A map of *cis*-regulatory modules and constituent transcription factor binding sites in 77.5% regions of the human genome

Pengyu Ni[1]

Zhengchang Su [1,*]

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

Contact Information:

Zhengchang Su

Email:zcsu@uncc.edu

Tel: +01-704-687-7996

Fax: +01-704-687-8667

Running title: A map of *cis*-regulatory modules in the human genome

ABSTRACT

Annotating all cis-regulatory modules (CRMs) in genomes is essential to understand genome functions, however, it remains an uncompleted task despite great progresses made since the development of ChIP-seq techniques. As a continued effort, we developed a new algorithm dePCRM2 for predicting CRMs and constituent transcription factor (TF) binding sites (TFBSs) by integrating numerous TF ChIP-seq datasets based on a new ultra-fast, accurate motif-finding algorithm and an efficient combinatory motif pattern discovery method. dePCRM2 partitions genome regions covered by extended binding peaks in the datasets into a CRM candidates (CRMCs) set and a non-CRMCs set, and predicts CRMs and constituent TFBSs by evaluating each CRMC using a novel score. Applying dePCRM2 to 6,092 datasets covering 77.47% of the human genome, we predicted 201 unique TF binding motif families and 1,404,973 CRMCs. Intriguingly, the CRMCs are under stronger evolutionary constraints than the non-CRMCs, and the higher a CRMC can score, the stronger evolutionary constraint it receives and the more likely it is a full-length enhancer. When evaluated on functionally validated VISTA enhancers and causal ClinVar mutants, dePCRM2 achieves 97.43~100.00% sensitivity at p-value ≤ 0.05. dePCRM2 also largely outperforms existing methods in sensitivity and specificity, as well as by the evaluation of evolution constraints. Based on our predictions and evolutionary behaviors of the genome, we estimated that about 21.95% and 54.87% of the genome might code for TFBSs and CRMs, respectively, for which we predicted 80.21%.

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 recognized and bound by specific transcription factors (TFs), and their functional states are responsible for specific transcriptomes in various cell types in multi-cellular eukaryotes. A growing body of evidence indicates that it is mainly

CRMs, rather than coding sequences (CDSs), that mainly account for inter-species divergence and intra-species diversity (King and Wilson 1975; Wittkopp and Kalay 2012; Haraksingh and Snyder 2013; Rubinstein and de Souza 2013; Siepel and Arbiza 2014; Reilly et al. 2015). Recent genome-wide association studies (GWAS) have found that most complex trait-associated single nucleotide polymorphisms (SNPs) do not reside in CDSs, but rather lie in non-coding sequences (NCSs) (Hindorff et al. 2009; Ramos et al. 2014), and often overlap with or are in linkage disequilibrium (LD) with TF binding sites (TFBSs) in CRMs (Maurano et al. 2012b). It has also been shown that complex trait-associated variants systematically disrupt TFBSs of TFs related to the traits (Maurano et al. 2012a), and that variation in TFBSs affects DNA binding, chromatin modification, transcription (Kasowski et al. 2013; Kilpinen et al. 2013; McVicker et al. 2013; Wu et al. 2013; Huang and Ovcharenko 2015), and susceptibility to complex diseases (Majewski and Pastinen 2011; Attanasio et al. 2013; Fu et al. 2013; Spielmann and Mundlos 2013; Smith and Shilatifard 2014; Mathelier et al. 2015) including cancer (Herz et al. 2014; Ongen et al. 2014; HD 2015; Heyn 2016; Heyn et al. 2016; Khurana et al. 2016; Zhou et al. 2016; Li et al. 2018). In principle, variation in a CRM may result in changes in the affinity and interactions between TFs and their binding sites, leading to alterations in histone modifications and target gene expressions in relevant cells. These alterations in molecular phenotypes can lead to changes in cellular and organ-related phenotypes among individuals of a species (Ward and Kellis 2012; Pai et al. 2015). However, it has been difficult to link non-coding variants to phenotypes (Collins 2010; Burga and Lehner 2012; Paul et al. 2014; Albert and Kruglyak 2015; Whitaker et al. 2015), largely because of our lack of a good understanding of all CRMs and their constituent TFBSs in genomes.

Fortunately, the recent development of ChIP-seq techniques for locating TFBSs of a TF in the genomes of specific cell/tissue types (Johnson et al. 2007; Robertson et al. 2007; Chen et al. 2008) has led to the generation of enormous amounts of data by large consortia such as ENCODE (Consortium 2004; Consortium 2011; Stamatoyannopoulos et al. 2012) Roadmap Epigenomics (Bernstein et al. 2010;

Kundaje et al. 2015) and Genotype-Tissue Expression (GTEx) (Consortium 2013; Consortium 2015). These increasing amounts of ChIP-seq data for 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 at various levels (Paul et al. 2014; Kleftogiannis et al. 2015). At the lowest level, as the large number of binding-peaks in a ChIP-seq dataset dwarf earlier motif-finding tools (e.g., MEME(Bailey and Elkan 1994), WEEDER (Pavesi et al. 2001; Pavesi et al. 2004), Seeder (Fauteux et al. 2008) and BioProspector (Liu et al. 2001), new motif-finders (e.g., Trawler (Ettwiller et al. 2007), ChIPMunk (Kulakovskiy et al. 2010), HMS (Hu et al. 2010), CMF (Mason et al. 2010), STEME (Reid and Wernisch 2011), DREME (Bailey 2011), MEME-ChIP (Machanick and Bailey 2011), MICSA (Boeva et al. 2010), DECOD (Huggins et al. 2011), RSAT (Thomas-Chollier et al. 2012), POSMO (Ma et al. 2012), XXmotif (Hartmann et al. 2013), EXTREME (Quang and Xie 2014), FastMotif (Colombo and Vlassis 2015) and Homer (Heinz et al. 2010)) have been designed. However, some of these tools (e.g. MEME-ChIP, MICSA and CMF) were designed to find motifs in very short sequences (~200bp) around the binding-peak summits in a limited number of selected binding peaks due to their slow speed. Some faster tools (e.g., Seeder, ChIPMunk, DECOD, RSAT, POSMO, Homer, DREME, and XXmotif) are based on the discriminative motif-finding schema(Sinha and Tompa 2003) by finding overrepresented k-mers in a ChIP-seq dataset, but they often fail to identify TFBSs with subtle degeneracy. As TFBSs form CRMs for combinatory regulation (Gerstein et al. 2010; Chen and Zhou 2011; Negre et al. 2011; Whitington et al. 2011; Zhang et al. 2011; Bailey and Machanick 2012; Sun et al. 2012) tools (such as SpaMo (Whitington et al. 2011), CPModule (Sun et al. 2012), COPS (Ha et al. 2012), INSECT (Rohr et al. 2013), CCAT (Jiang and Singh 2014) and others (Muller-Molina et al. 2012; Vandenbon et al. 2012; Mathelier and Wasserman 2013)) 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

(Wingender et al. 2001) or JASPAR (Vlieghe et al. 2006)) to scan for cooperative TFBSs in binding peaks.

Methods for predicting CRMs based on multiple epigenetic marks have been developed using hidden Markov models (Won et al. 2008; Won et al. 2009; Ernst and Kellis 2010; Ernst et al. 2011; Ernst and Kellis 2012), dynamic Bayesian networks(Hoffman et al. 2012; Hoffman et al. 2013), time-delay neural networks random forest (Firpi et al. 2010; Rajagopal et al. 2013), and support vector machines (SVMs) (Villarroel et al. 2011; Kleftogiannis et al. 2015). Sequence features have also been used to predict tissue-specific enhancers using SVM (Ghandi et al. 2014; Kleftogiannis et al. 2015). Several enhancer databases have also been compiled either by combining results of multiple methods (Ashoor et al. 2015; Zerbino et al. 2015; Fishilevich et al. 2017) or by identifying overlapping regions of multiple epigenetic data tracks in a cell type (Gao et al. 2016; Cheneby et al. 2018; Zhang et al. 2018; Kang et al. 2019; Chen et al. 2020; Gao and Qian 2020). 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. Further, the results of these methods are quite inconsistent (Kwasnieski et al. 2014; Dogan et al. 2015; Fishilevich et al. 2017; Catarino and Stark 2018; Arbel et al. 2019), 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 DNase I hypersensitivity sites (DHSs) in Hela cells (Kleftogiannis et al. 2015), and although 1.3 million enhancers have been predicted using epigenetic marks (Wang et al. 2018a), few disease-associated non-coding SNPs map to them (Handel et al. 2017). Although some predictions provide TFBSs information by finding matches to known motifs in predicted CRMs (Ashoor et al. 2015; Fishilevich et al. 2017; Gao and Qian 2020), these methods were unable to identify TFBSs of novel motifs in CRMs.

Surprisingly, while TF ChIP-seq data provide more accurate information for TF-binding locations and their combinatory patterns in CRMs than epigenetic marks (Niu et al. 2014; Dogan et al. 2015; Kleftogiannis et al. 2015; Niu et al. 2018; Arbel et al. 2019), few efforts have been made to fully explore

the increasing volume of datasets due to technical difficulty to integrate them (Yip et al. 2012; Kheradpour and Kellis 2014; Niu et al. 2014; Niu et al. 2018). With this recognition, we have previously proposed a different strategy to first predict a catalog or a static map of CRMs and constituent TFBBs in the genome by integrating all available TF ChIP-seq datasets for different TFs in various cell/tissue types (Niu et al. 2014; Niu et al. 2018) as has been done for identifying all genes encoded in the genome (Allen et al. 2004). Once a map of CRMs and constituent TFBSs is available, the specificity of CRMs in any cell/tissue type can be determined using a single dataset collected in the cell/tissue type, reflecting their functional states. Although very promising results have been obtained using even insufficient datasets available then (Niu et al. 2014; Niu et al. 2018), there are three limitations in our earlier algorithm dePCRM to integrate much larger datasets available now and in the future. First, although existing motif-finders such as DREME used in dePCRM work well for relatively small ChIP-seq datasets from organisms with smaller genomes such as the fly (Niu et al. 2014), they are too slow for very large datasets from human cells/tissues, so we had to split large datasets into smaller ones for the motif finding (Niu et al. 2018), which may compromise the accuracy of motif finding and complicate subsequent data integration. Second, the distance and interactions of TFBSs in a CRM were not explicitly considered (Niu et al. 2018), potentially limiting the accuracy. Third, the original "branch-and-bound" approach to integrate motifs is not efficient enough to handle much larger number of motifs found in ever increasing number of large ChIP-seq datasets from human cells/tissues. To overcome these drawbacks, we developed a new CRM predictor dePCRM2 that combines an ultrafast, accurate motif-finder ProSampler (Li et al. 2019) with a novel effective combinatory motif pattern discovery method. Applying dePCRM2 to available 6,092 ChIP-seq datasets covering 77.47% of the human genome, we predicted 201 unique TF binding motif families and 1,404,973 CRM candidates (CRMCs). Both evolutionary and independent experimental data indicate that dePCRM2 achieves very high sensitivity and specificity in predicting CRMs and constituent TFBSs.

RESULTS

Design of the dePCRM2 algorithm

We overcome the aforementioned shortcomings of dePCRM (Niu et al. 2014; Niu et al. 2018) as follows. First, using an ultrafast, accurate motif-finder ProSampler (Li et al. 2019), dePCRM2 can find significant motifs in any size of available ChIP-seq datasets without the need to split large datasets into small ones (Figures 1A, 1B). Second, after identifying highly co-occurring motifs pairs (CPs) in the extended binding peaks in each dataset (Figure 1C), dePCRM2 clusters highly similar motifs in the CPs and finds a unique motif (UM) in each resulting cluster (Figure D). Third, dePCRM2 models interactions among cognate TFs of the binding sites in a CRM by constructing interaction networks of the UMs based on the distance between the binding sites and the extent to which biding sites in the UMs cooccur (Figure 1E). Fourth, dePCRM2 identifies 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 in the datasets into a CRMCs set and a non-CRMCs set. Fifth, dePCRM2 evaluates each CRMC using a novel score that considers the quality of binding sites as well as the strength of interactions among the corresponding UMs defined in the interaction networks (Figure 1G). Lastly, dePCRM2 computes 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, thus dePCRM2 is highly scalable.

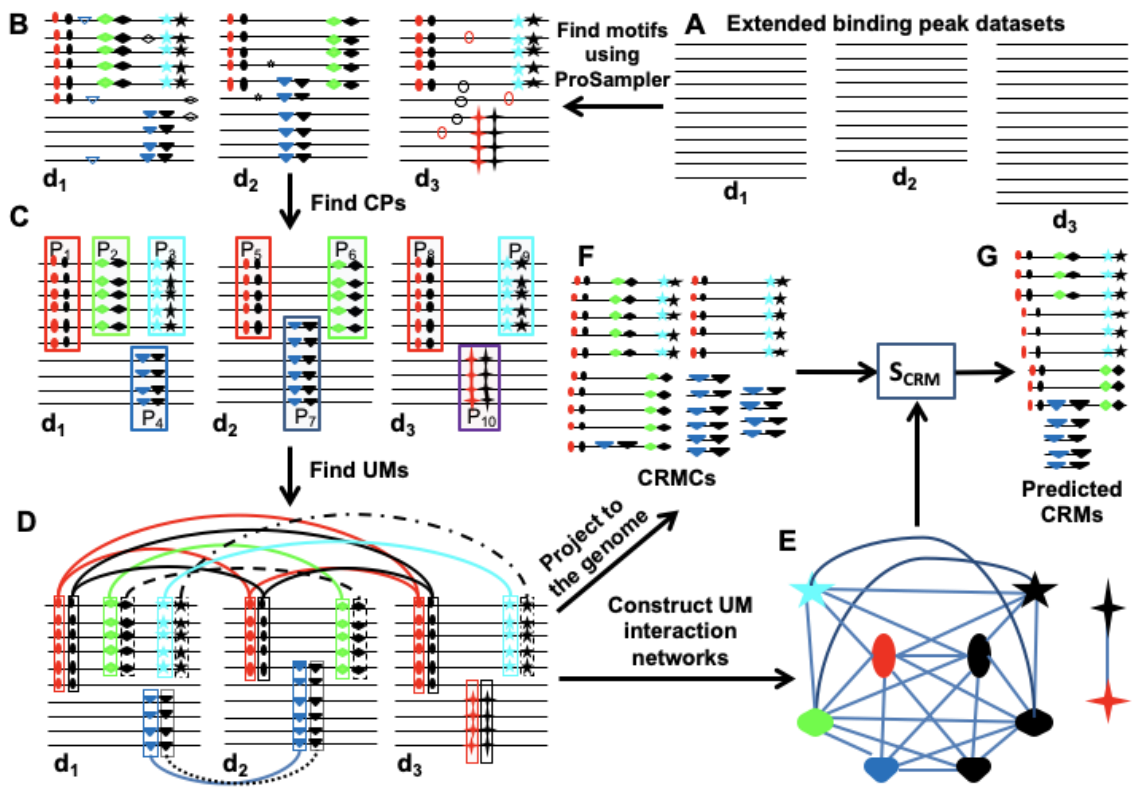

Figure 1. Schematic of the dePCRM2 algorithm. A. Extend each binding peak in each dataset to its two ends to reach a preset length, e.g., 1,000bp. B. Find motifs in each dataset using ProSampler. C. Find CPs in each dataset. For clarity, only the indicated CPs are shown, while those formed between motifs in pairs P1 and P2 in d1, 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 represents a motif, weights on the edges are omitted for clarity. Clusters are connected by edges of the same color and line type. E. Construct UM interaction networks. Each node in the networks represents a UM, weights on the edges are omitted for clarity. F. Project binding sites in the UMs back to the genome and identify CRMCs along the genome. G. Evaluate each CRMC by computing its S_{CRM} score and the associated p-value.

Extended binding peaks in the datasets cover 77.5% of the genome and have extensive overlaps

After filtering out low-quality peaks in the 6,092 ChIP-seq datasets, we ended up with 6,070 non-empty datasets for 779 TFs in 2,631 cell/tissue/organ types. The datasets are strongly biased to few cell types (Figure 2A). For example, 532, 475 and 309 datasets were collected from mammary gland epithelium, colon epithelium, and bone marrow erythroblast, respectively, while only one dataset was generated from 129 cell/tissue types, including heart embryonic fibroblast, fetal skin fibroblast, and bone marrow haematopoietic progenitor, and so on (Figure 2A). The datasets also are strongly biased to few TFs

(Figure 2B). For example, 682, 370 and 263 datasets were collected for TFs POLR2A, CTCF and ESR1, respectively, while just one dataset was produced for 324 TFs, such as MSX2, RAX2, and MYNN, and so on. The number of called binding peaks in a dataset is highly varying, ranging from 2 to 100,539, with an average of 19,314 (Figure 2C). For instance, datasets for STAT1, POLR2A, and NR3C1 have the smallest number of 2 binding peaks in HeLa-S3, HeLa-S3, and HEK293 cells, respectively, while datasets for CEBPB, BRD4 and FOXA2 have the largest number of 115,776, 99,646, and 99,512 binding peaks in HepG2, U87, and Mesenchymal Stem Cells, respectively. The highly varying numbers of binding peaks in the datasets suggest that different TFs might bind a highly varying number of sites in the genomes of cells. However, some datasets with very few binding peaks might be resulted from technical artifacts, thus are of low quality (see below), even though they passed our quality filtering. The lengths of binding peaks in the datasets range from 75 to 10,143bp, with a mean of 300bp (Figure 2D), and 99.12% of binding peaks are shorter than 1,000pb. All the binding peaks in the 6,070 datasets cover a total of 1,265,474,520bp (40.98%) of the genome (3,088,269,832bp). For each binding peak in each dataset, we extracted a 1,000bp genome sequence centering on the middle of its binding summit, thereby extending the lengths for most binding peaks. The extended binding peaks contain a total of 115,710,048,000bp, which is 37.5 times the size of the genome, but cover only 2,392,488,699bp (77.47%) of the genome, leaving the remaining 22.53% of the genome uncovered, indicating that they have extensive overlaps. Nonetheless, by extending the original binding peaks, we increased the coverage of genome by 89.04% (77.47% vs 40.98%). Notably, we may not know the functional states of some predicted binding sites in extended parts of the original binding peaks from a cell/tissue type if no binding peaks for other TFs tested in the cell/tissue type overlap the extended parts. We trade this drawback for a more complete prediction of the catalogs/map of CRMs and TFBSs in the genome (Niu et al. 2018). The incomplete coverage of the genome suggests that more diverse, less biased data for untested TFs in untested cell/tissue types are needed in the future to cover the entire functional genome. As dePCRM2 predicts

CRMs and constituent TFBSs based on overlapping patterns between datasets of cooperative TFs (Niu et al. 2014; Niu et al. 2018) (Methods and Materials), we evaluated the extent to which the extended binding peaks in different datasets overlap one another. To this end, we hierarchically clustered the 6,070 datasets using an overlap score between each pair of the datasets (formula 1). As shown in Figure 2E, there are extensive distinct overlapping patterns among the datasets. As expected, clusters are formed by datasets for largely the same TF in different cell/tissue types, and/or by datasets for different TFs that are known or potential collaborators in transcriptional regulation. For instance, a cluster is formed by 1, 1 and 48 datasets for RAD21, SMC3 and CTCF, respectively, in various cell/tissue types (Figure 2F). It is well-known that RAD21, SMC3 and CTCF are core subunits of the cohensin complex, and are widely colocalized in mammalian genomes (Zuin et al. 2014). In another example (Figure 2G), a cluster is formed by 50 datasets for 45 TFs in various cell/tissue types. Multiple sources of evidence indicate that these 45 TFs have extensive physical interactions for DNA binding and transcriptional regulation (Figure 2H) (Orchard et al. 2014; Oughtred et al. 2019). For example, it has been shown that LEF1 interacts with the TFG beta activating regulator SMAD4 (Labbé et al. 2000), E2F4 helps to recruit PML to the TBX2 promoter (Martin et al. 2012), and FOXK 1 and TP53 can form a distinct protein complex on and off chromatin (TPLi et al. 2015). These overlapping patterns between the extended binding peaks in the datasets warrant us to predict CRMs and constituent TFBSs in the covered 77.47% of the genome (Niu et al. 2014; Niu et al. 2018)(Methods and Materials). In addition, 785 (80.43%) VISTA enhancers (Visel et al. 2007), 181,436 (98.38%) FANTOM5 promoters (Consortium et al. 2014), 32,029 (97.98%) FANTOM5 enhancers (Andersson et al. 2014), 402,730 (94.84%) of ClinVar single nucleotide polymorphisms (SNPs) (Landrum et al. 2016; Landrum and Kattman 2018), 82,378 (90.16%) of GWAS SNPs (Buniello et al. 2019) and 2,734,505 (94.68%) DNase I hypersensitive sites (DHSs) (MacArthur et al. 2017; Buniello et al. 2019) are located in the covered 77.47% genome regions, indicating that these experimentally determined sequence elements are highly enriched in these covered regions compared

to the uncovered 22.53% genome regions. We will evaluate the sensitivity of dePCRM2 to recall these elements at different S_{CRM} scores and associated p-value cutoffs.

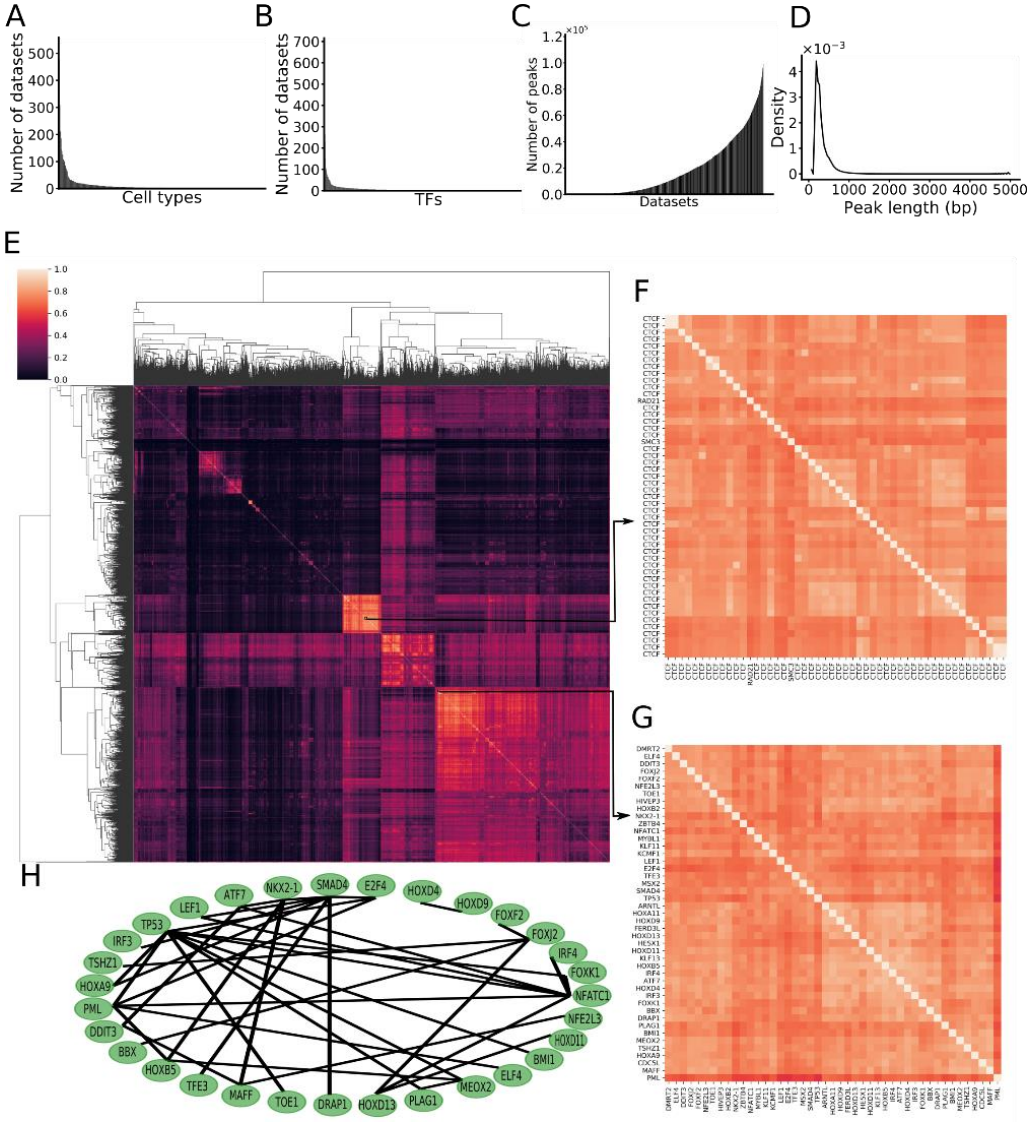

Figure 2. Properties of the datasets. A. Number of datasets collected in each cell/tissue types sorted in descending order. B. Number of datasets collected for each TF sorted in descending order. C. Number of peaks in each dataset sorted in ascending order. D. Distribution of the lengths of binding peaks in the entire datasets. E. Heatmap of overlaps of extended binding peaks between each pair of the datasets. F. A blowup view of the indicated cluster in E, formed by 48 datasets for CTCF in different cell/tissue types, as well as one dataset for each of its two collaborators, RAID21 and SMC3. G. A blowup view of the indicated cluster in E, formed by 50 datasets for 45 TFs. H. Known physical interactions between the 45 TFs whose 50 datasets form the cluster in E.

Unique motifs recover most known TF motifs families and have distinct patterns of cooccurrence.

ProSampler identified at least one motif in 5,991 (98.70%) datasets but failed to find any motifs in the remaining 79 (1.30%) datasets that all contain less than 310 binding peaks, indicating that they are likely of low quality. As shown in Figure 3A, 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 5,991 datasets with at least one motif found. dePCRM2 finds 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 2). As shown in Figure 3B, S_c scores show a three-mode distribution. dePCRM2 selects as CPs the motif pairs that account for the mode with the highest S_c scores, and discards those that account for the other two modes with lower S_c scores (by default, S_c<0.7), because these low-scoring motif pairs are likely to co-occur by chance. In total, dePCRM2 identified 4,455,838 CPs containing 226,355 (26.4%) motifs from 5,578 (93.11%) of the 5,991 datasets. Therefore, we filtered out 413 (6.89%) of the 5,991 datasets because each had a low S_c score compared with other datasets. Clearly, more and balanced datasets are needed to rescue their use in the future for more complete predictions. Clustering the 226,355 motifs in the CPs resulted in 245 clusters consisting of 2~72,849 motifs, most of which form a complete similarity graph or clique, indicating that member motifs in a cluster are highly similar to each other (Supplementary Figure S1A). dePCRM2 found a UM in 201 of the 245 clusters (Supplementary Figure S1B and Table S1) but failed to do so in 44 clusters due to the low similarity between some member motifs (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 retains 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 an average of 905,288 (Figure 3C and Table S1), reminiscent of highly varying number of binding peaks in the datasets (Figure 2D). The

lengths of the UMs range from 10bp to 21pb with a mean of 11pb (Figure 3D), which are in the range of the lengths of known TF binding motifs, although they are biased to 10bp due to the limitation of the motif-finder to find longer motifs. As expected, a UM is highly similar to its member motifs that are highly similar to each other (Supplementary Figure S1A). For example, UM44 contains 250 highly similar member motifs (Figure 3E). Of the 201 UMs, 117 (58.2%) match at least one of the 856 annotated non-redundant motifs in the HOCOMOCO (Kulakovskiy et al. 2013; Kulakovskiy et al. 2018) and JASPAR (Mathelier et al. 2016) databases (supplementary Table S1), suggesting that these UMs might be recognized by cognate TFs of matched known motifs. The remaining 84 of the 201 UMs might be novel ones recognized by unknown TFs (Supplementary Table S1). Of the 117 UMs matching at least a known motif, 92 (78.63%) match at least two, 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 (Weirauch et al. 2014; Lambert et al. 2018; Ambrosini et al. 2020). Thus, 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 3 adjacent zinc finger factors” family (Figure 3F and Supplementary Table S1). The high similarity of these motifs suggest that they might form a superfamily. On the other hand, 64 (71.91%) of the 89 annotated motifs TF families match one of the 201 UMs (Supplementary Table S1), thus our predicted UMs include most of the known TF motif families. To model interactions between cognate TFs of the UMs, dePCRM2 computes interaction scores between the UMs (formula 3). As shown in Figure 3G, there are extensive interactions between the UMs, which indeed reflect the interactions among their cognate TFs or TF families in transcriptional regulation. For example, in a cluster formed by 10 UMs (Figure 3H), 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 cooperates with ZEB1 for repressing expression of CDH1 gene in human melanoma cells via direct binding two close binding sites at CDH promoter (Perrot et al. 2013) , ZIC and GLI cooperatively regulate neural and skeletal development through physical interactions between their zinc finger domains (Koyabu et al. 2001), and ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate WNT target gene expression (Sánchez-Tilló et al. 2015), to name a few.

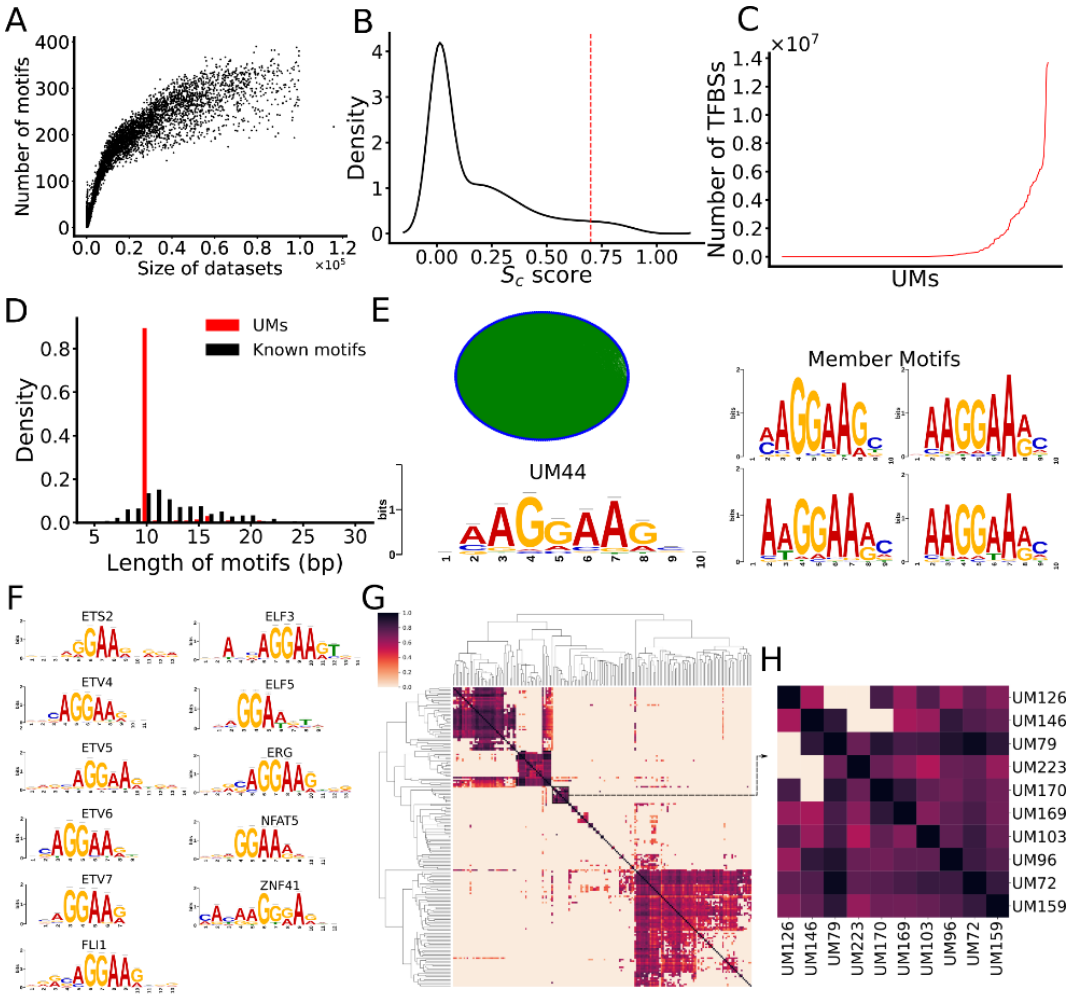

Figure 3. Prediction of UMs. A. Relationship between the number of predicted motifs in a dataset and the size of the dataset (number of binding peaks in the dataset). The datasets are sorted in ascending order of their sizes. B. Distribution of cooccurrence scores (Sc) of motif pairs found in a dataset. The dotted vertical line indicates the cutoff value of Sc 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. The 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 left panel. F. UM44 matches known motifs of nine TFs of the "ETS", "NFAT-related factor", and "more than 3 adjacent zinc finger factors" families. G. Heatmap of the interaction networks of the 201 UMs, names of most 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 the TFs are known collaborators in transcriptional regulation.

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

By concatenating closely located binding sites of the UMs along the genome, dePCRM2 partitions the extended binding peak-covered genome regions (2,392,488,699bp) in two exclusive sets, the CRMCs set containing 1,404,973 CRMCs with a total length of 1,359,824,275bp (56.84%) and the non-CRMCs set with a total length of 1,032,664,424bp (43.16%), covering 44.03% and 33.44% of the genome, respectively. Interestingly, 57.88% (776,999,862bp) of nucleotide positions of the CRMCs overlap those of the original binding peaks, while the remaining 42.12% (565,448,583bp) nucleotide positions are covered by the extended parts of the original binding peaks. Hence, in predicting CRMCs dePCRM2 used only 61.40% of nucleotide positions and abandoned the remaining 38.60% covered by the original binding peaks (1,265,512,389bp), while extending them by 44.68% (5654,48,583pb) covered the extended parts. If a CRMC overlaps an original binding peak in a cell/tissue type, we predict the CRMC to be active in the cell/tissue type. Thus, we could predict functional states of about 57.88% of the CRMCs in at least one of the cell/tissue types, from which data were used in the prediction. However, we could not predict functional states of the remaining 42.12% of the CRMCs that do not overlap any original binding peaks.

As shown in Figure 4A, the distribution of the S_{CRM} scores of the CRMCs is strongly right-skewed relative to that of the same number of Null CRMCs (Materials and Methods), indicating that the CRMCs generally score higher than NULL CRMCs, and thus are unlikely produced by chance. Based on the distribution of the S_{CRM} scores of Null CRMCs, dePCRM2 computes a p-value for each CRMC (Figure 4A).

With the increase in the S_{CRM} cutoff α ($S_{CRM} \geq \alpha$), the associated p-value drops rapidly, while both the number of predicted CRMs and the proportion of the genome predicted to be CRMs decrease slowly, indicating that dePCRM2 might achieve high prediction specificity (Figure 4B). Specifically, with α increasing from 56 to 922, p-value drops precipitously from 0.05 to 1.00×10^{-6}, while the number of predicted CRMs decreases from 1,155,151 to 327,396, and the proportion of the genome predicted to be CRMs decreases from 43.47% to 27.82% (Figure 4B). Predicted CRMs contain from 20,835,542 (p-value ≤ 1×10^{-6}) to 31,811,310 (p-value ≤ 0.05) non-overlapping putative TFBSs that consist of from 11.47% (p-value≤ 1×10^{-6}) to 16.54% (p-value ≤ 0.05) of the genome (Figure 4C). In other words, 38.05~41.23% of predicted nucleotide positions in the predicted CRMs are made of putative TFBSs (Figure 4C). As expected, most of the predicted CRMs (93.99~95.46%) and constituent TFBSs (93.20~94.67%) are located in non-exonic sequences (NESs) (Figure 4C), comprising 26.66~42.47% and 10.94~16.03% of NESs, respectively (Figure 4D). Surprisingly, the remaining 4.54~6.01% and 5.33~6.80% of the predicted CRMs and constituent TFBSs, respectively, are in an exon (Figure 4C), comprising 76.82%~85.50% and 35.42~38.17% of exonic sequences (ESs, including CDSs, 5'- and 3'-untranslated regions), respectively (Figure 4D).

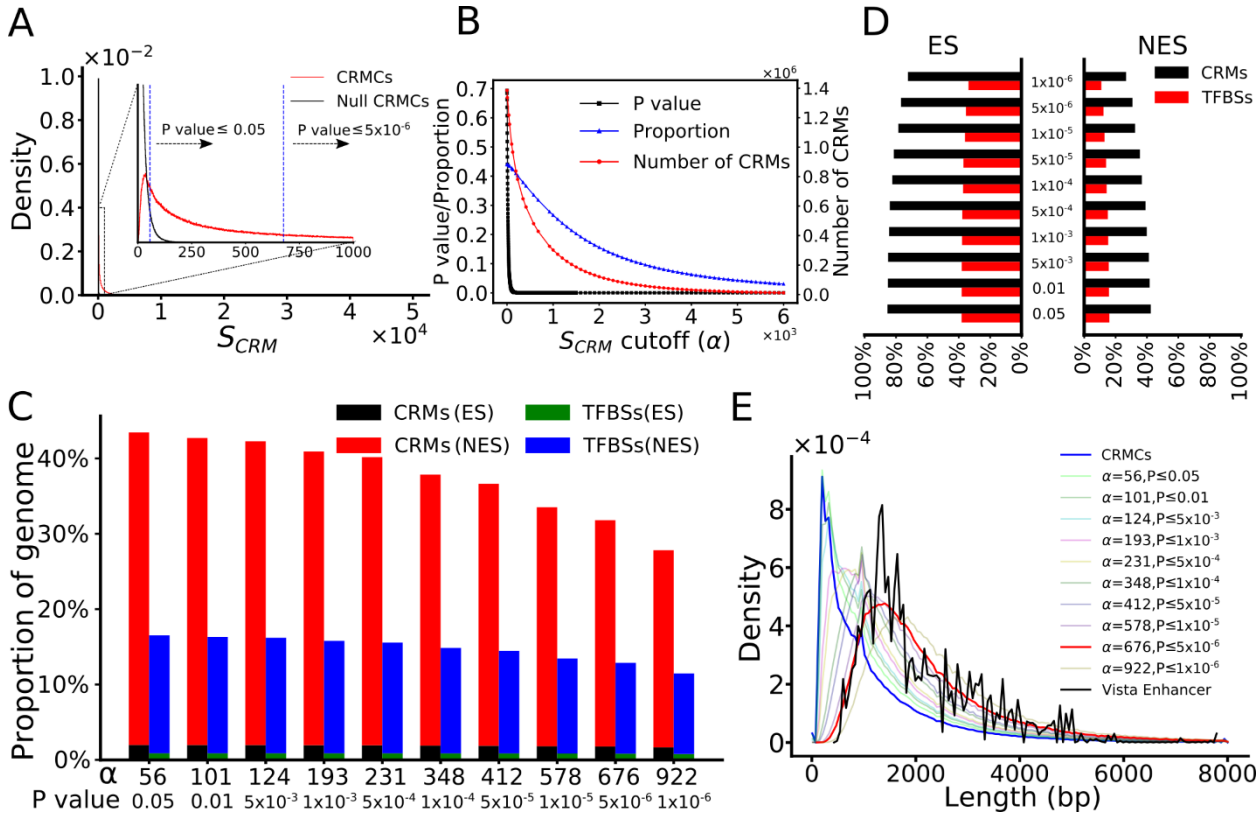

Figure 4. Prediction of CRMs using different S_{CRM} cutoffs. A. Distribution of S_{CRM} scores of the CRMCs and Null CRMCs. B. Number of the predicted CRMs, proportion of the genome predicted to be CRMs and the associated p-value as functions of the S_{CRM} cutoff α. C. Percentage of the genome that are predicted to be CRMs and TFBSs in ESs and NESs using various S_{CRM} cutoffs and associated p-values. D. Percentage of NESs and ESs that are predicted to be CRMs and TFBSs using various S_{CRM} cutoffs and associated p-values. E. Distribution of the lengths of CRMs predicted using different S_{CRM} cutoffs and associated p-values.

We designed the S_{CRM} score to capture essential features of enhancers including their lengths. To see whether it achieves this goal, we compared the distribution of the lengths of predicted CRMCs at different S_{CRM} cutoffs α and associated p-values with those of functionally verified VISTA enhancers. As shown in Figure 4E, most of CRMCs are quite short (average length 981bp) compared to VISTA enhancers (average length 2,049bp). Specifically, almost half (621,842 or 44.26%) of the 1,404,973 CRMCs are shorter than the shortest VISTA enhancer (428bp), suggesting that most of these short CRMCs are likely short CRMs (such as promoters) or parts/components of long enhancers. However, these short CRMCs (< 428bp, 44.26%) consist of only 7.42% of the total length of the CRMCs, while the

remaining 733,132 (55.74%, >428bp) CRMCs comprise 92.58% of the total length of the CRMCs. With the increase in α (decrease in p-value cutoff), the distribution of the lengths of the predicted CRMs shifts to right and gradually approaches to that of VISTA enhancers (Figure 4E). Intriguingly, at α=676 (p-value ≤ $5x10^{-6}$), the distribution fits very well to that of VISTA enhancers, indicating that the corresponding 428,628 predicted CRMs have similar length distribution (average length 2,292bp) to those of VISTA enhancers (average length 2,049bp) (Figure 4E), and thus they are likely full-length VISTA-like enhancers. These results demonstrate that short CRMs and CRM components tend to have smaller S_{CRM} scores than full-length enhancers, and can be effectively filtered out by a higher S_{CRM} cutoff α (a smaller p-value). Therefore, S_{CRM} indeed captures the length property of full-length enhancers. Although these 428,628 putative full-length VISTA-like CRMs consist of only 30.51% of the 1,404,973 CRMCs, they comprise 72.25% (982,470,181bp) of the total length (1,359,824275bp) of the CRMCs, while the remaining 976,345 (69.49%) short CRMCs consist of only 27.75% of the total length of the CRMCs, indicating that full-length VISTA-like enhancers dominate the CRMCs in length. The 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 (average length 182bp) using a much smaller number and less diverse 670 datasets (Niu et al. 2018). As we argued earlier (Niu et al. 2018) and confirmed here by the much longer CRMCs (average length 982bp) predicted using the much larger and more diverse datasets albeit still strongly biased to a few TFs and cell/tissue types. We anticipate that full-length CRMs of these short CRM components can be predicted using even larger and more diverse TF ChIP-seq data when available in the future.

Predicted CRMs tend to be under either positive selection or negative selection

To see how effective that dePCRM2 partitions the covered genome regions into the CRMCs set and the non-CRMCs set, we compared their evolutionary behaviors using the GERP (Cooper et al. 2010) and

phyloP (Pollard et al. 2010) scores of their nucleotide positions in the human genome. Both the GERP and the phyloP scores quantify conservation levels of nucleotide positions in the genome 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 selection; the smaller a negative GERP or phyloP score of a position, the more likely it is under positive selection; and a GERP or phyloP score around zero means that the position is selectively neutral or nearly so. For convenience of discussion, we arbitrarily consider a position with a GERP or phyloP score within [-1, 1] to be selectively neutral or nearly so, and a position with a score greater than 1 or less than –1 to be under negative or positive selection, respectively. 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 the scores of the positions within the window [-1, 1]. Because ESs evolve quite differently from NESs, we focused on the CRMCs and constituent TFBSs in NESs and left those that at least partially overlap ESs in another analysis. Intriguingly, 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 (Figure 5A), suggesting that most of the non-CRMCs are selectively neutral or nearly so. In sharp contrast, the distribution of the GERP scores of the 1,292,356 CRMCs (1,298,719,954bp) in NESs has a blunt peak around score 0, with high right and left shoulders (Figure 5A), suggesting that most of the CRMCs are under either strong negative selection or positive selections. More specifically, the CRMCs have a much smaller proportion of neutrality (0.3105) than the non-CRMCs (0.7053) and a much larger proportion (0.6995) under evolutionary constraints than the non-CRMCs (0.2947) (Figure 5A). Similar results are seen using the phyloP scores (supplementary Figure S3A). Therefore, at least most of the CRMCs are likely to be functional while at least most of the non-CRMCs might not be functional. In other words, dePCRM2 effectively partitions the covered genome regions into a functional set and a non-functional set.

Remarkably, the abandoned nucleotide positions covered by the original binding peaks have a GERP score distribution almost identical to those in the non-CRMCs, thus they are unlikely to be functional, indicating that not all parts of the original binding peaks are functional. Thus, dePCRM2 is able to accurately distinguish functional and non-functional parts in both the original binding peaks and their extended parts. Interestingly, the uncovered 22.53% genome regions have a GERP score distribution and a proportion of neutrality (0.5922) in between those of the covered regions (0.4864) and those of the non-CRMCs (0.7053) (Figure 5A). These results indicate that the uncovered regions are more evolutionarily selected than the non-CRMCs, but less evolutionary selected 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 density of CRMs is proportional to the size of evolutionarily selected regions, the density of CRMs in the uncovered regions is (1-0.5922)/(1-0.4864)=79.40% of the covered regions. Similar results could be obtained using the phyloP scores (supplementary Figure S3A).

We then investigated the relationship between the conservation scores of the CRMCs and their S_{CRM} scores. To this end, we plotted the distributions of the conservation scores of the CRMCs with a S_{CRM} score in different non-overlapping intervals. Remarkably, the CRMCs with S_{CRM} scores in the lowest interval [0, 1) have a much smaller proportion of neutrality (0.5639) than the non-CRMCs (0.7053) (Figure 5B), indicating that even these low-scoring CRMCs with short lengths (Figure 4E) are more likely to be under strong evolutionary constraints and functional than the non-CRMCs. With the increase in the lower bound of S_{CRM} intervals, the proportion of neutrality of the CRMCs in the intervals drops rapidly, followed by a slow linear decrease around the interval [1000,1400) (Figure 5B). Therefore, the higher the S_{CRM} score of a CRMC, the more likely it is under strong evolutionary constraint, suggesting that the S_{CRM} score captures the evolutionary behavior of a CRM as a functional

element, in addition to its length feature (Figure 4E). The same conclusion can be drawn from the phyloP scores (supplementary Figure S3B).

We next examined the relationship between the conservation scores of the predicted CRMs and S_{CRM} score cutoffs α (or p-values) used for their perditions. As shown in Figure 5C, even the CRMs predicted at the low S_{CRM} cutoffs (e.g., α=0, i.e., the CRMCs) have a much smaller proportion of neutrality (0.3105) than the non-CRMCs (0.7053), suggesting that the specificity of the predicted CRMs is high, and that the non-CRMCs might contain few undiscovered CRMCs. With the increase in the S_{CRM} cutoff α (decrease in p-value), the proportion of neutrality of the predicted CRMs decreases further but slowly (Figure 5C). Thus, the higher the S_{CRM} cutoff α, the more likely the predicted CRMs are under strong evolutionary constraints, suggesting again that the S_{CRM} scores capture the essence of CRMs as functional elements. The infinitesimal difference in the proportion of neutrality of the CRMs predicted at a wide range of S_{CRM} cutoffs (Figure 5C) indicates that the predicted CRMs are under similarly strong evolutionary constraints, and thus, the specificity of predictions might approach a saturated high level. Similar results are observed using the phyloP scores (supplementary Figure 3C).

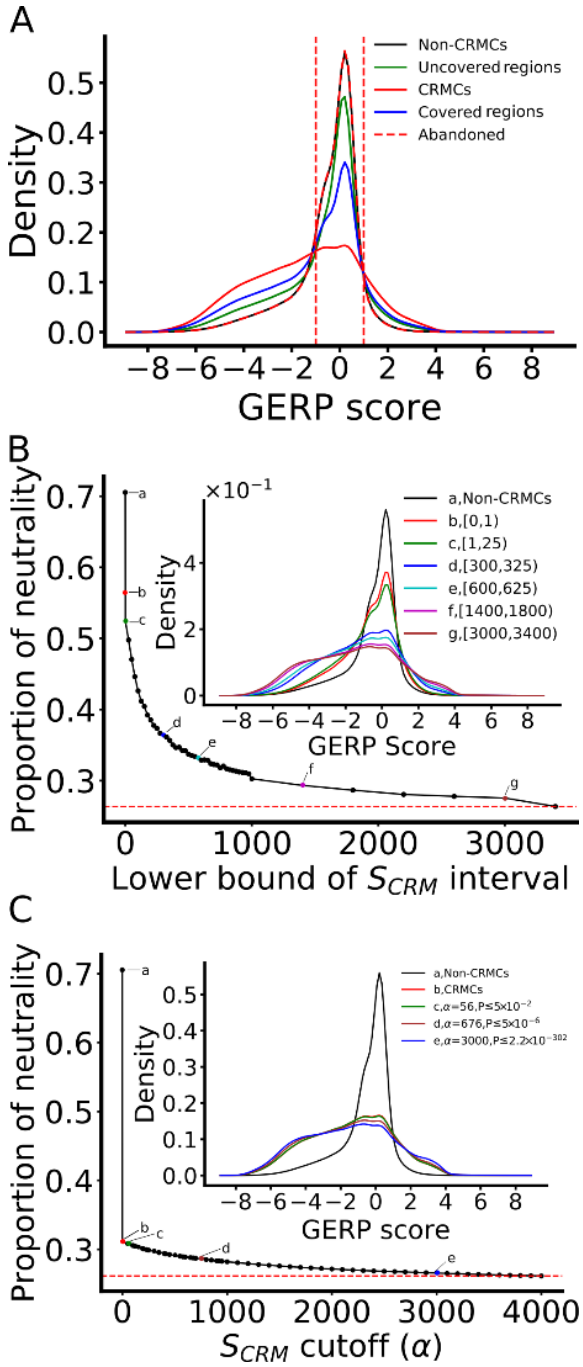

Figure 5. Different strengths of evolutionary constraints on the predicted CRMCs and non-CRMCs in NESs. A. Distributions of the GERP scores of nucleotides of the predicted CRMCs, non-CRMCs, abandoned genome regions covered by the original binding peaks, genome regions covered by the extended binding peaks and genome regions uncovered by the extended binding peaks. The area under the density curves in the score interval [-1, 1] is defined as the proportion of neutrality of the sequences. B. 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 the GERP scores of the non-CRMCs and CRMCs with S_{CRM} scores in the intervals indicted by color and letters. C. Proportion of neutrality of CRMs predicted using different S_{CRM} score cutoffs and associated p-values in comparison with those of the non-CRMCs (a) and CRMCs (b). The inset shows the distributions of the GERP scores of the non-CRMCs, CRMCs and the predicted CRMs using the S_{CRM} score cutoffs and p-values indicated by color and letters.

dePCRM2 achieves high sensitivity and specificity for recovering functionally validated CRMs and SNPs

To further evaluate the accuracy of the predicted CRMs, we calculated the sensitivity (true positive rate (TPR)) of CRMs predicted at different S_{CRM} cutoffs α and associated p-values for recovering experimentally determined 785 VISTA enhancers, 181,436 FANTOM5 promoters, 32,029 FANTOM5 enhancers, 402,730 of ClinVar SNPs, 2,734,505 DHSs (94.68%) and 82,378 (90.16%) of GWAS SNPs, located in the covered genome regions. Here, if a predicted CRM and a sequence element overlap each other by at least 50% of the length of the shorter one, we say that the CRM recovers the element, and we consider the p-value cutoff for a prediction to be the false positive rate (FPR), and (1 - p-value) to be the specificity. As shown in Figure 6, with the increase in the p-value cutoff (decrease in specificity), the sensitivity increases very rapidly and becomes saturated at p-value ≤ 0.05 (α ≥ 56, specificity=0.95) to 100.00%, 88.77%, 81.90%, 97.43%, 74.68%, 64.50% for recovering 785 VISTA enhancers, 161,065 FANTOM5 promoters, 26,233 FANTOM5 enhancers, 392,394 ClinVar SNPs, 2,042,355 DHSs and 53,130 GWAS SNPs, respectively. Supplementary Figures S4A~S4F show examples of the predicted CRMs that recover these elements. Interestingly, many recovered ClinVar SNPs (42.59%) and GWAS SNPs (38.18%) are located in critical positions in predicted binding sites of the UMs (e.g., Supplementary Figures S4D and S4F). In contrast, the matched randomly selected genome sequences could only recover about 40~50% of these elements as expected by chance (Figure 6). Notably, dePCRM2 achieves very high specificity (~95.00%) yet highly varying sensitivity (64.50~100.00%) for recovering these elements determined by various experimental methods, although results are significant in all the cases (Figure 6). Specifically, it achieves very high sensitivity for functionally validated VISTA enhancers (~100.00%) and causal ClinVar SNPs (~97.43%), high sensitivity for FANTOM5 promoters (~88.77%) and FANTOM5 enhancers (~81.90%), both of which were determined by reliable high throughput methods, and intermediate sensitivity for DHSs (~74.68%) and GWAS SNPs (~64.50%). The relatively lower sensitivity for recovering DHSs (~74.68%) might be due to the high noise in the data produced by less reliable

DNase-seq techniques (Dogan et al. 2015; Arbel et al. 2019), while that for GWAS SNPs (64.49%) might be due to the fact that a lead SNP associated with a trait may not necessarily be causal and located in a CRM; rather, some variants in a CRM, which are in LD with the lead SNP, are the culprits (Supplementary Figures S4G and S4H).

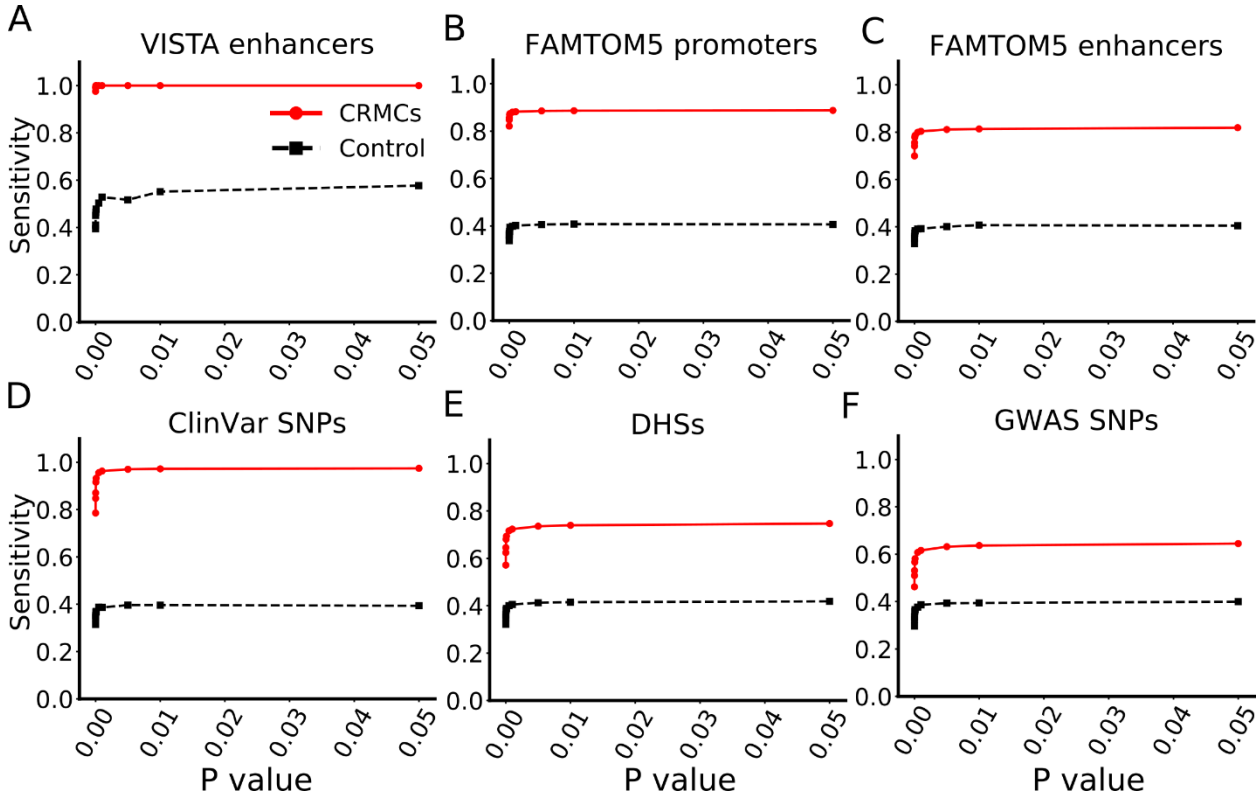

Figure 6. Validation of the predicted CRMs by multiple sets of the experimentally determined sequence elements. Sensitivity (TPR) of the predicted CRMs as a function of p-value cutoff (FPR) for recovering VISTA enhancers (A), FANTOM5 promoters (B), FANTOM5 enhancers (C), ClinVar SNPs (D), DHSs (E), and GWAS SNPs (F).

We also evaluated the accuracy of dePCRM2 for predicting the lengths of CRMs at p-value cutoff 0.05. To this end, we compared the distributions of the lengths of the 785 recovered VISTA enhancers and the 26,233 recovered FANTOM5 enhancers with those of the 836 and 22,235 recovering CRMs, respectively. As shown in Figure 7A, although the recovering CRMs have longer median length (3,799bp) than the recovered VISTA enhancers (1,613bp), by and large, they have similar length distributions. Moreover, even though a few (7.92%) VISTA enhancers are recovered by multiple short CRMCs,

suggesting that some of our predicted CRMCs are indeed only components of long CRMCs. Nonetheless, these results suggest that dePCRM2 is able to largely correctly predict the length of VISTA enhancers, albeit not perfect. By contrast, the recovered FANTOM5 enhancers are generally shorter (median length 293bp) than the recovering CRMs (median length 2,371bp) (Figure 7B), and a considerable proportion (26.69%) of the recovered FANTOM5 enhancers overlap the same CRMs (Figures 7D and 7E). To see whether the FANTOM5 enhancers also are generally shorter than their overlapping VITSTA enhancers, we compared the distribution of the lengths of 113 FANTOM5 enhancers with that of their 91 overlapping VISTA enhancers. Indeed, the 113 FANTOM5 enhancers (median length 319bp) also are much shorter than the 91 VISTA enhancers (median length 2,686bp), and in a few cases, multiple FANTOM5 enhancers overlap different parts of the same VISTA enhancers (Figures 7D and 7E). Thus, it appears that the eRNA-seq based method for determining FANTOM5 enhancers tends to identify short CRMs or CRM components of otherwise long CRMs, instead of full-length enhancers. In good agreement with these results, we noted that the decrease in the p-value cutoff (increase in the S_{CRM} cutoff α and specificity, thus filtering out shorter CRMCs) has little effects on the sensitivity for recovering VISTA enhancers that are generally long (median length 1,677 bp) (Figure 6A), but reduces the sensitivity to various levels for recovering other short elements, i.e., FANTOM5 promoters (median length of 15bp) (Figure 6B), FANTOM5 enhancers (median length of 288bp) (Figure 6C), ClinVar SNPs (Figures 6D), DHSs (median length of 150bp) (Figure 6E) and GWAS SNPs (6F), indicating that VISTA enhancers are largely recovered by long CRMs, and that shorter CRMCs that are filtered out with a lower p-value cutoff (a higher S_{CRM} cutoff α) could be short CRMs like promoters or CRM components like short FANTOM5 enhancers. For instance, the 428,628 putative full-length CRMs predicted at p-value ≤ $5x10^{-6}$ (α = 676, Figure 4E) have almost the same sensitivity as the 1,155,151 CRMs predicted at p-value ≤0.05 (α = 56) for recovering VISTA enhancers (99.10% vs 100.00%), but have lower sensitivity for recovering FANTOM5 promoters (84.74% vs 88.77%) and FANTOM5 enhancers (74.12% vs 81.90%), ClinVar SNPs

(84.75% vs 97.43%), DHSs (62.42% vs 74.68%), and GWAS SNPs (50.98% vs 64.49%). Taken together, these results demonstrate that dePCRM2 can accurately predict both the locations and lengths of CRMs at higher S_{CRM} cutoffs, in particular, VISTA-like long enhancers.

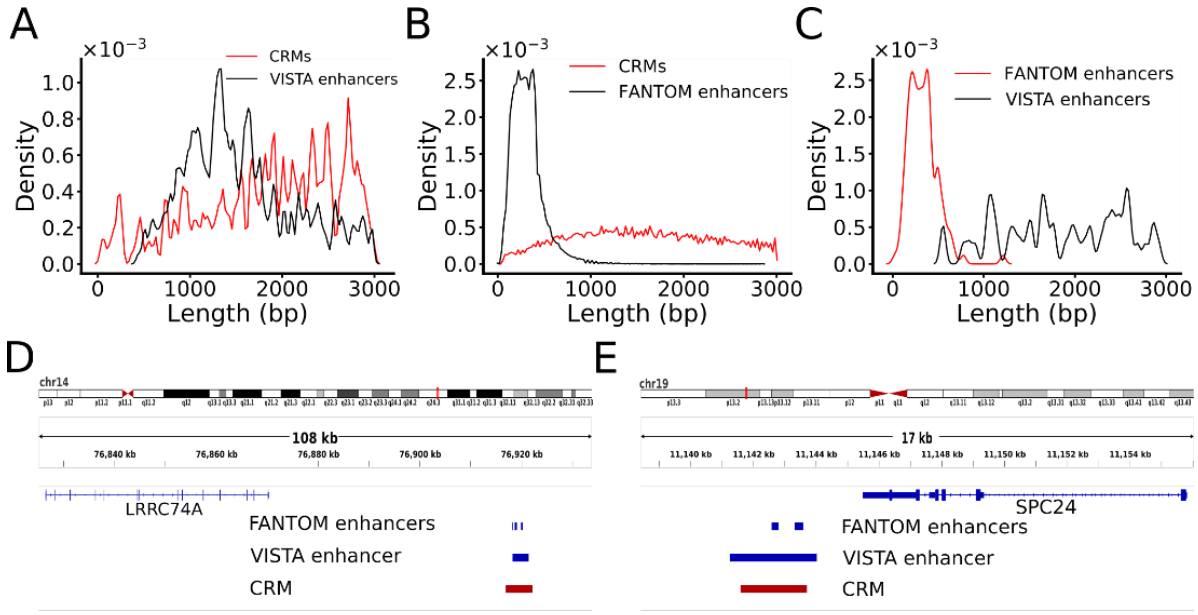

Figure 7. dePCRM2 can largely correctly predict the lengths of VISTA enhancers. A. Distributions of the lengths of the recovered VISTA enhancers and the recovering CRMs. B. Distributions of the lengths of the recovered FANTOM5 enhancers and the recovering CRMs. C. Distributions of the lengths of FANTOM5 enhancers and the overlapped VISTA enhancers. D. Three different FANTOM5 enhancers (chr14:76918231-76918434, chr14:76918637-76919122, chr14:76919959-76920278) overlap VISTA enhancer 1466 (chr14:76918298-76921354) and a predicted CRM (chr14:76916942-76922129) downstream of gene *LRRC74A.* E. Two FANTOM enhancers (chr19:11142586-11142806, chr19:11143315-11143595) overlap VISTA enhancer 1754 (chr19:11141255-11144027) and a predicted CRM (chr19:11141600-11143708) downstream of gene *SPC24.*

dePCRM2 outperforms state-of-the-art algorithms

We compared our predicted CRMs at p-value ≤ 0.05 ($S_{CRM} \geq 56$) with two most comprehensive sets of predicted enhancers, i.e., GeneHancer 4.14 (Fishilevich et al. 2017) and EnhancerAtals2.0 (Gao and Qian 2020). GeneHancer 4.14 is the most updated version containing 394,086 non-overlapping enhancers covering 18.99% (586,582,674bp) of the genome (Figure 8A). These enhancers were predicted by integrating multiple sources of both predicted and experimentally determined enhancers, including ENCODE Enhancer-like regions (Bernstein et al. 2012), ENSEMBL regulatory build (Zerbino et al. 2015),

dbSUPER (Khan and Zhang 2015), EPDnew promoters (Dreos et al. 2013), UCNEbase (Dimitrieva and Bucher 2013), CraniofacialAtlas (Wilderman et al. 2018), VISTA enhancers (Visel et al. 2007), and FANTOM5 active enhancers (Andersson et al. 2014). Enhancers from ENCODE and ESEMBL were predicted based on multiple tracks of epigenetic marks using the well-regarded tools ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman et al. 2012). Of the 394,086 GeneHancer enhancers, 388,407 (98.56%) have a at least one nucleotide located in the covered 77.47% genome regions, covering 18.89% of the genome (Figure 8A). EnhancerAtlas 2.0 contains 7,433,367 overlapping cell/tissue-specific enhancers in 277 cell/tissue types, covering 58.99% (1,821,795,020bp) of the genome (Figure 8A). EnhancerAtlas enhancers were predicted by integrating 4,159 TF ChIP-seq, 1,580 epigenetic, 1,113 DHS-seq, and 1153 other enhancer function related datasets, such as FANTOM5 enhancers (Gao et al. 2016; Gao and Qian 2020). After removing redundancy (identical enhancers in difference cell/tissues), we ended up with 3,452,739 EnhancerAtlas enhancers that may have overlaps. Of the 3,452,739 enhancers, 3,417,629 (98.98%) have at least one nucleotide located in the covered 77.47% genome regions, covering 58.78% of the genome (Figure 8A). Thus, both the number (1,155,151) and percentage of genome coverage (43.47%) of our predicted CRMs are larger than those of GeneHancer enhancers (388,407 and 18.89%) that at least partially overlap the covered regions, but smaller than those of EnhancerAtlas enhancers (3,417,629 and 58.78%) that at least partially overlap the covered regions.

To make the comparison relatively fair, we first computed recall rates of these enhancers that at least partially overlap the covered 77.47% genome regions, for recovering VISTA enhancers, ClinVar SNPs and GWAS SNPs. We omitted FANTOM5 enhancers and DHSs for the valuation as they are already parts of both GeneHancer 4.14 and EnhancerAtlas 2.0. We included VISTA enhancers for the evaluation as they were not included in EnhancerAtlas 2.0, although they were parts of GeneHancer 4.14. Remarkably, our predicted CRMs outperform EnhancerAtlas enhancers for recovering VISTA enhancers

(100.00% vs 94.01%) (Figure 8B) and ClinVar SNPs (97.43% vs 7.03%) (Figure 8C), even though our CRMs cover a smaller percentage of the genome (43.47% vs 58.78%) (Figure 8A), indicating that dePCRM2 has both higher sensitivity and specificity than the method behind EnhancerAtlas 2.0. Our predicted CRMs underperform EnhancerAtlas enhancers for recovering GWAS SNPS (64.50% vs 69.36%) (Figure 8D). However, as we argued earlier, the lower sensitivity of dePCRM2 for recalling GWAS SNPs might be due to the fact that an associated SNP may not necessarily be causal (Figure 6F and Supplementary Figure S4). The higher recall rate of EnhancerAtlas enhancers for GWAS SNPs might be simply thanks to their higher coverage of the genome (58.78%) than our predicted CRMs' (43.47%) (Figure 8A). Our predicted CRMs also outperform GeneHancer enhancers for recovering both ClinVar SNPs (97.43% vs 33.16%) (Figure 8C) and GWAS SNPs (64.50% vs 34.11%) (Figure 8D). However, no conclusion can be drawn because our predicted CRMs cover a higher proportion of the genome (43.47%) than GeneHancer enhancers (18.89%) (Figure 8A). Interestingly, the GeneHancer enhancers outperform EnhancerAtlas enhancers for recovering ClinVar SNPs (33.16% vs 7.03%) (Figure 8C), even though the former latter have less than a quarter genome coverage than the latter (18.89% vs 58.78%) (Figure 8A).

GeneHancer enhancers and EnhancerAtlas enhancers share 414,806,711bp (70.72%) and 926,396,395bp (50.85%) of their nucleotide positions with our predicted CRMs, respectively, corresponding to 30.90% and 69.01% of nucleotide positions of our CRMs (Figure 8E). Therefore, we next compared evolutionary behaviors of the nucleotide positions that GeneHancer enhancers and EnhancerAtlas enhancers shared with our CRMs, as well as of their respective unshared 29.28% and 49.14% of nucleotide positions. As expected, as two subsets of nucleotides of our predicted CRMs, the shared nucleotides of both GeneHancer enhancers and EnhancerAtlas enhancers evolve similarly to the entire set of our predicted CRMs (at <0.05), particularly, those of EnhancerAtlas enhancers (Figure 8F). Notably, the shared nucleotides of GeneHancer enhancers are under slightly higher evolutionary

constraints than the entire CRM set (Figure 8F). This subset evolves similarly to the CRMs predicted at a more stringent p-value cutoff $5x10^{-6}$ (Figures 5C and 9F). By stark contrast, the remaining 29.28% and 49.14% unshared nucleotides of GeneHancer enhancers and EnhancerAtlas enhancers, respectively, are under much less evolutionary constraints, with a GERP score distribution similar to that of our predicted non-CRMCs (Figure 8F). These results strongly suggest that the predicted enhancers consisting of the shared nucleotides in GeneHancer and EnhancerAtlas are likely to be authentic, while most of those made up of unshared nucleotides might be false positives. Compared with our predicted CRMs, it appears that GeneHancer 4.14 largely under-predicted enhancers even with a near 29.28% false positive rate (Figure 8F), while EnhancerAtlas 2.0 largely over-predicted enhancers with a near 49.14% false positive rate (Figure 8F). Taken together, these results unequivocally indicate that our predicted CRMs are much more accurate than both GeneHancer 4.14 enhancers and EnhancerAtlas 2.0 enhancers. Moreover, GeneHancer enhancers share more nucleotide positions (70.72%) with our predicted CRMs than EnhancerAtlas enhancers (50.85%), thus might have higher specificity than the latter, in addition to having higher sensitivity for recovering ClinVar SNPs (33.16% vs 7.03%) (Figure 8C). Therefore, GeneHancer enhancers might be more accurate than EnhancerAtlas enhancers.

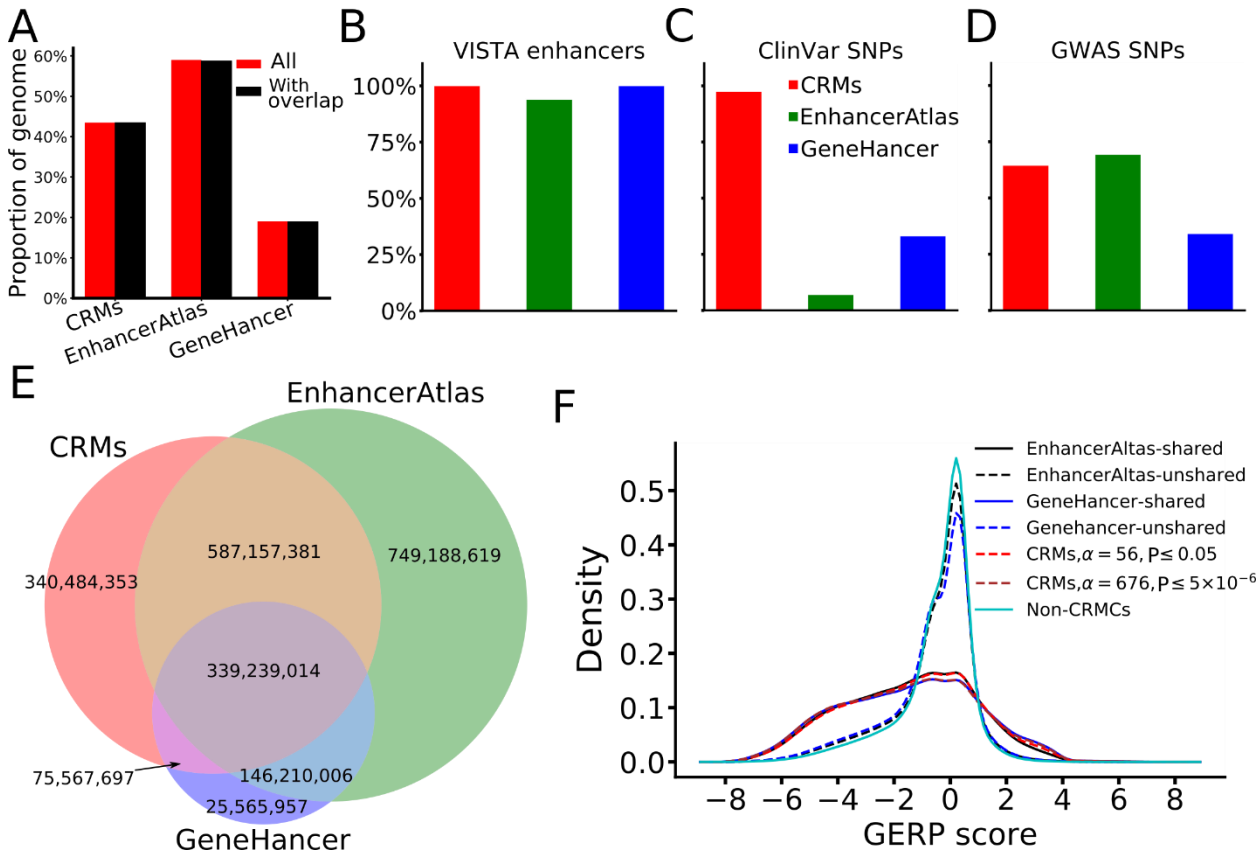

Figure 8. Comparison of the performance of dePCRM2 and two state-of-the-art methods. A. Percentage of genome regions covered by all CRMs/enhancers predicted by different methods, and percentage of genome regions covered by predicted CRMs/enhancers that at least partially overlap the covered genome regions. B, C, and D. Recall rates for recovering VISTA enhancers, ClinVar SNPs and GWAS SNPs, by the predicted CRMs/enhancers that at least partially overlap the covered genome regions. E. Ven diagram showing numbers of nucleotide positions shared among the predicted CRMs, GeneHancer enhancers and EnhancerAtlas enhancers. F. Distributions of GERP scores of nucleotide positions of CRMs predicted at p-value ≤ 0.05 and p-value ≤ $5X10^{-6}$, and the non-CRMCs, as well as of nucleotide positions that GeneHancer enhancers and EnhancerAtlas enhancers share or do not share with the predicted CRMs at p-value ≤ 0.05.

At least half of the human genome might code for CRMs

What is the proportion of the human genome coding for CRMs? The high accuracy of our results might allow us to more accurately address this interesting, yet unanswered question (Pennacchio et al. 2013; Kellis et al. 2014). To this end, we calculated the expected number of true positive CRMCs and false positive 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 9A), yielding 1,383,152 (98.45%) expected true positives and 21,821 (1.55%) expected false positives in the CRMCs (Figure 9B). The vast

majority of the 21,821 expected false positive CRMCs have low a S_{CRM} score < 4 (inset in Figure 9A) with an average length of 28 pb (Figure 4D), making up only 0.02% (21,821x28/3,088,269,832) and 0.045% of the genome and the CRMCs in length, respectively (Figure 9C). On the other hand, as the CRMCs covering 44.03% of the genome recover 100.00% of VISTA enhancers and 97.51% of ClinVar SNPs in the covered genome regions, the average sensitivity of the CRMCs for recovering these functional elements would be 98.76%. Thus, we estimate the false negatives of the CRMCs in the covered regions to be 0.55% (0.4403/0.9876-0.4403) and 1.27% of the genome and the non-CMCs in length, respectively (Figure 9C). Hence, the CRMCs in the covered regions make up 44.56% (44.03%-0.02%+0.55%) of the genome (Figure 9C). If all short authentic CRM components could be expanded and connected to form full-length CRMs when more data are available, CRMs in the covered regions would be 44.58 (44.56% +0.02%) of the genome. In addition, as we argued earlier, the uncovered 22.53% genome regions have a 79.4% CRMC density as in the covered regions, then, CRMCs in the uncovered regions would be about 10.29% (22.53% x 44.58%x79.40%/77.47%) of the genome (Figure 9C). Taken together, we estimate that about 54.87% (44.58%+10.29%) of the genome might code for CRMs, for which we have predicted 80.21% [(44.03-0.02)/54.87]. Moreover, as we predict that about 40% of CRCs are made up of TFBSs (Figure 4C), we estimate that about 21.95% of the genome might encode TFBSs. Furthermore, assuming the mean length of CRMs is 2,049bp as VISTA enhancers, we estimate the human genome encodes about 827,005 CRMs (3,088,269,832x0.5487/2049).

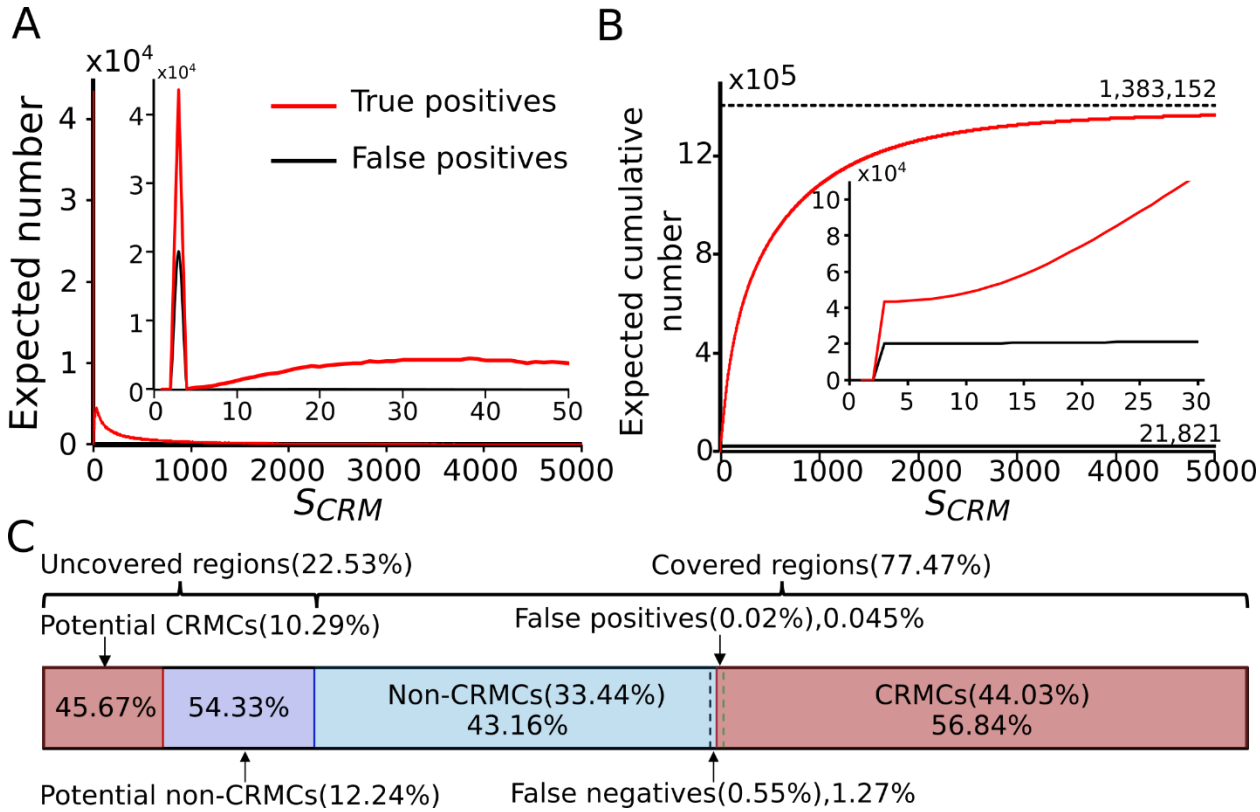

Figure 9. Estimation of the portion of the human genome encoding CRMs. A. Expected number of true positive and false positive CRMCs in the predicted CRMCs in each one-unit interval of the S_{CRM} score. The inset is a blow-up view of the axes defined region. B. Expected cumulative number of true positives and false positives with the increase in S_{CRM} score cutoffs for predicting CRMs. The inset is a blow-up view of the axes defined region. C. Proportions of the genome that are covered and uncovered by the extended binding peaks and estimated proportions of CRMCs in the regions. Numbers in the braces are the estimated proportions of the genome being the indicated sequence types, and numbers in the boxes are proportions of the indicated sequence types in the covered regions or the uncovered regions.

DISCUSSION

Identification of all functional elements, in particular, CRMs in genomes has been the central task in the postgenomic era (Bernstein et al. 2012; Gasperini et al. 2020). Although great progresses have been made to predict cell/tissue specific CRMs using multiple tracks of epigenetic marks of CRMs, in particular, enhancers (Zerbino et al. 2015; Gao et al. 2016; Fishilevich et al. 2017; Wang et al. 2018b; Gao and Qian 2020), these methods are limited by low resolution of predicted CRM boundaries, lack of constituent TFBS information, particularly, the novel ones, and few cell/tissue types for which the

required epigenetic data are available. To circumvent these limitations, we proposed a different strategy to first predict a static map of CRMs and constituent TFBBs in the genome (Niu et al. 2014; Niu et al. 2018), just as has been done for finding all genes encoded in the genome before knowing their functions specific cell/tissue types (Allen et al. 2004). Since it is mainly TFBSs in a CRM that define its structure and function, and since TF ChIP-seq data contain information of TF-binding locations, we proposed to integrate all available TF ChIP-seq data for different TFs in various cell/tissue types to achieve the goal. One advantage of this approach is that because CRMs are often repeatedly used in multiple tissues, developmental stages and physiological condictiones, 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 TFBBSs in the genome. Instead, we might only need a large but limited number of datasets that are less-biased and cover the entire functional genome to achieve the goal. In particular, by appropriately extending the called binding peaks in each dataset, we can largely increase the coverage of the genomes, and reduce the number of datasets needed. Our earlier application of the approach resulted in very promising results in the fly and human genomes even using a relatively small number of strongly biased datasets available then (Niu et al. 2014; Niu et al. 2018). However, we were limited by the drawbacks of the earlier algorithm, as well as the lack of a sufficiently large number of ChIP-seq datasets for highly diverse TFs in various cell/tissue types (Niu et al. 2014; Niu et al. 2018). In this study, we developed dePCRM2, which largely circumvents the shortcomings of the earlier version, based on an ultrafast, accurate motif finder ProSampler (Li et al. 2019) and a novel efficient method for identifying over-represented combinatory patterns of motifs found in the extended binding peaks of numerous TF ChIP-seq datasets. Although the available 6,092 ChIP-seq datasets for 779 TFs in 2,631 cell/tissue/organ types are enormous, they are still strongly biased to a few TF and cell types (Figures 2A and 2B), and the originally called binding peaks cover only 40.96% of the genome. However, after extending each short binding peak to 1,000bp, the extended binding peaks cover 77.47% of the genome, an 89.14 % increase. Using

these extended binding peaks, dePCRM2 predicts a total of 201 UM families and partitions the covered genome regions into the non-CRMCs set (43.16%) and the CRMCs set (56.84%) containing 1,404,973 CRMCs covering 44.03% of the genome (Figure 9C). In predicting the CRMCs, dePCRM2 abandoned 38.60 % of genome regions covered by the original binding peaks, while extending them by 44.68%. The extended parts of the original binding peaks contribute 42.12% nucleotide positions of the CRMCs. These results strongly suggest that our approach can largely increase the power of the existing datasets, and reduce the amount of data needed to predict most, if not all, of the CRMs and constituent TFBSs encoded in the genome.

Remarkably, the predicted CRMCs are under strong evolutionary constraints while the non-CRMCs are not (Figure 5A), strongly suggesting that at least most of the CRMCs might be functional, while at least most non-CRMCs are not. This conclusion is further supported by the estimated low false positive positions (0.045%) in the CRMCs and low false negative position (1.27%) in the non-CRMCs (Figure 9C). Thus, dePCRM2 is able to effectively divide the covered genome regions into the largely functional CRMCs set and the largely non-functional non-CRMCs set. Intriguingly, dePCRM2 is able to accurately distinguish CRMCs and non-CRMCs in both the original binding peaks and their extended parts, as evidenced by the aforementioned results and the finding that the abandoned genome regions covered by the original binding peaks evolve almost identically to those of non-CRMCs. However, some CRMCs might not be predicted as full-length CRMs, as 621,842 (44.26%) of the 1,404,973 predicted CRMCs are shorter than the shortest VISTA enhancer (428bp). Nevertheless, these CRMCs shorter than 428bp consist of only 7.42% of the total length of the CRMCs set, whereas the remaining 733,132 (55.74%) CRMCs (≥428bp) comprise 92.58% of the total length of the CRMCs. Therefore, the CRMCs are dominated by long and likely full-length CRMs, particularly, in length. Possible false positive CRMCs and short CRM components can be effectively filtered out by an appropriate S_{CRM} score or the associated p-

value cutoff, because the S_{CRM} score computed for each CRMC captures the essential features of full-length enhancers. More specifically, the higher the S_{CRM} score of a CRMC, the more likely it is a full-length enhancer (Figure 4E) and the stronger evolutionary constraint it receives (Figure 5B). For instance, at $S_{CRM} \geq 676$ or p-value ≤ 5×10^{-6}, the distribution of the lengths of the predicted 428,628 CRMCs fits nicely to that of VISTA enhancers (Figure 4E), suggesting that these CRMCs might be full-length VISTA-like enhances. In agreement with this, these CRMCs recover vast majority (99.10%) of VISTA enhancers in the covered regions (Figure 6A). The failure to predict some CRMCs in full-length CRMs might be due to the insufficient coverage of the genome by the current strongly biased datasets albeit in a large quantity (Figures 2A and 2B). Thus, efforts should be made in the future to increase the genome coverage and reduce data biases by including more untested TFs and untested cell types in the ChIP-seq data generation. Once the entire functional genome is well covered by extended binding peaks, additional data might be redundant for the purpose.

It has been estimated that the human genome encodes from 2,000 to 3,000 TFs belonging to at least 100 protein families (Fulton et al. 2009; Lambert et al. 2018). However, the exact number of TFs and TF families encoded in the genome remains unknown (Vaquerizas et al. 2009; Lambert et al. 2018). Our prediction of the 201 UM families in the covered genome 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 almost indistinguishable similar motifs (Jolma et al. 2013; Ambrosini et al. 2020), 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 remaining 84 (41.79%) UMs might be motifs of novel TF families that remain to be elucidated. On the other hand, as most (75.2%) of the known motif families match one of our UMs, we have recovered most known motif families. Therefore, 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 77.47% regions of the genome.

Importantly, our results indicate that dePCRM2 achieves very high specificity and sensitivity in predicting CRMs and constituent TFBSs. First, even the CRMCs with the lowest S_{CRM} scores ((0,1]) are under stronger evolutionary constraints than non-CRMCs (Figure 5B), suggesting that even these low-scoring CRMCs are still likely to be functional, not to mention CRMCs with higher S_{CRM} scores. Second, with the increase in the S_{CRM} cutoff (e.g., from 0 to 56), the associated p-value decreases rapidly (from 0.694 to 0.05) (Figure 4B), and prediction specificity (i.e., 1 - p-value) increases rapidly (from 0.306 to 0.95); meanwhile, the number of the predicted CRMs (from 1,4327,396 to 1,155,151) and their total length (from 44.03% to 43.47% of the genome) only decreases slowly (Figure 4B). Thus, high specificity can be achieved by filtering out only a small portion of low-scoring short CRMCs. Third, with the increase in the S_{CRM} cutoff (decrease in p-value), strength of evolutionary constraints on the predicted CRMs saturates quickly followed by small increments (Figures 5C and S3C), indicating that the prediction specificity is saturated to a high level at even less stringent S_{CRM} cutoffs. Finally, with the increase in the S_{CRM} cutoff (decrease in p-value and increase in specificity), the sensitivity of the predicted CRMs for recalling the experimentally determined elements increases rapidly and becomes saturated to varying levels at low p-values (≤0.05) (specificity≥0.95) (Figure 6). Remarkably, at p-value ≤ 0.05 and specificity=0.95, dePCRM2 achieves perfect (100.00%) and very high sensitivity (97.43%) for recovering functionally validated VISTA enhancers (Visel et al. 2007) and carefully compiled causal ClinVAR SNPs (Landrum and Kattman 2018; Landrum et al. 2018); high sensitivity for recovering FANTOM5 promoters (88.77%) and FANTOM5 enhancers (81.90%), determined by relatively reliable high throughput CAGE (CAP analysis of gene expression) techniques (Kanamori-Katayama et al. 2011; Andersson et al. 2014; Consortium et al. 2014); and intermediate sensitivity for recovering DHSs (74.68%) and GWAS SNPs

(64.49%). The relatively low sensitivity for recovering DHSs might be due to the less reliability of the DNase-seq techniques used for their determination (Dogan et al. 2015; Arbel et al. 2019), while that for GWAS SNPs might be due to the fact that a lead SNP may not necessarily be causal (MacArthur et al. 2017; Buniello et al. 2019).

dePCRM2 also largely outperforms the state-of-the-art method behind EnhancerAtlas 2.0 enhancers (Gao et al. 2016; Gao and Qian 2020) for recovering functionally validated ClinVar SNPs and VISTA enhancers, even though our CRMs cover a smaller proportion (43.47%) of the genome than EnhancerAtlas enhancers (58.78%). To our surprise, we found that nucleotide positions that are not shared with our predicted CRMs (at p-value ≤ 0.05), in enhancers predicted in both EnhancerAtlas 2.0 and GeneHancer 4.14 (Fishilevich et al. 2017), evolve largely like our predicted non-CRMCs (Figure 8F), indicating that they might have near 29.28% and 49.15% false positive rates, respectively. Thus, our predicted CRMs are much more accurate than those in either EnhancerAtlas 2.0 or GeneHancer 4.14. This result is remarkable, because our CRMs were predicted using only 5,578 TF ChIP-seq datasets (512 of the 6,090 datasets were filtered out due to their low quality or lack of overlaps with other datasets). In contrast, EnhancerAtlas enhancers were predicted using 8,005 datasets, including 4,159 TF ChIP-seq and 1,580 epigenetic datasets, as well as experimentally determined potential enhancers (Gao and Qian 2020). And GeneHancer 4.14 is a meta-prediction based on the results of multiple algorithms including the well-regarded tools ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman et al. 2012). Although dePCRM2 can predict the functional states of CRMs in a cell/tissue type that have available original binding peaks overlapping the CRMs, it cannot predict the functional states of CRMs in the extended parts of the original binding peaks in a cell/tissue if the CRMs do not overlap any binding peaks of all TFs tested in the cell/tissue type. However, once a map of CRMs in the gnome as we predicted in this study is available, the functional state of each CRM in the map in any cell/tissue type could be assigned using a single dataset collected from the very cell/tissue type using nucleosome-free region profiling or

H3K27Ac histone mark profiling techniques, which reflects the functional states of CRMs. Thus, our 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.

The proportion of the human genome that is functional is a topic under hot debate (Bernstein et al. 2012; Stamatoyannopoulos 2012) and a wide range from 5% to 80% of the genome has been suggested to be functional based on difference sources of evidence (King et al. 2005; Elizabeth 2012; Stamatoyannopoulos 2012; Thurman et al. 2012; Pennacchio et al. 2013; Gasperini et al. 2020). The major disagreement is for the proportion of functional NCSs in the genome, mainly CRMs, which has been coarsely estimated to be from 8% to 40% of the genome (Bernstein et al. 2012; Stamatoyannopoulos 2012). Moreover, a wide range of CRM numbers from 400,000 (Bernstein et al. 2012) to more than a few million (Gao and Qian 2020; Gasperini et al. 2020) encoded in the human genome have been suggested. However, to our knowledge, no estimate has been made rigorously. Here, we estimate that about 54.87% of the genome might code for CRMs, for which we have predicted 80.21% in length, covering 44.01 of the genome, based on the expected false positives and false negatives for predicting the 1,404,973 CRMCs in the covered 77.47% genome regions as well as the estimate of the density of CRMs in the uncovered 22.53% regions relative to the covered regions (Figure 9C). The proportion of the genome covered by our predicted CRMCs (44.03%) is higher than those covered by CRMs predicted in GeneHancer v.4.14 (18.99%) (Fishilevich et al. 2017), REMAP (13.0%) (Cheneby et al. 2018) and DENdb (43.0%) (Ashoor et al. 2015; Fishilevich et al. 2017), but it is lower than those covered by EnhancerAtlas 2.0 enhancers (58.99%) (Gao and Qian 2020). The much higher accuracy of our predicted CRMs suggests that most of these methods might make underpredictions, whereas EnhancerAtlas 2.0 might make overpredictions. Assuming the mean length of CRMs is about the same as VISTA enhancers (2,049bp), we estimate that the human genome encodes about 827,005 CRMs, which is more than twice an earlier estimate of 400,000 (Bernstein et al. 2012). Of the 1,404,973 predicted

CRMCs, 733,132 (55.74%) comprising 92.58% of the total length of the CRMCs are longer than the shortest VISTA enhancer (428bp), thus at least most of them are likely to be full-length CRMs. In this sense, we might have predicted 88.65% (733,132/827,005) of the CRM numbers encoded in the genome, which is consistent with the estimated 80.21% (44.03/54.87) in length. As about 40% of nucleotide positions of CMCs are made up of TFBSs (Figures 4C and 11A), about 21.95% of the genome might encode TFBSs.

In conclusion, we have developed a new highly accurate and scalable algorithm dePCRM2 for predicting CRMs and constituent TFBSs in a large genome by integrating a large number of TF ChIP-seq datasets for various TFs in a variety of cell/tissue types of the organism. Applying dePCRM2 to all available more than 6,000 TF ChIP-seq datasets, we predicted a comprehensive map of CRMs and constituent TFBSs in 77.47% of the human genome, which are covered by the extended binding peaks of the datasets. Evolutionary and experimental data suggest that dePCRM2 achieves very high prediction sensitivity and specificity. With more diverse and balanced data covering the whole genomes becoming available in the future, it is possible to predict more complete maps of CRMs and constituent TFBSs in the human and other important genomes.

MATERIALS AND METHODS

Datasets

We downloaded 6,092 TF ChIP-seq datasets from the Cistrome database (Mei et al. 2017). The binding peaks in each dataset were called using a pipeline for uniform processing (Mei et al. 2017). We filtered out binding peaks with a read depth score less than 20. For each binding peak in each dataset, we extracted a 1,000 bp genome sequence centering on the middle of the summit of the binding peak. We downloaded 976 experimentally verified enhancers from the VISTA Enhancer database (Visel et al. 2007), and 32,689 enhancers and 184,424 promoters from the FANTOM5 project website (Andersson et

al. 2014; Consortium et al. 2014). We downloaded a total of 424,622 ClinVar SNPs from the ClinVar database (Landrum et al. 2016), and 2,888,031 of DHS from the Regulatory Elements Database (Sheffield et al. 2013) and 91,369 GWAS SNPs from GWAS Catalog (MacArthur et al. 2017; Buniello et al. 2019).

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,

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

The dePCRM2 algorithm

Step 1: Find motifs in each dataset using ProSampler (Li et al. 2019)(Figures.1A, 1B).

Step 2. Compute pairwise motif co-occurring scores and find co-occurring motif pairs: As True motifs are more likely to co-occur in the same sequence than spurious ones, to filter out false positive motifs, we find overrepresented co-occurring motif pairs (CPs) in each dataset (Figs.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,

$$S_c\left(M_i(i), M_j(j)\right) = \frac{o\left(M_d(i), M_d(j)\right)}{max\{|M_d(i)|, |M_d(i)|\}}, \quad (2)$$

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 $o\left(M_d(i), M_d(j)\right)$ the number of binding peaks containing TFBSs of both the motifs in d. We identify CPs with an $S_c \geq 0.7$ (by default) (Figs.1C).

Step 3. Construct a motif similarity graph and find unique motifs: We combine highly similar motifs in the CPs from different datasets to form a unique motif (UM) presumably recognized by a TF or highly similar TFs of the same family/superfamily (Weirauch et al. 2014). Specifically, for each pair of motifs $M_d(i)$ and

$M_b(i)$ from different datasets *a* and *b,* respectively, we compute their similarity score S_s using our SPIC (Zhang et al. 2009) metric. We then build a motif similarity graph using motifs in the CPs as nodes and connecting two motifs with their S_s being the weight on the edge, if and only if (iff) $S_s > \beta$ (by default, β =0.8, Fig.1D). We apply the Markov cluster (MCL) algorithm (van Dongen and Abreu-Goodger 2012) to the graph to identify dense subgraphs as clusters. For each cluster, we merge overlapping sequences, extend each sequence to a length of 30bp 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 (Fig.1D).

Step 4. Construct the integration networks of UMs: TFs tend to repetitively cooperate with each other to regulate genes in different contexts by binding to their cognate TFBSs in CRMs. The relative distances between TFBSs in a CRM often do not matter (billboard model) (Arnosti and Kulkarni 2005; Yanez-Cuna et al. 2013; Vockley et al. 2017) but sometimes they are constrained by the interactions between cognate TFs. To model both scenarios, we compute an interaction score between each pair of UMs,

$$S_{\mathrm{INTER}}(U_i, U_j) = \frac{1}{|D(U_i,U_j|}\sum_{d\in D(U_i,U_j)}\left(\frac{1}{|U_i(d)|}\sum_{s\in S(U_i(d),U_j(d))}\frac{150}{r(s)} + \frac{1}{|U_j(d)|}\sum_{s\in S(U_i(d),U_j(d))}\frac{150}{r(s)}\right), \quad (3)$$

where $D(U_i, U_j)$ is the datasets in which TFBSs of motifs U_i and U_j are found, $U_k(d)$ the subset of dataset d, containing at least one TFBS of U_k, $\mathrm{S}(U_i(d), U_j(d))$ the subset of d containing TFBSs of both U_i and U_j, and $r(s)$ the shortest distance between a TFBS of U_i and a TFBS of U_j in a sequence s. We construct UM/TF interaction networks using the UMs as nodes and connecting two nodes with their S_{INTER} being the weight on the edge (Fig.1E). Therefore, S_{INTER} allows flexible adjacency and orientation of TFBSs in a CRM and at the same time, it rewards motifs with binding sites co-occurring frequently in a shorter distance in a CRM (Arnosti and Kulkarni 2005; Yanez-Cuna et al. 2013; Vockley et al. 2017).

Step 5. Evaluate CRM candidates: We project TFBSs of each UM back to the genome, and link two adjacent TFBSs if their distance $d \leq 300$bp (roughly the length of DNA in two nucleosome). The resulting linked DNA segments are CRM candidates (CRMCs) (Fig.1F). We evaluate each CRMC containing n TFBSs ($b_1, b_2, \ldots, b_n$) by computing a CRM score, defined as,

$$S_{CRM}(b_1, b_2 \cdots, b_n) = \frac{2}{n-1} \times \sum_{i=1}^{n} \sum_{j>i} W\left[U(b_i), U(b_j)\right] \times \left[S(b_i) + S(b_j)\right], \qquad (4)$$

where $U(b_k)$ is the UM of TFBS b_k, $W\left[U(b_i),\ U\left(b_j\right)\right]$ the weight on the edge between the motifs of $U(b_i)$ and $U\left(b_j\right)$, 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.

Step 6. Predict CRMs: We create the Null interaction networks by randomly reconnecting the nodes with the edges in 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 (Li et al. 2019). 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 a p-value for each CRMC, and predict those with a p-value smaller than a preset cutoff as CRMs in the genome (Fig.1G).

Step 7. Prediction of the functional states of CRMs in a given cell type: For each predicted CRM, we predict it to be active in a cell/tissue type, if its constituent binding sites of the UMs whose cognate TFs were tested in the cell/tissue type overlap original binding peaks of the TFs; otherwise, we predict the CRM to be inactive in the cell/tissue type. If the CRM does not overlap any binding peaks of the TFs tested in the cell/tissue type, we assign its functional state in the cell/tissue type “to be determined” (TBD).

ACKNOWNLEDGES

The work was partially supported by US National Science Foundation (DBI- 1661332).

Conflict of interest statement.

The authors declare no competing financial interests.

REFERENCES

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197-212.

Allen JE, Pertea M, Salzberg SL. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res* **14**: 142-148.

Ambrosini G, Vorontsov I, Penzar D, Groux R, Fornes O, Nikolaeva DD, 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.

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.

Arbel H, Basu S, Fisher WW, Hammonds AS, Wan KH, Park S, Weiszmann R, Booth BW, Keranen SV, 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.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890-898.

Ashoor H, Kleftogiannis D, Radovanovic A, Bajic VB. 2015. DENdb: database of integrated human enhancers. *Database : the journal of biological databases and curation* **2015**.

Attanasio C, Nord AS, Zhu Y, Blow MJ, Li Z, Liberton DK, Morrison H, Plajzer-Frick I, Holt A, Hosseini R et al. 2013. Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**: 1241006.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**: 1653-1659.

Bailey TL, 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.

Bailey TL, Machanick P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**: e128.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045-1048.

Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, Delattre O, Barillot E. 2010. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res* **38**: e126.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, 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.
Burga A, Lehner B. 2012. Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *The FEBS journal* **279**: 3765-3775.
Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32**: 202-223.
Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, Xing J, Wang H, Zhang Y. 2020. SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. *Nucleic Acids Res* **48**: D198-d203.
Chen G, Zhou Q. 2011. Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells. *BMC Genomics* **12**: 515.
Chen X, Guo L, Fan Z, Jiang T. 2008. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics* **24**: 1121-1128.
Cheneby J, Gheorghe M, Artufel M, Mathelier A, 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.
Collins F. 2010. Has the revolution arrived? *Nature* **464**: 674-675.
Colombo N, Vlassis N. 2015. FastMotif: spectral sequence motif discovery. *Bioinformatics* **31**: 2623-2631.
Consortium EP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**: e1001046.
Consortium F the RP Clst Forrest AR Kawaji H Rehli M Baillie JK de Hoon MJ Haberle V Lassmann T et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462-470.
Consortium G. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**: 648-660.
Consortium GT. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580-585.
Consortium TEP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636-640.
Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* **7**: 250-251. doi: 210.1038/nmeth0410-1250.
Dimitrieva S, Bucher P. 2013. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res* **41**: D101-109.
Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, Keller CA, 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. *Epigenetics & chromatin* **8**: 16.
Dreos R, Ambrosini G, Cavin Perier R, Bucher P. 2013. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res* **41**: D157-164.
Elizabeth P. 2012. ENCODE Project Writes Eulogy For Junk DNA. *Science* **337**: 1159-1160.
Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.
Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215-216.
Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43-49.
Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J. 2007. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods* **4**: 563-565.

Fauteux F, Blanchette M, Stromvik MV. 2008. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics* **24**: 2303-2307.
Firpi HA, Ucar D, Tan K. 2010. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* **26**: 1579-1586.
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 : the journal of biological databases and curation* **2017**.
Fu W, O'Connor TD, Akey JM. 2013. Genetic architecture of quantitative traits and complex diseases. *Curr Opin Genet Dev* **23**: 678-683.
Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R. 2009. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* **10**: R29.
Gao T, He B, Liu S, Zhu H, Tan K, Qian J. 2016. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**: 3543-3551.
Gao T, 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.
Gasperini M, Tome JM, Shendure J. 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**: 292-310.
Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775-1787.
Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**: e1003711.
Ha N, Polychronidou M, Lohmann I. 2012. COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PLoS One* **7**: e52055.
Handel AE, Gallone G, Zameel Cader M, Ponting CP. 2017. Most brain disease-associated and eQTL haplotypes are not located within transcription factor DNase-seq footprints in brain. *Hum Mol Genet* **26**: 79-89.
Haraksingh RR, Snyder MP. 2013. Impacts of variation in the human genome on gene regulation. *J Mol Biol* **425**: 3970-3977.
Hartmann H, Guthohrlein EW, Siebert M, Luehr S, Soding J. 2013. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res* **23**: 181-194.
HD L. 2015. - The Maternal-to-Zygotic Transition. Preface. *Curr Top Dev Biol*: 00076-00079.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576-589.
Herz HM, Hu D, Shilatifard A. 2014. Enhancer malfunction in cancer. *Mol Cell* **53**: 859-866.
Heyn H. 2016. Quantitative Trait Loci Identify Functional Noncoding Variation in Cancer. *PLoS Genet* **12**: e1005826.
Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A et al. 2016. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* **17**: 11.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 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.
Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473-476.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827-841.

Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* **38**: 2154-2167.

Huang D, Ovcharenko I. 2015. Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res* **43**: 225-236.

Huggins P, Zhong S, Shiff I, Beckerman R, Laptenko O, Prives C, Schulz MH, Simon I, Bar-Joseph Z. 2011. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics* **27**: 2361-2367.

Jiang P, Singh M. 2014. CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Res* **42**: 2833-2847.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497-1502.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome research* **21**: 1150-1159.

Kang R, Zhang Y, Huang Q, Meng J, Ding R, Chang Y, Xiong L, Guo Z. 2019. EnhancerDB: a resource of transcriptional regulation in the context of enhancers. *Database : the journal of biological databases and curation* **2019**.

Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV et al. 2013. Extensive variation in chromatin states across humans. *Science* **342**: 750-752.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J. 2014. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* **111**: 6131-6138.

Khan A, Zhang X. 2015. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* doi:10.1093/nar/gkv1002.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**: 2976-2987.

Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. 2016. Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**: 93-108.

Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI et al. 2013. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**: 744-747.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research* **15**: 1051-1060.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107-116.

Kleftogiannis D, Kalnis P, Bajic VB. 2015. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res* **43**: e6.

Koyabu Y, Nakata K, Mizugishi K, Aruga J, Mikoshiba K. 2001. Physical and functional interactions between Zic and Gli proteins. *Journal of Biological Chemistry* **276**: 6889-6892.

Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. 2010. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* **26**: 2622-2623.

Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ. 2013. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41**: D195-202.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA 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.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330. doi: 310.1038/nature14248.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**: 1595-1602.

Labbé E, Letamendia A, Attisano L. 2000. Association of Smads with lymphoid enhancer binding factor 1/T cell-specific factor mediates cooperative signaling by the transforming growth factor-β and Wnt pathways. *Proceedings of the National Academy of Sciences* **97**: 8358-8363.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **175**: 598-599.

Landrum MJ, Kattman BL. 2018. ClinVar at five years: delivering on the promise. *Human mutation* **39**: 1623-1630.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**: D862-868.

Landrum MJ, Lee JM, Benson M, Brown GR, 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.

Li X, Shi L, Wang Y, Zhong J, Zhao X, Teng H, Shi X, Yang H, Ruan S, Li M et al. 2018. OncoBase: a platform for decoding regulatory somatic mutations in human cancers. *Nucleic Acids Res* doi:10.1093/nar/gky1139.

Li Y, Ni P, Zhang S, Li G, Su Z. 2019. ProSampler: an ultra-fast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics* doi:10.1093/bioinformatics/btz290.

Liu X, Brutlag DL, Liu JS. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127-138.

Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ. 2012. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res* **40**: e50.

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.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696-1697.

Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**: 72-79.

Martin N, Benhamed M, Nacerddine K, Demarque MD, Van Lohuizen M, Dejean A, Bischof O. 2012. Physical and functional interaction between PML and TBX2 in the establishment of cellular senescence. *The EMBO journal* **31**: 95-109.

Mason MJ, Plath K, Zhou Q. 2010. Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics* **26**: 2826-2832.

Mathelier A, Fornes O, Arenillas DJ, Chen CY, 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**: D110-115.

Mathelier A, Shi W, Wasserman WW. 2015. Identification of altered cis-regulatory elements in human disease. *Trends Genet* **31**: 67-76.

Mathelier A, Wasserman WW. 2013. The next generation of transcription factor binding site prediction. *PLoS Comput Biol* **9**: e1003214.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J et al. 2012a. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**: 1190-1195.

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. 2012b. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet* **8**: e1002599.

McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK. 2013. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**: 747-749.

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.

Muller-Molina AJ, Scholer HR, Arauzo-Bravo MJ. 2012. Comprehensive human transcription factor binding site map for combinatory binding motifs discovery. *PLoS One* **7**: e49086.

Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R et al. 2011. A cis-regulatory map of the Drosophila genome. *Nature* **471**: 527-531.

Niu M, Tabari E, Ni P, Su Z. 2018. Towards a map of cis-regulatory sequences in the human genome. *Nucleic Acids Res* **46**: 5395-5409.

Niu M, Tabari ES, 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. doi: 1010.1186/1471-2164-1015-1047.

Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, Sandoval J, Vidal E, Whiffin N, Planchon A et al. 2014. Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**: 87-90. doi: 10.1038/nature13602. Epub 12014 Jul 13623.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N. 2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* **42**: D358-D363.

Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R. 2019. The BioGRID interaction database: 2019 update. *Nucleic acids research* **47**: D529-D541.

Pai AA, Pritchard JK, Gilad Y. 2015. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* **11**: e1004857. doi: 1004810.1001371/journal.pgen.1004857. eCollection 1002015 Jan.

Paul DS, Soranzo N, Beck S. 2014. Functional interpretation of non-coding sequence variation: concepts and challenges. *Bioessays* **36**: 191-199. doi: 110.1002/bies.201300126. Epub 201302013 Dec 201300125.

Pavesi G, Mauri G, Pesole G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17 Suppl 1**: S207-214.

Pavesi G, Mereghetti P, Mauri G, Pesole G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32**: W199-203.

Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. 2013. Enhancers: five essential questions. *Nat Rev Genet* **14**: 288-295.

Perrot CY, Gilbert C, Marsaud V, Postigo A, Javelaud D, Mauviel A. 2013. GLI 2 cooperates with ZEB 1 for transcriptional repression of CDH1 expression in human melanoma cells. *Pigment cell & melanoma research* **26**: 861-873.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110-121.

Quang D, Xie X. 2014. EXTREME: an online EM algorithm for motif discovery. *Bioinformatics* **30**: 1667-1673.

Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. 2013. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* **9**: e1002968.

Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA. 2014. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet* **22**: 144-147.

Reid JE, Wernisch L. 2011. STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Res* **39**: e126.

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**: 1155-1159. doi: 1110.1126/science.1260943.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651-657.

Rohr CO, Parra RG, Yankilevich P, Perez-Castro C. 2013. INSECT: IN-silico SEarch for Co-occurring Transcription factors. *Bioinformatics* **29**: 2852-2858.

Rubinstein M, de Souza FS. 2013. Evolution of transcriptional enhancers and animal diversity. *Philos Trans R Soc Lond B Biol Sci* **368**: 20130017.

Sánchez-Tilló E, De Barrios O, Valls E, Darling D, Castells A, Postigo A. 2015. ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene* **34**: 5760-5770.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23**: 777-788.

Siepel A, Arbiza L. 2014. Cis-regulatory elements and human evolution. *Curr Opin Genet Dev* **29**: 81-89.

Sinha S, Tompa M. 2003. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* **31**: 3586-3588.

Smith E, Shilatifard A. 2014. Enhancer biology and enhanceropathies. *Nat Struct Mol Biol* **21**: 210-219. doi: 210.1038/nsmb.2784.

Spielmann M, Mundlos S. 2013. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays* **35**: 533-543.

Stamatoyannopoulos JA. 2012. What does our genome encode? *Genome Res* **22**: 1602-1611.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Sun H, Guns T, Fierro AC, Thorrez L, Nijssen S, Marchal K. 2012. Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res* **40**: e90.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* **40**: e31.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

TPLi X, Wang W, Wang J, Malovannaya A, Xi Y, Li W, Guerra R, Hawke DH, Qin J, Chen J. 2015. Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Molecular systems biology* **11**.

van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks. *Methods Mol Biol* **804**: 281-295.

Vandenbon A, Kumagai Y, Akira S, Standley DM. 2012. A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription. *BMC Genomics* **13 Suppl 7**: S11.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252-263.

Villarroel MC, Rajeshkumar NV, Garrido-Laguna I, De Jesus-Acosta A, Jones S, Maitra A, Hruban RH, Eshleman JR, Klein A, Laheru D et al. 2011. Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer. *Mol Cancer Ther* **10**: 3-8.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88-92.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* **34**: D95-97.

Vockley CM, McDowell IC, D'Ippolito AM, Reddy TE. 2017. A long-range flexible billboard model of gene activation. *Transcription* **8**: 261-267.

Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT et al. 2018a. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**.

Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, Kellis M. 2018b. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**: 5380.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095-1106.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431-1443.

Whitaker JW, Chen Z, Wang W. 2015. Predicting the human epigenome from DNA motifs. *Nat Methods* **12**: 265-272, 267 p following 272.

Whitington T, Frith MC, Johnson J, Bailey TL. 2011. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* **39**: e98.

Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. 2018. High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell Rep* **23**: 1581-1597.

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.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59-69.

Won KJ, Agarwal S, Shen L, Shoemaker R, Ren B, Wang W. 2009. An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS One* **4**: e5501.

Won KJ, Chepelev I, Ren B, Wang W. 2008. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* **9**: 547.

Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H, Snyder M. 2013. Variation and genetic control of protein abundance in humans. *Nature* **499**: 79-82.

Yanez-Cuna JO, Kvon EZ, Stark A. 2013. Deciphering the transcriptional cis-regulatory code. *Trends Genet* **29**: 11-22. doi: 10.1016/j.tig.2012.1009.1007. Epub 2012 Oct 1023.

Yip KY, Cheng C, Bhardwaj N, Brown JB, 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.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. *Genome Biol* **16:56.**: 10.1186/s13059-13015-10621-13055.

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.

Zhang S, Li S, Niu M, Pham PT, Su Z. 2011. MotifClick: prediction of cis-regulatory binding sites via merging cliques. *BMC Bioinformatics* **12**: 238.

Zhang S, Xu M, Li S, Su Z. 2009. Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res* **37**: e72.

Zhou S, Treloar AE, Lupien M. 2016. Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. *Cancer discovery* **6**: 1215-1229.

Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IJcken WF. 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences* **111**: 996-1001.