissue also performs well on other cell/tissue types. To assess this, we applied the models trained on human embryonic stem cell (H1) data to four H1-derived human cell lines and the models trained on mESCs to eight tissues from E11.5 mouse embryo. In human cell types, REPTILE and DELTA show the highest validation rate and the lowest misclassification rate compared with other methods, whereas REPTILE performed the best for mouse enhancer prediction (Fig. 2 *D–G* and Figs. S4 and S5). REPTILE predictions in E11.5 mouse tissues recapitulated several newly in vivo validated enhancers in E11.5 mouse embryo (Fig. 2*H*, Table S1, and *SI Methods*). We then tested REPTILE on in vivo experimentally validated regions and found it achieved the best performance for all test datasets, except in E11.5 midbrain and heart where it ranked second (Fig. 2*C*). Taken together, these results demonstrate REPTILE's superior prediction accuracy in both human and mouse cell/tissue types over existing methods, when training and prediction were performed on different samples.

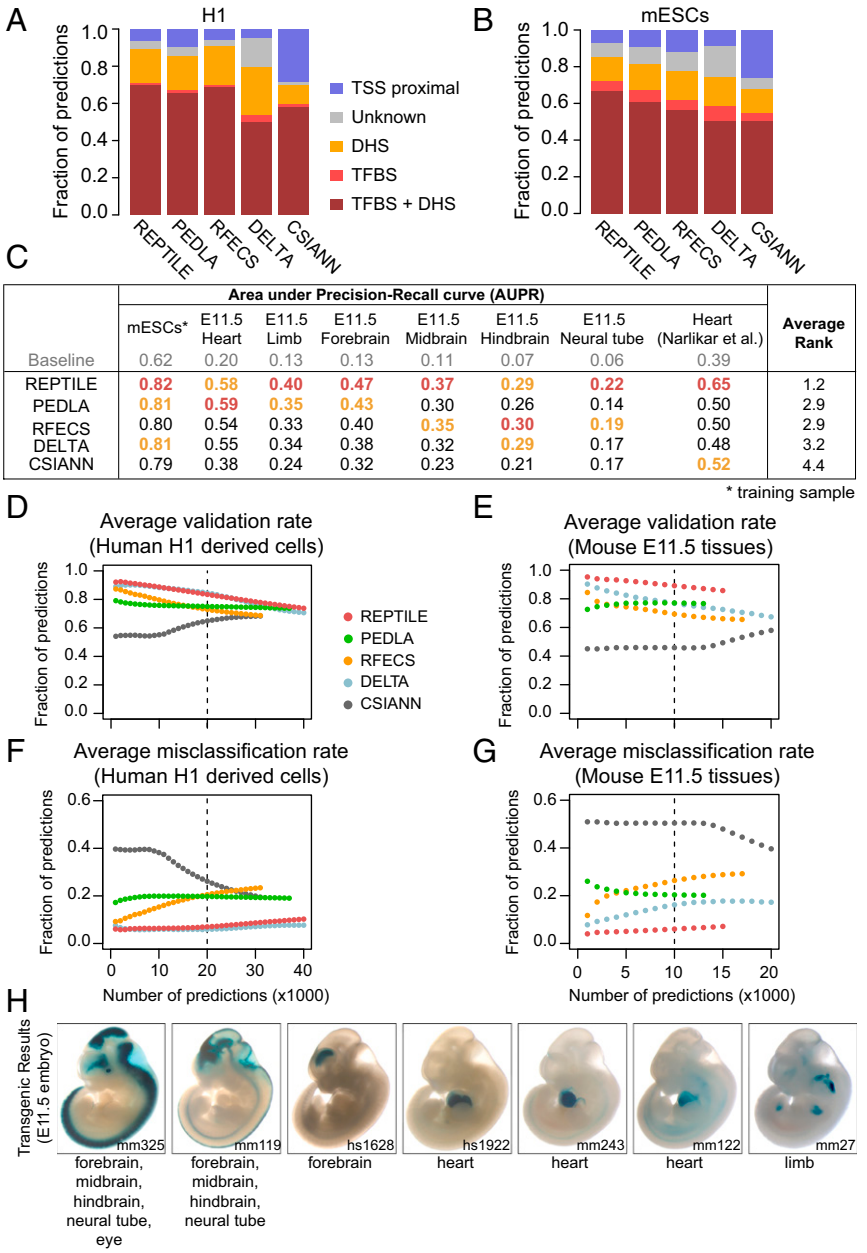
The Resolution of REPTILE Predictions Is Better than Existing Methods. Next, to measure the resolution of enhancer prediction methods, we calculated the average distance between the center of each prediction and the nearest distal DHS (*Methods*). We found a higher percentage (82%) of REPTILE mESCs predictions had distal DHS nearby (within 1 kb) compared with all other methods (77%; Fig. S3*E*). For H1 cells, its overlap (90%) ranked second, which is only slightly lower than RFECS predictions (91%) (Fig. S3*F*). Among these predictions, the centers of RFECS predictions are, on average, 36 bp (H1) and 44 bp (mESCs) closer to the nearest distal DHSs than REPTILE predictions, which ranked second (Fig. S3 *G* and *H*). The results highlight RFECS's superior prediction resolution in the training cell lines (H1 and mESCs), whereas REPTILE's performance is comparable; both outperformed all other methods.

However, we found that REPTILE achieved much better prediction resolution than all other methods when applied to cell/tissue types different from the training data. In H1-derived human cells, the enhancer predictions made by REPTILE are, on average, over 24 bp closer to the nearest distal DHSs compared with other methods, including RFECS (Fig. 3*A*). On average, 85% of REPTILE predictions are supported by nearby distal DHSs, which ranked second, only slightly lower than DELTA (86%; Fig. 3*B*). In tissues from E11.5 mouse embryo, REPTILE predictions are, on average, over 58 bp closer to the nearest distal DHSs than the other methods, and 92% of the REPTILE predictions are close to distal open chromatin regions, outperforming all other methods (84%; Fig. 3 *C* and *D*).

Identifying the Transcription Factors Functionally Related to Each Cell Type Using REPTILE Enhancers. Enhancers are frequently bound by TFs that are critical to the function of cells and tissues. In H1 and H1-derived cell lineages, we found that the predicted enhancers from REPTILE and other methods are enriched for the DNA motifs that are bound by the TFs (or complex) known to function in these cell lines (Fig. 4, Table S2, and *SI Methods*). Motif analysis of REPTILE enhancers recapitulated the enrichment of TF binding motifs in 25 out of the 27 cases (92.6%). Furthermore, in most cases (21 of 27, 77.8%), the TF binding motif showed stronger enrichment in REPTILE enhancers than in the putative enhancers from other methods. Notably, in the trophoblast-like cell lineage (TRO), the average fold enrichment of the TF motifs nearly doubled in enhancers from REPTILE compared with other methods (2.5-fold versus 1.3-fold; Fig. 4). These results indicate that REPTILE enhancer predictions facilitate the discovery of functionally related TFs in a given cell type by accurately pinpointing the location of their binding motifs.

REPTILE Enhancers Are Enriched for Noncoding GWAS SNPs and Associated with Increased Expression of Target Genes. Noncoding disease-associated genetic variants are enriched in the regulatory elements of related cell types and tissues (7). Stronger tissue-specific enrichment of such variants in putative enhancers of related tissues or cell types is likely indicative of better prediction accuracy and resolution. Therefore, we used enrichment as a metric for the evaluation of enhancer prediction methods.

First, we applied all methods to identify enhancers in human heart left ventricle. Because data are available for only some of the epigenetic marks in this tissue, we retrained all methods to generate the enhancer predictions (see *SI Methods* for more details). Then, we tested the enrichment of noncoding GWAS SNPs in these putative enhancers. Consistent with previous findings, only SNPs associated with traits in "Cardiovascular" category showed significant enrichment, indicating that the predicted enhancers are of reasonable quality (Fig. S6*A*). However, we found that these SNPs were most enriched in REPTILE predicted enhancers, suggesting its better resolution and accuracy compared with other methods (Fig. S6 *A* and *B*).



	Area under Precision-Recall curve (AUPR)								Average Rank
	mESCs*	E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural tube	Heart (Narlikar et al.)	
Baseline	0.62	0.20	0.13	0.13	0.11	0.07	0.06	0.39	
REPTILE	**0.82**	**0.58**	**0.40**	**0.47**	**0.37**	**0.29**	**0.22**	**0.65**	1.2
PEDLA	**0.81**	**0.59**	**0.35**	**0.43**	0.30	0.26	0.14	0.50	2.9
RFECS	0.80	0.54	0.33	0.40	**0.35**	**0.30**	**0.19**	0.50	2.9
DELTA	**0.81**	0.55	0.34	0.38	0.32	**0.29**	0.17	0.48	3.2
CSIANN	0.79	0.38	0.24	0.32	0.23	0.21	0.17	**0.52**	4.4

* training sample

Fig. 2. REPTILE shows better enhancer prediction accuracy than existing methods. (*A* and *B*) In H1 (*A*) and mESCs (*B*), the fractions of enhancers with their centers within 1 kb to TFBS+DHS (dark red, both distal TFBSs and distal DHSs), TFBS (red, only distal TFBSs), DHS (orange, only distal DHSs), TSS proximal (overriding all other categories), or none of the above (gray, labeled as Unknown). Distal TFBS (DHSs) are defined as TFBSs (DHSs) that are at least 1 kb away from any TSSs. "TFBS," "DHS," and "TFBS+DHS" are considered as true positives, whereas "TSS proximal" is considered as false positive and misclassification. (*C*) Performances of all methods in eight test datasets that contain experimentally validated enhancers. Performances are measured by the area under precision-recall curve (AUPR). Best results in each test dataset are highlighted in red, and second best results are marked in orange. The enhancer models used to make predictions in all samples were trained on data of mESCs. The baselines (AUPRs achieved using random guessing) for these datasets are shown in gray. Note that the AUPRs in different datasets cannot be compared because the fractions of validated enhancers are different. See Fig. S1*D* for basic statistics of each dataset. (*D* and *E*) The validation rate of each method in human cell lines derived from H1 (*D*) and mouse tissues from E11.5 embryo (*E*), at different numbers of predictions. Validation rate is defined as the fraction of predictions whose centers are within 1 kb from distal DHSs and are at least 1 kb away from TSSs. (*F* and *G*) The misclassification rate of each method in human cell lines derived from H1 (*F*) and mouse tissues from E11.5 embryo (*G*). Misclassification rate is the fraction of predictions whose centers are within 1 kb to TSSs. Vertical dashed lines show the cutoffs used to get the final putative enhancer sets. (*H*) Examples of newly validated enhancers recapitulated by REPTILE enhancer predictions. Candidate enhancers were tested in transgenic mouse assays at E11.5. The enhancer name (mm or hs number), a representative transgenic embryo, and the tissues showing reproducible reporter gene expression (blue staining) are shown for each enhancer. DHS, DNase hypersensitivity sites; mESCs, mouse embryonic stem cells; TFBS, transcription factor binding site; TSS, transcription start site. See also *SI Methods* for details.

Enhancers are expected to increase the transcription of target genes. To test this, we linked REPTILE putative enhancers to their target genes using expression quantitative trait loci (eQTLs) data of left ventricle tissue from Genotype–Tissue Expression (GTEx) Project (*SI Methods*). We found that indeed genes linked to REPTILE enhancers showed significantly higher expression than genes linked to other genomic loci (Fig. S6*C*).

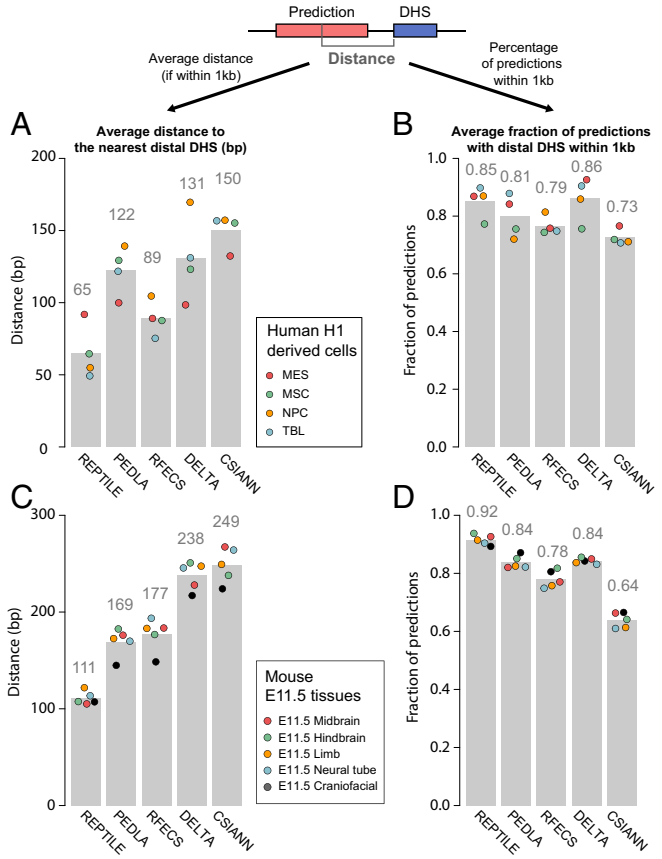


Fig. 3. The resolution of REPTILE predictions exceeds existing methods. (*A*) Average distance between the centers of predictions and the closest distal DHSs in four human cell types derived from H1. Predictions whose centers are beyond 1 kb away from the nearest distal DHS were considered as lack of support from open chromatin data and were not included in the calculation. Distal DHSs are at least 1 kb away from any TSSs. (*B*) Average percentage of predictions whose centers are within 1 kb to the closest distal DHS, in human cells derived from H1. (*C*) Average distance between the centers of predictions and the closest distal DHSs in mouse tissues from E11.5 embryo. (*D*) Average percentage of predictions whose centers are within 1 kb to the closest distal DHS, in mouse tissues from E11.5 embryo. The metric value in each individual cell/tissue is shown as a point in the bar chart. DHS, DNase hypersensitivity sites; Mes, mesendoderm; MSC, mesenchymal stem cells; NPC, neural progenitor cells; TRO, trophoblast-like cells; TSS, transcription start site. See also *SI Methods* for details.

REPTILE Score Correlates Better with in Vivo Enhancer Activity than Open Chromatin. Although open chromatin signatures using DNase-sEq (39)/ATAC-sEq (40) were used for validation in this study, we found that REPTILE score is more predictive of the in vivo activity of DNA elements from VISTA database than open chromatin data (Fig. 5*A* and *SI Methods*). Two recent studies showed that low CG methylation in candidates of regulatory regions is an indicator of enhancers (41, 42). To test this idea, we implemented an approach to predict enhancers based on the CG methylation level in DHSs (DHS+mCG; *SI Methods*). Although useful, this approach does not provide better performance than REPTILE predictions (Fig. 5*A*). We further tested other single histone marks as well as the H3K27ac signal in DHSs and found that none of these is as predictive as the REPTILE score (Fig. 5*A*). Consistently, the enhancer predictions based on REPTILE score consistently achieved the best precision given different score cutoffs (Fig. 5 *B*–*E* and *SI Methods*). These results highlight the value of a method that uses integrative data. At the same time, it suggests that open chromatin regions may not be the ideal data type to validate predicted enhancers.

Discussion

In this study, we describe an algorithm, REPTILE, which is able to predict active enhancers by integrating tissue-specific histone modification data and base-resolution mCG data. We found that the overall accuracy and resolution of REPTILE predictions exceeds other methods, especially when applied to cell/tissue types different from the training data. Further benchmarking revealed that REPTILE's performance is robust to different DMR inputs and reference choices (Figs. S7 and S8; *SI Notes*, *Performance of REPTILE Is Robust to Choice of Reference and Suboptimal Differentially Methylated Region Calling*). In summary, by incorporating DNA methylation data produced by whole-genome bisulfite sequencing and using information of cell/tissue type-specific variation of epigenetic marks, REPTILE greatly improves upon current methods for annotation of enhancers across a variety of cell and tissue types (see also Figs. S7 and S9; *SI Notes*, *Epigenomic Variation Information Improves Enhancer Prediction Resolution and Accuracy*).

Although some methods showed better performance in a few tests, REPTILE's performance was superior in most tests. Although we tried to evaluate the prediction accuracy of all methods in an unbiased manner, we should point out that these benchmarks might be further improved in several ways. First, the validated regions in mESCs were originally selected based on

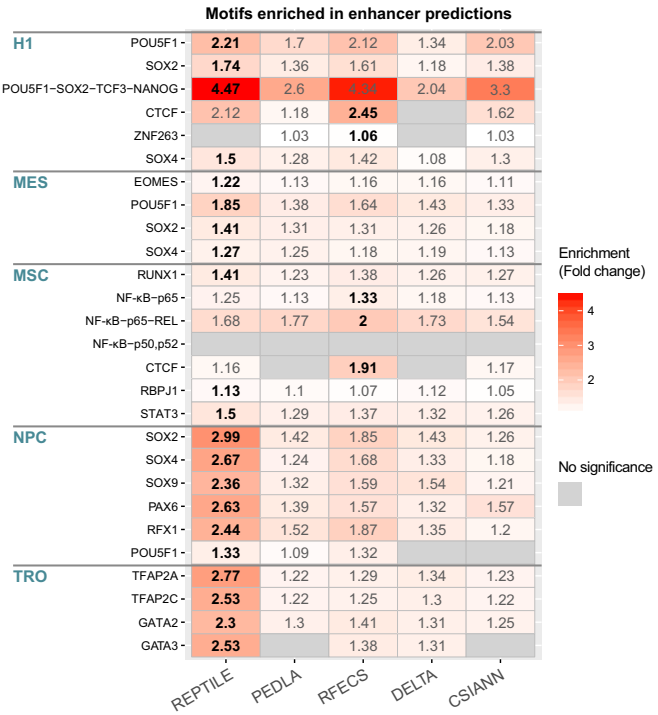


Fig. 4. REPTILE enhancers improve the detection of known transcriptional regulators for each cell type. Enrichment of transcription factor–binding-site motifs in the putative enhancers in H1 and H1 derived cells, respectively. Motif enrichments in each cell type were calculated on the predicted enhancers in matched cell type. Enrichment fold change is the fraction of predicted enhancers (target sequences) that contain a certain motif divided by the fraction of background sequences that contain the same motif. Highest enrichment of each motif in each cell type is marked in bold. Not significant enrichment (q value > 0.05) is shown in gray. The transcription factors (complex) listed under each cell type are known to function in that cell type, which were based on the list from Xie et al. (47). See *SI Methods* for details. H1, H1 human embryonic stem cells; Mes, mesendoderm; MSC, mesenchymal stem cells; NPC, neural progenitor cells; TRO, trophoblast-like cells.

A

Method	AUPR			
	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural tube	E11.5 Limb
REPTILE	**0.37**	**0.29**	**0.22**	**0.40**
DHS	0.22	0.20	**0.18**	**0.33**
DHS+H3K27ac	0.28	0.17	0.12	0.30
DNase-seq Signal	0.21	0.16	0.15	**0.33**
H3K27ac	0.26	0.18	0.12	0.31
mCG	0.20	0.14	0.10	0.23
DHS+mCG	**0.32**	**0.28**	0.17	0.32
H3K4me1	0.21	0.11	0.08	0.24
H3K4me2	0.18	0.13	0.09	0.19
H3K4me3	0.15	0.09	0.07	0.15
H3K27me3	0.08	0.07	0.06	0.15
H3K9ac	0.16	0.10	0.07	0.15

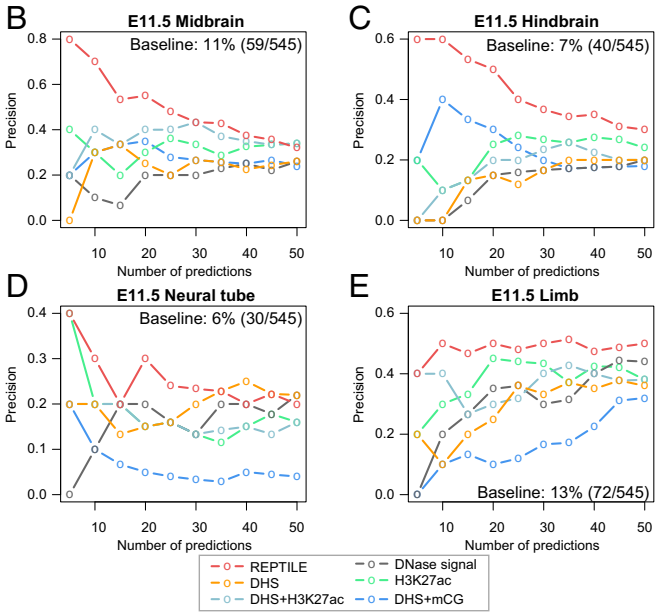


Fig. 5. REPTILE enhancer confidence score is more predictive of enhancer activity than open chromatin or any single epigenetic mark. (*A*) Performance of REPTILE and several enhancer prediction methods that are based on open chromatin, single epigenetic mark, or the H3K27ac signal or mCG in open chromatin regions. The benchmark was done in four test datasets, where DNase-seq data are available in the corresponding samples. Performance is measured by the area under precision-recall curve (AUPR). For each test dataset, the best performance(s) are highlighted in red, and the second best are marked in orange. "REPTILE" generated scores on elements based on the enhancer model trained on data of mouse embryonic stem cells. "DHS" method assigned score to each element as the maximum normalized DNase-seq read count across all overlapping DHSs. The score is 0 if the region contains no overlapping DHS. "DHS+H3K27ac" and "DHS+mCG" are similar to "DHS," but instead of DHS signal, it uses H3K27ac fold enrichment or CG methylation level as signal. The rest of the methods except mCG, DHS+mCG, and H3K27me3 methods use the fold enrichment in whole elements as score. In contrast, mCG, DHS+mCG, and H3K27me3 methods uses the signal values with reversed sign (i.e., depletion) because mCG and H3K27me3 are known to be repressive. (*B–E*) Precision of predicted enhancers that is based on the scores from REPTILE (red), DHS (orange), DHS+H3K27ac (light blue), DNase signal (gray), H3K27ac (green), and DHS+mCG (blue) in E11.5 midbrain (*B*), hindbrain (*C*), neural tube (*D*), and limb (*E*). Precision is defined as the percentage of enhancer predictions that showed enhancer activity in vivo. DHS, DNase hypersensitivity sites. See also *SI Methods* for details.

RFECS predictions, which introduces a potential bias. However, if this bias alters the performance of prediction algorithms, it is likely to inflate the performance of RFECS more than REPTILE. Second, the number of validated enhancer elements is currently limited, although this issue may be resolved in the near future, as more elements will be tested for in vivo function. Third, the negative datasets obtained from the VISTA enhancer database were mostly "putative" enhancer elements from previous studies and therefore may be very similar to true enhancers in many aspects, such as the degree of evolutionary conservation (43). As a result, the prediction accuracy on VISTA enhancer dataset is likely to be lower than the accuracy of whole-genome prediction because many of the "negatives" in the VISTA database actually have some enhancer-like characteristics, which likely makes them harder to differentiate from true positives. Although improvements are possible (such as benchmarking of methods on genomic regions tested in high-throughput enhance assay and incorporating more sophisticated features in the REPTILE model), our results show that REPTILE outperforms existing enhancer prediction methods, especially for samples where training data are unavailable.

As epigenomic information of a larger number of cell/tissue types continues to be comprehensively profiled by the effort of Encyclopedia of DNA Elements (ENCODE) (32, 44, 45), Roadmap Epigenomics Mapping Consortium (REMC) (46), International Human Epigenome Consortium (IHEC), and other consortia, we envision that REPTILE will be a valuable tool to generate accurate enhancer annotations for these datasets, facilitating better regulatory DNA predictions and fueling biological insights.

Methods

Overview of Data Acquisition. To systematically benchmark REPTILE, we collected epigenomic data of various human and mouse cells and tissues. These epigenetic marks included base-resolution DNA methylation data (WGBS) and six histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, and H3K9ac (Fig. S1 *B* and *C*, and Tables S3 and S4). We downloaded data of five human cell lines from Xie et al. (47): H1 human embryonic stem cells (H1), mesendoderm (Mes), mesenchymal stem cells (MSCs), neural progenitor cells (NPCs), and trophoblast-like cells (TRO). Human data also contain WGBS of heart left ventricle from Schultz et al. (19) and histone modification data of the same tissue from Leung et al. (48). In addition, we included data of nine mouse samples: mESCs and eight mouse tissues from E11.5 embryo (*SI Methods*).

Next, to train the computational enhancer prediction methods, we obtained EP300 binding data from mouse and human ESCs (*SI Methods*). It has been shown that EP300 binding is a key feature of a fraction of active enhancers but computational approaches are able to learn the chromatin signatures of these enhancers and predicts other active enhancers without EP300 binding (10, 11). In this regard, we used EP300 binding sites as putative active enhancers in the training datasets.

To validate the enhancer predictions from these methods, we downloaded in vivo enhancer validation data in E11.5 embryonic mouse tissues from the VISTA enhancer browser (31) as well as high-throughput report assay data in mESCs from Yue et al. (32). We also included in vivo validated embryonic heart enhancers from Narlikar et al. (49). In total, eight test datasets were used (Fig. S1*D*). In addition, in all five human cell lines, mESCs, and five E11.5 mouse tissues, we downloaded publicly available DNase-seq data to validate enhancer predictions, assuming the actual location of enhancers to coincide with distal DHSs in the corresponding cell/tissue types. See also *SI Methods* for more details.

REPTILE. REPTILE is an algorithm that generates high-resolution prediction of active enhancers genome-wide by integrating mCG and histone modification data. REPTILE uses the DMRs that are identified across all samples as high-resolution enhancer candidates, and it is able to capture local epigenomic signatures that may otherwise be washed out in the signal of larger region. In addition, it takes into account the tissue specificity of enhancers as features to further improve its performance; REPTILE predicts enhancers based on epigenomic data of not only the target sample (where enhancer predictions are generated) but also additional reference samples to exploit the useful information in variation between cells and tissues.

The overview of REPTILE workflow is shown in Fig. 1*C*, which includes four major steps:

i) DMR calling: DMRs are identified by comparing the DNA methylomes of input samples. We first called differentially methylated sites (DMSs). Next, we merged DMSs into blocks if they both show similar sample-specific methylation patterns and are within 250 bp. These two steps were performed as previously described (19) (see also *SI Methods* for details). We then filtered out the blocks that contain only one DMS. The remaining blocks were then extended 150 bp from each side to include the two regions covered by first upstream and first downstream nucleosomes, respectively. These extended blocks are defined as DMRs, which were used in later steps.

ii) Data integration: Then, REPTILE integrates various types of input data to obtain the epigenomic signatures of DMRs and query regions, in preparing for the next two steps: enhancer model training and prediction generation. Specifically, each DMR or query region is represented as a feature vector and each variable in the vector corresponds to the intensity or intensity deviation of one epigenetic mark (Fig. 1*D*). In this study, the intensity of each histone modification is defined as the logtwofold change in reads per million mapped reads (RPM) relative to control and the intensity of mCG is the CG methylation level. Note that different definitions of intensity can also be used, such as the RPM with subtraction of control or simply RPM of ChIP-seq itself. It makes REPTILE more flexible and allows various way of normalization to be imposed on the input data. Intensity deviation of an epigenetic mark is defined as the intensity in target sample subtracted by its mean intensity in reference samples (i.e., reference epigenome), and this type of feature quantifies the tissue specificity of the epigenetic mark (Fig. S1*A*). Because the data of reference samples is only used to calculate the mean signal value, REPTILE does not require that all epigenetic marks are available in all reference samples, that is, missing data are allowed. However, the target sample, where enhancer predictions are generated, must contain the data of all of the epigenetic marks. In this study, we used seven epigenetic marks (DNA methylation and six histone modifications), and thus the complete REPTILE model contains in total 14 features (two features, intensity and intensity deviation, for each mark; Fig. S9).

The input data vary according to the next step. (*i*) The training step requires data of known/putative enhancers (such as EP300 binding sites) and known negative regions as well as the DMR list and the epigenomic data of target sample and reference samples. (*ii*) Prediction generation takes the enhancer model obtained from the training step, together with the DMRs, the epigenomic data, as input. It also requires query regions. The query regions can be 2-kb sliding windows with step size 100 bp across the genome for generating genome-wide enhancer predictions (see below). They can also be predefined regions, such as conserved elements in the genome, where their enhancer activity is of interest. More details about REPTILE input preparation are available at https://github.com/yupenghe/REPTILE/.

iii) Model training: In the next step, REPTILE enhancer model are trained by learning the epigenomic signatures of query regions, including known enhancers and negatives, as well as the DMRs within them. Specifically, one random forest classifier is trained to learn the epigenomic profiles of the labeled query regions, whereas another random forest classifier is trained to learn epigenomic features in the DMRs that overlap with the query regions. Both classifiers use same 14 features, but the values of these features are calculated differently. The classifier for query regions computes feature values based on the epigenomic data of whole query regions, whereas the classifier for DMRs is trained and applied on the data of DMRs.

The random forest classifier for query regions can be trained on data of known active enhancers and negative regions. However, the classifier for DMRs cannot be trained in such a straightforward way due to the lack of labels for DMRs. To circumvent this, we label all DMRs that are within known enhancers as active, and we label the ones that are within negative regions as inactive. Then, we use these labels to train the random forest classifier for DMRs in a similar fashion as in the training of classifier for query regions. The rationale behind this is that (we assume that) DMRs within negative regions are inactive and part of the DMRs within active enhancers can be inactive. In the training dataset where negative regions greatly outnumber active enhancers, we expect that there are many more DMRs labeled as inactive than active. Therefore, although the inactive DMRs within active enhancers might be incorrectly labeled as active, they only compose a small portion of DMRs. In this paper, the ratio of negatives to positives in the training datasets is at least 7:1 (*SI Methods*). The random forest model can be successfully trained on such data with a small fraction of instances incorrect labeled, which has been demonstrated by the better performance of REPTILE than existing methods. The implementation of random forest model is built on the R (version 3.2.1) package "randomForest" (version 4.6.12) with parameter "ntree=2000, nodesize=1."

iv) Prediction generation: Last, we apply the enhancer model learned in the training step to generate enhancer predictions. Specifically, for every query region or DMR, the corresponding random forest classifier will generate an enhancer confidence score, which is defined as the fraction of decision trees in the random forest model that vote in favor of the active enhancer class.

Given a set of regions of interest, REPTILE is able to predict their enhancer activity. First, REPTILE generates one enhancer confidence score based on the epigenomic signature of certain query region and also multiple scores based on the data of DMRs within it. Then, the maximum is assigned as the final score for this region. In this design, data of DMRs are used to complement the prediction based on query regions. We found that, with correct enhancer model, even if the DMRs were not correctly identified, the prediction performance did not decrease much (see REPTILE w/ shuf DMR in Fig. S7). It is because the incorrect DMRs are not likely to show enhancer-like epigenomic signatures and low enhancer confidence scores will be assigned to them. In this case, the prediction will be dominated by the enhancer confidence score calculated based on the data of whole query regions (see REPTILE w/o DMR in Fig. S7).

REPTILE can also generate enhancer predictions across the genome. In this study, we used REPTILE to first calculate enhancer scores for all DMRs in the genome as well as all 2-kb sliding windows with 100-bp step size across the whole genome. The empirical choices of window size of 2 kb and step size of 100 bp are based on the benchmark results from previous study (35, 50). Then, DMRs with score higher than a given cutoff (0.5 is used in this study) are predicted to be enhancers (termed "enhancer-like DMRs"). To generate nonoverlapping enhancer predictions, overlapping enhancer-like DMRs are merged into single prediction and its score is the highest score of all enhancer-like DMRs that are merged to form this prediction. Next, to capture the enhancers with no detectable mCG variation, REPTILE calls peaks of the enhancer scores across the sliding windows that pass the given score cutoff using the following procedure: (*i*) All sliding windows that pass the cutoff are labeled as enhancer candidates. Candidates that are within 1 kb to each other are grouped into clusters. (*ii*) For each cluster, the candidate with maximum score is set as a peak. If multiple candidates share the highest score, we randomly select one of them as the peak. (*iii*) For each cluster, the peak and all candidates that are within 1 kb of the peak are excluded from the candidate list. (*iv*) Steps 2 and 3 are repeated until the candidate list in each cluster is empty.

After this process, all sliding windows that have score greater than threshold are either peaks or within 1 kb to peaks. The rationale behind this is that the sliding windows adjacent to a peak are part of the peak. Last, the final predictions are the union of the enhancer-like DMRs and the sliding windows that are called as peaks but have no overlap with any enhancer-like DMRs. Similar to the prediction on given regions, this procedure is robust to incorrect DMRs because the enhancers that can be identified using the epigenomic mark of sliding windows will still be called.

Software Availability. The REPTILE software is published under the BSD 2-Clause License. It was written in R and Python. The R code was submitted as an independent R package, called "REPTILE," in the Comprehensive R Archive Network (CRAN). The source code, pretrained enhancer models, use, and further details of the complete pipeline are available in https://github.com/yupenghe/REPTILE.

ACKNOWLEDGMENTS. We thank Dr. John A. Stamatoyannopoulos for generously sharing the DNase-seq data of five E11.5 mouse tissues. We specifically thank Dr. Nisha Rajagopal for kindly helping with the data from MERA. We thank Drs. Shao-shan Carol Huang, Chongyuan Luo, and Manoj Hariharan for their critical comments. Y.H. is supported by the H. A. and Mary K. Chapman Charitable Trust. D.U.G. is supported by the A. P. Giannini Foundation and NIH Institutional Research and Academic Career Development Award K12 GM068524. Transgenic mouse work was conducted at the E. O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. J.R.E. is an Investigator of the Howard Hughes Medical Institute and is supported by grants from the Gordon and Betty Moore Foundation (GBMF3034), the NIH (R01 MH094670 and U01 MH105985), and the California Institute for Regenerative Medicine (GC1R-06673-B). This work was funded by NIH Grant U54 HG006997.

1. Lettice LA, et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12(14):1725–1735.
2. Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T (2005) Elimination of a long-range *cis*-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* 132(4):797–803.
3. Pomerantz MM, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 41(8):882–884.
4. Harismendy O, et al. (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* 470(7333):264–268.
5. Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am J Hum Genet* 76(1):8–32.

6. Sakabe NJ, Savic D, Nobrega MA (2012) Transcriptional enhancers in development and disease. *Genome Biol* 13(1):238.
7. Maurano MT, Humbert R, Rynes E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190–1195.
8. Tak YG, Farnham PJ (2015) Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* 8:57.
9. Merika M, Williams AJ, Chen G, Collins T, Thanos D (1998) Recruitment of CBP/p300 by the IFN β enhanceosome is required for synergistic activation of transcription. *Mol Cell* 1(2):277–287.
10. Heintzman ND, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459(7243):108–112.
11. Heintzman ND, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39(3):311–318.
12. Creyghton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936.
13. Kleftogiannis D, Kalnis P, Bajic VB (2015) Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* 17(6):967–979.
14. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1): 6–21.
15. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220.
16. Jones PA (2012) Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13(7):484–492.
17. Smith ZD, Meissner A (2013) DNA methylation: Roles in mammalian development. *Nat Rev Genet* 14(3):204–220.
18. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
19. Schultz MD, et al. (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 523(7559):212–216.
20. Ziller MJ, et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500(7463):477–481.
21. Varley KE, et al. (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* 23(3):555–567.
22. He Y, Ecker JR (2015) Non-CG methylation in the human genome. *Annu Rev Genomics Hum Genet* 16:55–77.
23. Stadler MB, et al. (2012) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 484:550.
24. Sayeed SK, Zhao J, Sathyanarayana BK, Golla JP, Vinson C (2015) C/EBPβ (CEBPB) protein binding to the C/EBP|CRE DNA 8-mer TTGC|GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. *Biochim Biophys Acta* 1849(6):583–589.
25. O'Malley RC, et al. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 165(5):1280–1292.
26. Stephens DC, Poon GMK (2016) Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains. *Nucleic Acids Res* 44(18):8671–8681.
27. Xu T, et al. (2015) Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res* 43(5):2757–2766.
28. Hwang W, Oliver VF, Merbs SL, Zhu H, Qian J (2015) Prediction of promoters and enhancers using multiple DNA methylation-associated features. *BMC Genomics* 16(Suppl 7):S11.
29. Park PJ (2009) ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genet* 10(10):669–680.
30. Hon GC, et al. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* 45(10):1198–1206.
31. Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* 35(Database issue): D88–D92.
32. Yue F, et al.; Mouse ENCODE Consortium (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515(7527):355–364.
33. Breiman L (2001) Random forests. *Mach Learn* 45:5–32.
34. Liu F, Li H, Ren C, Bo X, Shu W (2016) PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. *Sci Rep* 6:28517.
35. Rajagopal N, et al. (2013) RFECS: A random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* 9(3):e1002968.
36. Lu Y, Qu W, Shan G, Zhang C (2015) DELTA: A distal enhancer locating tool based on AdaBoost algorithm and shape features of chromatin modifications. *PLoS One* 10(6): e0130622.
37. Firpi HA, Ucar D, Tan K (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* 26(13):1579–1586.
38. Rajagopal N, et al. (2016) High-throughput mapping of regulatory DNA. *Nat Biotechnol* 34(2):167–174.
39. Boyle AP, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132(2):311–322.
40. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213–1218.
41. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP (2015) Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol* 16:105.
42. Rhie SK, et al. (2016) Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits. *Epigenetics Chromatin* 9:50.
43. Erwin GD, et al. (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* 10(6):e1003677.
44. Birney E, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816.
45. Consortium EP, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
46. Kundaje A, et al.; Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330.
47. Xie W, et al. (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153(5):1134–1148.
48. Leung D, et al. (2015) Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* 518(7539):350–354.
49. Narlikar L, et al. (2010) Genome-wide discovery of human heart enhancers. *Genome Res* 20(3):381–392.
50. Won K-J, Chepelev I, Ren B, Wang W (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* 9:547.
51. Robinson MD, et al. (2014) Statistical methods for detecting differentially methylated loci and regions. *Front Genet* 5:324.
52. Ma H, et al. (2014) Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* 511(7508):177–183.
53. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12(6): 996–1006.
54. Schultz MD, Schmitz RJ, Ecker JR (2012) "Leveling" the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet* 28(12):583–585.
55. Perkins W, Tygert M, Ward R (2011) Computing the confidence levels for a root-mean-square test of goodness-of-fit. *Appl Math Comput* 217:9072–9084.
56. Bancroft T, Du C, Nettleton D (2013) Estimation of false discovery rate using sequential permutation p-values. *Biometrics* 69(1):1–7.
57. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
58. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
59. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9): R137.
60. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
61. Ong C-T, Corces VG (2014) CTCF: An architectural protein bridging genome topology and function. *Nat Rev Genet* 15(4):234–246.
62. Neph S, et al. (2012) BEDOPS: High-performance genomic feature operations. *Bioinformatics* 28(14):1919–1920.
63. Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory* (Springer, Berlin), pp 23–37.
64. Ernst J, Kellis M (2012) ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* 9(3):215–216.
65. Hoffman MM, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9(5):473–476.
66. Pennacchio LA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444(7118):499–502.
67. Kothary R, et al. (1989) Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* 105(4):707–714.
68. Heinz S, et al. (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576–589.
69. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22.