Published online 17 June 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
https://doi.org/10.1093/nargab/lqab052

Accurate prediction of *cis*-regulatory modules reveals a prevalent regulatory genome of humans

Pengyu Ni and Zhengchang Su*

Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28223, USA

Received February 26, 2021; Revised May 01, 2021; Editorial Decision May 16, 2021; Accepted June 14, 2021

ABSTRACT

***cis*-regulatory modules(CRMs) formed by clusters of transcription factor (TF) binding sites (TFBSs) are as important as coding sequences in specifying phenotypes of humans. It is essential to categorize all CRMs and constituent TFBSs in the genome. In contrast to most existing methods that predict CRMs in specific cell types using epigenetic marks, we predict a largely cell type agonistic but more comprehensive map of CRMs and constituent TFBSs in the gnome by integrating all available TF ChIP-seq datasets. Our method is able to partition 77.47% of genome regions covered by available 6092 datasets into a CRM candidate (CRMC) set (56.84%) and a non-CRMC set (43.16%). Intriguingly, the predicted CRMCs are under strong evolutionary constraints, while the non-CRMCs are largely selectively neutral, strongly suggesting that the CRMCs are likely *cis*-regulatory, while the non-CRMCs are not. Our predicted CRMs are under stronger evolutionary constraints than three state-of-the-art predictions (GeneHancer, EnhancerAtlas and ENCODE phase 3) and substantially outperform them for recalling VISTA enhancers and non-coding ClinVar variants. We estimated that the human genome might encode about 1.47M CRMs and 68M TFBSs, comprising about 55% and 22% of the genome, respectively; for both of which, we predicted 80%. Therefore, the *cis*-regulatory genome appears to be more prevalent than originally thought.**

INTRODUCTION

cis-regulatory sequences, also known as *cis*-regulatory modules (CRMs) (i.e. promoters, enhancers, silencers and insulators), are made of clusters of short DNA sequences that are recognized and bound by specific transcription factors (TFs) (1). CRMs display different functional states in different cell types in multicellular eukaryotes during development and physiological homeostasis, and are responsible for specific transcriptomes of cell types (2). A growing body of evidence indicates that CRMs are at least as important as coding sequences (CDSs) to account for inter-species divergence (3,4) and intra-species diversity (5), in complex traits. Recent genome-wide association studies (GWAS) found that most complex trait-associated single nucleotide variants (SNVs) do not reside in CDSs, but rather lie in non-coding sequences (NCSs) (6,7), and often overlap or are in linkage disequilibrium (LD) with TF binding sites (TFBSs) in CRMs (8). It has been shown that GWAS SNVs systematically disrupt binding sites of TFs related to the traits (8), and that variation in TFBSs affects DNA binding, chromatin modification, transcription (9–11), and susceptibility to complex diseases (12,13) including cancer (14–17). In principle, variation in a CRM may result in changes in the affinity and interactions between TFs and cognate binding sites, thereby altering histone modifications and target gene expressions in relevant cells (18,19). Such alterations in molecular phenotypes can change cellular and organ-related phenotypes (20,21). However, it has been difficult to link non-coding variants to complex traits (18,22), largely because of our lack of a good understanding of all CRMs, their constituent TFBSs and target genes in genomes (23).

Fortunately, the recent development of ChIP-seq techniques for locating histone marks (24) and TF bindings in genomes in specific cell/tissue types (25) has led to the generation of enormous amount of data by large consortia such as ENCODE (26), Roadmap Epigenomics (27) and Genotype-Tissue Expression (GTEx) (28), as well as individual labs worldwide (29). These increasing amounts of ChIP-seq data for relevant histone marks and various TFs in a wide spectrum of cell/tissue types provide an unprecedented opportunity to predict a map of CRMs and constituent TFBSs in the human genome. Many computational methods have been developed to explore these data individually or jointly (30). For instance, as the large number of binding-peaks in a typical TF ChIP-seq dataset dwarfs earlier motif-finding tools (e.g. MEME (31) and BioProspector (32)) to find TFBSs of the ChIP-ed TF, new tools (e.g. DREME (33), MEME-ChIP (34), XXmotif (35) and Homer (36)) have been developed. However, some of these tools (e.g. MEME-ChIP) were designed to find primary

*To whom correspondence should be addressed. Tel: +1 704 687 7996; Fax: +1 704 687 8667; Email: zcsu@uncc.edu

motifs of the ChIP-ed TF in short sequences (~200 bp) around the binding-peak summits in a small number of selected binding peaks in a dataset due to their slow speed. Some faster tools (e.g. Homer, DREME and XXmotif) are based on the discriminative motif-finding schema (37) to find overrepresented *k*-mers in a ChIP-seq dataset, but they often fail to identify TFBSs with subtle degeneracy. As TFBSs form clusters in a CRM for combinatory regulation in higher eukaryotes (1,38,39), tools such as SpaMo (40), CPModule (41) and CCAT (42) have been developed to identify multiple closely located motifs as CRMs in a single ChIP-seq dataset. However, these tools cannot predict CRMs containing novel TFBSs, because they all depend on a library of known motifs (e.g. TRANSFAC (43) or JASPAR (44)) to scan for collaborative TFBSs in binding peaks. Due probably to the difficulty to find TFBS motifs in a mammalian TF ChIP-seq dataset that may contain tens of thousands of binding peaks, few efforts have been made to explore entire sets of an increasing number of TF ChIP-seq datasets to simultaneously predict CRMs and constituent TFBSs (45–48).

On the other hand, as a single histone mark is not a reliable CRM predictor, a great deal of efforts have been made to predict CRMs based on multiple histone marks and chromatin accessibility (CA) data from the same cell/tissue types using various machine-learning methods, including hidden Markov models (49), dynamic Bayesian networks (50), time-delay neural networks (51), random forest (52) and support vector machines (SVMs) (53). Although CRMs predicted by these methods are often cell/tissue type-specific, their applications are limited to cell/tissue types for which the required datasets are available (26,49,50,54). Many enhancer databases have also been created either by combining results of multiple such methods (55–57), or by identifying overlapping regions of CA and histone mark tracks in the same cell/tissue types (58–62). For example, the ENCODE phase 3 consortium (26) recently identified 926 535 candidate *cis*-regulatory elements (cCREs) based on overlaps between millions of DNase I hypersensitivity sites (DHSs) (63) and transposase accessible sites (TASs) (64), active promoter histone mark H3K4me3 (65) peaks, active enhancer mark H3K27ac (66) peaks and insulator mark CTCT (67) peaks, in a large number of cell/tissue types. The resolution of these predictions also low (49,50,54) and their predicted CRMs often lacks TFBSs information (26,49,50,54), particularly for novel motifs, although some predictions provide TFBSs locations by finding matches to known motifs (56,57,61). Moreover, results of these methods are often inconsistent (68–71), e.g. even the best-performing tools (DEEP and CSI-ANN) have only 49.8% and 45.2%, respectively, of their predicted CRMs overlap with the DHSs in Hela cells (53); and only 26% of predicted ENCODE enhancers in K562 cells can be experimentally verified (68). The low accuracy of these methods might be due to the fact that CA and histone marks alone are not reliable predictors of active CRMs (53,68,69,71).

It has been shown that TF binding data are more reliable for predicting CRMs than CA and histone mark data, particularly, when multiple closely located binding sites for key TFs were used (53,68,69,71). Moreover, although primary binding sites of a ChIP-ed TF tend to be enriched around the summits of binding peaks, TFBSs of collaborator of the ChIP-ed TFs often appear at the two ends of binding peaks that are parts of a CRM (72,73). With this recognition, instead of predicting cell/tissue type specific CRMs using CA and histone marks data, we proposed to first predict a largely cell-type agnostic or static map of CRMs and constituent TFBSs in the genome by integrating all available TF ChIP-seq datasets for different TFs in various cell/tissue types (47,48), just as has been done to find all genes encoded in the genome using gene expression data from all cell/tissue types (74). We proposed to appropriately extend short binding peaks to the typical length of enhancers, so that more TFBSs for collaborators of the ChIP-ed TF could be included (72,73) in extended parts, and full-length CRMs could be identified (47,75). Although we still need a large number of datasets for diverse TFs from diverse cell type to predict most, if not all, of CRMs and TFBSs in the genome, we do not need the volume of data for all TFs from all cell types due to the extensive reutilizations of CRMs in different cell types. In fact, the coverage of the genome by the growing number of datasets is already in the saturation phase, particularly, if binding peaks are appropriately extended (47). Once a map of CRMs and constituent TFBSs in the gnome is available, functional states of CRMs and constituent TFBSs in cell/tissue types could be predicted and studied in a more cost-effective way. Although our earlier implementation of this strategy, dePCRM, resulted in promising results using even insufficient datasets available then (47,75), we were limited by three technical hurdles. First, although existing motif-finders such as DREME used in dePCRM worked well for relatively small ChIP-seq datasets from organisms with smaller genomes such as the fly (48), they were unable to handle very large datasets from mammalian cells/tissues, so we had to split a large dataset into smaller ones for motif finding in the entire dataset (47), which might compromise the accuracy of motif finding and complicate subsequent data integration. Second, although the distances and interactions between TFBSs in a CRM are critical, both were not considered in our earlier scoring functions (47,48), potentially limiting the accuracy of predicted CRMs. Third, the earlier 'branch-and-bound' approach to integrate motifs found in different datasets was not efficient enough to handle a much larger number of motifs found in an ever-increasing number of large ChIP-seq datasets from human cells/tissues (47,48). To overcome these hurdles, we developed dePCRM2 based on an ultrafast, accurate motif-finder ProSampler (73), a novel effective combinatory motif pattern discovery method, and scoring functions that model essentials of both the enhanceosome and billboard models of CRMs (76–78). Using available 6,092 ChIP-seq datasets, dePCRM2 was able to partition the genome regions covered by extended binding peaks into a CRM candidate (CRMC) set and a non-CRMC set, and to predict 201 unique TF binding motif families in the CRMCs. Both evolutionary and independent experimental data indicate that at least the vast majority of the predicted 1,404 973 CRMCs might be *cis*-regulatory, while at least the vast majority of the predicted non-CRMCs might not be.

MATERIALS AND METHODS

Datasets

We downloaded 6092 TF ChIP-seq datasets from the Cistrome database (29) (Supplementary Table S1). The binding peaks in each dataset were called using a pipeline for uniform processing (29). We filtered out binding peaks with a read depth score <20. For each binding peak in each dataset, we extracted a 1000 bp genome sequence centering on the middle of the summit of the binding peak, as this length yielded the best results among all the lengths we tested (200, 500, 1000, 1500 and 3000 bp) for finding both the primary motif of the ChIP-ed TF and its collaborative motifs (73). We downloaded 976 experimentally verified enhancers from the VISTA database (79), 790,888 ClinVar SNVs from the ClinVar database (80), 32 689 enhancers (81) and 184 424 promoters (82) from the FANTOM5 project website, 255 937 GWAS SNVs from GWAS Catalog (83), and 122 468 173 DHSs in 1353 datasets (Supplementary Table S2), 29 520 736 transposase-accessible sites (TASs) in 1059 datasets (Supplementary Table S3), 99 974 447 H3K27ac peaks in 2539 datasets (Supplementary Table S4), 77 500 232 H3K4me1 peaks in 1210 datasets (Supplementary Table S5), and 70 591 888 H3K4me3 peaks in 2317 datasets (Supplementary Table S6) from the Cistrome database (29). For ClinVar and GWAS SNVs, we excluded those in CDs, leaving 208 065 (26.31%) and 234 016 (91.44%) SNVs, respectively, in NCSs for the analysis. We downloaded human protein–protein interaction datasets from the BioGRID (84) and reactome (85) databases.

Measurement of the overlap between two different datasets

To evaluate the extent to which the binding peaks in two datasets overlap with each other, we calculate an overlap score $S_0(d_i, d_j)$ between each pair of datasets d_i and d_j, defined as,

$$S_0\ (d_i, d_j) = \frac{1}{2} \times \left(\frac{o\,(d_i, d_j)}{|d_i|} + \frac{o\,(d_i, d_j)}{|d_j|} \right), \qquad (1)$$

where $o(d_i, d_j)$ is the number of binding peaks in d_i and d_j that overlap each other by at least one bp.

Identification of collaborative TF modules

We construct a graph using the TFs as the nodes and connecting two nodes with an edge if the two corresponding TFs physically interact with each other according to the BioGRID (84) and reactome (85) databases. Then we cut the graph into smaller community using the 'node perception' program in CDlib, which identifies overlapping network communities in a graph using local group information (86). We consider each resulting community with at least three TFs as physical interacting module.

Parameters for accuracy evaluation

We use the following definitions to evaluate the accuracy of predictions. Sensitivity = recall rate = TPR (true positive rate) = $\frac{TP}{TP+FN}$, FNR (false negative rate) = $\frac{FN}{TP+FN}$, Specificity = $\frac{TN}{FP+TN}$, FPR (false positive rate) = $\frac{FP}{FP+TN}$, FDR (false discorery rate) = $\frac{FP}{TP+FP}$, precision = $\frac{TP}{TP+FP}$, and FOR (false ommision rate) = $\frac{FN}{FN+TN}$, where TP is true positives; FN, false negatives; FP, false positives; and TN, true negatives.

Statistical tests

We used Mann–Whitney U test to evaluate the significance of difference between the means of two samples, χ^2-test to evaluate the significance of difference between ratios/proportions in samples, Kolmogorov–Smirnov (K–S) test to evaluate the significance of difference between the distributions of two random variables, and hypergeometric test to evaluate the overlap between two sets of elements such as TFs. To evaluate the statistical significance of clusters of overlapping datasets in Figure N1E (Supplementary Note), we compute a p value to reject the null hypothesis that a cluster for k ChIP-ed TFs is generated by chance, using a multinomial distribution,

$$p(x_1,\ x_i, \ldots,\ x_k | p_1,\ p_i, \ldots,\ p_k,\ N) = 1 - \sum_{j_1=0}^{x_1-1} \cdots \sum_{j_i=0}^{x_i-1} \cdots \sum_{j_k=0}^{x_k-1} \frac{N!}{j_1! \ldots j_i! \ldots j_k!} p_1^{j_1} \cdots p_i^{j_i} \cdots p_k^{j_k}, \qquad (2)$$

where x_i is the number of datasets for the ith ChIP-ed TF in the cluster, and p_i is the probability/frequency of datasets of the ith ChIP-ed TF in the entire set of N datasets. We used the Benjamini–Hochberg (BH) procedure to correct p values for multi-hypothesis tests.

The dePCRM2 pipeline

Step 1: Find motifs in each dataset using ProSampler (73) (Figure 1A and B).

Step 2. Compute pairwise motif co-occurring scores and find co-occurring motif pairs (CPs): As true TFBSs are more likely to co-occur in a binding peak than are spurious ones, to filter out false positive sites, we find overrepresented CPs in each dataset (Figure 1C). Specifically, for each pair of motifs $M_d(i)$ and $M_d(j)$ in each data set d, we compute their co-occurring scores S_c defined as,

$$S_c\ \left(M_i\,(i)\,,\ M_j\,(j)\right) = \frac{o\,(M_d\,(i)\,,\ M_d\,(j))}{max\,\{|M_d\,(i)|\,,\ |M_d\,(j)|\,\}}, \qquad (3)$$

where $|M_d(i)|$ and $|M_d(j)|$ are the number of binding peaks containing TFBSs of motifs $M_d(i)$ and $M_d(j)$, respectively; and $o(M_d(i),\ M_d(j))$ the number of binding peaks containing TFBSs of both the motifs in d. We identify CPs with an $S_c \geq \beta$. We choose β such that the component with the highest scores in the trimodal distribution S_c is kept (Figures 1C and 2B) (by default $\beta = 0.7$).

Step 3. Construct a motif similarity graph and find unique motifs (UMs): We combine highly similar motifs in

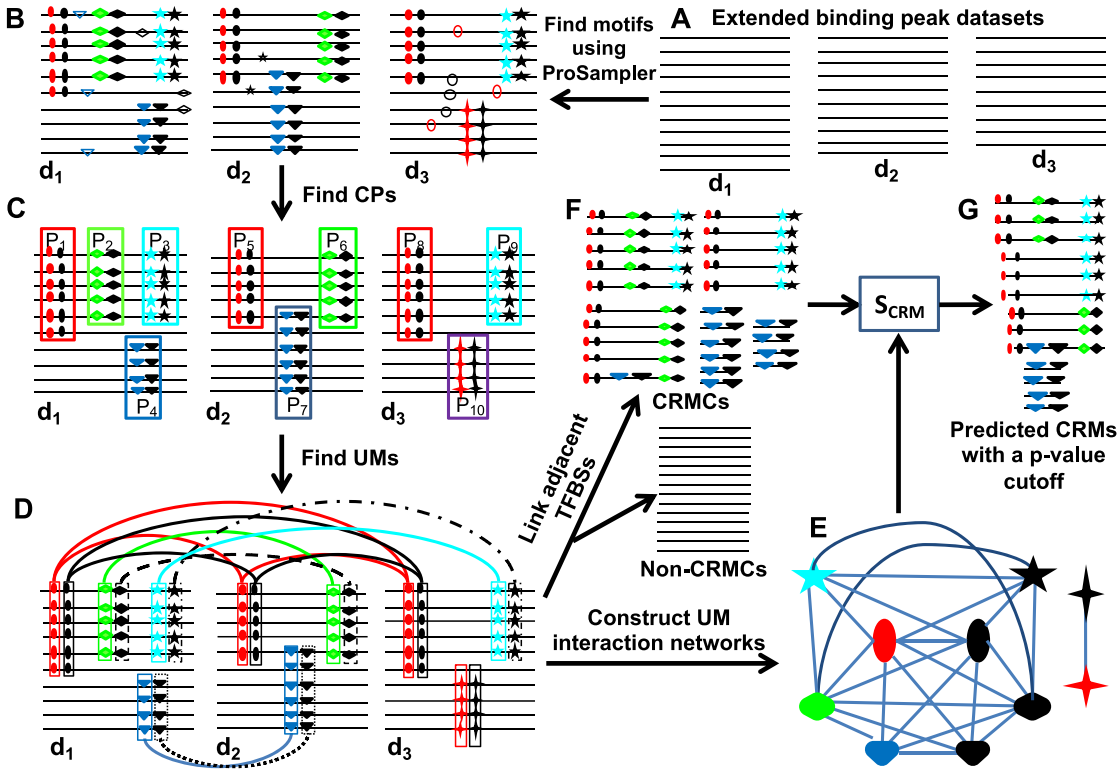

Figure 1. Schematic of the dePCRM2 pipeline. (**A**) Extend each binding peak in each dataset (d_1, d_2 and d_3) to its two ends to reach a preset length, i.e. 1000 bp. Each line represents an extended binding peak. (**B**) Find motifs in each dataset using ProSampler. The labels on a line represents an identified TFBS in the peak, and labels with the same shape and color represent an identified motifs in the dataset (**C**). Find CPs in each dataset. A pair of motifs in a rectangle is an identified CPs in the dataset. For clarity, only the indicated CPs are shown, while those formed between motifs in pairs P_1 and P_2 in dataset d_1, and so on, are omitted. (**D**) Construct the motif similarity graph, cluster similar motifs and find UMs in the resulting motif clusters. Each node in the graph is a motif, weights on the edges are omitted for clarity. Identified motif clusters are connected by edges of the same color and line type. (**E**) Construct UM interaction networks. Each node in the networks is a UM, weights on the edges are omitted for clarity. (**F**) Project binding sites in the UMs back to the genome and link adjacent TFBSs along the genome, thereby identifying CRMCs and non-CRMCs. (**G**) Evaluate each CRMC by computing its S_{CRM} score and the associated *P*-value.

the CPs from different datasets to form a UM that is presumably recognized by a TF or highly similar TFs of the same family/superfamily (87). Specifically, for each pair of motifs $M_a(i)$ and $M_b(j)$ from different datasets a and b, respectively, we compute their similarity score S_s using the SPIC metric (88). We then build a motif similarity graph using motifs in the CPs as the nodes and connecting two motifs with their S_s being the weight on the edge, if and only if (iff) $S_s > \beta$ (Figure 1D). By default, we set $\beta = 0.8$, at which the similarity between the motifs is highly significant (TOMTOM q-value < 0.05) (47,75,88). We apply the Markov cluster (MCL) algorithm (89) to the graph to identify dense subgraphs as clusters. For each cluster, we merge overlapping sequences, extend each sequence to a length of 30 bp by padding the same number of nucleotides from the genome to the two ends, and then realign the sequences to form a UM using ProSampler (73) (Figure 1D).

Step 4. Construct the interaction networks of the UMs/TFs: TFs tend to repetitively collaborate with each other to regulate genes in different contexts by binding to cognate TFBSs in CRMs. The relative distances between TFBSs in a CRM often do not matter (billboard model), but sometimes they are constrained by the interactions between cognate TFs (enhanceosome model) (76–78). To model essential features of both the scenarios, we compute an interaction score between each pair of UMs, U_i and U_j, defined as,

$$S_{\text{INTER}}\left(U_i, U_j\right) = \frac{1}{\left|D\left(U_i, U_j\right)\right|} \sum_{d \in D\left(U_i, U_j\right)} \left(\frac{1}{\left|d\left(U_i\right)\right|} + \frac{1}{\left|d\left(U_j\right)\right|}\right) \sum_{s \in S\left(d\left(U_i\right), \left(d\left(U_j\right)\right)\right)} \frac{150}{r\left(s\right)}, \quad (4)$$

where $D(U_i, U_j)$ is the datasets in which TFBSs of both U_i and U_j occur, $d(U_k)$ the subset of dataset d, containing at least one TFBS of U_k, $S(d(U_i), (d(U_j))$ the subset of d containing TFBSs of both U_i and U_j, and $r(s)$ the shortest distance between any TFBS of U_i and any TFBS of U_j in a sequence $s \in S(d(U_i), (d(U_j))$. We construct UM/TF interaction networks using the UMs as the nodes and connecting two nodes with their S_{INTER} being the weight on the edge (Figure 1E). Therefore, the S_{INTER} score allows flexible adjacency and orientation of TFBSs

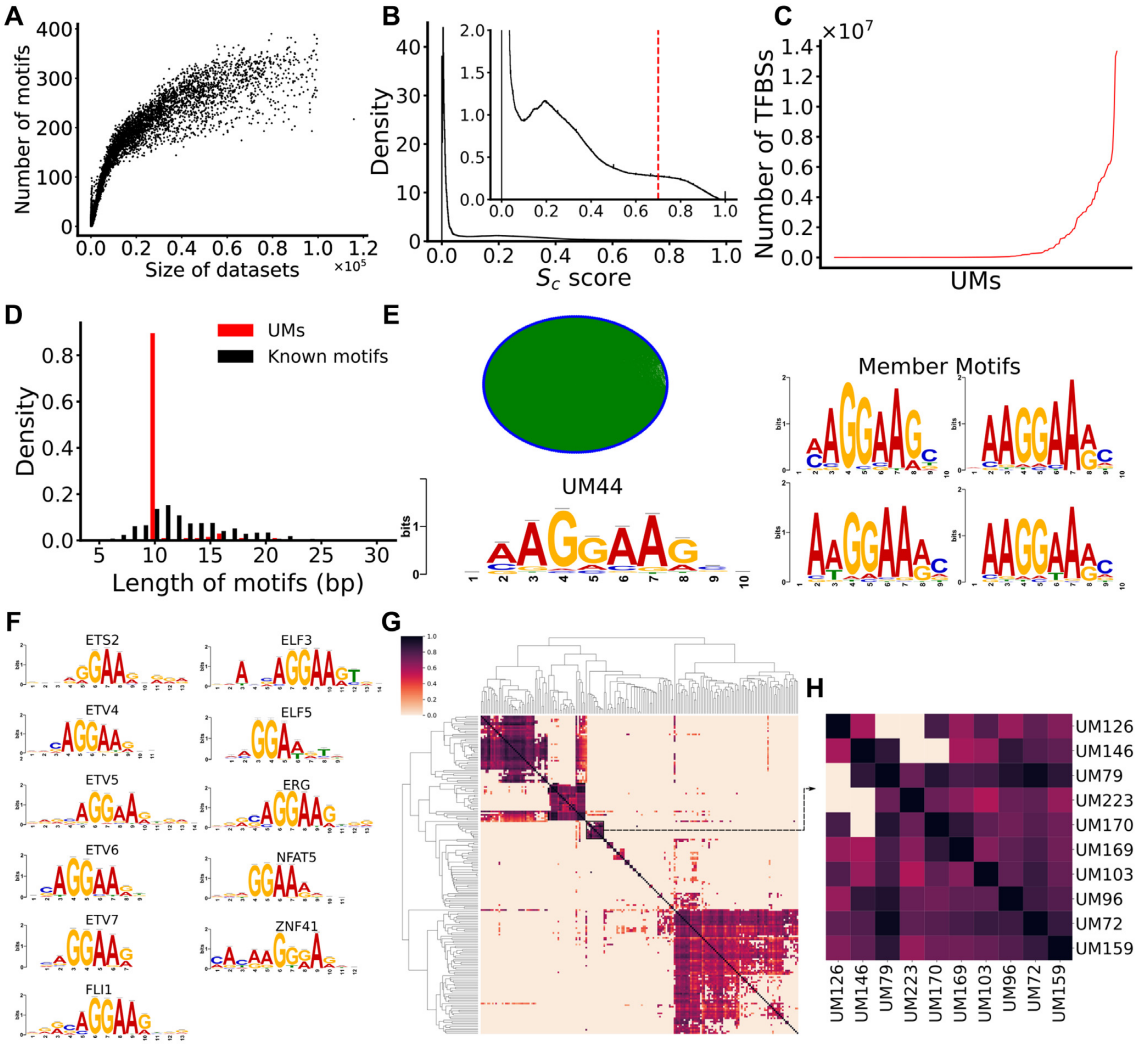

Figure 2. Prediction of UMs. (**A**) Relationship between the number of predicted motifs in a dataset and the size of the dataset (number of binding peaks in the dataset). The datasets are sorted in ascending order of their sizes. (**B**) Distribution of cooccurrence scores (S_c) of motif pairs found in a dataset. The inset is a blowup view of the region defined by the vertical axis. The dashed vertical line indicates the cutoff value (0.7) of S_c for predicting cooccurring pairs (CPs). (**C**) Number of putative binding sites in each of the UMs sorted in ascending order. (**D**) Distribution of the lengths of the UMs and known motifs in the HOCOMOCO and JASPAR databases. (**E**) Logo and similarity graph of the 250 member motifs of UM44. In the graph, each node in blue represents a member motif, and two member motifs are connected by an edge in green if their similarity is greater than 0.8 (SPIC score). Four examples of member motifs are shown in the right panel. (**F**) UM44 matches known motifs of nine TFs of the 'ETS', 'NFAT-related factor', and 'more than three adjacent zinc finger factors' families. (**G**) Heatmap of the interaction networks of the 201 UMs, names of the UMs are omitted for clarity. (**H**) A blowup view of the indicated cluster in G, formed by 10 UMs, of which UM126, UM146, UM79, UM223, UM170, UM103 and UM159 match known motifs of MESP1|ZEB1, TAL1::TCF3, ZNF740, MEIS1|TGIF1|MEIS2|MEIS3, TCF4|ZEB1|CTCFL|ZIC1|ZIC4|SNAI1, GLI2|GLI3 and KLF8, respectively. Some of these TFs are known collaborators in transcriptional regulation.

in a CRM (billboard model) and at the same time, it rewards motifs with binding sites co-occurring frequently in a shorter distance in a CRM (enhanceosome model), particularly within a nucleosome with a length of about 150 bp (76,77,90).

Step 5. Partition the covered regions into a CRM candidate (CRMC) set and a non-CRMC set: We project TFBSs of each UM back to the genome, and link two adjacent TFBSs if their distance $d \leq \varepsilon$. The resulting linked sequence segments are CRMCs, while sequence segments in the covered regions that cannot be linked are non-CRMCs (Figure 1F). By default, $\varepsilon = 300$ bp (roughly the length of two nucleosomes).

Step 6. Evaluate each CRMC: We compute a CRM score for a CRMC containing n TFBSs ($b_1, b_2, \ldots, b_n$), defined as,

$$S_{\mathrm{CRM}}(b_1, b_2 \cdots, b_n) = \frac{2}{n-1} \times \sum_{i-1}^{n} \sum_{j>i} S_{\mathrm{INTER}} \left[U(b_i), U(b_j)\right] \times \left[S(b_i) + S(b_j)\right], \quad (5)$$

where $U(b_k)$ is the UM of TFBS b_k, $S_{\mathrm{INTER}}[U(b_i), U(b_j)]$ the weight on the edge between $U(b_i)$ and $U(b_j)$, in the interaction networks, and $S(b_k)$ the score of b_k based on the position weight matrix (PWM) of $U(b_k)$. Only TFBSs with a positive score are considered. Thus, S_{CRM} considers the

number of TFBSs in a CRMC, as well as their quality and strength of all pairwise interactions.

Step 7. Predict CRMs: We create Null interaction networks by randomly rewiring the interaction networks constructed in Step 4. For each CRMC, we generate a Null CRMC that has the same length and nucleotide compositions as the CRMC using a third order Markov chain model (73). We compute a S_{CRM} score for each Null CRMC using the Null interaction networks, and the binding site positions and PWMs of the UMs in the corresponding CRMC. Based on the distribution of the S_{CRM} scores of the Null CRMCs, we compute an empirical *P*-value for each CRMC, and predict those with a *P*-value smaller than a preset cutoff as CRMs in the genome (Figure 1G).

Step 8. Predict functional states of CRMs in a given cell type: For each predicted CRM, we predict it in a cell type to be: (i) active (TF binding), if it overlaps the summit of a called binding peak containing at least a TFBS of a ChIP-ed TF in the cell type; (ii) non-active (no TF binding), if it contains at least a TFBS of a ChIP-ed TFs in the cell type, but the TFBS does not overlap the summit of any binding peaks of the ChIP-ed TFs and (iii) undetermined (UD), if it does not overlap the summit of any binding peak available in the cell type, because it might be bound by some TFs that have not been ChIP-ed in the cell/tissue type.

Generation of control sequences for validation of predicted CRMs

To create a set of control sequences for validating the predicted CRMs using experimentally determined elements used in Figure 5A, for each predicted CRMC, we randomly selected a sequence segment with the same length as the CRMC from the genome regions covered by the extended binding peaks. To calculate the S_{CRM} score of a control sequence, we assigned it the TFBS positions and their UMs according to those in the counterpart CRMC. Thus, the control set contains the same number and length of sequences with similar nucleotide compositions as in the CRMCs, but with arbitrarily assigned TFBSs and UMs.

RESULTS

The dePCRM2 pipeline

TFs in higher eukaryotes tend to collaboratively bind to their TFBSs in CRMs (1). Different CRMs of the same gene are structurally similar and closely located (91). For example, in the locus control region (LCR) of the hemoglobin genes in the mouse genome, multiple enhancers with similar combinations of TFBSs regulate the expression of different hemoglobin genes in different tissues and developmental stages (92). Moreover, functionally related genes are often regulated by the same sets of TFs in different cell types during development and in maintaining physiological homeostasis (1). Due to the clustering nature of TFBSs of collaborative TFs in a CRM, if we extend the called short binding peaks in a TF ChIP-seq dataset from the two ends and reach the typical length of a CRM (500–2000 bp) (79), the extended peaks would have a great chance to include TFBSs of collaborative TFs (47,48,73). For instance, we have shown that extension of binding peaks to 500–1000 bp could substantially increase the chance of finding TFBSs of collaborative TFs of the ChIP-ed TF, while the introduced noise had a little effect on identifying the primary motif of ChIP-ed TF (73). Moreover, if some TFs collaboratively bind a set of CRMs in a cell/tissue type, or even in different cell/tissue types due to the reutilization of the CRMs, then at least some of the extended peaks of datasets for these TFs from these cell types might contain their cognate TFBSs, and even have some overlaps. Therefore, if we have a sufficiently large number of ChIP-seq datasets for diverse TFs and from diverse cell types, we are likely to include datasets for some collaborative TFs, and their TFBSs may co-occur in some extended peaks that are at least parts of CRMs. Based on these observations, we designed dePCRM (47,48) and dePCRM2 to predict CRMs and constituent TFBSs by identifying overrepresented co-occurring patterns of motifs found by a motif-finder in a large number of TF ChIP-seq datasets. dePCRM2 overcomes the aforementioned shortcomings of dePCRM as follows. First, using an ultrafast and accurate motif-finder ProSampler (73), we can find significant motifs in available ChIP-seq datasets of any size (Figure 1A and B) without the need to split large datasets into small ones (47). Second, after identifying highly co-occurring motifs pairs (CPs) in the extended binding peaks in each dataset (Figure 1C), we cluster highly similar motifs in the CPs and find a unique motif (UM) in each resulting cluster (Figure 1D). Third, to improve prediction accuracy, we model distances and interactions among cognate TFs of the binding sites in a CRM by constructing interaction networks of the TFs/UMs based on the cooccurrence of their binding sites and the distance between them (Figure 1E). Fourth, we identify as CRMCs closely located clusters of binding sites of the UMs along the genome (Figure 1F), thereby partitioning genome regions covered by the extended binding peaks (covered regions, hereafter) into a CRMCs set and a non-CRMCs set. Fifth, we evaluate each CRMC using a novel score that considers not only the number of TFBSs in a CRM, but also the distances between the TFBSs, their quality scores and all pairwise cooccurring frequencies between their motifs (Figure 1G). Lastly, we compute a *P*-value for each S_{CRM} score, so that CRMs and constituent TFBSs can be predicted at different significant levels using different S_{CRM} score or *P*-value cutoffs. Clearly, as the number of UMs is a small constant number constrained by the number of TF families encoded in the genome, the downstream computation based on the set of UMs runs in a constant time, dePCRM2 is highly scalable. The source code of dePCRM2 is available at http://github.com/zhengchangsulab/pcrm2

Unique motifs recall most known TF motif families and have distinct patterns of interactions

ProSampler identified at least one motif in 5991 (98.70%) of the 6092 ChIP-seq datasets (Supplementary Note) but failed to find any motifs in the remaining 101 (1.66%) datasets that all contain <310 binding peaks (Supplementary Table S1), indicating that they are likely of low quality. As shown in Figure 2A, the number of motifs found in a dataset generally increases with the increase in the number of binding peaks in the dataset, but enters a saturation

phase and stabilizes around 250 motifs when the number of binding peaks is beyond 40 000. In total, ProSampler identified 856 793 motifs in the 5991 datasets. dePCRM2 found co-occurring motif pairs (CPs) in each dataset (Figure 1C) by computing a cooccurring score S_c for each pair of motifs in the dataset (formula 3). As shown in Figure 2B, S_c scores show a trimodal distribution. We found that the first and second low-scoring components were largely due to low-scoring spurious motifs with low information content, or due probably to by-chance cooccurrence of motifs as indicated by their low S_c scores. The high scoring third component is likely due to cooccurrence of collaborative motifs, although there is no clear-cut valley between it and the second component. To find the optimal S_c cutoff between the second and third components, thereby largely separating true CPs from spurious ones, we tested different values ranging from 0.6 to 0.8. We found that at the S_c cutoff of 0.7, dePCRM2 identified 4 455 838 CPs containing 226 355 (26 42%) motifs, and these motifs had the highest proportion (24%) matching one of the 856 annotated motifs in the HOCOMOCO (93) and JASPAR (94) databases using TOMTOM (95) (q-value $\leq$ 0.1). Using this optimal S_c cutoff, we filtered out 630 438 (73.58%) of possible spurious motifs found in the 5991 datasets. Clustering the 226 355 motifs in the CPs resulted in 245 clusters, each consisting of 2∼72 849 motifs, most of which form a complete similarity graph or a clique (Supplementary Figure S1A), indicating that member motifs in a cluster are highly similar to each other. dePCRM2 found a UM in 201 (82.04%) of the 245 clusters (Supplementary Figure S1B and Table S7), but failed to do so in 44 clusters due to the low similarity between some member motifs (Supplementary Figure S1A). Binding sites of the 201 UMs were found in 39.87–100% of the sequences in the corresponding clusters, and in only 1.49% of the clusters binding sites were not found in more than 50% of the sequences due to the low quality of member motifs (Supplementary Figure S2). Thus, this step retained most of putative binding sites in most clusters. The UMs contain highly varying numbers of binding sites ranging from 64 to 13 672 868 with a mean of 905 288 (Figure 2C and Supplementary Table S7), reminiscent of highly varying number of binding peaks in the datasets (Supplementary Note). The lengths of the UMs range from 10 to 21 bp with a mean of 11 bp (Figure 2D), which are in the range of the lengths of known TF binding motifs, although they are biased to 10 bp due to the limitation of the motif-finder to find longer motifs. As expected, a UM is highly similar to its member motifs, which also are highly similar to each other (Supplementary Figure S1A). For example, UM44 contains 250 highly similar member motifs (Figure 2E). Of the 201 UMs, 117 (58.2%) match (TOMTOM q-value $\leq$ 0.05) at least one of the 856 annotated motifs, and 92 (78.63%) match at least two (Supplementary Table S7), suggesting that most UMs might consist of motifs of different TFs of the same TF family/superfamily that recognize highly similar motifs, a well-known phenomenon (96,97), and a UM might represent a motif family/superfamily for the cognate TF family/superfamily. For instance, UM44 matches known motifs of nine TFs of the 'ETS' family ETV4∼7, ERG, ELF3, ELF5, ETS2 and FLI1, a known motif of NFAT5 of the 'NFAT-related factor' family, and a known motif of ZNF41 of the 'more than three adjacent zinc finger factors' family (Figure 2F and Supplementary Table S7). The high similarity of these motifs suggest that they might form a superfamily. The remaining 84 (43.28%) of the 201 UMs might be novel motifs recognized by unknown cognate TFs (Supplementary Figure S1B and Table S7). On the other hand, 64 (71.91%) of the 89 annotated TF motif families match (TOMTOM q-value $\leq$ 0.05) one of the 201 UMs (Supplementary Table S8), thus, our predicted UMs include most of the known TF motif families.

To model interactions between cognate TFs of the UMs, we computed an interaction score S_{INTER} based on cooccurrence levels and distances between binding sites of two UMs (formula 4), which largely improves our earlier score (data not shown) that only considers cooccurring frequencies of binding sites in two motifs (47,48). As shown in Figure 2G, there are clear interaction patterns between putative cognate TFs of many UMs, many of which are supported by experimental evidence. For example, in a cluster formed by 10 UMs (Figure 2H), seven of them (UM126, UM146, UM79, UM223, UM170, UM103 and UM159) match known motifs of MESP1/ZEB1, TAL1::TCF3, ZNF740, MEIS1/TGIF1/MEIS2/MEIS3, TCF4/ZEB1/CTCFL/ZIC1/ZIC4/SNAI1, GLI2/GLI3 and KLF8, respectively. At least a few of them are known collaborators in transcriptional regulation. For example, GLI2 collaborates with ZEB1 to repress the expression of CDH1 in human melanoma cells via directly binding to two close binding sites in the CDH1 promoter (98); ZIC and GLI collaboratively regulate neural and skeletal development through physical interactions between their zinc finger domains (99); and ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate the expression of WNT (100), to name a few.

Appropriate extension of original binding peaks greatly increases the power of datasets

By connecting closely located binding sites of the UMs along the genome, dePCRM2 partitioned the covered regions that is 77.47% of the mappable genome (Supplementary Note) in two exclusive sets (Figure 1F). To find the optimal minimal length ε for linking adjacent putative TFBSs (MATERIALS AND METODS), we tested ε = 100, 150, 250, 300 and 350 bp, and found that 150–300 bp yielded similar results in terms of the distinct evolutionary behaviors of the resulting CRMCs and non-CRMs (see below). Therefore, we chose the longest value, i.e. ε = 300 bp in the current application. The resulting CRMC set contain 1 404 973 CRMCs with a total length of 1 359 824 275 bp (56.84%) covering 44.03% of the genome, and the resulting non-CRMC set contain 1 957 936 sequence segments with a total length of 1 032 664 424 bp (43.16%) covering 33.44% of the genome (Figure 3A). Interestingly, dePCRM2 only predicts 62.18% of nucleotide positions covered by original binding peaks to be CRMs (kept original), while abandoning the remaining 37.82% of nucleotide positions as non-CRMCs (abandoned original) (Figure 3A). The kept original positions account for 57.87% (776 999 862 bp) of genome positions of the predicted CRMCs (Figure 3A). The abandoned original positions might be not en-

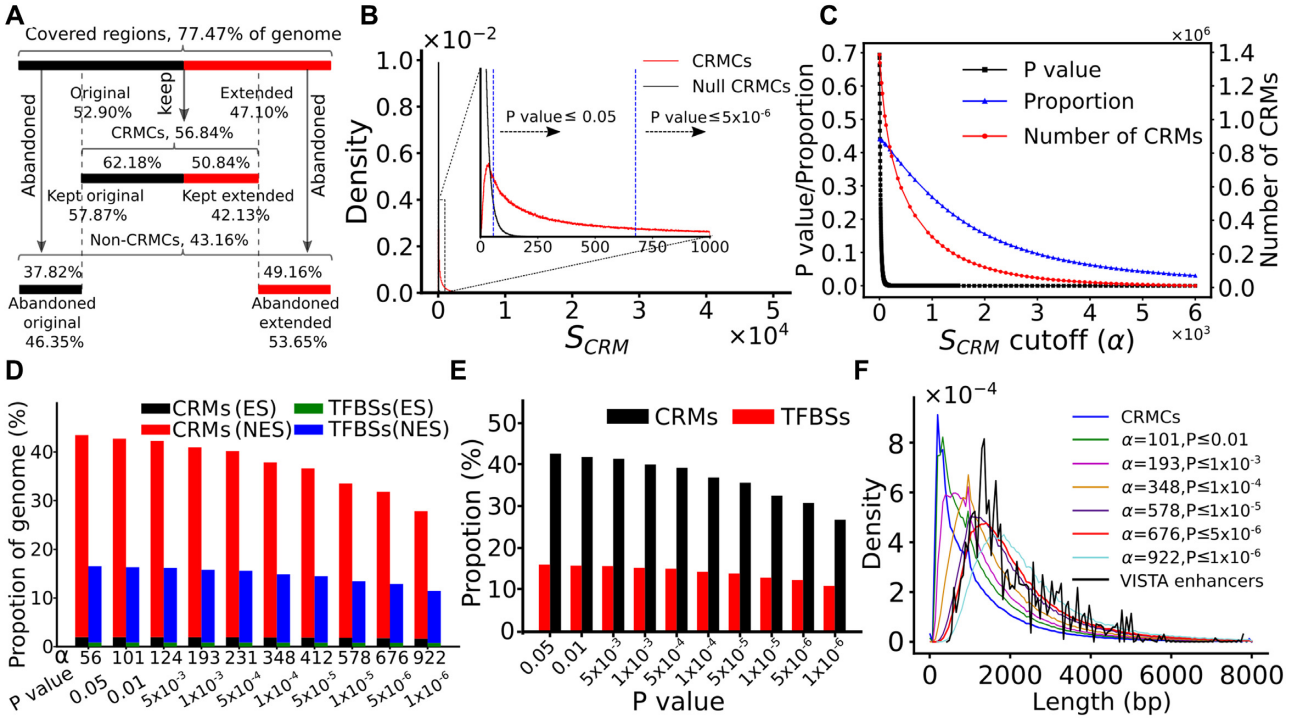

Figure 3. Prediction of CRMCs and CRMs. (**A**) Cartoon shows the proportions of the 77.47% of genome covered by originally called binding peaks and their extended parts as well as their relative contributions to the predicted CRMs (kept original and kept extended, respectively) and non-CRMCs (abandoned original and abandoned extended, respectively) (percentages below the lines). Percentage above the lines are the proportion of originally called binding peak and their extended parts that are predicted to be CRMCs and non-CRMCs. (**B**) Distribution of S_{CRM} scores of the CRMCs and the Null CRMCs. The inset is a blowup view of the indicated region. The vertical dashed lines indicate the associated *P*-values of the S_{CRM} cutoffs mentioned in the main text. (**C**) Number of CRMs predicted, proportion of the genome predicted to be CRMs and the associated *P*-value as functions of the S_{CRM} cutoff α. (**D**) Proportion of the genome predicted to be CRMs and TFBSs in ESs and NESs using various S_{CRM} cutoffs and associated *P*-values. (**E**) Proportion of NESs that are predicted to be CRMs and TFBSs using various S_{CRM} cutoffs and associated *P*-values. (**F**) Distribution of the lengths of CRMs predicted using different S_{CRM} cutoffs and associated *P*-values.

riched for TFBSs, which is in agreement with earlier findings about the noisy nature of TF ChIP-seq data (101–103). On the other hand, dePCRM2 predicts 50.84% of nucleotide positions only covered by extended parts of original binding peaks to be CRMCs (kept extended), while discarding the remaining 49.16% of nucleotide positions as non-CRMCs (abandoned extended) (Figure 3A). The kept extended positions account for the remaining 42.12% (565 448 583 bp) genome positions of the predicted CRMCs (Figure 3A), suggesting that TFBSs of collaborative TFs might be indeed enriched in some extended parts as has been shown earlier (47,48,72,73), and that dePCRM2 is able predict CRMs that are not covered by any binding peaks. Therefore, by appropriately extending original binding peaks, we could greatly increase the power of datasets. Based on the overlap between a CRMC and original binding peaks containing a binding site of ChIP-ed TFs in a cell/tissue type (Materials and Methods), dePCRM2 predicts functional states of 57.88% of the CRMCs in at least one of the cell/tissue types from which binding peaks were available in the datasets. However, dePCRM2 is not able to predict the functional states of the remaining 42.12% of the CRMCs that do not overlap any original binding peaks in the datasets. The predicted CRMCs and constituent TFBSs are available at https://cci-bioinfo.uncc.edu/

The CRMCs are unlikely predicted by chance

To further evaluate the predicted CRMCs, we computed a S_{CRM} score for each CRMC (formula 5). As shown in Figure 3B, the distribution of the S_{CRM} scores of the CRMCs is strongly right-skewed relative to that of the Null CRMCs (Materials and Methods), indicating that the scores of CRMCs are generally much higher than those of the Null CRMCs, thus, the CRMCs are unlikely produced by chance. Based on the distribution of the S_{CRM} scores of the Null CRMCs, dePCRM2 computes a *P*-value for each CRMC (Figure 3B). With the increase in the S_{CRM} cutoff α ($S_{CRM} \geq \alpha$), the associated *P*-value cutoff drops rapidly, while both the number of predicted CRMs and the proportion of the genome covered by the predicted CRMs decrease slowly (Figure 3C), indicating that most CRMCs have low *P*-values. For instance, with α increasing from 56 to 922, *P*-value drops precipitously from 0.05 to 1.00×10^{-6} (5×10^5 fold), while the number of predicted CRMs decreases from 1,155,151 to 327,396 (3.53 fold), and the proportion of the genome covered by predicted CRMs decreases from 43.47% to 27.82% (1.56 fold) (Figure 3C). Predicted CRMs contain from 20,835,542 (*P*-value $\leq 1 \times 10^{-6}$) to 31,811,310 (*P*-value ≤ 0.05) non-overlapping putative TFBSs that consist of from 11.47% (*P*-value$\leq 1 \times 10^{-6}$) to 16.54% (*P*-value ≤ 0.05) of the genome (Figure 3D).

In other words, dependent on *P*-value cutoffs (1×10^{-6}–0.05), 38.05–41.23% of nucleotide positions of the predicted CRMs are made of putative TFBSs (Figure 3D), and most of predicted CRMs (93.99∼95.46%) and constituent TFBSs (93.20–94.67%) are located in non-exonic sequences (NESs), comprising 26.66–42.47% and 10.94–16.03% of NESs, respectively (Figure 3E). Surprisingly, dependent on *P*-value cutoffs (1×10^{-6}–0.05), the remaining 5.33–6.80% and 4.54–6.01% of the predicted CRMs and constituent TFBSs, respectively, are in exonic sequences (ESs, including CDSs, 5′- and 3′-untranslated regions), respectively (Figure 3D), in agreement with an earlier report (104).

The S_{CRM} score captures the length feature of enhancers

As shown in Figure 3F, the CRMCs with a mean length of 981bp are generally shorter than VISTA enhancers with a mean length of 2,049bp. Specifically, 621 842 (44.26%) of the 1 404 973 CRMCs are shorter than the shortest VISTA enhancer (428 bp), suggesting that they might be short CRMs (such as promoters or short enhancers) or components of long CRMs. However, these shorter CRMCs (<428 bp) comprise only 7.42% of the total length of the CRMCs. As remaining 733 132 (55.74%) CRMCs comprising 92.58% of the total length of the CRMCs are longer than the shortest VISTA enhancer (428 bp), most of them are likely full-length CRMs, and CRMC positions are mainly covered by full-length or longer CRMCs. As expected, with the increase in α (decrease in *P*-value cutoff), the distribution of the lengths of predicted CRMs shifts to right and even surpass that for VISTA enhancers (Figure 3F), indicating shorter CRMCs can be effectively filtered out by a higher S_{CRM} cutoff α (a smaller *P*-value). For instance, at a rather stringent S_{CRM}cutoff $\alpha = 676$ ($P = 5 \times 10^{-6}$), we filtered out 976 345 (69.49%) shorter CRMCs with a mean length of 387 bp (Figure 3F), and the remaining 428,628 (30.51%) CRMCs have similar length distribution (mean length of 2292 bp) to that of VISTA enhancers (mean length of 2049 bp) (Figure 3F). VISTA enhancers are mainly involved in development-related functions and are generally longer than other types of enhancers (105). However, it is worth noting that a VISTA enhancer may not necessarily be in its full-length form, because even a portion of an enhancer could be still partially functional (1,106), and it is still technically difficult to validate very long enhancers in transgene animal models in a large scale. Therefore, it is not surprising that with even more stringent S_{CRM} cutoffs, the predicted CRMs could be longer than VISTA enhancers (Figure 3F, and see later). Taken together, these results suggest that the S_{CRM} score captures the length feature of enhancers.

The CRMCs and non-CRMCs show dramatically distinct evolutionary behaviors

To see how effectively dePCRM2 partitions the covered regions into the CRMC set and the non-CRMC set, we compared their evolutionary behaviors with those of the entire set of 976 VISTA enhancers using the GERP (107) and phyloP (108) scores of their nucleotide positions in the genome. Both the GERP and the phyloP scores quantify conservation levels of genome positions based on nucleotide substitutions in alignments of multiple vertebrate genomes. The larger a positive GERP or phyloP score of a position, the more likely it is under negative/purifying selection; and a GERP or phyloP score around zero means that the position is selectively neutral or nearly so (107,108). Although a negative phyloP score is related to positive selection (108), a negative GERP score is cautiously so (109). For convenience of discussion, we consider a position with a GERP or phyloP score within an interval centering on 0 [–δ,+ δ] ($\delta > 0$) to be selectively neutral or nearly so, and a position with a score greater than δ to be under negative selection. We define proportion of neutrality of a set of positions to be the size of the area under the density curve of the distribution of their scores within the window [–δ,+ δ]. Because ESs evolve quite differently from NESs, we focused on the CRMCs and constituent TFBSs in NESs, and left those that overlap ESs in another analysis (Jing Chen, Pengyu Ni, Jun-tao Guo and Zhengchang Su). The choice of $\delta =$ 0.1, 0.2, 0.3, 0.4,0.5, 1 and 2 gave similar results (data not shown), so we chose $\delta = 1$ in the subsequent analyses. As shown in Figure 4A, GERP scores of VISTA enhancers show a trimodal distribution with a small peak around score -5, a blunt peak around score 0, a sharp peak around score 3.5, and a small proportion of neutrality of 0.23, indicating that most nucleotide positions of VISTA enhancers are under strong evolutionary selection, particularly, 37% of the positions are under strong purifying selection with a GERP score >1. This result is consistent with the fact that VISTA enhancers are mostly ultra-conserved (110), development-related enhancers (111,112). The 0.23 proportion of neutrality of the VISTA enhancer positions indicates that this proportion of positions might simply serve as non-functional spacers between adjacent TFBSs. In contrast, the distribution of the GERP scores of the non-CRMCs (1 034 985 426 bp) in NESs displays a sharp peak around score 0, with low right and left shoulders, and a higher proportion of neutrality of 0.71 than do VISTA enhancers (2.2×10^{-302}, χ^2-test) (Figure 4A), suggesting that most positions of the non-CRMCs are selectively neutral or nearly so, and thus are likely to lack *cis*-regulatory functions. The remaining 0.29 portion of positions of the non-CRMCs seem to be under varying levels of selection with 5% of positions under purifying selection with a GERP score >1 (Figure 4A), suggesting that they might have other functions than *cis*-regulation. Intriguingly, the distribution of the GERP scores of the 1 292 356 CRMCs (1 298 719 954 bp) in NESs has a blunt peak around score 0, with high right and left shoulders, and a smaller proportion of neutrality of 0.31 than the non-CRMCs (0.71) (2.2×10^{-302}, χ^2-test) (Figure 4A), indicating that most positions of the CRMCs are under strong evolutionary selection, and thus, are likely to be functional, while the small proportion (0.31) of neutrality suggests that this proportion of positions in the CRMCs might serve as non-functional spacers between TFBSs. Notably, unlike the distribution for VISTA enhancers, that for the CRMCs lack obvious peaks around scores –5 and 3.5 (Figure 4A), indicating the average selection strength on the CRMCs is weaker than that on VISTA enhancers. For instance, only 14% of CRMC positions are under purifying

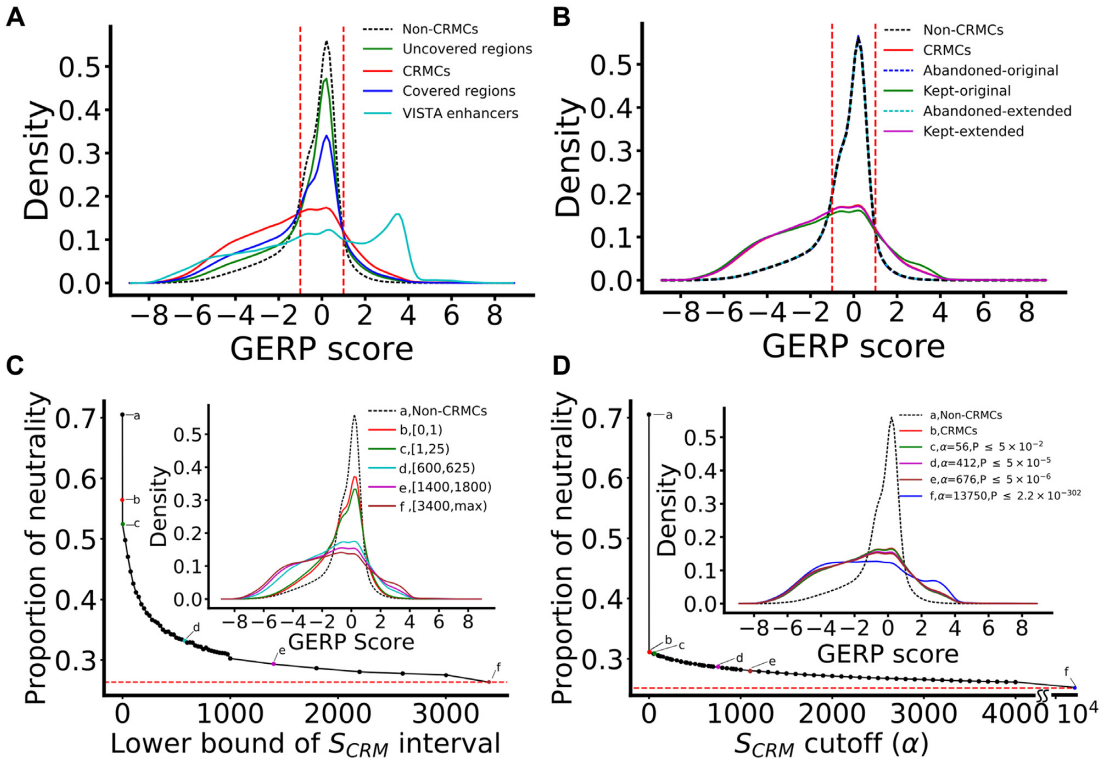

Figure 4. CRMCs and non-CRMCs in NESs show different evolutionary behaviors measured by GERP scores. (**A**) Distributions of GERP scores of nucleotide positions of VISTA enhancers, CRMCs, non-CRMCs, covered regions and uncovered regions. The area under the density curves in the score interval [–1, 1] is defined as proportion of neutrality of the sequences. The distribution for CRMCs is significantly different from that for non-CRMCs, $P < 2.2 \times 10^{-302}$ (K–S test). (**B**) Distributions of GERP scores of nucleotide positions of CRMCs, non-CRMCs, CRMCs, the kept-original, the kept-extended, the abandoned-original and the abandoned-extended. The distributions for kept original positions and kept extended positions are significantly different from those of abandoned original positions and abandoned extended positions, respectively, $P < 2.2 \times 10^{-302}$ (K–S test). (**C**) Proportion of neutrality of CRMCs with a S_{CRM} score in different intervals in comparison with that of the non-CRMCs (a). The inset shows the distributions of GERP scores of the non CRMCs and CRMCs with S_{CRM} scores in the intervals indicted by the colored curves and letters. (**D**) Proportion of neutrality of CRMs predicted using different S_{CRM} score and associated *P*-value cutoffs in comparison with those of the non-CRMCs (a) and CRMCs (b). The inset shows the distributions of GERP scores of the non-CRMCs, CRMCs and the predicted CRMs using the S_{CRM} score and *P*-value cutoffs indicated by the colored curves and letters. The dashed lines in B and C indicate the saturation levels.

selection with a GERP score >1, which is less than half of that (37%) for VISTA enhancers (but see the section 'The higher the S_{CRM} score of a CRMC, the stronger evolutionary constraint it is under'). Nonetheless, this is expected considering the ultra-conversation nature of the small set of development-related VASTA enhancers (110–112). K-S test indicates that the CRMCs and the non-CRMCs have significantly different GERP score distributions (Figure 4A, $P < 2.2 \times 10^{-302}$). By contrast, two sets of sequences generated by randomly partitioning the covered regions with matched numbers and sizes of the CRMCs (CRMC-control) and of the non-CRMCs (non-CRMC-control) with 100 repeats have indistinguishable distributions of GERP scores (Supplementary Figure S3A). These results once again strongly suggest that CRMCs and non-CRMCs cannot be predicted by any chance factors, and that dePCRM2 is able to partition the covered regions into the CRMC set that is likely enriched with sequences with *cis*-regulatory functions, and the non-CRMC set that is likely enriched with sequences without such functions. Similar results were obtained using the phyloP scores, although they display quite different distributions than the GERP scores (Supplementary Figures S3B, S4A).

Interestingly, the uncovered regions have a GERP score distribution and a proportion of neutrality (0.59) in between those of the covered regions (0.49) and those of the non-CRMCs (0.71) (Figure 4A), indicating that the uncovered regions are more evolutionarily selected than the non-CRMCs, but less so than the covered regions. This implies that the uncovered regions contain functional elements such as CRMs, but their density could be lower than that of the covered regions. Assuming that the total length of CRMs in a region is proportional to the total length of evolutionarily constrained parts in the region, we estimate the proportion of uncovered regions that might be CRMs to be (1–0.59)/(1–0.49) = 80.04% of that in the covered regions. Therefore, existing studies and resulting datasets are strongly biased to more evolutionary constrained regions due probably to their large effect sizes and more critical functions that more easily brought the attention of researchers. Similar results were obtained using the phyloP scores (Supplementary Figure S4A).

As we indicated earlier, dePCRM2 predicts 50.84% of nucleotide positions that are only covered by the extended parts of original binding peaks to be CRMCs (kept extended), accounting for 42.12% of CRMC position, while

it predicts 37.82% of nucleotide positions covered by original binding peaks to be non-CRMCs (abandoned original) (Figure 3A). To see why these results are possibly true, we compared the distributions of conservation scores of nucleotide positions of relevant sets. As shown in Figure 4B and Supplementary Figure S4B, kept extended positions have conservation score distributions almost identical to both those of kept original positions (Figure 3A) and those of the entire set of CRMC positions (Figure 4B and Supplementary Figure S4B). As kept extended positions are largely under strong evolutionary constraints, they likely to be *cis*-regulatory. On the other hand, abandoned original positions have conservation score distributions almost identical to both those of abandoned extended positions (Figure 3A) and those of the entire set of non-CRMC positions (Figure 4B and Supplementary Figure S4B). Since abandoned original positions are largely selectively neutral, they are unlikely to be *cis*-regulatory. Taken together, these results indicate that dePCRM2 is able to accurately distinguish *cis*-regulatory and non-*cis*-regulatory sequences in the genome covered by original binding peaks as well as by their extended parts, and further endorse the merit of appropriately extending original binding peaks for more complete prediction of CRMs and TFBSs.

The higher the S_{CRM} score of a CRMC, the stronger evolutionary constraint it is under

To see whether the S_{CRM} score of a CRMC captures the strength of evolutionary selection that it is under, we plotted the distributions of conservation scores of subsets of CRMCs with a S_{CRM} score in different non-overlapping intervals. Remarkably, even the subset with S_{CRM} scores in the lowest interval [0, 1) has a significantly smaller proportion of neutrality (0.56) than the non-CRMCs (0.71) ($P < 2.2 \times 10^{-302}$, χ^2-test) (Figure 4C), indicating that even these low-scoring CRMCs with short lengths (Figure 3F) are more likely to be under strong evolutionary constraints than the non-CRMCs, and thus might be more likely *cis*-regulatory. With the increase in the lower bound of S_{CRM} intervals, proportion of neutrality of the corresponding subsets of CRMCS drops rapidly, followed by a slow linear decrease around the interval [1000, 1400) (Figure 4C). The higher the S_{CRM} score of a CRMC, the more likely it is under strong evolutionary constraint, suggesting that the S_{CRM} score indeed captures the evolutionary behavior of a CRM as a functional element, in addition to its length feature (Figure 3F). The same conclusion can be drawn from the phyloP scores (Supplementary Figure S4C).

We next examined the relationship between conservation scores of the predicted CRMs and the S_{CRM} score cutoffs α (or *P*-value cutoffs) used for their predictions. As shown in Figure 4D, even the CRMs predicted at a low α have much smaller proportion of neutrality (e.g. 0.31 for the smallest $\alpha = 0$, i.e. the entire CRMC set) than the non-CRMCs (0.71) ($P < 2.2 \times 10^{-302}$, χ^2-test), suggesting that most of the predicted CRMs might be authentic although some short ones may not be in full-length, while the non-CRMCs might contain few false negative CRMCs. With the increase in α (decrease in *P*-value cutoff), proportion of neutrality of the predicted CRMs decreases slowly, as it is already in the saturation phase (Figure 4D). Interestingly, at very high α values, the predicted CRMs evolve like VISTA enhancers (Figure 4A), with a trimodal GERP score distribution, and thus might be involved in more conserved functions such as development (113,114). For instance, at $\alpha = 13{,}750$, the distribution of GERP scores of the predicted CRMs displays a peak around score -5 and a peak around score 3.5, with a small proportion of neutrality of 0.24 (Figure 4D) (it is 0.23 for VISTA enhancers, Figure 4A). The infinitesimal decrease in proportion of neutrality of predicted CRMs with the increase in S_{CRM} cutoffs (Figure 4D) strongly suggests that the predicted CRMs, particularly those at a low *P*-value cutoff, are under similarly strong evolutionary constraints to those on the VISTA enhancers. Similar results are observed using the phyloP scores (Supplementary Figure S4D). The results suggest that the false discovery rates(FDRs) of our predicted CRMs might be very low. However, without the availability of a gold standard negative CRM set in the genome (23), we could not calculate the false positive rates (FPRs) of the predicted CRMs at different *P*-value cutoffs.

dePCRM2 achieves high sensitivity for recalling functionally validated CRMs and non-coding SNVs

To further evaluate the accuracy of dePCRM2, we calculated the sensitivity (recall rate or true positive rate (TPR)) of CRMs predicted at different S_{CRM} cutoffs α and associated *P*-values for recalling a variety of CRM function-related elements located in the covered regions (Supplementary Tables S2–S6). As a control, we also calculated the sensitivity of a set of control sequences that are randomly selected from the covered regions with the matched numbers and sizes of predicted CRMs (Materials and Methods) for recalling these elements. We define that a sequence (a predicted CRM or a control sequence) recalls an element, if the sequence and the element overlap each other by at least 50% of the length of the shorter one. As shown in Figure 5A, with the increase in the *P*-value cutoff, the sensitivity for recalling the elements in all the 10 datasets increases rapidly and becomes saturated well before *P*-value increases to 0.05 ($\alpha \geq 56$). Supplementary Figure S5A–J show examples of the predicted CRMs overlapping and recalling the elements in the 10 datasets. Particularly, at *P*-value cutoff 5×10^{-5} ($\alpha = 412$), the predicted 593 731 CRMs (Figure 3C) recall 100% of VISTA enhancers (79) and 89.26% of non-coding ClinVar SNVs (79) (Figure 5A), located in the covered regions (Supplementary Note). The rapid saturation of sensitivity for recalling these two types of validated functional elements at such a low *P*-value cutoff once again strongly suggests that true CRMs might be highly enriched in our predicted CRMs, particularly those with a low *P*-value or a high S_{CRM} score. On the other hand, even at a higher *P*-value cutoff 0.05 ($\alpha = 56$), the predicted 1 155 151 CRMs only achieve varying intermediate levels of sensitivity for recalling FANTOM5 promoters (FPs) (88.77%) (82), FANTOM5 enhancers (FEs) (81.90%) (81), DHSs (74.68%) (63), TASs (84.32%) (29), H3K27ac (82.96%) (29), H3K4me1 (76.77%) (29), H3K4me3 (86.96%) (29) and GWAS SNVs (64.60%) (83). However, in all the cases, the control sequences (MATERIALS AND METHODS) only

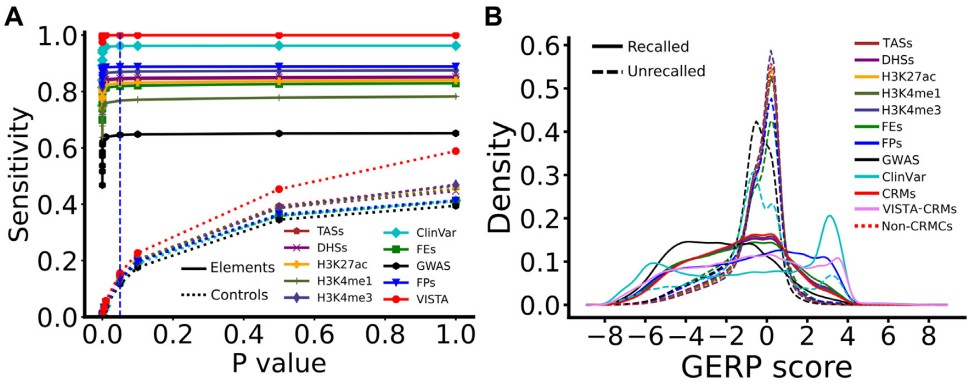

Figure 5. Validation of the predicted CRMs by 10 experimentally determined sequence elements datasets. (**A**) Sensitivity (recall rate or TPR) of the predicted CRMs and control sequences as a function of *P*-value cutoff for recalling the elements in the datasets. The dashed vertical lines indicate the *P*-value ≤0.05 cutoff. The sensitivity of CRMs predicted at all the indicated *P*-value cutoffs are significantly higher ($P < 2.2 \times 10^{-302}$, χ^2-test) than the control sequences for recalling all the ten types of elements. (**B**) Distributions of GERP scores of the recalled and unrecalled elements in each dataset in comparison with those of the predicted CRMs at $P \leq 0.05$ and non-CRMCs. The difference between the distributions of recalled elements and unrecalled elements in each dataset is significant, $P < 2.2 \times 10^{-302}$ (K–S test). Note that there are no unrecalled VISTA enhancers. The curve labeled by VISTA-CRMs is the distribution of CRMs that overlap and recall the 785 VISTA enhancers.

recall much smaller portions of the 10 type of elements at all the *P*-value cutoffs ($P < 2.2 \times 10^{-302}$, χ^2-test) (Figure 5A). For example, at *P*-value $\leq 5 \times 10^{-5}$ and 0.05, the control sequences recall about 7% and 15%, respectively, of all the 10 types of elements.

To find out the reasons for such varying sensitivity of dePCRM2 for recalling the 10 types of elements, we plotted the distribution of GERP scores of the recalled and uncalled elements in each dataset by our predicted CRMs at *P*-value ≤0.05. Since we have already plotted the distribution of the entire set of the 976 VISTA enhancers (Figure 4A), to avoid redundancy, we instead plotted the distribution for the CRMs that recall the 785 VISTA enhancers located in the covered regions (VISTA-CRMs). As shown in Figure 5B, like the predicted CRMs, the recalled elements in all the datasets are under strong evolutionary selections, thus are likely functional. However, VISTA-CRMs, recalled ClinVar SNVs and recalled FPs evolve more like VISTA enhancers, all with a trimodal GERP score distribution (Figure 5B), suggesting that they are under stronger evolution constraints than the other recalled element types. These results are not surprising, as we mentioned earlier, VISTA enhancers are mostly ultra-conserved (110–112), while ClinVar SNVs were identified for their larger effect sizes duo to conserved critical functions (80), and promoters are well-known to be more conserved than are enhancers (115). In stark contrast, like the non-CRMCs, all unrecalled elements in the 10 datasets are largely selectively neutral, and thus, are unlikely to be functional, with the exception that the 74 222 (3.82%) unrecalled ClinVar SNVs display a trimodal distribution, and that there are no unrecalled VISTA enhancers (Figure 5B). Notably, proportion of neutrality of unrecalled PEs (0.59) and PFs (0.63) are smaller than that of the non-CRMCs (0.71) (Figure 5B), suggesting we might miss a small portion of authentic PEs and PFs (see below for false negative rate (FNR) estimations of our CRMs). Assuming that at least most of unrecalled elements in the datasets except the VISTA and ClinVar datasets, are non-*cis*-regulatory, we estimated that the FDR of the remaining eight datasets might be up to 11.23% (FP), 18.10% (FEs), 25.32 (DHSs), 15.68% (TASs), 13.04% (H3K4m3), 23.23% (H3K4m1), 17.04% (H3K27ac) and 35.40% (GWAS SNVs). The high FDRs for CA (DHSs and TASs) and histone marks are consistent with an earlier study (69). Interestingly, the trimodal distribution of GERP scores of the 3.82% of unrecalled ClinVar SNVs displays a large peak around score 0 and two small peaks around –5 and 3.5, with a proportion of neutrality 0.45 (Figure 5B), indicating that about 45% of the relevant SNVs might be selectively neutral, and thus non-functional. We therefore estimated the FDR of the ClinVar SNV dataset to be about 0.45 × 3.82% = 1.72%. Therefore, Like VISTA enhancers, non-coding ClinVar SNVs are a reliable set for evaluating CRM predictions. The peak of the unrecalled ClinVar SNVs around score 3.5 (Figure 5B), indicates that the relevant SNVs are under strong purifying selection, and thus might be functional, but were missed by dePCRM2. We thus estimate our predicted CRMs (at *P*-value ≤0.05) might have a FNR < 3.82%-1.72% = 2.10%. In other words, the estimated real sensitivity (= 1 – FNR = 97.9%) for dePCRM2 to recall authentic causal ClinVar SNVs might be slightly higher than the calculated 96.18% (Figure 5A). These estimates are supported by the zero FNR and 100% sensitivity for our predicted CRMs to recall VISTA enhancers (Figure 5A) and a simulation to be described later.

The zero, very low (<1.72%) and low (11.23%) FDRs of VISTA enhancers, ClinVar SNVs and FPs datasets, respectively, are clearly related to the high reliability of the experimental methods used to characterize them. However, these low FDRs might also be related to the highly conserved nature of these elements (Figure 5B), as their critical functions and large effect sizes may facilitate their correct characterization. In this regard, we note that the intermediately high FDRs of the FEs (18.10%), DHSs (25.32), TASs (15.68%), H3K4m3 (13.04%), H3K4m1 (23.23%) and H3K27ac (17.04%) datasets might be due to the facts that bidirectional transcription (116), CA (69,71,117) and histone marks (69,71) are not unique to active enhancers. The

very high FDR of GWAS SNVs (35.5%) might be due to the fact that a lead SNV associated with a trait may not necessarily be located in a CRM and causal; rather, some variants in a CRM, which are in LD with the lead SNV, are the culprits (83,118). Example of GWAS SNVs in LD with positions in a CRM are shown in Supplementary Figure S5K and S5L. Interestingly, many recalled ClinVar SNVs (42.59%) and GWAS SNVs (38.18%) are located in critical positions in predicted binding sites of the UMs (e.g., Supplementary Figure S5D and F).

dePCRM2 outperforms state-of-the-art methods for predicting both CRM positions and lengths

We compared our predicted CRMs at P-value $\leq$ 0.05 ($S_{CRM} \leq 56$) with three most comprehensive sets of predicted enhancers/promoters, i.e. GeneHancer 4.14 (57), EnhancerAtals2.0 (61) and cCREs (26) predicted by their respective methods. The GeneHancer set is the most updated prediction containing 394,086 non-overlapping enhancers covering 18.99% of the genome (Figure 6A). These enhancers were predicted by integrating multiple sources of both predicted and experimentally determined CRMs using a voting schema. The sources of data include VISTA enhancers (79), ENCODE phase 2 enhancer-like regions (119), ENSEMBL regulatory build (55), dbSUPER (120), EPDnew promoters (121), UCNEbase (122), CraniofacialAtlas (123), FPs (82) and FEs (81). Enhancers from ENCODE phase 2 and ENSEMBL were predicted based on multiple tracks of epigenetic marks using the well-regarded tools ChromHMM (49) and Segway (124). Of the GeneHancer enhancers, 388 407 (98.56%) have at least one nucleotide located in the covered regions, covering 18.89% of the genome (Figure 6A). The EnhancerAtlas set contains 7 433 367 overlapping cell/tissue-specific enhancers in 277 cell/tissue types, which were predicted using an unsupervised machine-learning method based on 4159 TF ChIP-seq, 1580 histone mark, 1,113 DHS-seq, and 1153 other enhancer function-related datasets, such as FEs (125). After removing redundancy (identical enhancers in difference cell/tissues), we ended up with 3 452 739 EnhancerAtlas enhancers that may still have overlaps, covering 58.99% of the genome (Figure 6A), and 3 417 629 (98.98%) of which have at least one nucleotide located in the covered regions, covering 58.78% of the genome (Figure 6A). The cCRE set represents the most updated prediction of CRMs by the ENCODE phase 3 consortium (26), containing 926 535 non-overlapping cell type agnostic enhancers and promoters covering 8.20% of the genome. cCREs were predicted based on overlaps among 703 DHS, 46 TAS and 2,091 histone mark datasets in various cell/tissue types produced by ENCODE phases 2 and 3, as well as by the Roadmap Epigenomics project (26). Of these cCREs, 917 618 (99.04%) have at least one nucleotide located in the covered regions, covering 8.13% of the genome (Figure 6A). Both the number (1,155,151) and genome coverage (43.47%) of our predicted CRMs (P-value$\leq$0.05) are larger than those of GeneHancer enhancers (388 407 and 18.89%) and of cCREs (917 618 and 8.12%), but smaller than those of EnhancerAtlas enhancers (3 417 629 and 58.78%), in the covered regions.

To make the comparisons fair, we first computed the sensitivity of these three sets of enhancers for recalling VISTA enhancers, ClinVar SNVs and GWAS SNVs in the covered regions. We omitted FPs, FEs, DHSs, TASs and the three histone marks for the valuation as they were used in predicting CRMs by GeneHancer 4.14, EnhancerAtlas 2.0 or ENCODE phase 3 consortium. We also excluded VISTA enhancers for evaluating GeneHancer enhancers as the former were compiled in the latter (57). Remarkably, our predicted CRMs outperform EnhancerAtlas enhancers for recalling VISTA enhancers (100.00% versus 94.01%) and ClinVar SNVs (96.18% versus 23.35%) (Figure 6B), even though our CRMs cover a smaller proportion of the genome (43.47% versus 58.78%) (Figure 6A). However, EnhancerAtlas enhancers outperform our CRMs for recalling GWAS SNVS (73.53% versus 64.60%) (Figure 6B). As we indicated earlier, the lower sensitivity of our CRMs for recalling GWAS SNVs might be due to the fact that an associated SNV may not necessarily be causal (Supplementary Figure S5K and S5L). The higher sensitivity of EnhancerAtlas enhancers for recalling GWAS SNVs might be simply thanks to their higher coverage of the genome (58.78%) than that of our predicted CRMs (43.47%) (Figure 6A). Our predicted CRMs outperform cCREs for recalling VISTA enhancers (100% versus 85.99%), ClinVar SNVs (96.18% versus 17.50%) and GWAS SNVs (64.60% versus 18.21%) (Figure 6B). Our predicted CRMs also outperform GeneHancer enhancers for recalling ClinVar SNVs (96.18% versus 30.93%) and GWAS SNVs (64.60% versus 38.05%) (Figure 6B). However, no conclusion can be drawn from these results, because our predicted CRMs cover a higher proportion of the genome than both (43.47% versus 18.89% and 8.20%).

As shown in Figure 6C, overlaps between nucleotide positions of the four sets of predicted CRMs/enhancers/cCREs are quite low. For instance, EnhancerAtlas enhancers, GeneHancer enhancers and cCREs share 50.85%, 70.72% and 76.86% of their positions with our predicted CRMs, corresponding to 69.01%, 30.90% and 14.51% of the positions of our CRMs (Figure 6C), respectively. The positions shared by all the four sets make up only 5.80%, 18.00%, 41.69% and 7.87% of positions of EnhancerAtlas enhancers, GeneHancer enhancers, cCREs, and our CRMs, respectively. As expected, the 50.85%, 70.72% and 76.86% of positions of EnhancerAtlas enhancers, GeneHancer enhancers and cCREs, which they share with our CRMs, respectively, evolve similarly to our predicted CRMs, although those of GeneHancer enhancers and cCREs are under slightly higher evolutionary constraints than our CRMs (Figure 6D). However, at a higher S_{CRM} cutoff, e.g. $\alpha = 3{,}000$, our predicted CRMs are even under stronger evolutionary constraints than the shared GeneHancer enhancers and cCREs positions (Figure 6D). As the positions that GeneHancer enhancers or cCREs share with our predicted CRMs evolve like subsets of our CRMs predicted with higher S_{CRM} scores, they are likely functional. By stark contrast, like the non-CRMCs, the remaining 49.14%, 29.28% and 23.13% of positions of EnhancerAtlas enhancers, GeneHancer enhancers and cCREs that they do not share with our CRMs, respectively, are largely selectively neutral, although they all have slightly smaller

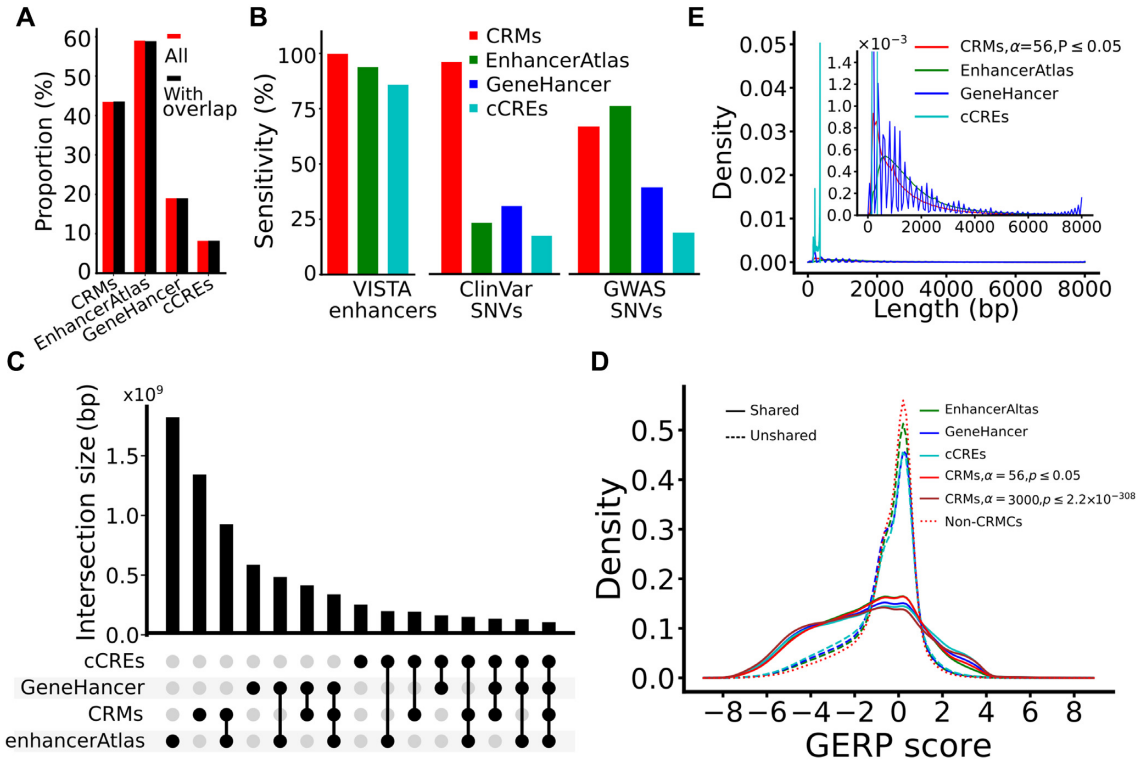

Figure 6. Comparison of the performance of dePCRM2 and three state-of-the-art methods. (**A**) Proportion of the genome that are covered by enhancers/CRMs/cCREs predicted by the four methods (all), and proportion of genome regions covered by predicted enhancers/CRMs/cCREs that at least partially overlap the covered regions (With overlap). (**B**) Sensitivity for recalling VISTA enhancers, ClinVar SNVs and GWAS SNVs, by the predicted enhancers/CRMs/cCREs that at least partially overlap the covered regions. (**C**) Upset plot showing numbers of nucleotide positions shared among the predicted CRMs, GeneHancer enhancers, EnhancerAtlas enhancers and cCREs. (**D**) Distributions of GERP scores of nucleotide positions of the CRMs predicted at *P*-value ≤ 0.05 and *P*-value $\leq 2.2 \times 10^{-308}$, and the non-CRMCs, as well as of nucleotide positions that GeneHancer enhancers, EnhancerAtlas enhancers and cCREs share and do not share with the predicted CRMs at *P*-value ≤ 0.05. The difference between the distributions of shared and unshared positions is significantly different for all the datasets, $P < 2.2 \times 10^{-302}$, K–S test. (**E**) Distributions of lengths of the four sets of predicted enhancers/CRMs/cCREs. The inset is a blow-up view of the axes defined region.

proportion of neutrality than that of the non-CRMCs (0.66, 0.63 and 0.61 versus 0.71, respectively) (Figure 6D), due probably to the small FNR (<2.10%) of our predicted CRMs. As the vast majority of unshared positions of the three sets of predicted enhancers/cCREs are selectively neutral or nearly so, they might be false positives. Based on the proportion of these unshared positions in the three sets, we estimate the FDR of EnhancerAtlas enhancers, GeneHancer enhancers and cCREs to be around 49.14%, 29.28% and 23.13%, respectively. With a much smaller of coverage of the genome by the GeneHancer enhancers (18.99%) and cCREs (8.13%) than by our predicted CRMs (43.47%, at *P*-value ≤0.05) (Figure 6A), it is highly likely that both GeneHancer and ENCODE phase 3 largely under-predict enhancers, even though they have rather high FDRs (29.28% and 23.12%, respectively). On the other hand, with a much higher coverage of the genome by EnhancerAtlas enhancers (58.99%) than by our predicted CRMs (43.47, at *P*-value ≤0.05) (Figure 6A), it is highly likely that EnhancerAtlas might largely over-predict enhancers due to a very high FDR (49.14%).

Next, we compared the lengths of the four sets of the predicted CRMs/enhancers/cCREs. As shown in Figure 6E, the distribution of the lengths of cCREs has a narrow high peak at 345 bp with a mean length of 273 bp and a maximal length of 350bp. Such short lengths of cCREs strongly suggest that the vast majority of even authentic cCREs are just components of long CRMs, because even the longest cCREs (350bp) is shorter and the shortest VISTA enhancer (428 bp). The highly uniform lengths of the predicted cCREs also indicate the limitation of the underlying prediction pipeline (26). The distribution of GeneHancer enhancers oscillates with a period of 166bp (Figure 6E), which might be an artifact of the underlying algorithm for combining results from multiple sources (57). Moreover, with a mean length of 1,488bp, GeneHancer enhancers are generally longer than our predicted CRMs at *P*-value ≤0.05 with a mean length of 1,162 bp, while the latter are generally longer than the EnhancerAtlas enhancers with a mean length of 680bp (Figure 6E). This high inconsistency in CRM length predictions highlights the difficulty of the problem.

Finally, we compared dePCRM2 with the EnhancerAtlas and the ENCODE phase 3 methods (26) for predicting the lengths of VISTA enhancers by computing a ratio of the length of a predicted CRM/enhancer/cCRE over that of its recalled VISTA enhancer. As show in Figure 7A, with a median ratio of 0.12, all recalling cCREs are only a fraction of

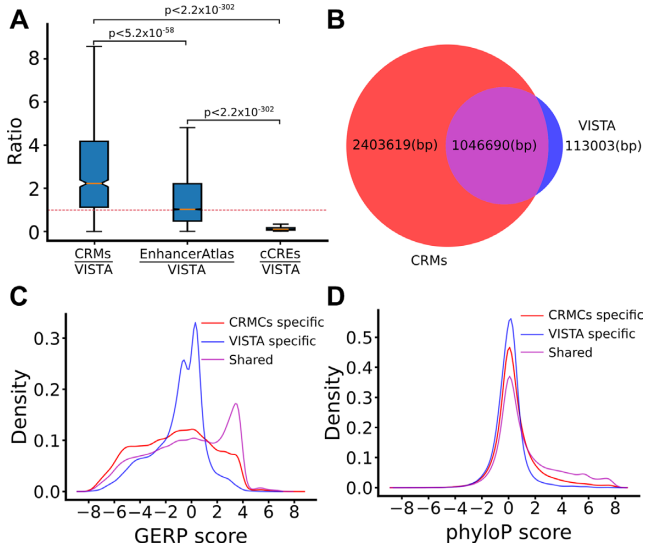

Figure 7. Comparison of the three methods for predicting the lengths of VISTA enhancers. (**A**) Boxplots of distributions of the ratio of the length of a CRM, an EnhancerAtlas enhancer or a cCRE over that of its recalled VISTA enhancer. The significant levels were calculated using the Mann–Whitney *U* test. (**A**) Venn diagrams showing the number of nucleotide positions shared by recalling CRMs and recalled VISTA enhancers, and the number of positions specific to the recalling CRMs and the recalled VISTA enhancers. (**C** and **D**) Distribution of GERP (C) and phyloP (D) scores of the positions shared by the CRMs and the VISTA enhancers, and specific to the CRMs or VISTA enhancers. The difference between the distributions of shared and VISTA specific positions is significantly different, $P < 2.2 \times 10^{-302}$, K–S test.

the lengths of their 675 (85.99%) recalled VISTA enhancers (Figure 6B), as expected from the aforementioned results. With a median ratio of 1.03 (Figure 7A), most recalling EnhancerAtlas enhancers largely have the similar lengths to those of their 738 (94.01%, Figure 6B) recalled VISTA enhancers. With a median ratio of 2.23 (Figure 7A), most of our recalling CRMs are at least twice as long as their 785 (100%, Figure 6B) recalled VISTA enhancers. However, the vast majority (90.26%) of VISTA enhancer positions are covered by our recalling CRMs (Figure 7B), and only 9.74% of VISTA enhancer positions are missed by our recalling CRMs. On average, our recalling CRMs (mean length 4395 bp) are 2.98 times as long as recalled VISTA enhancers (mean length 1477 bp). To see whether dePCRM2 over predicts the length of VISTA enhancers, or VISTA enhancers are actually only a part of otherwise longer super-enhancers whose parts can still have some enhancer activities (106), we compared conservation scores of positions shared by our recalling CRMs and recalled VISTA enhancers, of positions specific to the recalling CRMs, of positions specific to the recalled VISTA enhancers (Figure 7B). As expected, shared positions are under strong selection constrains measured either by GERP (Figure 7C) or by phyloP scores (Figure 7D). To our surprise, CRMs specific positions (69.66%) are also under strong evolutionary constraints though less conserved than shared positions, while VISTA specific positions (9.74%) are largely evolutionarily neutral or nearly so (Figure 7C and D). These results strongly suggest that VISTA enhancers might be the most conserved parts of residing longer super-enhancers, while a small portion (on average 9.47%) of some VISTA enhancers might be non-functional. Taken together, these results unequivocally indicate that dePCRM2 is much more accurate than the three state-of-the-art methods for predicting both the nucleotide positions and the lengths of CRMs.

At least half of the human genome might code for CRMs

What is the proportion of the human genome coding for CRMs and TFBSs? Our predicted CRMs and constituent TFBSs in the covered regions might position us to address this interesting and important, yet unanswered question (126,127). To this end, we took a semi-theoretic approach. Specifically, we calculated the expected number of true positives and false positives in the CRMCs in each non-overlapping S_{CRM} score interval based on the predicted number of CRMCs and the density of S_{CRM} scores of Null CRMCs in the interval (Figure 8A), yielding 1,383,152 (98.45%) expected true positives and 21 821 (1.55%) expected false positives in the CRMCs (Figure 8B). The vast majority of the 21,821 expected false positive CRMCs have a low S_{CRM} score < 4 (inset in Figure 8A) with a mean length of 28bp, comprising 0.02% (21 821 × 28 bp/3 088 269 832 bp) of the mappable genome and 0.05% (0.0002/0.4403) of the total length of the CRMCs, i.e. a FDR of 0.05% for nucleotide positions (Figure 8C). On the other hand, as the CRMCs miss 3.72% of noncoding ClinVar SNVs in the covered regions (the point at *P*-value = 1 in Figure 5A), the FNR of predicting CRMCs would be <3.72% × (1 – 0.45) = 2.05%, given the proportion of neutrality of 0.45 for the unrecalled ClinVar SNVs (Figure 5B). False negative CRMCs would be 2.05% × 44.03 = 0.90% of the genome, which is 0.090/33.44% = 3.02% of the total length of the non-CRMCs, meaning a false omission rate (FOR) of 3.02% for nucleotide positions (Figure 8C). Hence, true CRM positions in the covered regions would be 44.03% – 0.02% + 0.90% = 44.91% of the genome (Figure 8C). In addition, as we argued earlier, the CRMC density in the uncovered 22.53% genome regions is about 80.04% of that in the covered regions, CRMCs in the uncovered regions would be about 0.2253 × 0.4491 × 0.8004/0.7747 = 10.45% of the genome (Figure 8C). Taken together, we estimated about 44.91% + 10.45% = 55.36% of the genome to code for CRMs, for which we have predicted (44.03 – 0.02)/55.36 = 79.50%. Moreover, as we predicted that about 40% of CRCs are made up of TFBSs (Figure 3D), we estimated that about 0.4 × 55.36% = 22.14% of the genome might encode TFBSs. Furthermore, assuming a mean length 1162 bp for CRMs (the mean length of our predicted CRMs at *P*-value ≤0.05), and a mean length of 10 bp for TFBSs (Figure 2D), we estimated that the human genome would encode about 1 471 313 CRMs (3 088 269 832 × 0.55.36/1162) and 68 374 294 TFBSs (3 088 269 832 × 0.2214/10).

DISCUSSION

Identification of all functional elements, in particular, CRMs in genomes has been the central task in the postgenomic era, and enormous CRM function-related data have been produced to achieve the goal (23,128). Great efforts have been made to predict CRMs in the genomes

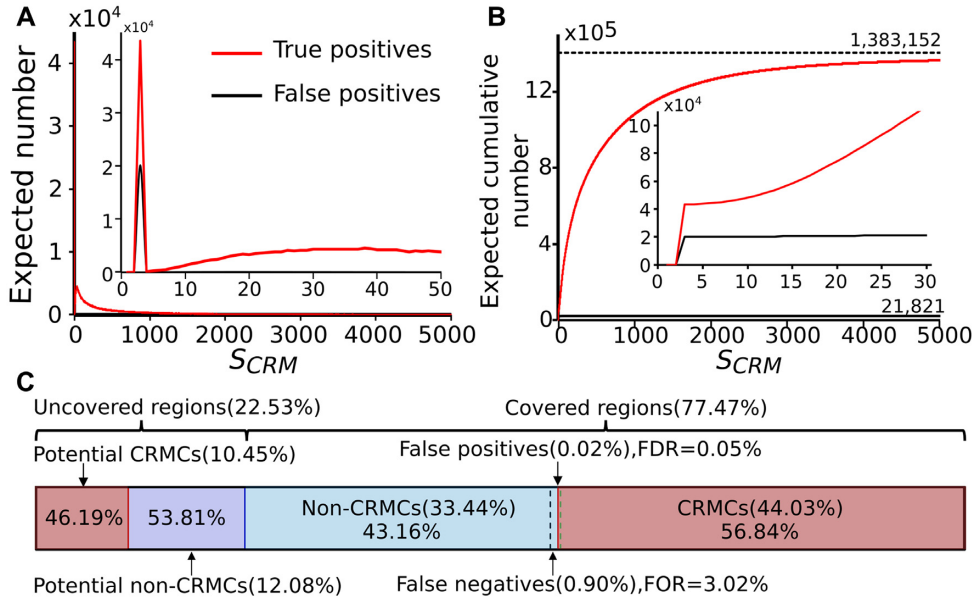

Figure 8. Estimation of the portion of the human genome encoding CRMs. (**A**) Expected number of true positive and false positive CRMCs in the predicted CRMCs in each one-unit interval of the S_{CRM} score. The inset is a blow-up view of the axes defined region. (**B**) Expected cumulative number of true positives and false positives with the increase in S_{CRM} score cutoff for predicting CRMs. The inset is a blow-up view of the axes defined region. (**C**) Proportions of the covered regions (77.47%) and uncovered regions (22.53%) in the genome and estimated proportions of CRMCs in them. Percentages in the braces are the proportions of the indicated sequence types in the genome, and percentages in the boxes are the proportions of the indicated sequence types in the covered regions or in the uncovered regions.

(26,55,57,61,129) using these data. Most existing methods attempt to predict cell/tissue specific CRMs using CA and multiple tracks of histone marks collected in the same cell/tissue types (26,49,57,61,124). Though conceptually attractive, these methods are limited by the scope of applications (26,49,124), low resolution of predicted CRMs (26,61), lack of constituent TFBS information (26,61), and high FDRs (69). To overcome these limitations, we proposed a different approach to predict a largely cell type agnostic or static map of CRMs and constituent TFBSs in the genome (47,48) by identifying repeatedly cooccurring patterns of motifs found in appropriately expanded binding peaks in a large number of TF ChIP-seq datasets for different TFs in various cell/tissue types. Since it is mainly TFBSs in a CRM that define its structure and function, it is not surprising that TF ChIP-seq data are a more accurate predictor of CRMs than CA and histone mark data (53,69,71). Another advantage of our approach is that we do not need to exhaust all TFs and all cell/tissue types of the organism in order to predict most, if not all, of CRMs and constituent TFBSs in the genome as we demonstrated earlier (47,75), because CRMs are often repeatedly used in different cell/tissue types, developmental stages and physiological homeostasis (1), and binding sites of collaborative TFs are closely located in CRMs. Moreover, by appropriately extending the called binding peaks in each dataset, we could largely increase the chance to identify collaborative motifs and full-length CRMs, thereby increasing the power of existing data, and further reducing the number of datasets needed as we have demonstrated in this and previous studies (47,48). We might only need a large, but limited, number of datasets to predict most, if not all, CRMs and TFBSs in the genome, as the numbers of predicted UMs and CRMs enter a saturation phase when more than few hundreds of datasets were used for the predictions as we showed earlier (47). Our earlier application of the approach resulted in very promising results in the fly (48) and human (47) genomes even using a relatively small number of strongly biased datasets available then. However, the earlier implementations were limited by computational inadequacies of underlying algorithms to find and integrate motifs in ever increasing number of large TF ChIP-seq datasets in mammalian cell/tissues (47,48). In this study, we circumvent the limitations by developing the new pipeline dePCRM2 based on an ultrafast and accurate motif finder ProSampler, an efficient motif pattern integration method, and a novel CRM scoring function that captures essential features of full-length CRMs.

A limitation of dePCRM2 is that although it can predict a CRM's functional state in a ChIP-probed cell type if the CRM overlaps original binding peaks of the ChIP-ed TFs in the cell type, the fraction of CRMs whose functional states can be so predicted in most cell types could be quite low or even 0, since only few or even no ChIP-seq datasets are available in most cell types due to the strong bias of current datasets (Supplemental Note). Moreover, to predict functional states of all CRMs in a cell type in this way, one might need ChIP-seq data for all TFs working in the cell type, and this can be too costly or is currently unfeasible. For this reason, our predicted CRMs are largely cell-type agnostic. Fortunately, it has been shown that when the locus of a CRM is correctly anchored by multiple TFs' binding peaks, one or few epigenetic marks at the locus can accurately predict the CRM's functional state (69). We are

in the process to develop such capability. Thus, this two-step approach might be more cost-effective for predicting both a static map of CRMs and constituent TFBSs in the genome and their functional states in various cell/tissue types.

It has been estimated that the human genome encodes from 2,000 to 3,000 TFs belonging to 100–200 protein families (96,130). However, the exact number of TFs and TF families encoded in the genome remains unknown (96,131). Our prediction of the 201 UM families in the covered regions provides us an opportunity to estimate the number of TFs families encoded in the genome. As different TFs of the same protein family/superfamily bind similar motifs (97,132), it is highly likely that a predicted UM is recognized by multiple TFs of the same family/superfamily. Indeed, 92 (78.63%) of the 117 (58.21%) UMs matching at least a known motif, match at least two. The cognate TF families of the remaining 84 (41.79%) UMs remain to be elucidated. On the other hand, as 64 (71.91%) of the 89 known motif families match one of our UMs, we might have recovered most known motif families. Based on these results, we estimate the lower bound of the number of TF/motif families encoded in the human genome to be around 200, considering that the uncovered 22.53% regions of the genome might harbor novel UMs that do not appear in the covered regions.

Remarkably, dePCRM2 enables us to partition the covered regions into two exclusive sets, i.e., the CRMCs and the non-CRMCs. Multiple pieces of evidence strongly suggest that the partition might be highly accurate. First, although evolutionary information is not used in our prediction of the CRMCs and the non-CRMCs, yet the two sets display dramatically different evolutionary behaviors (Figure 4A and Supplementary Figure S4A). More specifically, the vast majority of nucleotide positions of the CRMCs are under strongly evolutionary constraints (Figure 4B and Supplementary Figure S4B), and a subset of which with higher S_{CRM} scores are under even stronger evolutionary constraints that are comparable to the ultra-conserved, mostly developmentally related VISTA enhancers (Figure 4C and Supplementary Figure S4C). In stark contrast, positions of the non-CRMC positions are largely selectively neutral or nearly so, thus are likely to lack cis-regulatory functions (Figure 4A and Supplementary Figure S4A). Second, our control studies (Supplementary Figure S3) together with the small *P*-values for the vast majority of the CRMCs (Figure 3C), strongly suggest that the partition could not be generated by virtually any serendipity. Third, all experimentally validated VISTA enhancers and almost all (96.28%) of well-documented non-coding ClinVar SNVs in the covered regions are recalled by the CRMCs at a very stringent *P*-value 5×10^{-5} (Figure 5A), while the control sequences randomly selected from the covered regions could only recall a small portion (7%) of these elements (Figure 5A), indicating that the CRMCs have a very high sensitivity, and are highly enriched for true CRMs. Finally, our simulation studies indicate that the CRMCs have a very low FDR of 0.05% (or a high precision of 99.95%), and the non-CRMCs have a low FOR of 3.02% (Figure 8C). To the best of our knowledge, dePCRM2 is the first of its kind to partition a large portion (77.47%) of the genome into two sets such that one (CRMCs) are highly likely to be *cis*-regulatory, and the other (non-CRMCs) are not.

Accurate prediction of the length of CRMs is also critical, but this appears to be a difficult problem as evidenced by the peculiar distributions of the lengths of GeneHancer enhancers and cCREs (Figure 6E). The problem can be further complicated by the difficulty to accurately validate the lengths of predicted CRMs, because even experimentally validated VISTA enhancers may not necessarily be in their correct full-length forms, as a portion of an enhancer could be still partially functional (1,106). Interestingly, we found that our predicted CRMs that recall VISTA enhancers are on average twice longer than the recalled VISTA enhancers (Figure 7A), and the 'extra' sequences in the recalling CRMs are under strong evolutionary constraints (Figure 7C and D), suggesting that most VISTA enhancers might be only a part of longer enhancers. Meanwhile, as an average of 10% of VISTA enhancer positions, which do not overlap the recalling CRMs, are largely evolutionally neutral, they might not be *cis*-regulatory (Figure 7C and D). Therefore, it is highly likely that dePCRM2 is able to predict full-length CRMs. On the other hand, as 44.26% (621 841) of our predicted 1 404 973 CRMCs are shorter than the shortest (428bp) VISTA enhancer, they might not be in full length. However, these potential CRM components comprise only 7.42% of the total length of the CRMCs, while the remaining 55.74% (783 132) of the CRMCs that comprise 92.58% of the total length of the CRMCs are likely to be in full-length. However, as very short CRMCs tend to have small S_{CRM} scores and to be under weak evolutionary constraints, they can be effectively filtered out using more stringent S_{CRM} cutoffs (Figures 3F, 4D and Supplementary Figure S4D). It has been shown that an enhancer's length and evolutionary behavior are determined by its regulatory tasks (105), and conserved enhancers are active in development (113,114), while fragile enhancers are associated with evolutionary adaptation (113). CRMCs with different S_{CRM} scores might belong to different functional types as indicated by their different evolutionary behaviors (Figure 4A, C, Supplementary Figure S4A and C) and length distributions (Figure 3F). For example, as CRMs predicted at high S_{CRM} cutoffs tend to be longer (Figure 3F) and under stronger evolutionary constrains (Figure 4D and Supplementary Figure S4D), they might be mainly involved in development and other more conserved functions. Since CRMs predicted at lower S_{CRM} cutoffs tend to be shorter (Figure 3F) and under weaker evolutionary constrains (Figure 4D and Supplementary Figure S4D), they might be mainly involved in non-development related and other less conserved functions. This conclusion is further supported by our finding that the CRMs recalling VISTA enhancers that are mostly developmentally related, have a much longer mean length (4395 bp) than do the entire set of CRMs (1162 bp) predicted at *P*-value $\leq$0.05. Our failure to predict full-length CRMs of short CRM components might be due to insufficient data coverage on the relevant loci in the genome. This is reminiscent of our earlier predicted, even shorter CRMCs (mean length = 182 bp) using a much smaller number and less diverse 670 datasets (47). As we argued earlier (47) and confirmed here by the much longer CRMCs (mean

length = 981 bp) predicted using the much larger and more diverse datasets albeit still strongly biased to a few TFs and cell/tissue types (Supplementary Note). We anticipate that full-length CRMs of these short CRM components can be predicted using even larger and more diverse TF ChIP-seq data. Therefore, efforts should be made in the future to increase the genome coverage and reduce data biases by including more untested TFs and probed cell types in the TF ChIP-seq data generation.

Interestingly, our predicted CRMs even at a stringent *P*-value $\leq 5 \times 10^{-5}$ achieve perfect (100.00%) and very high (96.28) sensitivity for recalling VISTA enhancers (79) and noncoding ClinVar SNVs (80), respectively, but varying intermediate sensitivity ranging from 64.60% (for GWAS SNVs) to 88.77% (for FPs) for recalling other eight types of CRM function-related elements at a high *P*-value≤0.05 (Figure 5A). It appears that such varying sensitivity is due to varying FDRs ranging from 0% (for VISTA enhancers) to 35.4% (for GWAS SNVs) of the methods used to characterize the elements (Figure 5B). Our finding that DHSs, TASs, and histone mark (H3K4m1, H3K4m3 and H3K27ac) peaks have high FDRs for predicting CRMs is consistent with an earlier study showing that histone marks or CA were less accurate predictor of active enhancers than TF binding data (69). Thus, it is not surprising that our predicted CRMs substantially outperform the three state-of-the-art sets of predicted enhancers, i.e. GeneHancer (57), EnhancerAtlas (61) and cCREs (26), for recalling ClinVar SNVs and VISTA enhancers (we excluded GeneHancer enhancers for this evaluation since VISTA enhancers were a part of it) (Figure 6B), as well as for predicting the lengths of CRMs (Figures 6E and 7), because these three sets were mainly predicted based on overlaps between multiple tracks of CA and histone marks in various cell/tissue type.

Although originally called binding peaks is strongly biased to few cell types and TFs (Supplementary Note), and the 6,092 TF ChIP-seq datasets cover only 40.98% of the genome, after appropriately extending the binding peaks, we increased the genome coverage to 77.47%. Nucleotide positions of the extended parts of the called peaks contribute 42.13% positions of the predicted CRMCs (Figure 3A). Like the other 57.87% of CRMC positions covered by original binding peaks(Figure 3A), these 42.13% of CRMC positions covered by the extended parts also are under strong evolutionary constraints (Figure 4B), and thus are likely to be functional. Therefore, appropriate extension of called binding peaks in the datasets can indeed substantially increase the power of available data. On the other hand, we abandoned 37.82% of positions covered by the original binding peaks(Figure 3A), which might lack cis-regulatory functions, as like the non-CRMCs, they are largely selectively neutral (Figure 4B and Supplementary Figure S4B). This result is consistent with the earlier finding that called binding peaks cannot be equivalent to CRMs or parts of CRMs (101–103), and highlights the necessity to integrate a large number of diverse TF ChIP-seq datasets for accurate genome-wide prediction of CRMs and TFBSs as demonstrated in this study.

The proportion of the human genome that is functional is a topic under hot debate (127,133–135), and a wide range from 3% to 80% of the genome has been suggested to be functional based on different sources of evidence and definition of functions (126,136–138). For instance, the ENCODE consortium argued that 80% of the genome might be functional based on their biochemical activities, while Graur (138) mentioned that functional sequences cannot exceed 15% of the genome based on their evolutionary constrains. The major disagreement is for the proportion of functional NCSs in the genome, mainly CRMs, which have been estimated to comprise from 5% to 40% of the genome (126,136–138). Most of these estimates were based on proportions of the genome that were inferred to be under purifying selection using various evolutionary models (134,135,138), resulting in such highly varying estimates. Since some parts of a CRM and even the entire CRM might not be conserved, it is highly likely that these methods underestimated the regulatory genome (139). Moreover, a wide range of CRM numbers have been suggested to be encoded in the human genome, from tens of thousands to a few million (61,119,129). For example, ENCODE phase 2 identified 469 416 CRMs in the human genome, while ENCODE phase 3 recently updated the number to 926,535 using more datasets (26), and this number may increase further when the current ENCODE phase 4 is complete in the future. As we indicated earlier, GeneHancer (57) and EnhancerAtlas (57) predicted 394,086 and 3,452,739 enhancers, respectively, although both sets might suffer FDRs. Our predicted CRMCs cover 44.03% of the genome, which is higher than cCREs (7.9%) and GeneHancer enhancers (18.99%) do, but lower than EnhancerAtlas enhancers (58.99%) (61) do. The higher accuracy of our predicted CRMs suggests that cCREs and GeneHancer might underpredict CRMs, whereas EnhancerAtlas might overpredict them even using limited data. Based on the estimated FDR and FNR in predicting the CRMCs as well as the estimated density of CRMCs in the uncovered regions relative to the covered regions (Figure 8C), we estimated that about 55.36% and 22.14% of the genome might code for CRMs and TFBSs, respectively, which encode about 1.47 million CRMs and 68 million TFBSs. As about 14% of CRMC positions are under purifying selection (Figure 4A and Supplementary Figure S4A), they account for about 7.6% (= 0.14*55.36) of the genome. This result is consistent with the earlier estimates that about 5–15% of the genome positions are under purifying selection (23,126,136–138). Since 86% of CRMC positions are not under purifying selection (Figure 4A and Supplementary Figure S4A), it is not surprising that the *cis*-regulatory genome appears to be more prevalent (55.36%) than originally thought (8∼40%) (126,136–138). We estimated that our true positive CRMs cover 44.01% of the genome, thus, we might have predicted 79.5% (44.01/55.36) CRM positions encoded in the genome.

With the availability of more diverse and balanced data covering more regions of the genome in the future, it is possible to predict a more complete map of CRMs and constituent TFBSs in the genome. However, even in its current form, this unprecedentedly complete map of CRMs and constituent TFBSs in the human genome may facilitate the community's efforts to functionally characterize the regulatory genome and identify causal noncoding variants of complex diseases. To assist such usages of the map, we have

created database (https://cci-bioinfo.uncc.edu/), where various queries to the map can be made, such as what is the nearest CRMs to a gene and vice versa; what TFBSs are in a CRM; what are CRMs that contain binding sites of a motif, and so on. The map also enables genome-wide investigations of structures, landscape, evolution and functions of CRMs and TFBSs in various contexts.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Would like to thank all lab members for their inputs, and anonymous reviewers for their criticisms and suggestions which substantially improve the manuscript. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.
Authors' contributions: Z.S. conceived the project. Z.S. and P.N. developed the algorithms and P.N. carried out all computational experiments and analysis. Z.S. and P.N. wrote the manuscripts. All authors read and approved the final manuscript.

FUNDING

US National Science Foundation [DBI-1661332]. Funding for open access charge: US National Science Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Davidson,E.H. (2006) In: *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press.
2. Wilczynski,B. and Furlong,E.E. (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, **6**, 383.
3. King,M. and Wilson,A. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
4. Rubinstein,M. and de Souza,F.S. (2013) Evolution of transcriptional enhancers and animal diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **368**, 20130017.
5. Siepel,A. and Arbiza,L. (2014) Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.*, **29**, 81–89.
6. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
7. Ramos,E.M., Hoffman,D., Junkins,H.A., Maglott,D., Phan,L., Sherry,S.T., Feolo,M. and Hindorff,L.A. (2014) Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147.
8. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
9. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
10. Kilpinen,H., Waszak,S.M., Gschwind,A.R., Raghav,S.K., Witwicki,R.M., Orioli,A., Migliavacca,E., Wiederkehr,M., Gutierrez-Arcelus,M., Panousis,N.I. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.
11. McVicker,G., van de Geijn,B., Degner,J.F., Cain,C.E., Banovich,N.E., Raj,A., Lewellen,N., Myrthil,M., Gilad,Y. and Pritchard,J.K. (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–749.
12. Smith,E. and Shilatifard,A. (2014) Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.*, **21**, 210–219.
13. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
14. Herz,H.M., Hu,D. and Shilatifard,A. (2014) Enhancer malfunction in cancer. *Mol. Cell*, **53**, 859–866.
15. Ongen,H., Andersen,C.L., Bramsen,J.B., Oster,B., Rasmussen,M.H., Ferreira,P.G., Sandoval,J., Vidal,E., Whiffin,N., Planchon,A. *et al.* (2014) Putative cis-regulatory drivers in colorectal cancer. *Nature*, **512**, 87–90.
16. Khurana,E., Fu,Y., Chakravarty,D., Demichelis,F., Rubin,M.A. and Gerstein,M. (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**, 93–108.
17. Zhou,S., Treloar,A.E. and Lupien,M. (2016) Emergence of the noncoding cancer genome: a target of genetic and epigenetic alterations. *Cancer Discov.*, **6**, 1215–1229.
18. Whitaker,J.W., Chen,Z. and Wang,W. (2015) Predicting the human epigenome from DNA motifs. *Nat. Methods*, **12**, 265–272.
19. Wang,M., Zhang,K., Ngo,V., Liu,C., Fan,S., Whitaker,J.W., Chen,Y., Ai,R., Chen,Z., Wang,J. *et al.* (2019) Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res.*, **47**, 6753–6768.
20. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
21. Pai,A.A., Pritchard,J.K. and Gilad,Y. (2015) The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.*, **11**, e1004857.
22. Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
23. Gasperini,M., Tome,J.M. and Shendure,J. (2020) Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.*, **21**, 292–310.
24. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
25. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
26. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
27. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
28. Consortium,G. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
29. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
30. Kleftogiannis,D., Kalnis,P. and Bajic,V.B. (2016) Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.*, **17**, 967–979.
31. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
32. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **2001**, 127–138.
33. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
34. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.

35. Hartmann,H., Guthohrlein,E.W., Siebert,M., Luehr,S. and Soding,J. (2013) *P*-value-based regulatory motif discovery using positional weight matrices. *Genome Res.*, **23**, 181–194.
36. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
37. Sinha,S. (2003) Discriminative motifs. *J. Comput. Biol.*, **10**, 599–615.
38. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
39. Junion,G., Spivakov,M., Girardot,C., Braun,M., Gustafson,E.H., Birney,E. and Furlong,E.E. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.
40. Whitington,T., Frith,M.C., Johnson,J. and Bailey,T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39**, e98.
41. Sun,H., Guns,T., Fierro,A.C., Thorrez,L., Nijssen,S. and Marchal,K. (2012) Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res.*, **40**, e90.
42. Jiang,P. and Singh,M. (2014) CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Res.*, **42**, 2833–2847.
43. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
44. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
45. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
46. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.
47. Niu,M., Tabari,E., Ni,P. and Su,Z. (2018) Towards a map of cis-regulatory sequences in the human genome. *Nucleic Acids Res.*, **46**, 5395–5409.
48. Niu,M., Tabari,E.S. and Su,Z. (2014) De novo prediction of cis-regulatory elements and modules through integrative analysis of a large number of ChIP datasets. *BMC Genomics*, **15**, 1047.
49. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
50. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.
51. Firpi,H.A., Ucar,D. and Tan,K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579–1586.
52. Rajagopal,N., Xie,W., Li,Y., Wagner,U., Wang,W., Stamatoyannopoulos,J., Ernst,J., Kellis,M. and Ren,B. (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, **9**, e1002968.
53. Kleftogiannis,D., Kalnis,P. and Bajic,V.B. (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.*, **43**, e6.
54. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
55. Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T. and Flicek,P.R. (2015) The ensembl regulatory build. *Genome Biol.*, **16**, 56.
56. Ashoor,H., Kleftogiannis,D., Radovanovic,A. and Bajic,V.B. (2015) DENdb: database of integrated human enhancers. *Database*, **2015**, bav085.
57. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, **2017**, bax028.
58. Chen,C., Zhou,D., Gu,Y., Wang,C., Zhang,M., Lin,X., Xing,J., Wang,H. and Zhang,Y. (2020) SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. *Nucleic Acids Res.*, **48**, D198–D203.
59. Kang,R., Zhang,Y., Huang,Q., Meng,J., Ding,R., Chang,Y., Xiong,L. and Guo,Z. (2019) EnhancerDB: a resource of transcriptional regulation in the context of enhancers. *Database (Oxford)*, **2019**, bay141.
60. Zhang,G., Shi,J., Zhu,S., Lan,Y., Xu,L., Yuan,H., Liao,G., Liu,X., Zhang,Y., Xiao,Y. *et al.* (2018) DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res.*, **46**, D78–D84.
61. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
62. Cheneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
63. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
64. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
65. Aday,A.W., Zhu,L.J., Lakshmanan,A., Wang,J. and Lawson,N.D. (2011) Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. *Dev. Biol.*, **357**, 450–462.
66. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
67. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
68. Kwasnieski,J.C., Fiore,C., Chaudhari,H.G. and Cohen,B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, **24**, 1595–1602.
69. Dogan,N., Wu,W., Morrissey,C.S., Chen,K.B., Stonestrom,A., Long,M., Keller,C.A., Cheng,Y., Jain,D., Visel,A. *et al.* (2015) Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenet. Chromatin*, **8**, 16.
70. Catarino,R.R. and Stark,A. (2018) Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.*, **32**, 202–223.
71. Arbel,H., Basu,S., Fisher,W.W., Hammonds,A.S., Wan,K.H., Park,S., Weiszmann,R., Booth,B.W., Keranen,S.V., Henriquez,C. *et al.* (2019) Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 900–908.
72. Goi,C., Little,P. and Xie,C. (2013) Cell-type and transcription factor specific enrichment of transcriptional cofactor motifs in ENCODE ChIP-seq data. *BMC Genomics*, **14**, S2.
73. Li,Y., Ni,P., Zhang,S., Li,G. and Su,Z. (2019) ProSampler: an ultra-fast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics*, **35**, 4632–4639.
74. Allen,J.E., Pertea,M. and Salzberg,S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res.*, **14**, 142–148.
75. Niu,M., Tabari,E.S. and Su,Z. (2014) De novo prediction of cis-regulatory elements and modules through integrative analysis of a large number of ChIP datasets. *BMC Genomics*, **15**, 1047.

76. Arnosti,D.N. and Kulkarni,M.M. (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.*, **94**, 890–898.
77. Yanez-Cuna,J.O., Kvon,E.Z. and Stark,A. (2013) Deciphering the transcriptional cis-regulatory code. *Trends Genet.*, **29**, 11–22.
78. Vockley,C.M., Barrera,A. and Reddy,T.E. (2017) Decoding the role of regulatory element polymorphisms in complex disease. *Curr. Opin. Genet. Dev.*, **43**, 38–45.
79. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser – a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
80. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
81. 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.
82. Forrest,A.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
83. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
84. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B.J., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F. *et al.* (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.
85. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
86. Soundarajan,S. and Hopcroft,J.E. (2015) Use of Local Group Information to Identify Communities in Networks. *ACM Trans. Knowl. Discov. Data*, **9**, 21.
87. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
88. Zhang,S., Jiang,L., Du,C. and Su,Z. (2013) SPIC: A novel information contents based similarity metric for comparing transcription factor binding site motifs. *BMC Syst. Biol.*, **7**, S14.
89. van Dongen,S. and Abreu-Goodger,C. (2012) Using MCL to extract clusters from networks. *Methods Mol. Biol.*, **804**, 281–295.
90. Vockley,C.M., McDowell,I.C., D'Ippolito,A.M. and Reddy,T.E. (2017) A long-range flexible billboard model of gene activation. *Transcription*, **8**, 261–267.
91. Snetkova,V. and Skok,J.A. (2018) Enhancer talk. *Epigenomics*, **10**, 483–498.
92. Li,Q., Peterson,K.R., Fang,X. and Stamatoyannopoulos,G. (2002) Locus control regions. *Blood*, **100**, 3077–3086.
93. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
94. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
95. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
96. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **175**, 598–599.
97. Ambrosini,G., Vorontsov,I., Penzar,D., Groux,R., Fornes,O., Nikolaeva,D.D., Ballester,B., Grau,J., Grosse,I., Makeev,V. *et al.* (2020) Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.*, **21**, 114.
98. Perrot,C.Y., Gilbert,C., Marsaud,V., Postigo,A., Javelaud,D. and Mauviel,A. (2013) GLI2 cooperates with ZEB1 for transcriptional repression of CDH1 expression in human melanoma cells. *Pigment Cell Melanoma Res.*, **26**, 861–873.
99. Koyabu,Y., Nakata,K., Mizugishi,K., Aruga,J. and Mikoshiba,K. (2001) Physical and functional interactions between Zic and Gli proteins. *J. Biol. Chem.*, **276**, 6889–6892.
100. Sánchez-Tilló,E., de Barrios,O., Valls,E., Darling,D.S., Castells,A. and Postigo,A. (2015) ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene*, **34**, 5760–5770.
101. Mendoza-Parra,M.A., Van Gool,W., Mohamed Saleem,M.A., Ceschin,D.G. and Gronemeyer,H. (2013) A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res.*, **41**, e196.
102. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, **4**, 209–223.
103. Devailly,G., Mantsoki,A., Michoel,T. and Joshi,A. (2015) Variable reproducibility in genome-scale public data: a case study using ENCODE ChIP sequencing resource. *FEBS Lett.*, **589**, 3866–3870.
104. Stergachis,A.B., Neph,S., Reynolds,A., Humbert,R., Miller,B., Paige,S.L., Vernot,B., Cheng,J.B., Thurman,R.E., Sandstrom,R. *et al.* (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, **154**, 888–903.
105. Li,L. and Wunderlich,Z. (2017) An enhancer's length and composition are shaped by Its regulatory task. *Front Genet*, **8**, 63.
106. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
107. Cooper,G.M., Goode,D.L., Ng,S.B., Sidow,A., Bamshad,M.J., Shendure,J. and Nickerson,D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
108. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
109. Cooper,G.M., Goode,D.L., Ng,S.B., Sidow,A., Bamshad,M.J., Shendure,J. and Nickerson,D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
110. Visel,A., Prabhakar,S., Akiyama,J.A., Shoukry,M., Lewis,K.D., Holt,A., Plajzer-Frick,I., Afzal,V., Rubin,E.M. and Pennacchio,L.A. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, **40**, 158–160.
111. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
112. Katzman,S., Kern,A.D., Bejerano,G., Fewell,G., Fulton,L., Wilson,R.K., Salama,S.R. and Haussler,D. (2007) Human genome ultraconserved elements are ultraselected. *Science*, **317**, 915.
113. Li,S., Kvon,E.Z., Visel,A., Pennacchio,L.A. and Ovcharenko,I. (2019) Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation. *Genome Biol.*, **20**, 140.
114. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
115. Villar,D., Berthelot,C., Aldridge,S., Rayner,T.F., Lukk,M., Pignatelli,M., Park,T.J., Deaville,R., Erichsen,J.T., Jasinska,A.J. *et al.* (2015) Enhancer evolution across 20 mammalian species. *Cell*, **160**, 554–566.
116. Young,R.S., Kumar,Y., Bickmore,W.A. and Taylor,M.S. (2017) Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol.*, **18**, 242.
117. Chereji,R.V., Eriksson,P.R., Ocampo,J., Prajapati,H.K. and Clark,D.J. (2019) Accessibility of promoter DNA is not the primary

determinant of chromatin-mediated gene regulation. *Genome Res.*, **29**, 1985–1995.
118. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
119. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
120. Khan,A. and Zhang,X. (2015) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, **44**, D164–D71.
121. Dreos,R., Ambrosini,G., Cavin Perier,R. and Bucher,P. (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.*, **41**, D157–D164.
122. Dimitrieva,S. and Bucher,P. (2013) UCNEbase–a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, **41**, D101–D109.
123. Wilderman,A., VanOudenhove,J., Kron,J., Noonan,J.P. and Cotney,J. (2018) High-resolution epigenomic atlas of human embryonic craniofacial development. *Cell Rep.*, **23**, 1581–1597.
124. 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.
125. Gao,T., He,B., Liu,S., Zhu,H., Tan,K. and Qian,J. (2016) EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics*, **32**, 3543–3551.
126. Pennacchio,L.A., Bickmore,W., Dean,A., Nobrega,M.A. and Bejerano,G. (2013) Enhancers: five essential questions. *Nat. Rev. Genet.*, **14**, 288–295.
127. Kellis,M., Wold,B., Snyder,M.P., Bernstein,B.E., Kundaje,A., Marinov,G.K., Ward,L.D., Birney,E., Crawford,G.E., Dekker,J. *et al.* (2014) Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 6131–6138.
128. Snyder,M.P., Gingeras,T.R., Moore,J.E., Weng,Z., Gerstein,M.B., Ren,B., Hardison,R.C., Stamatoyannopoulos,J.A., Graveley,B.R., Feingold,E.A. *et al.* (2020) Perspectives on ENCODE. *Nature*, **583**, 693–698.
129. Wang,X., He,L., Goggin,S.M., Saadat,A., Wang,L., Sinnott-Armstrong,N., Claussnitzer,M. and Kellis,M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.*, **9**, 5380.
130. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
131. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
132. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
133. Graur,D., Zheng,Y., Price,N., Azevedo,R.B., Zufall,R.A. and Elhaik,E. (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.*, **5**, 578–590.
134. Galeota-Sprung,B., Sniegowski,P. and Ewens,W. (2020) Mutational load and the functional fraction of the human genome. *Genome Biol Evol*, **12**, 273–281.
135. Ponting,C.P. and Hardison,R.C. (2011) What fraction of the human genome is functional? *Genome Res.*, **21**, 1769–1776.
136. King,D.C., Taylor,J., Elnitski,L., Chiaromonte,F., Miller,W. and Hardison,R.C. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, **15**, 1051–1060.
137. Rands,C.M., Meader,S., Ponting,C.P. and Lunter,G. (2014) 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.*, **10**, e1004525.
138. Graur,D. (2017) An upper limit on the functional fraction of the human genome. *Genome Biol Evol*, **9**, 1880–1885.
139. Huber,C.D., Kim,B.Y. and Lohmueller,K.E. (2020) Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS Genet.*, **16**, e1008827.