Published online 29 July 2017

Nucleic Acids Research, 2017, Vol. 45, No. 17 ***9823–9836***
doi: 10.1093/nar/gkx659

Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation

Yu Zhang[1,*] and Ross C. Hardison[2]

[1]Department of Statistics, the Pennsylvania State University, University Park, PA 16802, USA and [2]Department of Biochemistry and Molecular Biology, the Pennsylvania State University, University Park, PA 16802, USA

Received March 21, 2017; Revised June 29, 2017; Editorial Decision July 17, 2017; Accepted July 25, 2017

ABSTRACT

The Roadmap Epigenomics Consortium has published whole-genome functional annotation maps in 127 human cell types by integrating data from studies of multiple epigenetic marks. These maps have been widely used for studying gene regulation in cell type-specific contexts and predicting the functional impact of DNA mutations on disease. Here, we present a new map of functional elements produced by applying a method called IDEAS on the same data. The method has several unique advantages and outperforms existing methods, including that used by the Roadmap Epigenomics Consortium. Using five categories of independent experimental datasets, we compared the IDEAS and Roadmap Epigenomics maps. While the overall concordance between the two maps is high, the maps differ substantially in the prediction details and in their consistency of annotation of a given genomic position across cell types. The annotation from IDEAS is uniformly more accurate than the Roadmap Epigenomics annotation and the improvement is substantial based on several criteria. We further introduce a pipeline that improves the reproducibility of functional annotation maps. Thus, we provide a high-quality map of candidate functional regions across 127 human cell types and compare the quality of different annotation methods in order to facilitate biomedical research in epigenomics.

INTRODUCTION

Thousands of epigenetics datasets have been released in hundreds of human cell types (1–3); this constitutes a rich source of information for studying epigenetic events and improving our understanding of human gene regulation. However, interpretation of the raw data generated by high-throughput sequencing technologies to infer function is difficult, not only because the signals are noisy, but also because different epigenetic marks may represent distinct regulatory functions in a combinatorial fashion. To facilitate the discovery and interpretation of functional elements in human genomes, computational algorithms such as genome segmentation (4,5) have been used to annotate the genome based on multiple epigenetic datasets. The principle is to identify *de novo* combinatorial patterns of multiple epigenetic marks, which are called epigenetic states, within intervals across the genome. The epigenetic states inferred by genome segmentation methods have been shown experimentally to correspond to unique functional elements and have impacts on phenotypes (6). The inferred epigenetic states are a low-dimensional, de-noised representation of the high-dimensional raw data, which are convenient for visualization, interpretation and testing in downstream analyses.

The Roadmap Epigenomics project has published a set of genome segmentation results on 127 human cell types including 16 cell lines from the encyclopedia of DNA elements (ENCODE) project (2). These results have been used to facilitate new biological insights, such as prioritizing and interpreting non-coding genetic variants in human complex diseases (7–12). The Roadmap Epigenomics segmentations were produced by a widely used algorithm called ChromHMM (4), which employs a hidden Markov model (HMM) with binary emission probability to identify epigenetic states. The algorithm works by first converting the raw signals in 200-bp windows to binary values based on a significance cut off in each dataset and then linearly concatenating the epigenomes of all cell types together for joint segmentation. The advantages of this approach include computational speed and simplified interpretation of results, as the method deals with binary outcomes and only models one-dimensional (1D) data dependence across the genome. However, the method has significant limitations. First, because quantitative values are converted to binary, the magnitude of the raw signals is lost and the results are sensitive to threshold choices. Second, the number of epigenetic states must be predetermined, which can easily miss important epigenetic states that are not evident when globally comparing with known biological functions. Third, ChromHMM does not account for the fact that all cell types

*To whom correspondence should be addressed. Tel: +1 814 867 0780; Fax: +1 814 863 7114; Email: yzz2@psu.edu

share the same underlying DNA sequences and hence loses the position-dependent information. Thus, ChromHMM is a '1D' genome segmentation method that is not optimized for jointly segmenting multiple epigenomes.

We recently introduced a new genome segmentation algorithm called IDEAS (13) (for 'integrative and discriminative epigenome annotation system') to tackle the above issues. The IDEAS method works on continuous quantitative data, such that epigenetic signatures of similar patterns but at different scales can be distinguished. IDEAS employs Bayesian non-parametric techniques to automatically choose the number of states from the data instead of requiring user input. This is done by treating the number of epigenetic states in the model as a variable, which is then updated from the data and regularized via Bayesian priors. The underlying principle is to explore different numbers of states and determine the simplest model that can sufficiently explain the variability in the data. If preferred, however, the user can still fix the number of states. Importantly, IDEAS is a '2D' genome segmentation method that, in addition to modeling data dependence along the genome, further accounts for position-wise dependence of regulatory events across cell types. Computationally, the time complexity of inference using the IDEAS model is linear with respect to the genome size and the number of cell types involved. The method is thus computationally efficient even for analyzing hundreds of cell types simultaneously. Finally, because all segmentation methods are sensitive to reduced reproducibility because of the impact of initial values for model parameters, we introduce a novel pipeline that greatly improves reproducibility of the epigenetic states produced.

In light of the advantages of IDEAS over existing genome segmentation tools, we used IDEAS to generate a new map of regulatory elements in the 127 Roadmap Epigenomics cell types. The map can be accessed via the hub link (REFERENCE for track hubs?) for the University of California, Santa Cruz (UCSC) genome browser (http://bx.psu.edu/~yuzhang/Roadmap_ideas/ideas_hub.txt). We used the same five histone marks employed by ChromHMM to produce the map, as they are available in all cell types. We then comprehensively evaluated the ability of the segmentation results produced by IDEAS and ChromHMM to predict or correlate with distinct aspects of genomic function. Our hypothesis is that the epigenetic states are predictive of regulatory function and thus more accurate or less noisy segmentations should match or correlate better with measures of or proxies for genome function. We chose five independent categories of experimental datasets, specifically RNA-seq data in 56 cell types from the Roadmap Epigenomics project, expression quantitative trait loci (eQTL) detected in 44 tissues by the Genotype-Tissue Expression (GTEx) project (14), enhancer usage data generated in 808 human cap-analysis gene expression (CAGE) libraries from the Functional Annotation of the Mammalian Genome 5 (FANTOM5) project (15), four sequence-based scores for functional impact of DNA mutations (16–19) and promoter capture Hi-C data in 17 blood cell types in the International Human Epigenome Consortium (IHEC) project (20). Each category illuminates a different aspect of genome function and all of them are independent of the data we used for generating the functional maps. Thus, correlation between genome segmentation results and these experimental results can be used to assess the relative prediction accuracy of epigenetic states between IDEAS and ChromHMM.

MATERIALS AND METHODS

Roadmap Epigenomics datasets

We downloaded the negative $\log_{10}$ of the Poisson *P*-value tracks of a core set of five chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3) assayed in all of the 127 epigenomes from http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/pval/. We processed the signal tracks of each mark by taking the mean per 200-bp window across the genome in hg19. We removed regions associated with repeats and blacklisted regions as given in (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDukeMapabilityRegionsExcludable.bed.gz) and (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz). The processed dataset contained 635 genome-wide tracks over 13.8 million windows, constituting 8.8 billion observations in total. We took $\log_2(x + 0.1)$ transformation of the data as input to IDEAS. Here, x denotes the negative $\log_{10}$ *P*-values that were provided by the Roadmap Epigenomics project. We additionally applied $\log_2$ transformation of values plus the constant 0.1 to reduce data skewness.

IDEAS model

We previously developed IDEAS (13) to perform joint segmentation of multiple epigenomes simultaneously. This method is motivated by the observation that epigenetic marks and the regulatory functions they represent are correlated both across the genome and among cell types. The latter correlation is due to the same underlying DNA sequences shared by all cell types. Most existing genome segmentation methods only account for data dependence along the genome, but not across cell types. Their methods are therefore '1D'. To produce segmentations in multiple epigenomes, existing methods use either concatenation or stacking techniques in order to fit multiple epigenomes through 1D models, which is suboptimal. In contrast, IDEAS explicitly models both directions of data dependence along the genome and across cell types, and hence it is a '2D' method that produces 2D segmentations with improved power. Direct modeling of high-dimensional data dependence is technically and computationally challenging, because the dependence structure among cell types is unknown. More critically, cell-type-dependence varies along the genome depending on the regulatory functions coded by the underlying DNA sequences.

The IDEAS method tackles these issues and achieves both power and computational scalability by using a novel Bayesian non-parametric hierarchical latent class model. There are two unique components in the IDEAS model. First, it utilizes Bayesian non-parametric hierarchical clustering to identify locally related cell types based on epigenetic similarity. The assumption is that cell types with

similar epigenetic landscapes in a local region should have similar functions. Neither the number of cell type clusters nor how and where they vary along the genome is known. We therefore used infinite-state hidden Markov chains to learn all unknown variables from the data, with one Markov chain per cell type. The state in each Markov chain denotes the cluster membership of cell types at each genomic position and cell types in the same cluster at each position have the same distribution of regulatory functions. The emission distributions in these Markov chains are position-specific and are modeled by Dirichlet processes to account for genomic background variation. Second, the method classifies genome positions into categories. Each category corresponds to a distinct functional profile for the underlying DNA sequences, which is unobserved and agnostic of cell types. We assume that the positions in the same category have a distinct prior distribution of regulatory functions in all cell types. Examples include positions potentially encoding transcription factor binding sites, enhancers or inactive regions. Because the number of possible categories in the genome and the prior distribution of regulatory functions associated with each category are unknown, all unknown variables are learned from the data using another Bayesian non-parametric HMM, where the states in the Markov chain denote the categories of genome positions and each state emits a prior distribution of regulatory functions at each position. Taken together, these two components in the IDEAS model provide us with a linear time solution (with respect to the number of cell types and the genome size) to account for the two directions of data dependence in multiple epigenomes. More details of the IDEAS model can be found in the original paper (13).

Improved reproducibility

Independent runs of genome segmentation may produce different results depending on the initial values of model parameters. We developed a simple but effective approach to substantially improve the reproducibility of genome segmentations between independent runs. First, we randomly selected K regions of 20 Mb each in the genome and ran IDEAS in each region independently. Second, we collected the inferred epigenetic states from the K runs and hierarchically clustered the states based on the means of epigenetic marks in each state. Third, we identified a largest number G and cut the hierarchical tree of the epigenetic states into G or more sub-trees, such that exactly G sub-trees contained epigenetic states from at least x% of the K runs. Finally, we generated G consolidated epigenetic states by averaging the state parameters in each of the G sub-trees. This approach can be intuitively understood as follows. To identify an unknown number (G) of states and their parameters from multiple independent training of IDEAS, if we merged all states produced by the K runs together by cutting the tree at the root, we would obtain perfect reproducibility of states between runs, but with no power; on the other hand, if we treated each state from all runs as a distinct state by cutting the tree at the leaves, we would have poor reproducibility between runs and obtain too many states. As we move down the tree from the root to the leaves, the number of sub-trees will increase, so that we can find a maximum number of sub-trees, in G of which we have states clustered together by their similarity (and hence reproducibility) from at least x% of the K runs. As we move further down the tree toward the leaves, the number of sub-trees satisfying this criterion will decrease, as the total number of sub-trees will increase and there will be fewer states clustered within each sub-tree. Using this approach, we can find a maximum number of states that satisfies the criterion and the states identified by this procedure are reproducible in the sense that they appear in at least x% of the K runs.

To further improve the robustness of this procedure, we determined the number of states (G) using a leave-one-out experiment, i.e. by leaving the states from each of the K runs out, respectively. Based on this, we calculated an average number of G, which is robust to outliers. Given G, we finally used the full tree on all states from the K runs to obtain a final set of consolidated states. We note that the full tree may have more than G sub-trees satisfying the criterion, for which we simply obtained the results from the first feasible solution nearest to the root of the tree.

Our approach for generating reproducible states only requires the user to specify one parameter, x% reproducibility. If x is too small, a large number of less reproducible epigenetic states may be obtained. If x is too large, a small number of highly reproducible states may be obtained but some important states may be missing from this set. Here, we chose $x = 90$, i.e. 90% reproducibility. Another parameter that may be determined by the user is K, the number of independent trainings. Our results showed that this procedure is not sensitive to the choice of K for $K \geq 10$ and hence we used $K = 15$.

Finally, given the reproducible states identified by the above procedure, we ran IDEAS to segment the whole genome of 127 Roadmap Epigenomics cell types using those state parameters as priors. To improve computational efficiency, we implemented parallelization. For both training and whole genome segmentation, we ran IDEAS in 20 iterations. We tested a run of IDEAS in 100 iterations as well, but the results were not substantially better than using 20 iterations. That is, our training pipeline not only improved reproducibility but also enabled shorter runs of IDEAS without sacrificing accuracy.

ChromHMM result

We downloaded the 15-state model by ChromHMM from the Roadmap Epigenomics project website (http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state). We mapped the ChromHMM states to the same set of windows used by IDEAS.

Evaluation method

We performed genome-wide evaluation of the IDEAS predicted epigenetic states and compared them with the map published by the Roadmap Epigenomics project calculated on the same data. The evaluation was based on independently generated experimental data on different biological features in the genome. These data included data on gene expression in 56 Roadmap Epigenomics cell types; enhancer usage in 808 tissues and cell types; eQTLs in 44 tis-

sues; sequence scores capturing various functional potentials of the genome; and chromatin interaction in 17 blood cell types. A common hypothesis underlying our evaluation was that if the predicted epigenetic states are more accurate, they should better correlate with the various signals in these independent experimental datasets, because we expect that the true epigenetic states are indicative of the genomic features being tested. We measured accuracy by adjusted r^2 using regression models, which measured linear correlation between two datasets while accounting for the different numbers of epigenetic states. In several experimental datasets, the tissues and cell types did not match with those in the Roadmap Epigenomics project and in these cases, we calculated the adjusted r^2 for every pair of cell types between the two studies. Although mismatch in cell types is less useful for evaluating accuracy, on average we still expect that better annotation will lead to better prediction and *vice versa*, as epigenetic states are highly correlated across cell types.

RNA-seq data analysis

For each gene (Gencode.v10 (21)) and each cell type, we calculated the proportions of epigenetic states in the regions from 110 kb upstream (relative to the strand of the gene) of the gene's transcription start site (TSS) to 110 kb downstream of the transcription termination site (TTS). We used weighted averages to calculate proportions of states, where weights were defined by 36 B-splines using degree 5 at 30 knots evenly spread over a [0,1] interval. The position 110 kb upstream of TSS corresponds to 0, the position of 110 kb downstream of TTS corresponds to 1, and the TSS and TTS positions correspond to 0.4 and 0.6 in the [0,1] interval, respectively. The positions 10^k upstream of TSS were mapped evenly to the interval [0, 0.4) in log10 scale with respect to k. Similarly, the positions 10^k downstream of TTS were mapped evenly to the interval (0.6,1] with respect to k. Finally, the positions within a gene were mapped evenly to the interval [0.4, 0.6] in their original scale. The weighted state proportions were calculated by using each B-spline as the weight separately, followed by $\log(x + 1e\text{-}5)$ transformation. This resulted in 36 sets of state proportions corresponding to 36 B-splines. We note that each B-spline is a unimode curve with its mass centered at a unique distance to gene and the 36 B-splines together cover the entire genomic interval around the gene. This allows us to not only evaluate an overall (additive) epigenetic effect on expression without regard to the number and the heterogeneous locations of regulatory elements near the gene, but also enables us to evaluate their distance effects. We downloaded RNA-seq reads per kilobase per million mapped reads (RPKM) data in 56 cell types (excluding E000) from the Roadmap Epigenomics Project (http://www.roadmapproject.org/).

We performed two types of regression analyses: (i) prediction of within-cell type gene expression at all genes; and (ii) prediction of across-cell type differential expression at each gene. In both analyses, RPKM values ($\boldsymbol{Y}$) were used as the response and the weighted state proportions ($\boldsymbol{X}$) were used as predictors. The regression model is in a general form of $\boldsymbol{Y} \sim \alpha + \boldsymbol{\beta X} + \varepsilon$. We calculated adjusted r^2 by regression to measure how much RPKM variability is explainable by epigenetic states. To predict within-cell type gene expression, we regressed RPKM of all genes within each cell type on the corresponding epigenetic states weighted by the 36 B-splines to calculate the overall predictive power of epigenetic states on expression. We note that although the state proportions add up to 1, in log-scale the predictor matrix is still in full rank in linear space and thus does not have technical issues in regression. We next regressed RPKM of all genes on the epigenetic states weighted by each B-spline separately (20 predictors for IDEAS and 15 predictors for ChromHMM) to evaluate the predictive power of epigenetic states at a fixed distance to the gene. To further estimate the contribution of each epigenetic state to expression, we obtained partial r^2 values by leaving out each epigenetic state one at a time and calculated the ratio between the partial r^2 of each state and the sum of partial r^2s of all states. To predict across-cell type differential gene expression, we regressed RPKM of each gene in 56 cell types on the corresponding epigenetic states weighted by each B-spline separately (20 predictors for IDEAS and 15 predictors for ChromHMM). The results from individual B-splines were then combined to obtain an estimated power curve for predicting differential gene expression as a function of distance to genes.

GTEx data analysis

We downloaded eQTLs in 44 tissues (v6p) from the GTEx Portal (http://www.gtexportal.org/home/). Within each tissue, we grouped together eQTLs with P-values < 1e-5 that were within 50 kb of each other, where overlapping groups of eQTLs were further merged. We then extended the interval containing each group of eQTLs by 1 kb to each side. Within each eQTL interval, we calculated a weighted state proportion, where the weight is given by $\Sigma_i\{-\log(p_i)\exp(-d_i)\}$ at each position at distance d_i to the ith eQTL and p_i is the P-value of the ith eQTL. In this way, all epigenetic states within the eQTL interval are combined, with more weights given to positions closer to stronger eQTLs. Since eQTLs are enriched in genic regions, instead of using random genomic background, we used the same eQTL interval as controls by calculating an inversely-weighted state proportion. We finally took $\log(x + 1e\text{-}5)$ transformation of the weighted state proportions. Using a logistic regression model, with the response being a binary variable indicating eQTL intervals and controls (I_{eQTL}), and the predictor being the log-transformed state proportions ($\boldsymbol{X}$) (20 predictors for IDEAS and 15 predictors for ChromHMM), we evaluated adjusted r^2 for each pair of GTEx tissue and Roadmap Epigenomics cell type, respectively. The logistic regression model is given by $\text{logit}(I_{eQTL}) = \alpha + \boldsymbol{\beta X} + \varepsilon$.

FANTOM5 data analysis

We downloaded the CAGE-based enhancer data (phase 1 and 2 combined) from the FANTOM5 website (http://fantom.gsc.riken.jp/5/data/). There are two types of data: the tag counts of expression data normalized as tags per million mapped reads (TPM) and the binary peaks reported by the FANTOM5 project at a significance threshold determined by contrasting with control data, both

of which are available in 808 human CAGE libraries. In the regression analysis, we used the log-transformed state proportions in each enhancer region as the predictors, where the enhancer region may or may not be active in the specific CAGE library. We used the TPM values in each CAGE library as the response variable. To estimate an overall proportion of epigenetic states in all CAGE peaks, we calculated the state proportions within each pair of CAGE library and Roadmap Epigenomics cell type and averaged the proportions across all pairs. Finally, we downloaded the pre-calculated enhancer–TSS association data from http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed and calculated the state proportions within each region of the enhancer–TSS pairs.

Sequence-based score analysis

We downloaded the Genomic Evolutionary Rate Profiling (GERP) elements in hg19 from http://mendel.stanford.edu/SidowLab/downloads/gerp/. We downloaded the combined annotation dependent depletion (CADD) score precalculated on 1000 Genome phase 3 variants from http://cadd.gs.washington.edu/download/ and used the scaled version (PHRED-like score). We downloaded the fitness consequence of functional annotation (fitCons) score integrated across the three ENCODE cell types from http://compgen.cshl.edu/fitCons/0downloads/tracks/current/i6/scores/, and we used the highly significant scores ($P < 0.003$) as defined by the authors. We obtained the contextual analysis of transcription factor occupancy (CATO) scores precalculated at 13.4 million single nucleotide polymorphisms (SNPs) overlapping with DNase Hypersensitivity Site from http://www.uwencode.org/proj/CATO/.

We mapped all scores to the 200-bp windows used in this study and obtained the maximum score (0 if not available) within each window. In the regression analysis, we used $\log((x + 1e\text{-}4)/(1 - x + 1e\text{-}4))$ transformed scores as the response and we used the epigenetic states in the corresponding windows as dummy predictors. We further shifted the window positions (up to 5 kb) to evaluate the location precision of epigenetic states. Except for FitCons, all other scores were calculated without using cell-type-specific information. We thus performed regression analysis in each of the 127 Roadmap cell types separately. We further calculated the enrichment of epigenetic states relative to genome average at genomic positions whose scores fall within an interval and we defined the score intervals by partitioning the scores into equal-sized bins (i.e. each bin has a fixed number of instances of scores).

Promoter-capture HiC data analysis

We obtained the promoter-interacting regions (PIR) (CHiCAGO interaction scores >5 by the authors) identified in 17 IHEC blood cell types from Data S1 in Javierre *et al*. (20). In the regression analysis, we used $\log(x + 1)$ transformed CHiCAGO interaction scores of all PIRs as the response. We calculated the state proportion in both bait and target regions in each PIR, followed by $\log(x + 1e\text{-}5)$ transformation, as the predictor. Let PIR denote the log transformed CHiCAGO score, X_{bait} denote the log transformed state proportions in bait (20 predictors for IDEAS and 15 predictors for ChromHMM) and X_{target} denote the log transformed state proportions in target. We compared an additive regression model, $PIR = \alpha + \boldsymbol{\beta_1 X_{bait}} + \boldsymbol{\beta_2 X_{target}} + \varepsilon$, with an interaction model, $PIR = \alpha + \boldsymbol{\beta_1 X_{bait}} + \boldsymbol{\beta_2 X_{target}} + \boldsymbol{\gamma X_{bait} * X_{target}} + \varepsilon$. In the interaction model, the state proportions between the bait and the target regions were multiplied between all state pairs. The regression analysis was done in each pair of IHEC blood cell type and Roadmap Epigenomics cell type.

Blood cell type-specificity was calculated based on RPKM values from the Roadmap Epigenomics RNA-seq data. The Roadmap Epigenomics blood cell types included E037 BLD.CD4.MPC, E038 BLD.CD4.NPC, E047 BLD.CD8.NPC, E050 BLD.MOB.CD34.PC.F, E062 BLD.PER.MONUC.PC and E123 BLD.K562.CNCR. At each gene, we calculated the mean and the variance of RPKM in the blood cell types, as well as in non-blood cell types, in $\log(x + 1)$ scale. The variance for each gene in each cell type group (blood and non-blood) was calculated by first applying a loess regression to fit the variance on the mean. The variance for each gene was then taken as the value on the loess curve or 0.25, whichever is greater, at the mean RPKM of the gene, within blood and non-blood cell types, respectively. We finally calculated a Z-score as the two-sample t-statistic between the blood and non-blood cell types for each gene.

Statistical significance

We used a paired t-test to evaluate the statistical significance of the adjusted r^2 difference between IDEAS and ChromHMM. We first calculated the difference in adjusted r^2 values between IDEAS and ChromHMM within each Roadmap Epigenomics cell type. We then performed a one-sample t-test on the difference to obtain a P-value. We used a paired t-test because there are potential cell type-specific effects on the predictions. We finally adjusted for multiple testing by the Bonferroni method.

RESULTS

Joint segmentation in 127 epigenomes

We ran IDEAS on the uniformly processed P-value tracks of five histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3) commonly available in the 127 epigenomes. The program automatically identified 20 epigenetic states from the quantitative signals. Comparing the mean signals of epigenetic states between the 20-state model from IDEAS and 15-state model from ChromHMM, many chromatin states were commonly identified with similar proportions in the genome, including active transcription start sites (TssA), enhancers (Enh), bivalent TSS (TssBiv) and bivalent enhancers (Enh Biv), heterochromatin (Het), repressed polycomb (ReprPC) and quiescent regions (Quies) (Figure 1A). We adopted the mnemonics used by the Roadmap Epigenomics Consortium on the 15-state model to assign labels to our states; a brief interpretation of the mnemonics assignment is given in Supplementary Table S1. In addition, IDEAS captured some novel patterns in

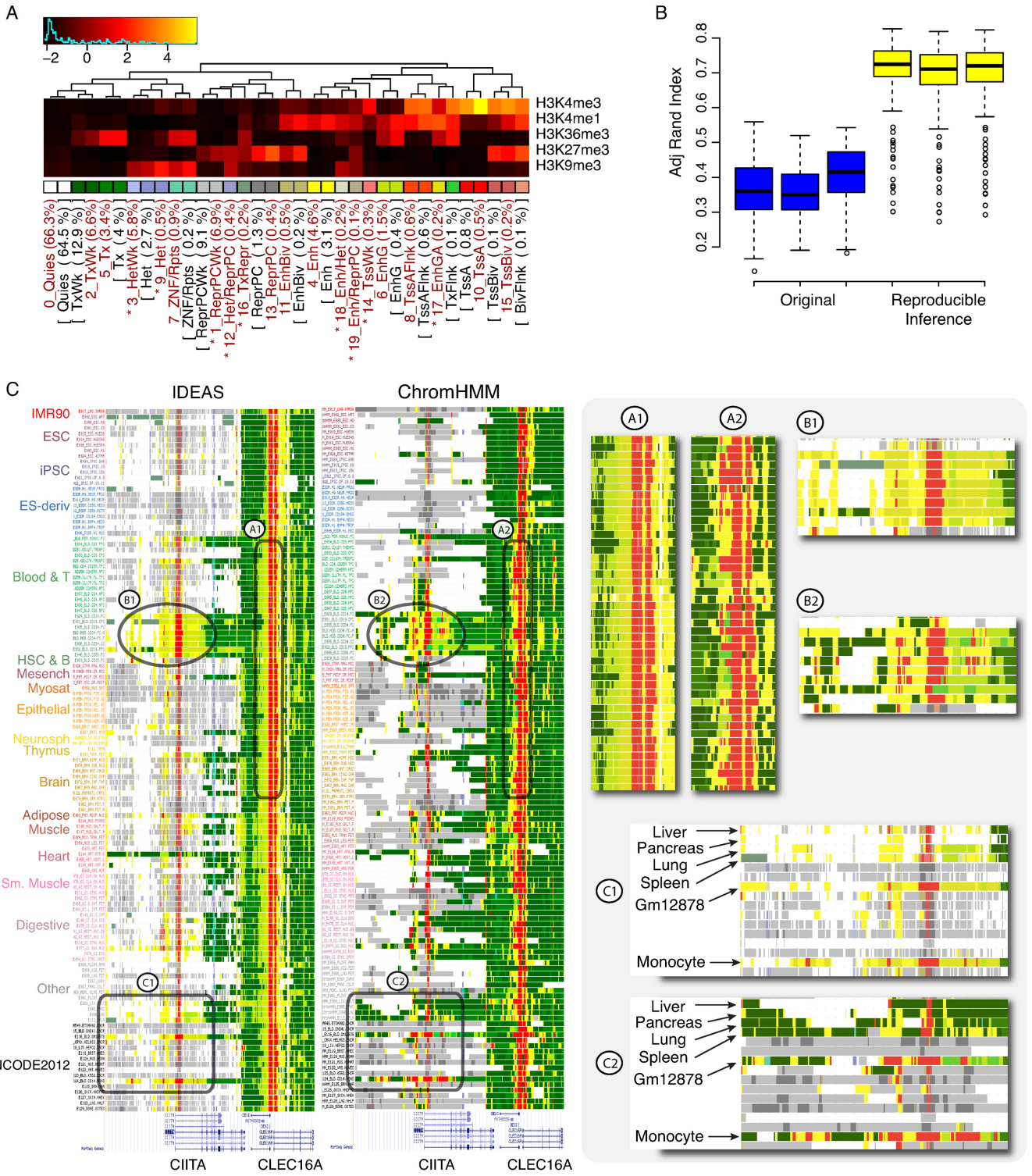

Figure 1. Inferred chromatin states in 127 cell types. (**A**) Mean epigenetic signal in the IDEAS inferred states (red labeled) and the ChromHMM inferred states (black labeled in brackets). Color key for each state is shown under the heatmap. Percentage of each state in the genome is shown in parenthesis. IDEAS states that do not have a one-to-one mapping with ChromHMM's states are marked by asterisk. (**B**) Reproducibility of segmentation by IDEAS between three independent runs using the original program (blue) and the proposed training pipeline (yellow). Each box shows the agreement of segmentation between two runs, measured by adjusted rand index between the inferred chromatin states within matched cell types. Adjusted rand index is a standardized statistics of similarity between two clustering results, which corrects for chance and accounts for different numbers of clusters. (**C**) Segmentation example by IDEAS and ChromHMM in 127 cell types at genes *CIITA* and *CLEC16A*. Blowups highlight some differences between the two maps. Color keys of chromatin states are defined in (A).

the quantitative signals of chromatin marks and their combinations, as observed in several novel states demonstrating a combination of signatures for enhancers, heterochromatin and repressive marks.

While we do not know if these computationally predicted *de novo* states denote unique biological functions, they were consistently recaptured in different runs of IDEAS. Here, we introduce a novel training pipeline of IDEAS that guarantees the generation of reproducible chromatin states. Briefly, we first performed mini-batch training of the IDEAS model to generate a collection of states. We then used these states to evaluate reproducibility and consolidate the similar states into a set of reproducible states. This simple pipeline empirically generated highly reproducible results between independent runs (Figure 1B) and hence the novel states we identified in this study are largely robust.

By visualizing our 20-state model in the UCSC genome browser and comparing with the Roadmap Epigenomics 15-state model, we observed an overall agreement between the two maps. As expected, we also observed some substantial differences. For example, at genes *CIITA* and *CLEC16A* (Figure 1C), our annotation is more consistent across cell types, and at most positions, the state boundaries are better aligned across cell types than those from ChromHMM. On the other hand, there are notable differences in state assignment between the two methods for enhancer, TSS and transcription states. For instance, ChromHMM had many more TSS states (red) assigned to positions away from known TSS than did IDEAS. ChromHMM annotated transcription states (green) in liver, pancreas, lung and spleen both within gene *CIITA* and its upstream non-coding regions. In contrast, IDEAS only annotated transcription states and enhancer states with transcription marks in lung and spleen within the *CIITA*, whereas liver and pancreas only had transcription-related states assigned toward the TTS of *CIITA*. This greater cell type-specificity in expression inferred by IDEAS was confirmed by examining independent gene expression data from both Roadmap Epigenomics and GTEx, as both showed that *CIITA* is expressed in lung and spleen but not in liver and pancreas.

Prediction of expression

We used RNA-seq data from Roadmap Epigenomics in 56 cell types to evaluate the accuracy of the predicted chromatin states by IDEAS. We hypothesize that better correlation between the inferred chromatin states and the RNA-seq data implies better accuracy. We used a functional regression model (22) to include all chromatin states within ±110 kb of each gene as predictors, where we assumed that the state's effects on expression could be modeled as a smooth curve with respect to their distances to genes. Whereas the states obtained by both methods are highly predictive of gene expression, IDEAS had consistently and significantly greater power in all cell types (Bonferroni adjusted *P*-value 4.3e-6 by paired *t*-test) (Figure 2A). Further investigation of the contribution of each state to expression, as a function of distance to genes, showed that the main difference between the two methods for predicting expression occurred near the TTSs of genes (Figure 2B). As expected, we observed predominant contributions from promoter-like states (TssA, TssAFlnk) near the TSS of genes and transcription states (Tx, TxWk) throughout (Figure 2B), to gene expression. We observed stronger contributions of several enhancer-related states inferred by IDEAS, specifically genic enhancers (EnhG) within gene bodies and other enhancer states (Enh, EnhBiv) before TSS and after TTS. The state effects on expression were uniformly positive or negative at all distances to genes for both methods, but the effect sizes were different (Supplementary Figure S1).

Orthogonal to predicting within cell-type expression, we also compared the two methods for predicting differential expression across cell types. Within each group of genes stratified by the levels of differential expression, IDEAS consistently outperformed ChromHMM in this prediction (Figure 2C). We observed three peaks of adjusted r^2 values, within genes near, 50 kb upstream and 50 kb downstream of TTS. The strongest peak is near TSS, which is likely due to promoters and enhancers near genes. The two peaks at a distance of 50 kb from the genes are much smaller in magnitudes, which could be either due to statistical artifacts, or perhaps in part reflecting regulatory activities in neighboring genes.

Prediction of validated enhancers

We next used the inferred chromatin states to predict experimentally validated enhancers. Although both IDEAS and ChromHMM explicitly predicted enhancer states, the state mnemonics were manually assigned and thus are subject to assignment bias. Instead, we calculated the correlation between the chromatin states and the FANTOM5 enhancer usage data by linear regression. The FANTOM5 project used enhancer RNA data derived from 808 CAGE libraries (normal tissues, cell types and cancer cell lines) to estimate enhancer usage in each cell type. Since the CAGE libraries do not match with the Roadmap Epigenomics cell types, we calculated the correlation between every pair of CAGE library and Roadmap Epigenomics cell type. As shown in Figure 3A, IDEAS states are significantly better correlated with the enhancer usage data than ChromHMM states. In particular, IDEAS significantly outperforms ChromHMM for predicting enhancers in 799 out of 808 CAGE libraries, with an average increase in adjusted r^2 by 25%. Further investigation of the relative predictive power of all pairs of CAGE libraries and Roadmap Epigenomics cell types revealed cell type-specific predictions (Figure 3B), including blood-, brain- and epithelial-specific enhancers. These results confirmed that the inferred chromatin states are predictive of cell type-specific enhancers.

The state compositions within significant FANTOM5 enhancer peaks (averaged over 808 CAGE libraries and 127 Roadmap epigenomes) were notably different between the two methods (Figure 3C). About 68% of enhancer peak regions were annotated as either enhancer or promoter like states by IDEAS, whereas this number drops to 54% when using ChromHMM. Notably, a larger proportion of enhancer states with transcription marks (EnhG, EnhGA) were annotated within enhancer peak regions by IDEAS than by ChromHMM, whereas the latter method assigns a larger proportion of weak transcription state (TxWk) within enhancer peak regions. The Tx- states have a strong

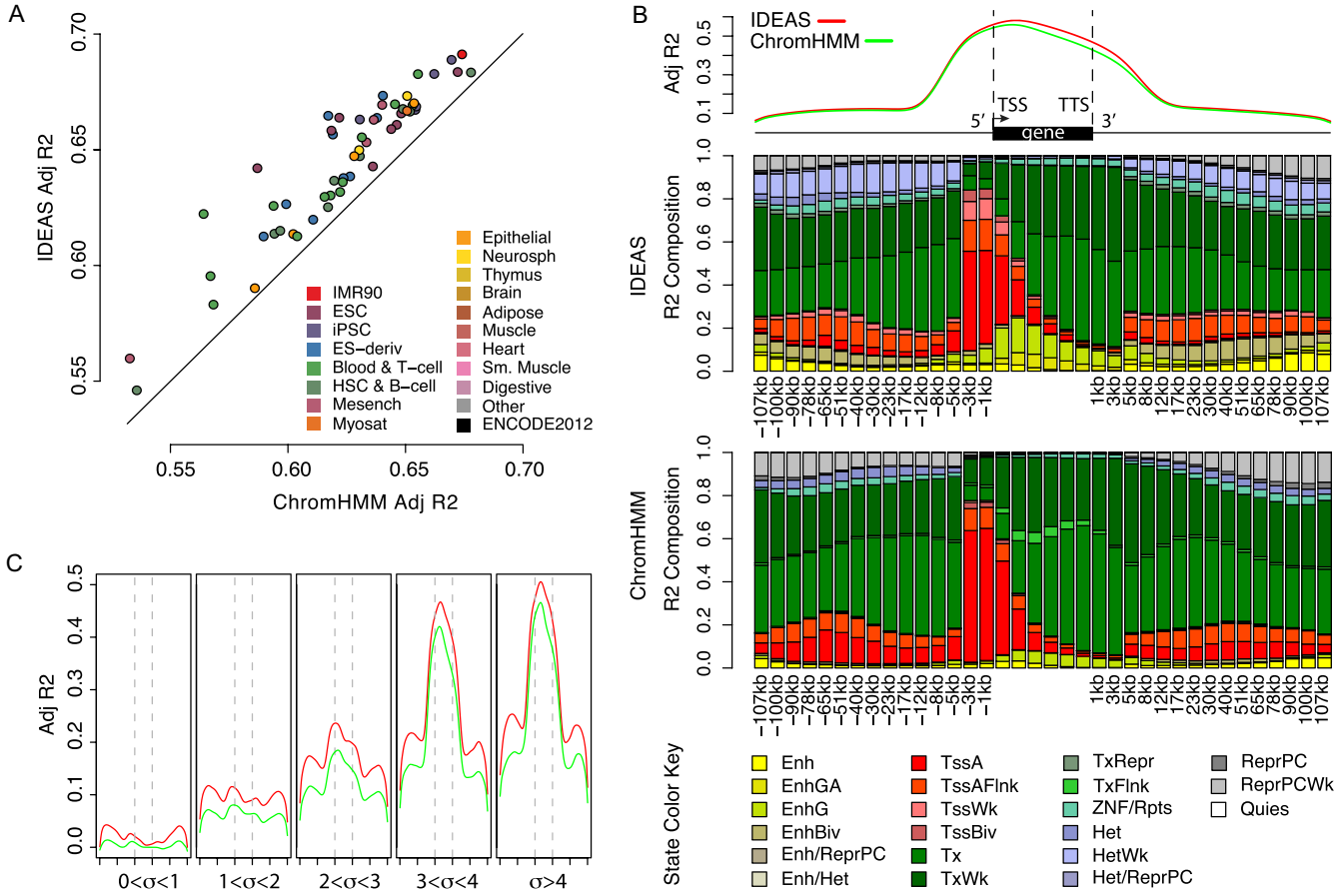

Figure 2. Evaluation by gene expression. (**A**) Within-cell type prediction of gene expression in 56 cell types. Each point shows one cell type, where color keys for cell type lineages are adopted from the Roadmap Epigenomics consortium (Supplementary Data). (**B**) State contribution to gene expression as a function of distance to genes. The panel on the top shows the overall predictive power of states on expression. The two barplots in the middle show the individual state contribution to expression. Color keys of states are shown at the bottom. (**C**) Prediction of differential gene expression across 56 cell types. Genes are stratified by their expression standard deviation across cell types. Each panel shows the adjusted r^2 for predicting differential expression by states as a function of distance to gene (x-axis, in the same scale as in (B) and the two vertical dashed lines in each panel show the transcription start site (TSS) and transcription termination site (TTS) locations, respectively). Red: IDEAS; green: ChromHMM.

signal for H3K36me3, a mark associated with transcriptional elongation after initiation. It is not expected that this histone modification would mark short nascent transcripts, such as the enhancer RNAs used by FANTOM5 to predict enhancers. Thus, the fact that IDEAS finds fewer Tx-states in FANTOM5 enhancer regions is an indication of improved performance. Further calculation of -fold enrichments confirmed that the inferred enhancer states are similarly enriched in the enhancer regions and the Tx- states are not enriched, by both methods. Taken together, these results suggest that IDEAS segmentation is more predictive of FANTOM5 enhancers than ChromHMM.

In addition, the FANTOM5 consortium has reported ~56 000 significant enhancer–TSS pairs showing correlated regulatory activities across the CAGE libraries. We investigated whether the states and their pairing between the enhancer–TSS regions are enriched. As shown in Figure 3D, the states generated by both methods showed a substantial enrichment of enhancer and TSS states in the enhancer–TSS regions. However, IDEAS annotated more enhancer states in the enhancer side of the paired regions and more TSS states in the TSS side of the paired regions. By accounting for the marginal enrichment of states within the enhancer–TSS regions, we further identified several pairwise combinations of chromatin states that are either enriched or depleted between the enhancer–TSS regions (Supplementary Figure S2). Consistent with our expectation, enhancer states were frequently paired with TSS states, repressed states tended to pair with low or repressed states and the enrichment pattern of state pairs depended on gene expression (Supplementary Figure S3). These results demonstrate that the pairing of the chromatin states between two remote regions is predictive of their potential trans-regulatory activities.

Prediction of eQTLs

Several studies have shown that regulatory elements are significantly enriched in eQTLs (23,24), and thus correlation between the inferred chromatin states and eQTLs can be used to assess accuracy. We analyzed the significant eQTLs (nominal P-value < 1e-5) in 44 tissues from the GTEx project. Due to linkage disequilibrium, most of the eQTLs

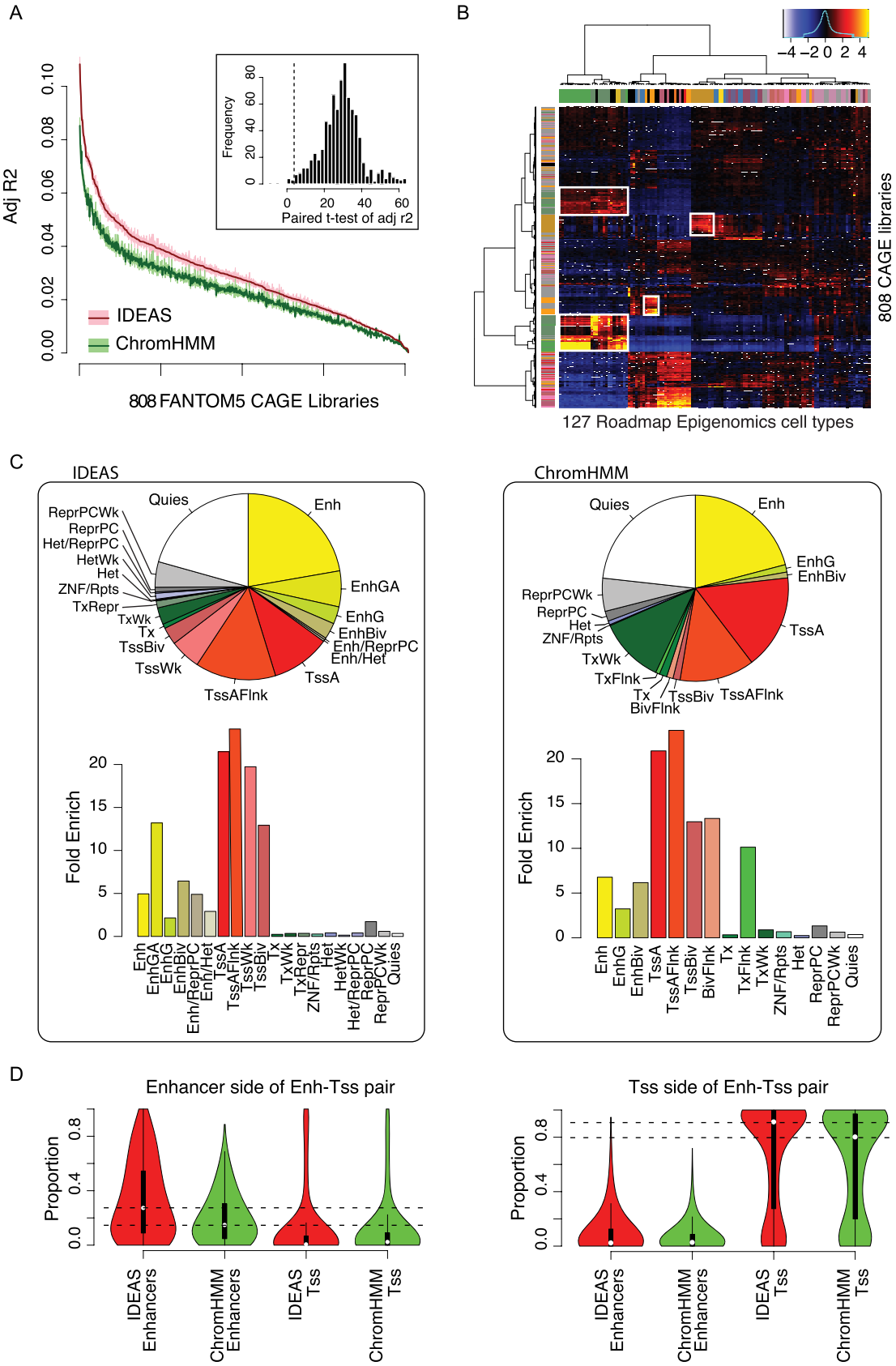

Figure 3. Evaluation by FANTOM5 enhancers. **(A)** Correlation between states- and tissue-specific enhancers in 808 FANTOM5 cap-analysis gene expression (CAGE) libraries. Dark lines show the mean adjusted r^2 over 127 cell types and the shaded areas show the 95% confidence intervals of mean. The

are likely non-causal but linked to causal SNP. We therefore clustered nearby eQTLs together to form an eQTL interval and we calculated weighted state proportions within each eQTL interval. The weights were positively correlated with the significance of eQTLs and negatively correlated with the distance to eQTLs in the interval. We also calculated inversely weighted state proportions within the same eQTL intervals as controls. In this way, the genomic background was the same between cases and controls. We used logistic regression to predict eQTL intervals against controls in each GTEx tissue by each Roadmap Epigenomics cell type. Our assumption was that more accurately inferred chromatin states can better separate eQTLs from controls. The states by both methods were predictive of eQTLs, but IDEAS significantly outperforms ChromHMM in all tissues (Bonferroni adjusted *P*-values 2.7e-61 ∼ 1.1e-15 by paired *t*-test) (Figure 4A). Enrichment analysis of chromatin states in eQTL intervals showed that the Tss-related states are the most strongly enriched states in eQTL intervals, followed by transcription and enhancer-like states, whereas heterochromatin states and repressed polycomb states are the most depleted states in eQTL intervals (Figure 4B).

Correlation with sequence-based scores

DNA sequences are predictive of regulatory functions and several sequence-based scores for predicting the function of nucleotides or SNPs have been computed in the human genome. This provides another way to evaluate the accuracy of chromatin states; specifically, the regulatory potential predicted by DNA sequences should be concordant with that predicted by epigenetic markers. Since neither IDEAS nor ChromHMM used DNA sequences as input, we expect that stronger correlation between sequence-based scores and the inferred states will suggest better accuracy. We included four different scores in this study: (i) the GERP score (16), which identifies functionally constrained elements in multiple alignments; (ii) the CADD score (17), which predicts deleterious effects of DNA mutations; (iii) the fitness consequence of functional annotation (fitCons) score (18), which integrates functional assays with selective pressure to score the fraction of genomic positions evincing a pattern of functional assays that are under selection and (iv) the CATO score (19), which quantifies effects of point mutations on transcription factor binding *in vivo*. Using these pre-computed scores for genome-wide mutations, we could assess how useful chromatin states would be for predicting and interpreting functional impacts of non-coding variants. The scores could also be used to evaluate the positional precision of our predictions, as the scores are calculated at a higher resolution than our 200-bp windows. Using linear regression on the log-transformed scores, the states generated by IDEAS were significantly (*P*-value 3.9e-48 ∼ 4.4e-82 by paired *t*-test) and substantially more predictive of all scores than the ChromHMM states were (Figure 5A). As we shifted the scores away from their original positions, the predictive power of both methods dropped quickly. The results thereby indicate that the IDEAS segmentation not only is more powerful for predicting functional potential of DNA sequences, but also has better positional precision than ChromHMM.

The enrichment patterns of chromatin states with respect to the scores are similar between the two methods (Supplementary Figure S4) but are different for different scores (Figure 5B). Overall, the active states such as enhancer-, Tss- and transcription-like states are enriched in higher scores, and the inactive states such as heterochromatin and quiescent states are enriched in lower scores. Interestingly, the repressed polycomb states (ReprPC, ReprPCWk) are slightly but consistently enriched in higher scores. This is likely due to the fact that we calculated state enrichment using all cell types combined and the repressed polycomb states co-occur with the bivalent TSS and enhancer states (TssBiv, EnhBiv) at the same positions but in different cell types.

Correlation with promoter-capture HiC

Chromatin looping is an important mechanism to enable distal regulation. We thus hypothesized that chromatin states were correlated with chromatin interaction and that better correlation implies more accurate prediction of states. We used the promoter-capture HiC data in 17 blood cell types from the IHEC project (20) to evaluate the ability of our states to predict chromatin interactions. We used the inferred chromatin states within both bait and target regions to predict the CHiCAGO (Capture HiC Analysis of Genomic Organisation) interaction scores (25). As we have shown in the FANTOM5 data, the state co-occurrence between two interacting regions was not random. This is also true in the promoter-capture HiC data. Indeed, we observed that using a pairwise interaction model between the states in bait and target regions consistently outperformed an additive model (Supplementary Figure S5). We thus used the interaction model to predict CHiCAGO interaction scores in each pair of IHEC blood cell type and Roadmap Epigenomics cell type. As shown in Figure 6A, IDEAS uniformly better predicted the CHiCAGO scores and as expected, the Roadmap Epigenomics blood cell types (e.g. Blood and T cells, hematopoietic stem cells (HSC) and B cells) had the

←

insert shows the paired *t*-test statistics for the mean difference of adjusted r^2 between IDEAS and ChromHMM in 808 CAGE libraries, where the dashed line marks the Bonferroni adjusted significance level of 0.05. (**B**) *Z*-scores of adjusted r^2 for predicting enhancers in 808 CAGE libraries (rows) by each Roadmap Epigenomics cell types (columns), calculated by removing row and column means and dividing an overall standard deviation. Library-specific predictions (similar cell types between Roadmap and FANTOM5) are highlighted in boxes, such as blood cell types (the two boxes on the left), brain tissues (the box in the middle) and epithelial cells (the box on the right). Color keys of the Roadmap Epigenomics cell types are the same as those defined in Figure 2A. Color keys for FANTOM5 libraries are manually assigned to match with those used by the Roadmap Epigenomics project. (**C**) State composition and enrichment within significant FANTOM5 enhancer peaks, averaged over 127 cell types and 808 CAGE libraries. The fold enrichment measures the frequency with which the specified segmentation state is found in the FANTOM5 enhancer peaks relative to the genome-wide state distribution. Color keys of states are the same as those given in Figure 2B. (**D**) Distribution of enhancer and TSS-related states in the FANTOM5 significant enhancer–TSS interacting regions.

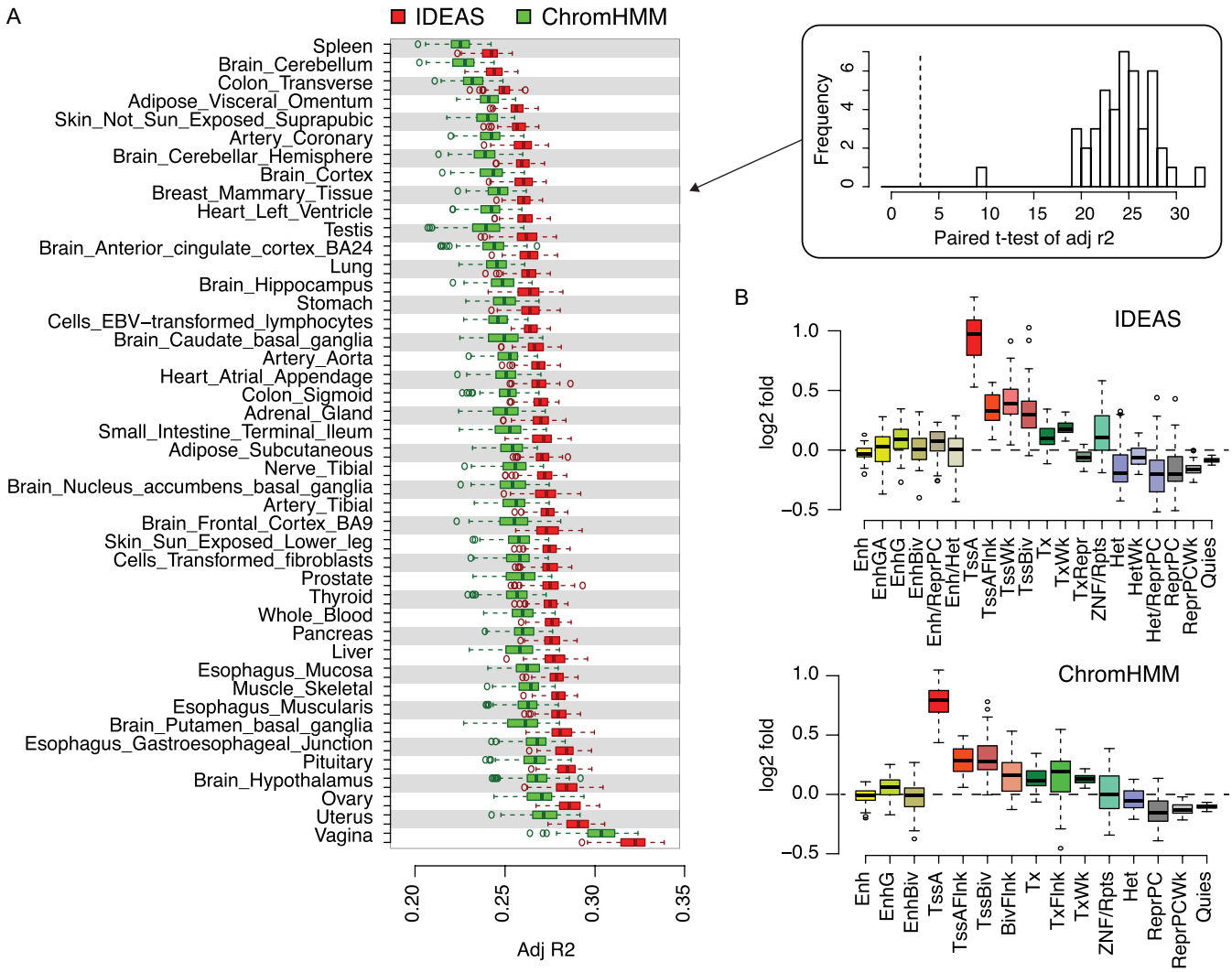

Figure 4. Evaluation by expression quantitative trait loci (eQTL). (**A**) Adjusted r^2 for predicting eQTL intervals in each of the 44 Genotype-Tissue Expression (GTEx) tissues. Each box shows the adjusted r^2 values by regressing eQTL intervals on the states in each of the 127 cell types individually. Paired *t*-test statistics on the difference of adjusted r^2 between IDEAS and ChromHMM in 44 GTEx tissues are shown at the upper right corner, where the dashed line marks the threshold for Bonferroni adjusted significance of 0.05. (**B**) Enrichment of states in eQTL intervals relative to local controls by IDEAS and ChromHMM. Each box shows the enrichment of states in 127 cell types. Color keys of states are the same as those given in Figure 2B.

best predictive power. In addition, within each IHEC blood cell type, the states by IDEAS in the Blood and T cells and HSC and B cells were significantly better correlated with chromatin interaction than the states by ChromHMM (Figure 6B).

Finally, we used RNA-seq data to evaluate whether the promoter-captured regions carry functional elements that affect gene expression. The states calculated by both methods within individual promoter-captured regions were in general correlated with the expression of the bait gene (Figure 6C), but the correlation was not strong. After adding the states in all promoter-captured regions together for the same bait, we observed a much stronger correlation with expression. In addition, the states of both methods at the bait regions were also strongly correlated with the expression of the bait gene. In all cases, the correlation was stronger for genes up- or downregulated specifically in the blood cell types. Comparing between the two methods, the IDEAS states have an overall greater correlation with expression than the ChromHMM states.

DISCUSSION

In this study, we present a new functional annotation map produced using information from 127 human cell types by our IDEAS approach. Using various independent experimental results, we show that the epigenetic states inferred by both IDEAS and ChromHMM are useful for predicting functional and structural information of the genome. We further demonstrate that the IDEAS map significantly and uniformly improves on the ChromHMM map in its ability to be used for predicting regulatory events both within and across cell types. At each genomic position, the IDEAS map shows notable consistency in state assignment across cell types. Simultaneously, it better captures epigenetic vari-

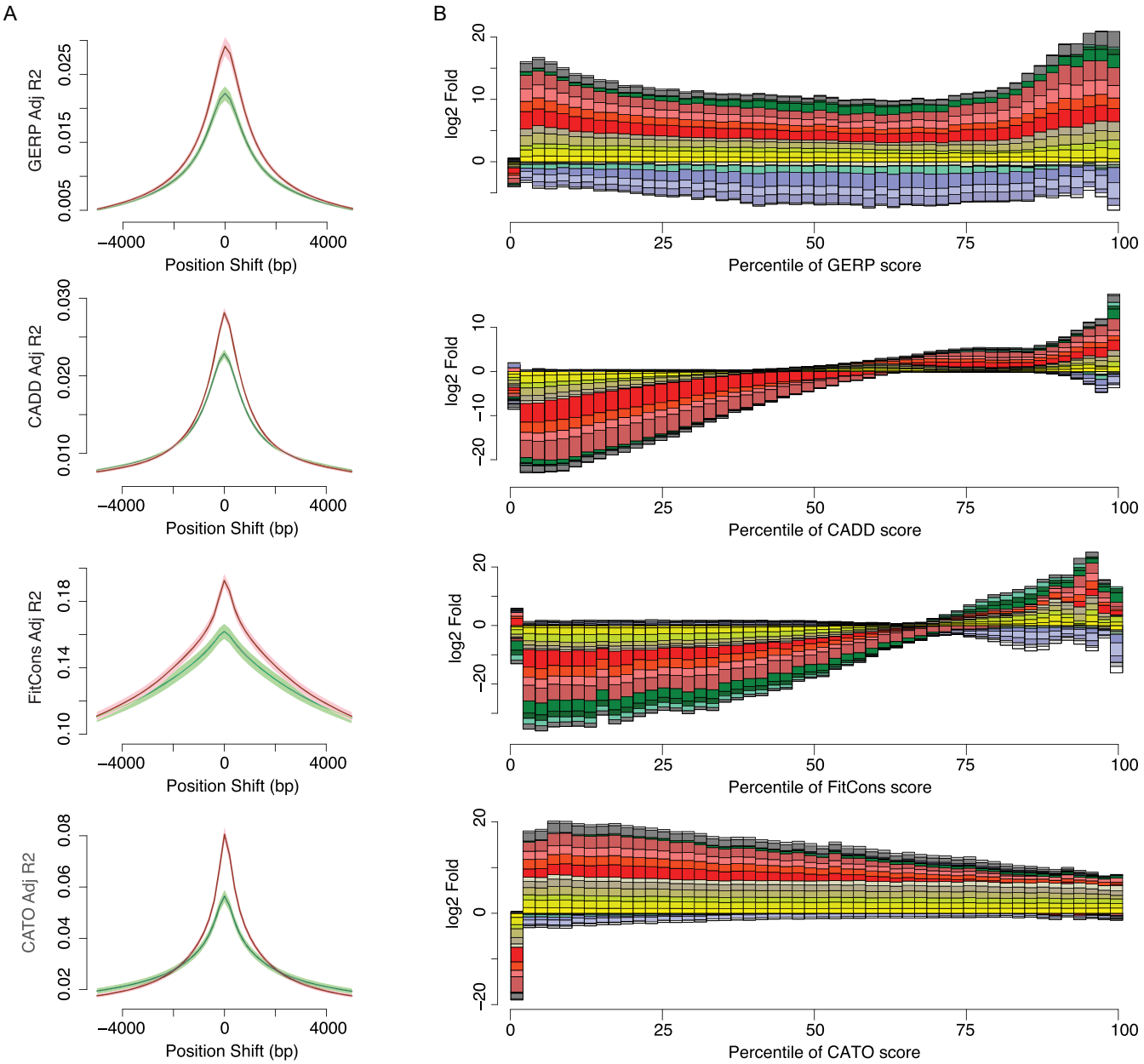

Figure 5. Evaluation by sequence-based scores. (**A**) Correlation with sequence-based scores by IDEAS (red) and ChromHMM (green). The positions of scores are shifted to show the position precision of annotations. (**B**) Cumulative enrichment of states by IDEAS within bins of scores, where the bins are determined by the ranks of score such that there are equal number of scores within each bin. The enrichment is relative to genome-wide average. State enrichment (positive values) and depletion (negative values) are stacked and shown in $\log_2$ scale. Color keys of states are the same as those given in Figure 2B.

ation across cell types as reflected by correlation with differential gene expression.

In this study, we tackled the important issue of state reproducibility in genome segmentation. States inferred by the same method under the same parameter settings must agree between independent runs in order to be useful. This is, however, a notoriously challenging problem, as no global optimum is guaranteed. Our experience with running existing genome segmentation tools showed that the inferred states can vary substantially between runs simply due to chance. We therefore developed an intuitive, simple and effective approach to substantially improve state reproducibility in our maps. The new map presented here thus offers the community an alternative, reliable and more accurate annotation of functional elements in a wealth of human cell types.

There are some limitations in this study. First, we only used five histone marks that are commonly available in the 127 cell types to produce the map. The marks are ideal for detecting basic functional elements such as enhancers, promoters and repressive states, but they do not provide sufficient power to capture more specific regulatory ele-

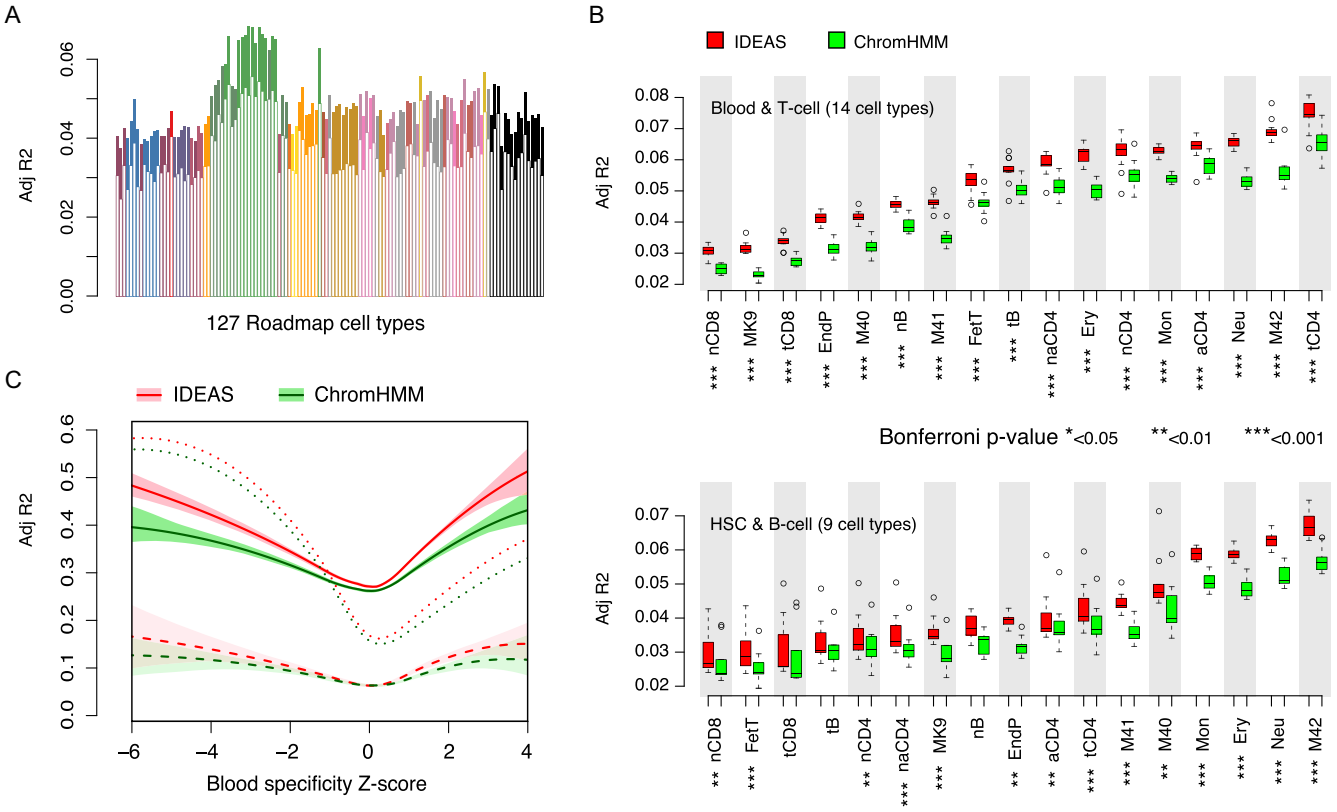

Figure 6. Evaluation by chromatin interaction. (**A**) Correlation between CHiCAGO interaction scores in 17 IHEC blood cell types and the inferred chromatin states in each of the 127 cell types. Adjusted r^2 is calculated from a regression model including interaction terms of states between bait and target regions. Hollow bars show the mean adjusted r^2 by ChromHMM states, averaged over 127 cell types and solid bars show the improvement in mean adjusted r^2 by IDEAS states. Color keys of cell types are the same as those given in Figure 2A, where green and dark green indicates Blood and T cells and HSC and B cell types, respectively. (**B**) Detailed comparison in each IHEC blood cell type using the states of Blood and T cells and HSC and B cell types. Bonferroni adjusted significance by paired *t*-test is indicated under each IHEC cell type. Red: IDEAS; green: ChromHMM. (**C**) Prediction of bait gene expression by the states in bait and target regions as a function of expression specificity for blood cell types relative to other cell types in Roadmap Epigenomics (*Z*-scores, *x*-axis). Dashed lines: mean adjusted r^2 of bait gene expression explained by the states in individual target regions. Solid lines: mean adjusted r^2 of bait gene expression explained by the sum of states in all target regions captured by the same bait. Dotted lines: mean adjusted r^2 of bait gene expression explained by the states in the same bait regions. Shaded area shows the 95% confidence intervals of means.

ments, such as insulators or transcription factor occupancy. The Roadmap Epigenomics project has released additional functional maps using more chromatin marks either in a subset of cell types or in all cell types after data imputation. We have yet to include those additional marks in this study. Second, the models we used in this study to correlate states with independent validation data are mostly linear. While we could have used non-linear models, linear models offer simple interpretation of the results and do not suffer as much from over-fitting the data. Third, interpretation of the inferred chromatin states, assignment of state mnemonics and visualization of genome segmentation remain challenging problems, particularly when many states are produced. Here, we adopted the mnemonics used in the Roadmap Epigenomics project, which may be subject to errors and bias. It will be desirable to further develop automatic-learning algorithms for *de novo* interpretation and visualization of the genome segmentation results.

Beyond generating functional maps, our 2D segmentation method enables new applications. Our method can be extended to leverage information from existing annotations in published cell types to detect functional elements in new cell types and experimental conditions. Our modeling of data dependence is unsupervised and local in the genome, such that information from distant and closely related cell types can both be integrated to make new predictions without cell type matching. Our joint model can be further extended to accommodate missing chromatin marks. New cell types with just one or two marks can still be annotated and benefit from the full spectrum of information provided by all chromatin marks that exist in the published results. This strategy does not require data imputation and thus avoids imputation bias and will save substantially on computing time and data storage. Functional maps produced in the genome of one species may also be applied to other species' conserved DNA sequences. Datasets generated in different species may be integrated and compared via our 2D modeling. Toward this end, we have applied the map in the human genome to the mouse genome in mm10 (http://bx.psu.edu/~yuzhang/Roadmap_ideas/mm10_hub.txt). This will be helpful in downstream analysis, because functional elements are largely conserved between human and mouse at the conserved DNA sequences (26). Finally, our method can in general be used to annotate any entities or subjects in

the broader scope of gene regulation studies, such as different cell types, experimental conditions, individuals, species or timepoints.

AVAILABILITY

The IDEAS software used in this study is available at http://stat.psu.edu/~yuzhang/IDEAS/, which includes the pipeline for generating reproducible segmentation. The R scripts used to produce the evaluation results in this study and the color code of cell type lineages, are available as Supplementary Data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National institute of health (NIH) Grant [1R24DK106766]. Funding for open access charge: NIH [R24DK106766].
Conflict of interest statement. None declared.

REFERENCES

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
3. Stunnenberg,H.G., Hirst,M. and International Human Epigenome Consortium (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell*, **167**, 1145–1149.
4. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
5. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
6. Hardison,R.C. (2012) Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J. Biol. Chem.*, **287**, 30932–30940.
7. Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.
8. Chung,D., Yang,C., Li,C., Gelernter,J. and Zhao,H. (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.*, **10**, e1004787.
9. Kichaev,G. and Pasaniuc,B. (2015) Leveraging functional annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.*, **97**, 260–271.
10. Kichaev,G., Yang,W.Y., Lindstrom,S., Hormozdiari,F., Eskin,E., Price,A.L., Kraft,P. and Pasaniuc,B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.*, **10**, e1004722.
11. Farh,K.K., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shoresh,N., Whitton,H., Ryan,R.J., Shishkin,A.A. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
12. Li,Y. and Kellis,M. (2016) Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.*, **44**, e144.
13. Zhang,Y., An,L., Yue,F. and Hardison,R.C. (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.
14. GTEx Consortium Human genomics (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
15. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
16. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
17. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
18. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
19. Maurano,M.T., Haugen,E., Sandstrom,R., Vierstra,J., Shafer,A., Kaul,R. and Stamatoyannopoulos,J.A. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, **47**, 1393–1401.
20. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S., Cairns,J., Wingett,S.W., Várnai,C., Thiecke,M.J. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
21. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
22. Ramsay,J.O. and Silverman,B.W. (1997) *Functional Data Analysis*. Springer-Verlag, NY.
23. Nicolae,D.L., Gamazon,E., Zhang,W., Duan,S., Dolan,M.E. and Cox,N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
24. Zhong,H., Beaulaurier,J., Lum,P.Y., Molony,C., Yang,X., Macneil,D.J., Weingarth,D.T., Zhang,B., Greenawalt,D., Dobrin,R. *et al.* (2010) Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.*, **6**, e1000932.
25. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Várnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.M., Osborne,C. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.*, **17**, 127–143.
26. Xiao,S., Xie,D., Cao,X., Yu,P., Xing,X., Chen,C.C., Musselman,M., Xie,M., West,F.D., Lewin,H.A. *et al.* (2012) Comparative epigenomic annotation of regulatory DNA. *Cell*, **149**, 1381–1392.