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

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: Accurate prediction of *cis*-regulatory modules in humans

**Abstract**

Annotating all *cis*-regulatory modules (CRMs) and constituent transcription factor (TF) binding sites (TFBSs) in genomes is essential to understand genome functions, however, the tasks remain highly challenging. Here, we tackle the tasks using a different strategy than common practice by integrating putative TFBSs motifs found in numerous TF ChIP-seq datasets. Using available 6,092 datasets covering 77.47% of the human genome, our compactional pipeline enables us to partition the covered genome regions into a CRM candidate (CRMC) set (56.84%) and a non-CRMC set (43.16%), consisting of 44.03% and 33.44% of the genome, respectively. Intriguingly, like experimentally validated VISTA enhancers and ClinVar mutants, the predicted 1,404,973 CRMCs are under strong evolutionary constraints, suggesting that at least the vast majority of them might be functional. In stark contrast, the non-CRMCs are largely selectively neutral, suggesting that at least the vast majority of them might be non-functional. When evaluated on VISTA enhancers and ClinVar mutants, the CRMCs recall all VISTA enhancers (100.00%) and almost all ClinVar mutants (97.43%) at p-values of $5x10^{-5}$ and 0.05, respectively, substantially outperforming three state-of-the-art methods tested. Based on our predictions and evolution of the genome, we estimate that the human genome might code for about 1.5 million CRMs and 68 million TFBSs, comprising about 55% and 22% of the genome, respectively; for both of which we predicted about 80%. Therefore, the *cis*-regulatory genome appears to be more prevalent than originally thought.

**Introduction**

*cis*-regulatory sequences, also known as *cis*-regulatory modules (CRMs) (i.e., promoters, enhancers, silencers and insulators), are made of clusters of short DNA sequences that are recognized and bound by specific transcription factors (TFs)(Davidson, 2006). CRMs display different functional states in different cell types in multicellular eukaryotes during development and physiological homeostasis, and are responsible for specific transcriptomes of the cell types(Wilczynski & Furlong, 2010). A growing body of

evidence indicates that CRMs are at least as important as coding sequences (CDSs) to account for inter-species divergence(M. King & Wilson, 1975; Rubinstein & de Souza, 2013) and intra-species diversity(Siepel & Arbiza, 2014), in complex traits. Recent genome-wide association studies (GWAS) 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 or are in linkage disequilibrium (LD) with TF binding sites (TFBSs) in CRMs(Maurano et al., 2012). It has been shown that complex trait-associated variants systematically disrupt TFBSs of TFs related to the traits (Maurano et al., 2012), and that variation in TFBSs affects DNA binding, chromatin modification, transcription(Kasowski et al., 2013; Kilpinen et al., 2013; McVicker et al., 2013), and susceptibility to complex diseases(Mathelier, Shi, & Wasserman, 2015; Smith & Shilatifard, 2014) including cancer(Herz, Hu, & Shilatifard, 2014; Khurana et al., 2016; Ongen et al., 2014; Zhou, Treloar, & Lupien, 2016). 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(M. Wang et al., 2019; Whitaker, Chen, & Wang, 2015). These alterations in molecular phenotypes can lead to changes in cellular and organ-related phenotypes among individuals of a species(Pai, Pritchard, & Gilad, 2015; Ward & Kellis, 2012). However, it has been difficult to link non-coding variants to complex traits(Albert & Kruglyak, 2015; Whitaker et al., 2015), largely because of our lack of a good understanding of all CRMs, their constituent TFBSs and target genes in genomes(Gasperini, Tome, & Shendure, 2020).

Fortunately, the recent development of ChIP-seq techniques for locating histone marks(Barski et al., 2007) and TF bindings in genomes in specific cell/tissue types(Johnson, Mortazavi, Myers, & Wold, 2007) has led to the generation of enormous amounts of data by large consortia such as ENCODE(Moore et al., 2020), Roadmap Epigenomics(Kundaje et al., 2015) and Genotype-Tissue Expression (GTEx)(Consortium, 2015), as well as individual labs worldwide(Mei et al., 2017). These increasing

amounts of ChIP-seq data for relevant histone marks and various TFs in a wide spectrum of cell/tissue types provide an unprecedented opportunity to predict a map of CRMs and constituent TFBSs in the human genome. Many computational methods have been developed to explore these data individually or jointly(Kleftogiannis, Kalnis, & Bajic, 2016). For instance, as the large number of binding-peaks in a typical TF ChIP-seq dataset dwarf earlier motif-finding tools (e.g., MEME(Bailey & Elkan, 1994) and BioProspector(Liu, Brutlag, & Liu, 2001)) to find TFBSs of the ChIP-ed TF, new tools (e.g., DREME(Bailey, 2011), MEME-ChIP(Machanick & Bailey, 2011), XXmotif (Hartmann, Guthohrlein, Siebert, Luehr, & Soding, 2013) and Homer(Heinz et al., 2010)) have been developed. However, some of these tools (e.g. MEME-ChIP) were designed to find primary motifs of the ChIP-ed TF in very short sequences (~200bp) around the binding-peak summits in a small number of selected binding peaks in a dataset due to their slow speed. Some faster tools (e.g. Homer, DREME, and XXmotif) are based on the discriminative motif-finding schema(Sinha, 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(Bailey & Machanick, 2012; Davidson, 2006), tools such as SpaMo (Whitington, Frith, Johnson, & Bailey, 2011), CPModule (Sun et al., 2012) and CCAT (Jiang & Singh, 2014) 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. Due probably to the difficulty in motif finding in all of tens of thousands of binding peaks in a typical mammalian TF ChIP-seq dataset, few efforts have been made to fully explore entire sets of increasing volumes of TF ChIP-seq datas to simultaneously predict CRMs and constituent TFBSs(Kheradpour & Kellis, 2014; Niu, Tabari, Ni, & Su, 2018; Niu, Tabari, & Su, 2014; Yip et al., 2012).

On the other hand, as a single histone mark is not a reliable CRM predictor, a great deal of

efforts have been made to predict CRMs based on multiple histone marks and chromatin accessibility (CA) data from the same cell/tissue types using various machine-learning methods, including hidden Markov models(Ernst & Kellis, 2012), dynamic Bayesian networks(Hoffman et al., 2013), time-delay neural networks(Firpi, Ucar, & Tan, 2010), random forest(Rajagopal et al., 2013), and support vector machines (SVMs)(Kleftogiannis, Kalnis, & Bajic, 2015). Many enhancer databases have also been created either by combining results of multiple such methods(Ashoor, Kleftogiannis, Radovanovic, & Bajic, 2015; Fishilevich et al., 2017; Zerbino, Wilder, Johnson, Juettemann, & Flicek, 2015), or by identifying overlapping regions of CA and histone mark tracks in the same cell/tissue types(Chen et al., 2020; Cheneby, Gheorghe, Artufel, Mathelier, & Ballester, 2018; Gao & Qian, 2020; Kang et al., 2019; G. Zhang et al., 2018). In particular, most recently, the ENCODE phase 3 consortium(Moore et al., 2020) identified 926,535 candidate *cis*-regulatory elements (cCREs) based on overlaps between millions of DNase I hypersensitivity sites (DHSs)(Thurman et al., 2012) and transposase accessible sites (TASs)(Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013), active promoter histone mark H3K4me3(Aday, Zhu, Lakshmanan, Wang, & Lawson, 2011) peaks, active enhancer mark H3K27ac(Creyghton et al., 2010) peaks, and insulator mark CTCT(Kim et al., 2007) peaks in a large number of cell/tissue types. 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(Ernst & Kellis, 2012; Hoffman et al., 2013; Moore et al., 2020; Won, Chepelev, Ren, & Wang, 2008). The resolution of these methods is low(Ernst & Kellis, 2012; Hoffman et al., 2013; Won et al., 2008) and often lacks TFBSs information(Ernst & Kellis, 2012; Hoffman et al., 2013; Moore et al., 2020; Won et al., 2008), particularly for novel motifs, although some predictions provide TFBSs locations by finding matches to known motifs(Ashoor et al., 2015; Fishilevich et al., 2017; Gao & Qian, 2020). Moreover, results of these methods are often inconsistent(Arbel et al., 2019; Catarino & Stark, 2018; Dogan et al., 2015; Kwasnieski, Fiore, Chaudhari, & Cohen, 2014), e.g., even the best-performing tools (DEEP and CSI-ANN) have only 49.8% and 45.2%,

respectively, of their predicted CRMs overlap with the DHSs in Hela cells(Kleftogiannis et al., 2015); and only 26% of predicted ENCODE enhancers in K562 cells can be experimentally verified(Kwasnieski et al., 2014). The low accurate of these methods might be due to the fact that CA and histone marks alone are not reliable predictors of active CRMs(Arbel et al., 2019; Dogan et al., 2015; Kleftogiannis et al., 2015; Kwasnieski et al., 2014).

It has been shown that TF binding data are more reliable for predicting CRMs than CA and histone mark data(Arbel et al., 2019; Dogan et al., 2015; Kleftogiannis et al., 2015; Kwasnieski et al., 2014). Moreover, in addition to the enrichment of primary binding sites of the ChIP-ed TF at the summit of called binding peaks, TFBSs of cooperators of the ChIP-ed TFs also tend to appear at the two ends of binding peaks(Goi, Little, & Xie, 2013; Y. Li, Ni, Zhang, Li, & Su, 2019). With this recognition, instead of predicting cell/tissue type specific CRMs using CA and histone marks data, we proposed to first predict a largely cell-type agnostic or static map of CRMs and constituent TFBBs in the genome by integrating all available TF ChIP-seq datasets for different TFs in various cell/tissue types(Niu et al., 2018; Niu et al., 2014), just as has been done for identifying all genes encoded in the genome using gene expression data from all cell/tissue types(Allen, Pertea, & Salzberg, 2004). We also proposed to appropriately extend short called TF binding peaks to the typical length of enhancers, so that more TFBSs for cooperator of the ChIP-ed TF could be included(Goi et al., 2013; Y. Li et al., 2019), and thus, full-length CRMs can be identified(Niu et al., 2018; Niu et al., 2014). Once a map of CRMs and constituent TFBSs is available, the specificity of CRMs in any cell/tissue type can be determined using one or few epigenetic mark datasets collected in the cell/tissue type(Moore et al., 2020), as when anchored by correctly predicted CRMs, the accuracy of epigenetic marks for predicting active CRMs could be largely improved(Dogan et al., 2015). Although our earlier implementation of this strategy, dePCRM, resulted in promising results using even insufficient datasets available then(Niu et al., 2018; Niu et al., 2014), we were limited by three technical

137 hurdles. First, although existing motif-finders such as DREME used in dePCRM worked well for relatively
138 small ChIP-seq datasets from organisms with smaller genomes such as the fly (Niu et al., 2014), they are
139 unable to work on very large entire datasets from mammalian cells/tissues, so we had to split a large
140 dataset into smaller ones for motif finding in the entire dataset(Niu et al., 2018), which may compromise
141 the accuracy of motif finding and complicate subsequent data integration. Second, although the
142 distances and interactions between TFBSs in a CRM are critical, both were not explicitly modeled(Niu et
143 al., 2018; Niu et al., 2014), potentially limiting the accuracy of predicted CRMs. Third, the earlier
144 "branch-and-bound" approach to integrate motifs found in different datasets is not efficient enough to
145 handle much larger number of motifs found in ever increasing number of large ChIP-seq datasets from
146 human cells/tissues(Niu et al., 2018; Niu et al., 2014). To overcome these hurdles, we developed
147 dePCRM2 based on an ultrafast, accurate motif-finder ProSampler(Y. Li et al., 2019), a novel effective
148 combinatory motif pattern discovery method, and a scoring function that considers distances and
149 interactions between TFBSs in a CRM. Using available 6,092 ChIP-seq datasets covering 77.47% of the
150 human genome, dePCRM2 enable us to partition the covered genome regions in a CRM candidate
151 (CRMC) set and a non-CRMC set, and predict 201 unique TF binding motif families in the CRMCs. Both
152 evolutionary and independent experimental data indicate that at least the vast majority of the predicted
153 1,404,973 CRMCs might be functional, while at least the vast majority of the predicted non-CRMCs
154 might be non-functional.
155
156 **Results**
157 **The dePCRM2 pipeline**
158 TFs in eukaryotes tend to cooperatively bind to their TFBSs in CRMs(Davidson, 2006). Different CRMs of
159 the same gene are structurally similar and closely located(Snetkova & Skok, 2018), and functionally
160 related genes are often regulated by the same sets of TFs in different cell types during development and

in maintaining physiological homeostasis(Davidson, 2006). Due to the clustering nature of TFBSs of cooperative TFs in a CRM, if we extend the short called binding peaks of a TF ChIP-seq dataset from the two ends and reach the typical size of a CRM (500~3000bp)(Visel, Minovitsky, Dubchak, & Pennacchio, 2007), the extended peaks have great chance to contain TFBSs of cooperative TFs(Y. Li et al., 2019; Niu et al., 2018; Niu et al., 2014). For instance, if two different TFs cooperatively regulate the same regulons in several cell types, then at least some of the extended peaks of datasets for the two TFs from these cell types should contain the TFBSs of both TFs, or even have some overlaps if the CRMs are reused in different cell types. Therefore, if we have a sufficient number of ChIP-seq datasets for different TFs from the same and different cell types, we are likely to include datasets for some cooperative TFs, and their TFBSs may co-occur in some extended peaks. Based on these observations, we designed dePCRM(Niu et al., 2018; Niu et al., 2014) and dePCRM2 to predict CRMs and constituent TFBSs by identifying overrepresented co-occurring patterns of motifs found by a motif-finder in a large number of TF ChIP-seq datasets. dePCRM2 overcomes the aforementioned shortcomings of dePCRM as follows. First, using an ultrafast, accurate motif-finder ProSampler(Y. Li et al., 2019), we can find significant motifs in any size of available ChIP-seq datasets (Figures 1A and 1B) without the need to split large datasets into small ones(Niu et al., 2018). Second, after identifying highly co-occurring motifs pairs (CPs) in the extended binding peaks in each dataset (Figure 1C), we cluster highly similar motifs in the CPs and finds a unique motif (UM) in each resulting cluster (Figure D). Third, we model distances and 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, we identify as CRMCs closely-located clusters of binding sites of the UMs along the genome (Figure 1F), thereby partitioning genome regions covered by the extended binding peaks into a CRMCs set and a non-CRMCs set. Fifth, we evaluate each CRMC using a novel score that incorporates the quality of binding sites and the strength of interactions among the corresponding UMs defined in the

interaction networks (Figure 1G). Lastly, we compute a p-value for each $S_{CRM}$ score, so that CRMs and constituent TFBSs can be predicted at different significant levels using different $S_{CRM}$ score or p-value cutoffs. Clearly, as the number of UMs is a small constant number constrained by the number of TF families encoded in the genome, the downstream computation based on the set of UMs runs in a constant time, thus dePCRM2 is highly scalable. The source code of dePCRM2 is available at http://github.com/zhengchangsulab/pcrm2

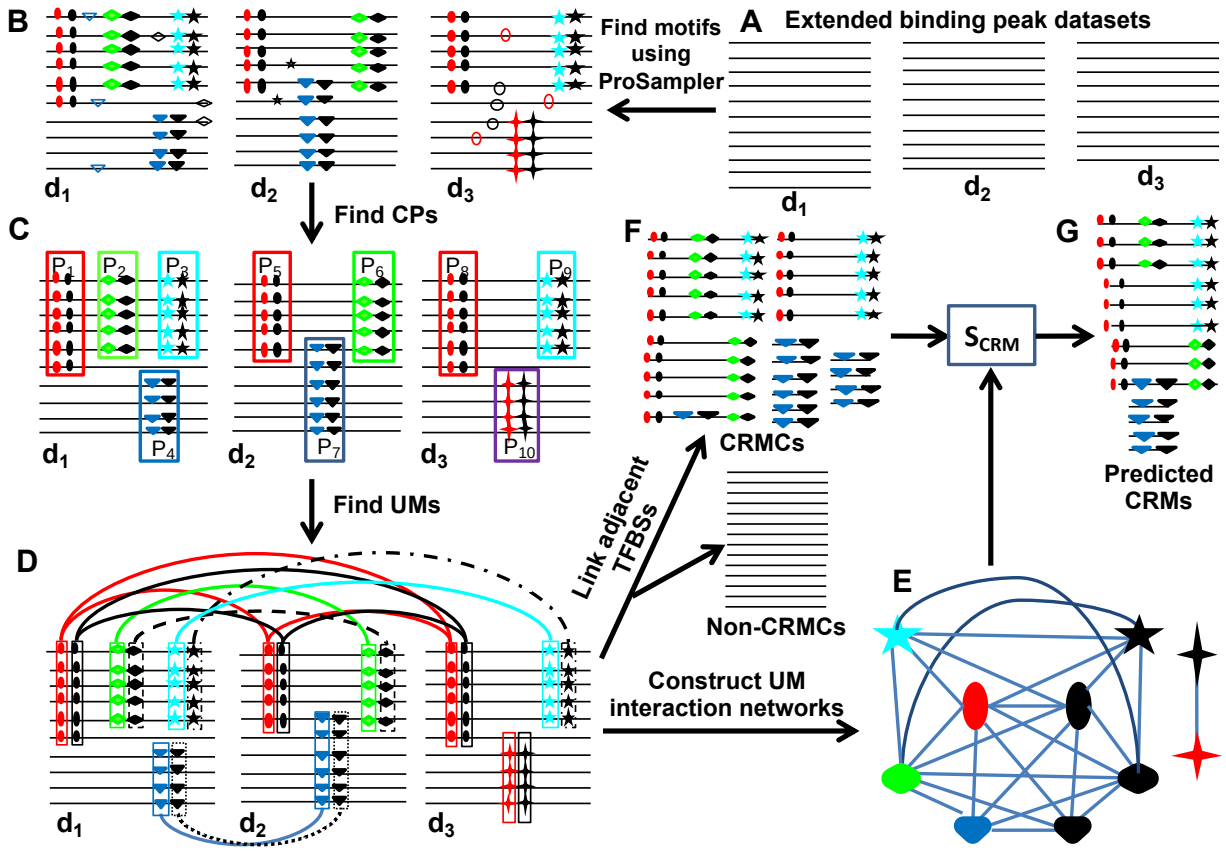


Figure 1. Schematic of the dePCRM2 pipeline. A. Extend each binding peak in each dataset to its two ends to reach a preset length, e.g., 1,000bp. B. Find motifs in each dataset using ProSampler. C. Find CPs in each dataset. For clarity, only the indicated CPs are shown, while those formed between motifs 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 link adjacent TFBSs along the genome, thereby identifying CRMCs and non-CRMCs. G. Evaluate each CRMC by computing its $S_{CRM}$ score and the associated p-value.

**Extended binding peaks in different datasets have extensive overlaps**

After filtering out low-quality peaks in the 6,092 ChIP-seq datasets (Table S1), we ended up with 6,070 non-empty datasets for 779 TFs in 2,631 cell/tissue 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 and Table S1). The datasets also are strongly biased to few TFs (Figure 2B and Table S1). For example, 370 and 263 datasets were collected for TFs 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 a mean of 19,314 (Figure 2C). For instance, datasets for STAT1 and NR3C1 have the smallest number of 2 binding peaks in 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 first quality filter. 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. We have shown that extension of binding peaks to 500~1,000bp could substantially increase the chance of finding TFBSs of cooperators of the ChIP-ed TFs, while the introduced noise had a little effect on identifying the primary motifs of ChIP-ed TFs(Y. Li et al., 2019). 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. Remarkably, by extending the originaly called binding peaks, we increased the coverage of genome by 89.04% (77.47% vs 40.98%). However, 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 CRMs and TFBSs in the genome(Niu et al., 2018). As stated earlier, we predict CRMs and constituent TFBSs based on overlapping patterns between datasets for ChIP-ed cooperative TFs, thus, 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 certain 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 in certain cell/tissue types (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, Letamendia, & Attisano, 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(X. Li 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 genome regions(Niu et al.,

2018; Niu et al., 2014). In addition, we collected 10 sets of experimentally determined CRM function-related elements (Materials and Methods) and found that they are enriched in the covered 77.47% genome regions relative to the uncovered 22.53% regions, including 785 (80.43%) VISTA enhancers(Visel et al., 2007), 402,730 (94.84%) of ClinVar SNPs(Landrum et al., 2018), 181,436 (98.38%) FANTOM5 promoters (FPs)(Forrest et al., 2014), 32,029 (97.98%) FANTOM5 enhancers (FEs)(Andersson et al., 2014), 82,378 (90.16%) of GWAS SNPs(Buniello et al., 2019), 121,075,184 (98.86%) DHSs(Mei et al., 2017) (Table S2), 29,195,778 (98.90%) TASs(Mei et al., 2017)(Table S3), 98,297,240(98.32%) H3K27ac peaks(Mei et al., 2017)(Table S4), 75,467,050 (97.38%) H3K4me1 peaks(Mei et al., 2017)(Table S5), and 69,282,044(98.14%) H3K4me3 peaks(Mei et al., 2017)(Table S6). These results indicate that the covered genome regions were better studied than the uncovered regions. We will evaluate the sensitivity of dePCRM2 to recall these elements at different $S_{CRM}$ scores and associated p-value cutoffs.

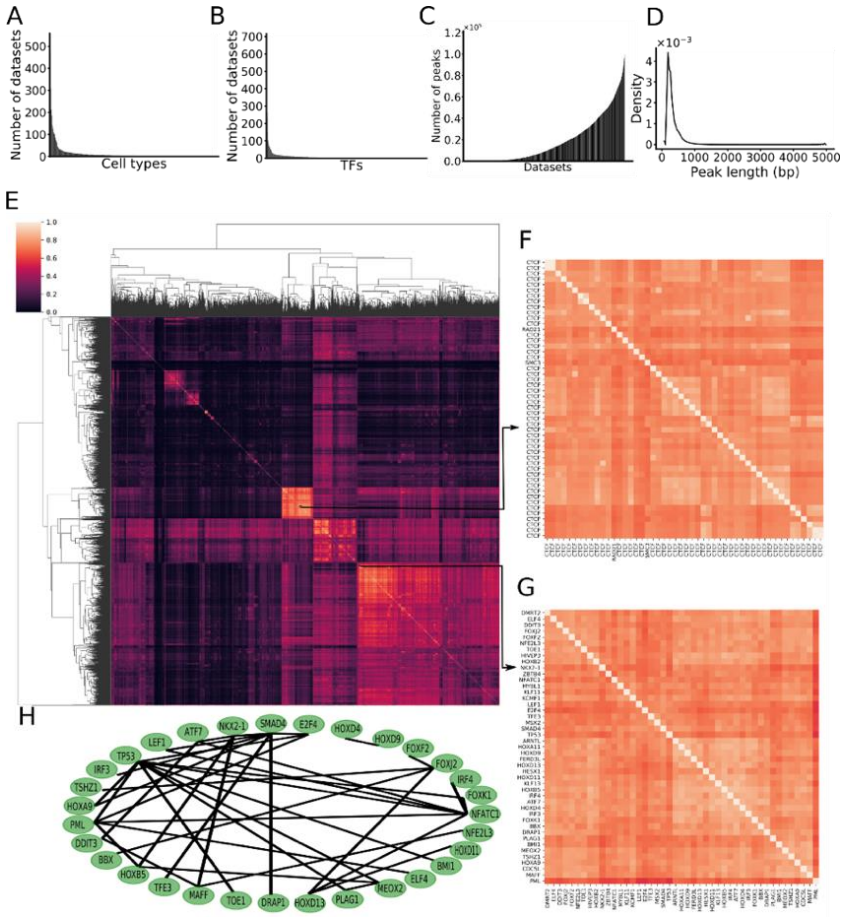

Figure 2. Properties of the datasets. A. Number of datasets collected in each cell/tissue types sorted in descending order. B. Number of datasets collected for each TF sorted in descending order. C. Number of peaks in each dataset sorted in ascending order. D. Distribution of the lengths of binding peaks 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 recall most known TF motifs families and have distinct patterns of interactions.**

ProSampler identified at least one motif in 5,991 (98.70%) of the 6070 datasets but failed to find any motifs in the remaining 79 (1.30%) datasets that all contain less than 310 binding peaks (Table S1), 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 trimodal 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, 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 further 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 less biased 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 (Figure S1A). dePCRM2 found a UM in 201 (82.04%) of the 245 clusters (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 (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 a mean of 905,288 (Figure 3C and Table S7), 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 (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 motifs in the HOCOMOCO(Kulakovskiy et al., 2018) and JASPAR(Mathelier et al., 2016) databases, and 92 (78.63%) match at least two(Table S7), suggesting that most UMs might consist of motifs of different TFs of the same TF family/superfamily that recognize highly similar motifs, a well-known phenomenon(Ambrosini et al., 2020; Lambert et al., 2018). 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 Table S7). 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 (Table S8), thus, our predicted UMs include most of the known TF motif families.

To model interactions between cognate TFs of the UMs, we compute an interaction score based on distances and cooccurrence levels between binding sites of two UMs (formula 3). As shown in Figure 3G, there are clear interaction patterns between putative cognate TFs of many UMs, many of which are supported by experimental evidence. For example, in a cluster formed by 10 UMs (Figure 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 to repress the expression of *CDH*1 in human melanoma cells via directly binding to two close binding sites at the *CDH*1 promoter(Perrot et al., 2013) ; ZIC and GLI cooperatively regulate neural and skeletal development through physical interactions between their zinc finger domains (Koyabu, Nakata, Mizugishi, Aruga, & Mikoshiba, 2001); and ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate the expression of *WNT*(Sánchez-Tilló et al., 2015), to name a few.

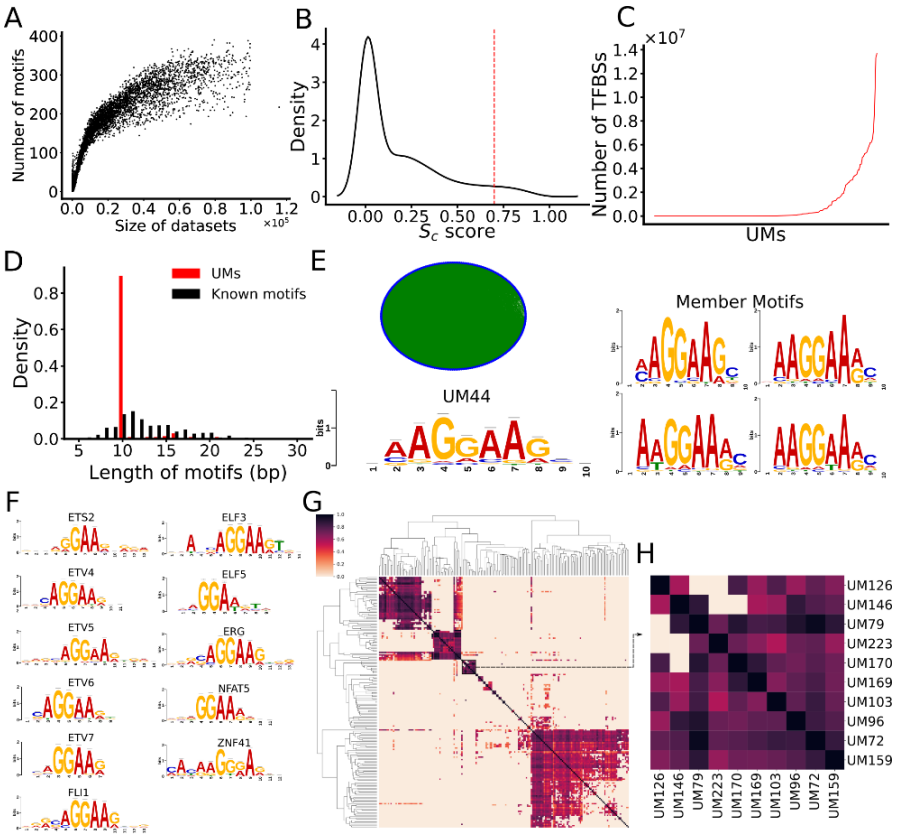


Figure 3. Prediction of UMs. A. Relationship between the number of predicted motifs in a dataset and the size of the dataset (number of binding peaks in the dataset). The datasets are sorted in ascending order of their sizes. B. Distribution of cooccurrence scores (Sc) of motif pairs found in a dataset. The dotted vertical line indicates the cutoff value (0.7) 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.

## An appropriate extension of original binding peaks could greatly increase the power of datasets

By concatenating closely located binding sites of the UMs along the genome, dePCRM2 partitions the extended binding peak-covered genome regions in two exclusive sets (Figure 1F ), the CRMC set containing 1,404,973 CRMCs with a total length of 1,359,824,275bp (56.84%) covering 44.03% of the genome, and the non-CRMC set containing 1,957,936 sequence segments with a total length of 1,032,664,424bp (43.16%) covering 33.44% of the genome. Interestingly, only 57.88% (776,999,862bp) of nucleotide positions of the CRMCs overlap those of the original peaks, while the remaining 42.12% (565,448,583bp) overlap those of the extended parts of the original peaks. Hence, dePCRM2 only retains 61.40% of positions covered by the original binding peaks, and abandoned the remaining 38.60% of position. These abandoned positions of called binding peak position might not enrich for TFBSs, which is in agreement with earlier studies(Devailly, Mantsoki, Michoel, & Joshi, 2015; Marinov, Kundaje, Park, & Wold, 2014; Mendoza-Parra, Van Gool, Mohamed Saleem, Ceschin, & Gronemeyer, 2013). Meanwhile, 565,448,583pb (42.12%) positions of the predicted CRMCs are covered by the extended parts of original binding peaks, suggesting that TFBSs of cooperative TFs are enriched in some extended parts as has been shown earlier(Goi et al., 2013; Y. Li et al., 2019; Niu et al., 2018; Niu et al., 2014). Thus by appropriately extending original binding peaks, we could greatly increase the power of datasets. We could predict functional states of about 57.88% of the CRMCs that are covered by original binding peaks 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. The predicted CRMCs and constituent TFBSs are available at https://cci-bioinfo.uncc.edu/

**Vast majority of the CRMCs have low p-values**

As shown in Figure 4A, the distribution of the $S_{CRM}$ scores of the CRMCs is strongly right-skewed relative to that of the Null CRMCs (Materials and Methods), indicating that the CRMCs generally score much higher than the Null CRMCs, and thus are unlikely produced by chance. Based on the distribution of the $S_{CRM}$ scores of the Null CRMCs, dePCRM2 computes a p-value for each CRMC (Figure 4A). With the increase in the $S_{CRM}$ cutoff α ($S_{CRM} \geq \alpha$), the associated p-value cutoff drops rapidly, while both the number of predicted CRMs and the proportion of the genome covered by the predicted CRMs decrease much slowly (Figure 4B), indicating that most CRMCs have low p-values. For instance, with α increasing from 56 to 922, p-value drops precipitously from 0.05 to $1.00 \times 10^{-6}$ ($5 \times 10^{5}$ fold), while the number of predicted CRMs decreases from 1,155,151 to 327,396 (3.53 fold), and the proportion of the genome covered by the predicted CRMs decreases from 43.47% to 27.82% (1.56 fold) (Figure 4B). Predicted CRMs contain from 20,835,542 (p-value ≤ $1 \times 10^{-6}$) to 31,811,310 (p-value ≤ 0.05) non-overlapping putative TFBSs that consist of from 11.47% (p-value≤ $1 \times 10^{-6}$) to 16.54% (p-value ≤ 0.05) of the genome (Figure 4C). In other words, dependent on p-value cutoffs ($1 \times 10^{-6}$ ~0.05), 38.05~41.23% of nucleotide positions of the predicted CRMs are made of putative TFBSs (Figure 4C), and most of 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, dependent on p-value cutoffs ($1 \times 10^{-6}$ ~0.05), the remaining 4.54~6.01% and 5.33~6.80% of the predicted CRMs and constituent TFBSs, respectively, are in exonic sequences (ESs, including CDSs, 5'- and 3'-untranslated regions), respectively (Figure 4C), in agreement with an earlier report(Stergachis et al., 2013).

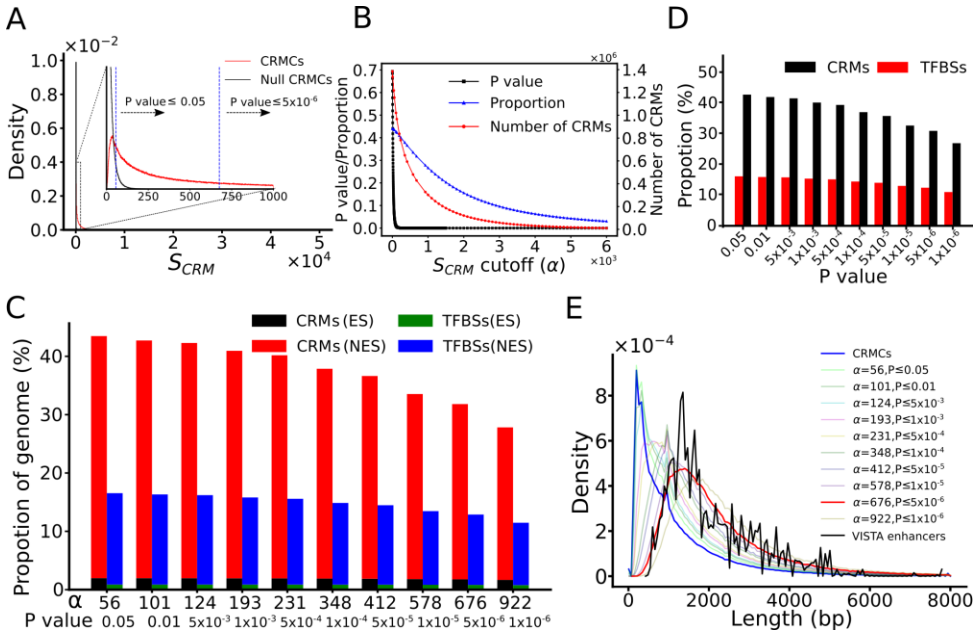


Figure 4. Prediction of CRMs using different $S_{CRM}$ cutoffs. A. Distribution of $S_{CRM}$ scores of the CRMCs and Null CRMCs. B. Number of the predicted CRMs, proportion of the genome predicted to be CRMs and the associated p-value as functions of the $S_{CRM}$ cutoff α. C. Percentage of the genome 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 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.

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

To see whether or not the $S_{CRM}$ score can capture the length features of CRMs, we compared the distribution of the lengths of predicted CRMCs at different $S_{CRM}$ cutoffs α and associated p-values with those of the entire set of 976 experimentally validated VISTA enhancers. As shown in Figure 4E, most of the CRMCs are shorter than VISTA enhancers (mean length 981bp vs 2,049bp). Specifically, There are 621,842 (44.26%) of the 1,404,973 CRMCs, which are shorter than the shortest VISTA enhancer (428bp), suggesting that most of these short CRMCs are likely short CRMs (such as promoters or short enhancers) or components of long enhancers. However, these short CRMCs (< 428bp) consist of only 7.42% of the total length of the CRMCs, while the remaining 733,132 (55.74%) CRMCs that are at least 428bp long comprise 92.58% of the total length of the CRMCs, suggesting that full-length CRMCs might dominate the predicted CRMC positions. Interestingly, 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 the VISTA enhancers, indicating that the corresponding 428,628 predicted CRMs have similar length distribution (mean length 2,292bp) to those of VISTA enhancers (mean length 2,049bp) (Figure 4E), and thus they are likely full-length VISTA-like enhancers, which are mostly development-related enhancers(Bejerano et al., 2004; Katzman et al., 2007; Visel et al., 2008) that tend be longer than other types of enhancers(L. Li & Wunderlich, 2017). As short CRMs and CRM components have smaller $S_{CRM}$ scores than full-length enhancers (Figure 4E), they 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. Token together, these results suggest that full-length, particularly, VISTA-like enhancers dominate the predicted CRMC positions.

**Predicted CRMs tend to be under strong evolutionary selections**

To see how effective that dePCRM2 partitions the covered genome regions into the CRMC set and the non-CRMC set, we compared their evolutionary behaviors with that of the entire set of VISTA enhancers using the GERP(Cooper et al., 2010) and phyloP(Pollard, Hubisz, Rosenbloom, & Siepel, 2010) scores of their nucleotide positions in the genome. Both the GERP and the phyloP scores quantify conservation levels of positions based on nucleotide substitutions in alignments of multiple vertebrate genomes. The larger a positive GERP or phyloP score of a position, the more likely it is under negative/purifying selection; and a GERP or phyloP score around zero means that the position is selectively neutral or nearly so(Cooper et al., 2010; Pollard et al., 2010). However, a negative a GERP or phyloP score is

427 cautiously related to positive selection(Cooper et al., 2010; Pollard et al., 2010). For convenience of
428 discussion, we consider a position with a GERP or phyloP score within an interval centering on 0 [-$\delta$,+ $\delta$]
429 ($\delta$ >0) to be selectively neutral or nearly so, and a position with a score greater than $\delta$ to be under
430 negative selection. We define proportion of neutrality of a set of positions to be the size of the area
431 under the density curve of the distribution of the scores of the positions within the window [-$\delta$,+ $\delta$].
432 Because ESs evolve quite differently from NESs, we focused on the CRMCs and constituent TFBSs in
433 NESs, and left those that overlap ESs in another analysis (Jing Chen, Pengyu Ni, Jun-tao Guo and
434 Zhengchang Su). The choice of $\delta$ = 0.1, 0.2, 0.3, 0.4 ,0.5, 1 and 2 gave similar results (data not shown), so
435 we choose $\delta$ =1 in the subsequent analyses. As shown in Figure 5A, GERP scores of VISTA enhancers
436 shows a trimodal distribution with a small peak around score -5, a blunt peak around score 0, a sharp
437 peak around score 3.5, and a small proportion of neutrality of 0.23, indicating that most positions of
438 VISTA enhancers are under strong evolutionary selection, particularly, negative selection. This result is
439 consistent with the fact that VISTAT enhancers are mostly ultra-conserved(Visel et al., 2008),
440 development related enhancers(Bejerano et al., 2004; Katzman et al., 2007). In contrast, the distribution
441 of the GERP scores of the non-CRMCs (1,034,985,426 bp) in NESs displays a sharp peak around score 0,
442 with low right and left shoulders, and a high proportion of neutrality of 0.71 (Figure 5A), suggesting that
443 most of the non-CRMCs are selectively neutral or nearly so, and thus are likely to be nonfunctional.
444 Intriguingly, the distribution of the GERP scores of the 1,292,356 CRMCs (1,298,719,954bp) in NESs has a
445 blunt peak around score 0, with high right and left shoulders, and a small proportion of neutrality of
446 0.31 (Figure 5A). Thus, like VISTA enhancers, most positions of the CRMCs are also under strong
447 evolutionary selections, and thus, are likely to be functional. However, notably, the distribution of GERP
448 scores of the CRMCs lack obvious peaks around scores -5 and 3.5 (Figure 5A), therefore, the average
449 selection strength on the CRMCs is weaker than that on VISTA enhancers, which is expected considering
450 the ultra-conversation nature of development-related VASTA enhancers(Bejerano et al., 2004; Katzman

et al., 2007; Visel et al., 2008). Nevertheless, the dramatic differences between the evolutionary behaviors of the non-CRMCs and those of the CRMCs strongly suggests that dePCRM2 largely partitions the covered genome regions into a functional CRMC set and a nonfunctional non-CRMC set. Similar results were obtained using the phyloP scores, although they display quite different distributions than the GERP scores (Figure S3A).

To see why dePCRM2 abandoned the 38.60% nucleotide positions covered by the original binding peaks in predicting the CRMCs, we plotted the distribution of their conservation scores. As shown in Figure 5A, these abandoned positions have a GERP score distribution almost identical to those in the non-CRMCs, indicating that, like the non-CRMCs, they are unlikely to be functional, strengthening our earlier argument that they might not contain TFBSs. Therefore, dePCRM2 is able to accurately distinguish functional and non-functional parts in both the original binding peaks and their extended parts. As we indicated earlier, the 10 CRM function-related elements datasets (Tables S2~S6) we collected for validating the predicted CRMSs (see below) are strongly biased to the covered genome regions relative to the uncovered regions, to see why this is possible, we plotted the distributions of conservation scores of the positions of the covered and uncovered regions in NESs. Interestingly, the uncovered regions have a GERP score distribution and a proportion of neutrality (0.59) in between those of the covered regions (0.49) and those of the non-CRMCs (0.71) (Figure 5A), indicating that the uncovered regions are more evolutionarily selected than the non-CRMCs as expected, 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 could be estimated to be (1-0.59)/(1-0.49)=79.40% of the covered regions. Therefore, it appears that existing studies are strongly biased to more evolutionary constrained regions due probably to their large effect sizes and more critical functions. Similar results

were obtained using the phyloP scores (Figure S3A).

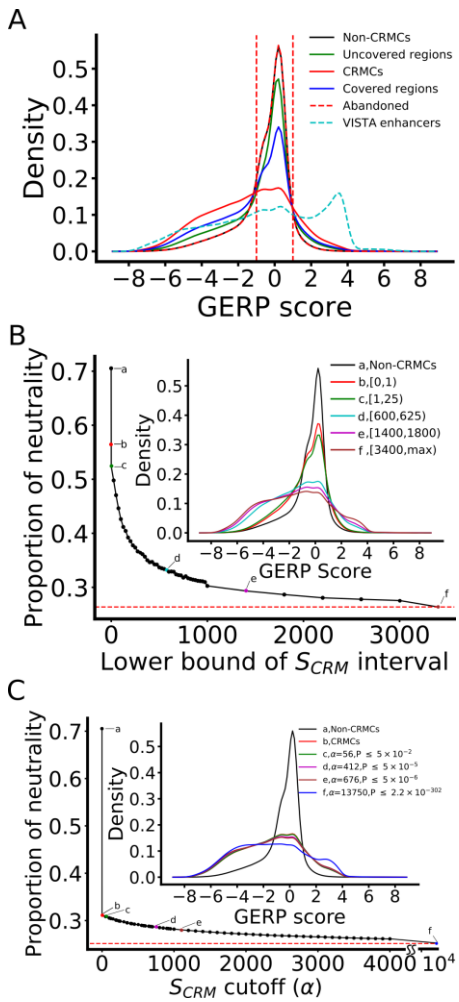


Figure 5. Different strengths of evolutionary constraints on the predicted CRMCs and non-CRMCs in NESs. A. Distributions of the GERP scores of nucleotides of the predicted CRMCs, non-CRMCs, abandoned genome regions covered by the original binding peaks, genome regions covered by the extended binding 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. The dotted lines in B and C indicate the saturation levels.

**The $S_{CRM}$ score captures the evolutionary feature of CRMs as functional elements**

To see whether the $S_{CRM}$ score of a CRMC captures its evolutionary feature, we plotted the distributions of the conservation scores of subsets of the CRMCs with a $S_{CRM}$ score in different non-overlapping intervals. Remarkably, even the subset with $S_{CRM}$ scores in the lowest interval [0, 1) have a smaller proportion of neutrality (0.56) than the non-CRMCs (0.71) (Figure 5B), indicating that these low-scoring CRMCs with short lengths (Figure 4E) are more likely to be under strong evolutionary constraints than the non-CRMCs, and thus might be more likely functional. With the increase in the lower bound of $S_{CRM}$ intervals, the proportion of neutrality of the subsets 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 indeed 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 (Figure S3B).

We next examined the relationship between the conservation scores of the predicted CRMs and $S_{CRM}$ score cutoffs α (or p-value cutoffs) used for their perditions. As shown in Figure 5C, even the CRMs predicted at a low α have a much smaller proportion of neutrality (e.g., 0.31 for the smallest α=0, i.e., the entire CRMC set) than the non-CRMCs (0.71), suggesting that most of the predicted CRMs might be authentic, while the non-CRMCs might contain few false negative CRMCs. With the increase in α (decrease in p-value cutoff), the proportion of neutrality of the predicted CRMs decreases but slowly, entering a saturation phase (Figure 5C). Interestingly, at high $S_{CRM}$ score cutoffs, the predicted CRMs evolve like VISTA enhancers. For instance, at α= 13,750, the distribution of GERP score of the predicted CRMs also displays a peak around score -5 and a peak around score 3.5, with a small proportion of neutrality of 0.24 (Figure 5C) (it is 0.23 for VISTA enhancers, Figure 5A). Thus, the higher α (i.e., the

smaller the p-value cutoff), the more likely the predicted CRMs are under strong evolutionary constraints, and thus functional. The infinitesimal decrease in the proportion of neutrality of predicted CRMs with the increase in $S_{CRM}$ cutoffs (Figure 5C) strongly suggests that the predicted CRMs, particularly those at a high p-value cutoff, are under similarly strong evolutionary constraints, close to the possibly highest saturation level to which ultra-conserved VISTA enhancers are subject. Therefore, it is highly likely that at high p-value cutoffs, specificity of the predicted CRMs might approach the possibly highest level that the VISTA enhancers achieve. However, without the availability of a gold standard negative CRM set in the genome(Gasperini et al., 2020), we could not explicitly calculate the specificity of the predicted CRMs at different p-value cutoffs. Similar results are observed using the phyloP scores (Figure S3C).

**dePCRM2 achieves high sensitivity and likely high specificity for recalling functionally validated CRMs and non-coding SNPs**

To further evaluate the accuracy of dePCRM2, we calculated the sensitivity (recall rate or true positive rate (TPR)) of CRMs predicted at different $S_{CRM}$ cutoffs α and associated p-values for recalling a variety of CRM function-related elements located in the covered genome regions in the 10 experimentally determined datasets in various cell/tissue types (Tables S2~S6, Materials and Methods). Here, if a predicted CRM and an element overlap each other by at least 50% of the length of the shorter one, we say that the CRM recalls the element. As shown in Figure 6A, with the increase in the p-value cutoff, the sensitivity for recalling the elements in all the 10 datasets increases rapidly and becomes saturated well before p-value increases to 0.05 (α ≥ 56). Figures S4A~S4J show examples of the predicted CRMs overlapping and recalling the elements in the 10 datasets. Particularly, at p-value cutoff $5x10^{-5}$ (α=412), the predicted 593,731 CRMs recall 100% of VISTA enhancers(Visel et al., 2007) and 91.61% of ClinVar SNPs(Visel et al., 2007) (Figure 6A). The very rapid saturation of sensitivity for recalling these two types of validated functional elements at very low p-value cutoffs once again strongly suggests that dePCRM2

also achieves very high specificity, although we could not explicitly compute it for the aforementioned reason. On the other hand, even at the higher p-value cutoff 0.05 (α=56), the predicted 1,155,151 CRMs only achieve varying intermediate levels of sensitivity for recalling FPs(Forrest et al., 2014) (88.77%), FEs(Andersson et al., 2014) (81.90%), DHSs(Thurman et al., 2012) (74.68%), TASs(Mei et al., 2017) (84.32%), H3K27ac(Mei et al., 2017) (82.96%) , H3K4me1(Mei et al., 2017) (76.77%), H3K4me3(Mei et al., 2017) (86.96%) and GWAS SNPs(Buniello et al., 2019) (64.50%), although all are significantly higher than that (15%) of randomly selected sequences with matched lengths from the covered genome regions (Figure 6A).

To find out the reasons for such varying sensitivity of dePCRM2 for recalling different types elements in the 10 datasets, we plotted the distribution of GERP scores of the recalled and uncalled elements in each dataset at p-value <0.05. Since we have already plotted the distribution of the entire set of VISTA enhancers (Figure 5A), to avoid redundancy, we instead plotted the distribution for the CRMs (VISTA-CRMs) that overlap and recall the 785 VISTA enhancers in the covered regions. As shown in Figure 6B, recalled elements in all the datasets evolve similarly to the predicted 1,155,151 CRMs (at p-value <0.05). However, VISTA-CRMs, recalled ClinVar SNPs and and recalled FPs evolve more like VISTA enhancers with a trimodal GERP score distribution (Figure 5A), suggesting that they are under stronger evolution constraints than the other recalled element types. These results are not surprising, as VISTA enhancers are mostly ultra-conserved, development related enhancers(Bejerano et al., 2004; Katzman et al., 2007; Visel et al., 2008), while ClinVar SNPs were identified for their conserved critical functions(Landrum et al., 2018), and promoters are well-known to be more conserved than are enhancers(Villar et al., 2015). In stark contrast, all unrecalled elements in the 10 datasets evolve similarly to the non-CRMCs, with the exception that the 10,350 (2.57%) unrecalled ClinVar SNPs display a trimodal distribution and there is no unrecalled VISTA enhancers (Figures 6B). Notably, proportions of neutrality of unrecalled Pes (0.59) and PFs(0.63) are smaller than that of the non-CRMCs (0.71) (Figure

6B), suggesting we might miss a small portion of authentic PEs and PFs (see below for false negative rate (FNR) estimations of our CRMs). Nevertheless, assuming that at least most of unrecalled elements in the datasets except the VISTA and ClinVar datasets are non-functional, we estimate that the false discovery rate (FDR) of the remaining eight datasets might be up to from 11.23% (1-0.8877) for FPs to 35.32% (1-0.6450) for GWAS SNPs. Such high FDRs for CA (DHSs and TASs) and histone marks are consistent with an earlier study(Dogan et al., 2015). Interestingly, the trimodal distribution of GERP scores of the 2.57% of unrecalled ClinVar SNPs displays a large peak around score 0 and two small peaks around -5 and 3.5, with a proportion of neutrality 0.40 (Figure 6B), indicating that about 40% of the relevant SNPs might be selectively neutral, and thus non-functional. We therefore estimate the FDR of the ClinVar SNP dataset to be about 0.40*2.57%=1.03%. Hence, like VISTA enhancers, ClinVar SNPs are a reliable set for evaluating CRM predictions. The peak of the unrecalled ClinVar SNPs around score 3.5 (Figure 6B), indicates that the relevant SNPs are under strong purifying selection, and thus might be functional, but were missed by dePCRM2. We therefore estimate our predictions (at p-value <0.05) might have a FNR < 2.57%-1.03%=1.54%. In other worlds, the real sensitivity (=1-FNR) for dePCRM2 to recall authentic ClinVar SNPs might be higher than the calculated 97.54% (Figure 6A). These estimates are supported by the zero FNR and 100% sensitivity for our predicted CRMs to recall VISTA enhancers (Figure 6A) and a simulation to be described.

The zero, very low (<1.03%) and low (11.23%) FDRs of VISTA enhancers, ClinVar SNPs and FPs datasets, respectively, are clearly related to the high reliability of the experimental methods used to characterize them. However, the low FDRs might also be related to the highly conserved nature of these elements (Figure 6B), as their critical functions and large effect sizes may facilitate their correct characterization. In this regarding, we note that the intermediately high FDRs of the FEs(18.10%), DHSs(25.32), TASs (15.68%), H3K4m3 (13.04%), H3K4m1 (23.23%) and H3K27ac (17.04%) datasets might be due to the facts that bidirectional transcription(Young, Kumar, Bickmore, & Taylor, 2017), CA(Arbel et

al., 2019; Chereji, Eriksson, Ocampo, Prajapati, & Clark, 2019; Dogan et al., 2015) and histone marks(Arbel et al., 2019; Dogan et al., 2015) are not unique to active enhancers. The very high FDR of GWAS SNPs (35.5%) might be due to the fact that a lead SNP associated with a trait may not necessarily be located in a CRM and causal; rather, some variants in a CRM, which are in LD with the lead SNP, are the culprits(Buniello et al., 2019; MacArthur et al., 2017). Example of GWAS SNPs in LD with positions in a CRM are shown in Figures S4K and S4L. Interestingly, many recalled ClinVar SNPs (42.59%) and GWAS SNPs (38.18%) are located in critical positions in predicted binding sites of the UMs (e.g., Figures S4D and S4F).

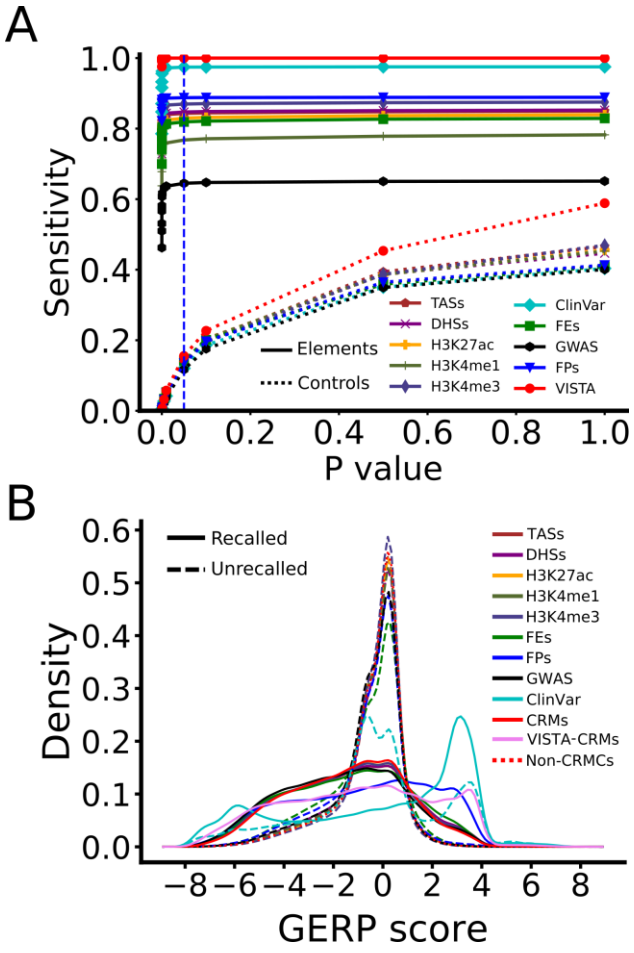


Figure 6. Validation of the predicted CRMs by 10 experimentally determined sequence elements datasets. A. Sensitivity (recall rate or TPR ) of the predicted CRMs and control sequences as a function of p-value cutoff for recalling the elements in the datasets. The dotted vertical lines indicate the p-value ≤0.05 cutoff. B. Distributions of the GERP scores of the recalled and unrecalled elements in each dataset in comparison with those of the predicted CRMs at p≤0.05 and non-CRMCs. Note that there is no unrecalled VISTA enhancers, and the distribution (not

shown) of recalled 785 VISTA enhancers in the covered genome regions is almost identical to the entire set of 976 VISTA enhancers (Figure 5A). The curve labeled by VISTA-CRMs is the distribution of CRMs that overlap and recall the 765 VISTA enhancers.

**dePCRM2 largely correctly predicts the lengths of CRMs**

To further evaluate the accuracy of dePCRM2 for predicting the lengths of CRMs, we compared the distributions of the lengths of the 785 recalled VISTA enhancers and the 26,233 recalled FEs with those of the 836 and 22,235 recalling CRMs (p-value < 0.05), respectively. As shown in Figure 7A, the recalling CRMs have a largely similar length distribution to the recalled VISTA enhancers, although the former have longer median length (3,799bp) than the latter (1,613bp). However, it is unclear whether or not the VISTA enhancers are in full-length, as a portion of an enhancer could be still partially functional(Davidson, 2006). On the other hand, a few (7.92%) VISTA enhancers are recalled by multiple short CRMCs (Figure 7A), suggesting that a small portion of our predicted CRMCs are not in full-length, but only components of long CRMCs as we pointed out earlier. It has been shown that development-related enhancers that most VISTA enhancers belong to tend to be longer than other types of enhancers(L. Li & Wunderlich, 2017) . In agreement with this, as mentioned earlier, all VISTA enhancers are recalled at a rather low p-value cutoff $5x10^{-5}$ ($\alpha$ = 412) (Figure 6A) that filters out most short CRMCs (Figure 4E). At even a lower p-value cutoff $5x10^{-6}$ ($\alpha$ = 676) the resulting 428,628 predicted CRMs with almost an identical length distribution as the entire sets of 976 VISTA enhancers (Figure 4E) still recall 99.10% of the VISTA enhancers. Taken together, these results suggest that dePCRM2 is able to largely correctly predict the length of VISTA enhancers.

By contrast, the recalled FEs at p-value <0.05 are generally shorter (median length 293bp) than the recalling CRMs (median length 2,371bp) (Figure 7B), and a considerable proportion (26.69%) of the recalled FEs overlap the same CRMs at different parts (Figures 7D and 7E). To see whether the FEs also are generally shorter than their overlapping VISTA enhancers, we identified a total of 113 FEs that

overlap 91 VISTA enhancers, indicating that overlapping rates between these two sets of elements are quite (11.59% for VISTA enhancers and 0.35% for FEs). The 113 FEs (median length 319bp) are much shorter than their 91 overlapping VISTA enhancers (median length 2,686bp)(Figure 7C), and in a few cases, multiple FEs overlap different parts of the same VISTA enhancers (Figures 7D and 7E). Thus, it is likely that the CRMs that match multiple FEs might be in full-length, while the eRNA-seq based method tends to identify short CRMs or CRM components of long CRMs, instead of full-length enhancers. Interestingly, the CRMs that recall EHs tend to be shorter than those that recall VISTA enhancers (median length 2,371bp vs 3,799pb) (Figures 7A and B), suggesting that EHs and VISTA enhancers might be different types of enhancers. This conclusion is consistent the low overlapping rates between VISTA enhancers and FEs, as well as the fact that most VISTA enhancers are development-related while most FEs are not as they were mainly determined in adult primary tissues and cell line(Andersson et al., 2014). Taken together, these results suggest that dePCRM2 is able to accurately predict the length of either short and long CRMs, although a small portion of the predicted CRMCs might be components of longer CRMs.

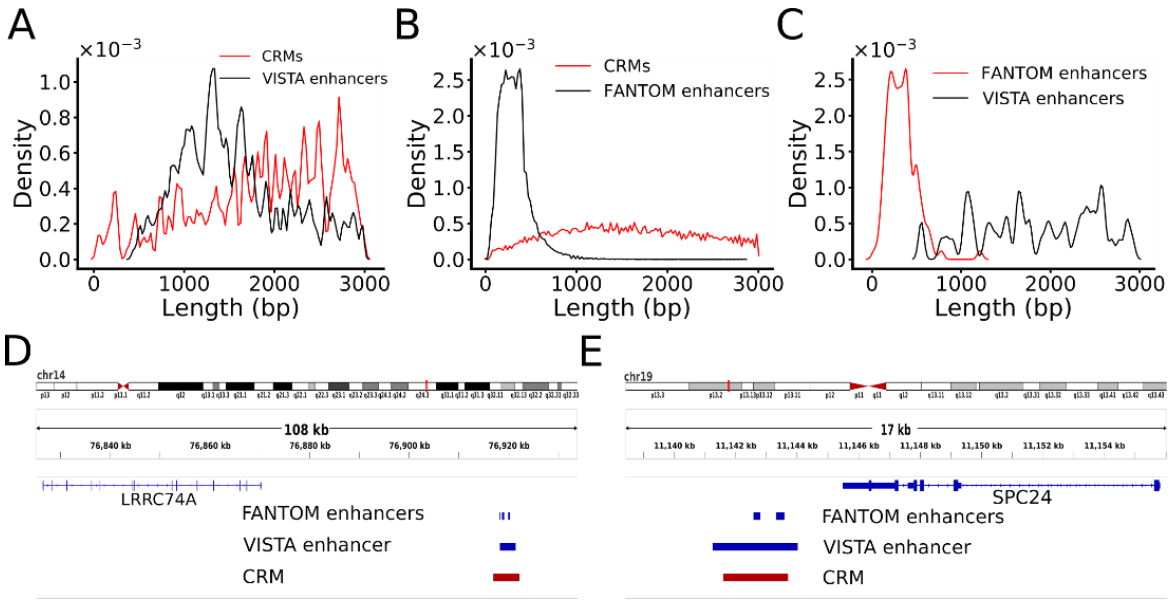


Figure 7. dePCRM2 can largely correctly predict the lengths of VISTA enhancers. A. Distributions of the lengths of the recalled VISTA enhancers and the recalling CRMs. B. Distributions of the lengths of the recalled FANTOM5 enhancers and the recalling CRMs. C. Distributions of the lengths of FANTOM5 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 methods**

We compared our predicted CRMs at p-value ≤ 0.05 ($S_{CRM}$ < 56) with three most comprehensive sets of predicted enhancers/promoters, i.e., GeneHancer 4.14(Fishilevich et al., 2017), EnhancerAtals2.0(Gao & Qian, 2020) and cCREs(Moore et al., 2020). For convivence of discussion, we call these three sets enhancers or cCREs. 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 CRMs, including ENCODE phase 2 enhancer-like regions(Bernstein et al., 2012), ENSEMBL regulatory build(Zerbino et al., 2015), dbSUPER(Khan & Zhang, 2015), EPDnew promoters(Dreos, Ambrosini, Cavin Perier, & Bucher, 2013), UCNEbase(Dimitrieva & Bucher, 2013), CraniofacialAtlas(Wilderman, VanOudenhove, Kron, Noonan, & Cotney, 2018), VISTA enhancers(Visel et al., 2007), FPs(Forrest et al., 2014) and FEs (Andersson et al., 2014). Enhancers from ENCODE phase 2 and ESEMBL were predicted based on multiple tracks of epigenetic marks using the well-regarded tools ChromHMM(Ernst & Kellis, 2012) and Segway(Hoffman et al., 2012). Of the GeneHancer enhancers, 388,407 (98.56%) have at least one nucleotide located in the covered 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, which were predicted by integrating 4,159 TF ChIP-seq, 1,580 histone mark, 1,113 DHS-seq, and 1,153 other enhancer function-related datasets, such as FEs(Gao et al., 2016). After removing redundancy (identical enhancers in difference cell/tissues), we ended up with 3,452,739 EnhancerAtlas enhancers that may still have overlaps, covering 58.99% (1,821,795,020bp) of the genome (Figure 8A), and 3,417,629 (98.98%) of which have at least one nucleotide located in the covered genome regions,

covering 58.78% (1,815,133,195bp) of the genome (Figure 8A). cCREs represents the most recent CRM prediction by the ENCODE phase 3 consortium(Moore et al., 2020), containing 926,535 non-overlapping cell type agnostic enhancers and promoters covering 8.20% (253,321,371bp) of the genome. The cCREs were predicted based on overlaps among 703 DHS, 46 TAS and 2,091 histone mark datasets in various cell/tissue types produced by ENCODE phases 2 and 3 as well as the Roadmap Epigenomics projects(Moore et al., 2020). Of these cCREs, 917,618 (99.04%) have at least one nucleotide located in the covered genome regions, covering 8.13% (251,078,466bp) of the genome (Figure 8A). Thus, due probably the aforenoted reasons, all the three sets of predicted enhancers and cCREs are strongly biased to the covered regions relative to the uncovered regions. Both the number (1,155,151) and genome coverage (43.47%) of our predicted CRMs (p-value<0.05) are larger than those of GeneHancer enhancers (388,407 and 18.89%) and of cCREs (917,618 and 8.12%), but smaller than those of EnhancerAtlas enhancers (3,417,629 and 58.78%), in the covered regions.

To make the comparisons fair, we first computed sensitivity of the three sets of enhancers and cCREs for recalling VISTA enhancers, ClinVar SNPs and GWAS SNPs in the covered regions. We omitted FPs, FEs, DHSs, TASs and the three histone marks for the valuation as they were used in predicting CRMs by GeneHancer 4.14, EnhancerAtlas 2.0 or ENCODE phase 3 consortium. We included VISTA enhancers for the evaluation as they were not included in EnhancerAtlas enhancers and cCREs, although they were a part of GeneHancer 4.14. Remarkably, our predicted CRMs outperform EnhancerAtlas enhancers for recalling VISTA enhancers (100.00% vs 94.01%) and ClinVar SNPs (97.43% vs 7.03%) (Figure 8B), even though our CRMs cover a smaller proportion of the genome (43.47% vs 58.78%, or 35.22% more) (Figure 8A), indicating that dePCRM2 has both higher sensitivity and specificity than the method behind EnhancerAtlas 2.0(Gao & Qian, 2020). However, our predicted CRMs underperform EnhancerAtlas enhancers for recalling GWAS SNPS (64.50% vs 69.36%, or 7.54% more) (Figure 8B). As we indicated 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 (Figures S4K and S4L). The higher sensitivity of EnhancerAtlas enhancers for recalling GWAS SNPs might be simply thanks to their 35.22% more coverage of the genome (58.78%) than that of our predicted CRMs (43.47%) (Figure 8A). Our predicted CRMs outperform cCREs for recalling VISTA enhancers (100% vs 85.99%), ClinVar SNPs (97.43% vs 18.28%) and GWAS SNPs (64.50% vs 15.74%) (Figure 8B). Our predicted CRMs also outperform GeneHancer enhancers for recalling ClinVar SNPs (97.43% vs 33.16%) and GWAS SNPs (64.50% vs 34.11%) (Figure 8B). However, no conclusion can be drawn from these results about the specificity of our predicted CRMs compared with GeneHancer enhancers and cCREs, because our predicted CRMs cover a higher proportion of the genome than both of them (43.47% vs 18.89% and 8.20%). On the other hand, both GeneHancer enhancers and cCREs outperform EnhancerAtlas enhancers for recalling ClinVar SNPs (33.16% and 18.28% vs 7.03%)(Figure 8B), even though they have a much smaller genome coverage than EnhancerAtlas enhancers (18.89% and 8.20% vs 58.78%) (Figure 8A), indicating that they have higher specificity than EnhancerAtlas enhancers.

As shown in Figure 8C, the intersections/overlaps between the four predicted CRMs/enhancer/cCREs sets are quite low. For instance, EnhancerAtlas enhancers, GeneHancer enhancers and cCREs share 926,396,395bp (50.85%), 414,806,711bp (70.72%), and 194,709,825bp (76.86%) of their nucleotide positions with our predicted CRMs, corresponding to 69.01%, 30.90% and 14.51% of the positions of our CRMs (Figure 8C), respectively. There are only 105,606,214pb shared by all the four sets, corresponding to 5.80%, 18.00%, 41.69% and 7.87% of nucleotide positions covered by EnhancerAtlas enhancers, GeneHancer enhancers, cCREs and our CRMs, respectively. To estimate the FDRs of EnhancerAtlas enhancers, GeneHancer enhancers and cCREs, we plotted the distributions of GERP scores of the positions that each of them share and do not share with our CRMs. As expected, the shared positions of the three predicted enhancers/cCERs sets all evolve similarly to our predicted CRMs, although those of GeneHancer enhancers and cCREs are under slightly higher evolutionary constraints

723 than our CRMs(Figure 8D). However, at a higher $S_{CRM}$ cutoff, e.g. α=3,000 ($p<2.2x10^{-302}$), our predicted
724 CRMs are even under stronger evolutionary constraints than the shared GeneHancer enhancers and
725 cCREs positions (Figures 5C and 8D). Therefore, these shared GeneHancer enhancers and cCREs
726 positions just evolve like subsets of our predicted CRMs with higher $S_{CRM}$ scores. By stark contrast, the
727 remaining 49.14%, 29.28% and 23.13% unshared positions of EnhancerAtlas enhancers, GeneHancer
728 enhancers and cCRES, respectively, evolve similarly to the non-CRMCs, although they all have slightly
729 smaller proportion of neutrality than that of the non-CRMCs (0.66, 0.63 and 0.61 vs. 0.71, respectively)
730 (Figure 8D), due probable to the small FNR (<1.54%) of our predicted CRMs. Nonetheless, these results
731 suggest that the vast majority of the unshared positions of the three sets of predicted enhancers/eCREs
732 are selectively neutral, and thus might be nonfunctional. Therefore, it appears that the predicted
733 enhancers/cCREs in the three sets that overlap our CRMs are likely to be authentic, while most of those
734 that do not might be false positives. Thus, we estimate the FDR of EnhancerAtlas enhancers,
735 GeneHancer enhancers and cCRES might be around 49.14%, 29.28% and 23.13%, respectively.
736 Therefore, it is highly likely that GeneHancer 4.14 and cCREs might largely under-predicted enhancers as
737 evidenced the fact that they are targeted evolutionarily more constrained elements (Figure 8D), even
738 though they have rather high FDRs around 29.28% and 23.12%, respectively (Figure 8D), while
739 EnhancerAtlas 2.0 might largely over-predicted enhancers with a very high FPR around 49.14% (Figure
740 8D).
741 Finally, we also compared the lengths of the four sets of predicted CRMs/enhancers/cCREs with
742 those of VISTA enhancers . As shown in Figure 8E, the distribution of the lengths of cCREs has a narrow
743 high peak at 345bp with a mean length of 273bp and a maximal length of 350bp. It is highly likely that
744 the vast majority of authentic cCREs are just components of long CRMs, because even the shortest
745 known enhancer in VISTA is 428bp long. The highly uniform lengths of the predicted cCREs indicate the
746 limitation of the underlying prediction pipeline. The distribution of GeneHancer enhancers oscillates

with a period of 166bp (Figure 8E), which might be an artifact of the underlying algorithm. Moreover, with a mean length of 1,488bp, GeneHancer enhancers are shorter than the VISTA enhancers (with mean length 2,049pb) (Figure 8E). EnhancerAtlas enhancers also have a shorter mean length (680bp) than VISTA enhancers (Figure 8E). Our predicted CRMs at p-value <0.05 have a mean length of 1,162pb, thus also are shorter than that of VISTA enhancers (Figure 8E). However, as we indicated earlier, with a higher p-value cutoff $5x10^{-6}$, the resulting 428,628 predicted CRMs have almost an identical length distribution as the VISTA enhancers (Figure 4E). Taken together, these results unequivocally indicate that our predicted CRMs are much more accurate than the three state-of-the-art predicted enhancer/cCRE sets for both the positions and lengths of CRMs/enhancers/cCREs.

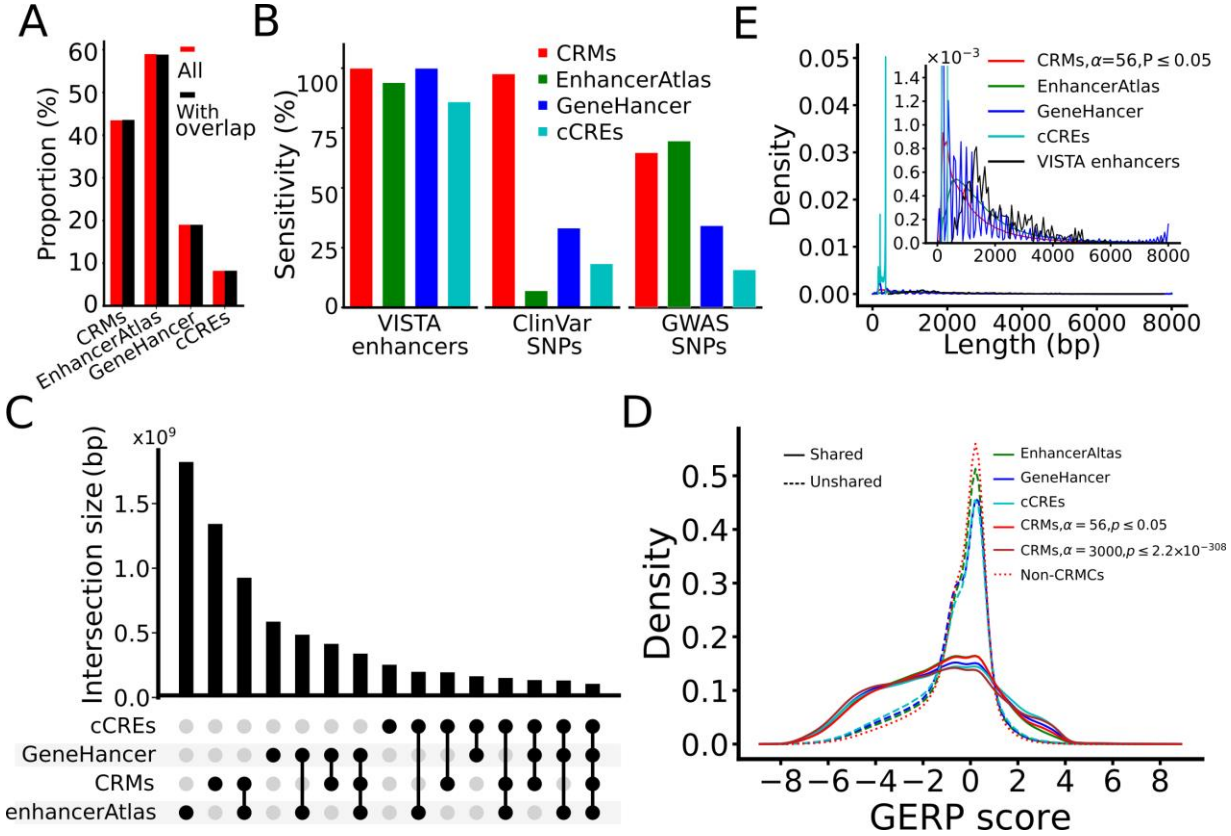


Figure 8. Comparison of the performance of dePCRM2 and three state-of-the-art methods. A. Proportion of the genome that are covered by enhancers/CRMs predicted by the four methods, and proportion of genome regions covered by predicted enhancers/CRMs that at least partially overlap the covered genome regions. B. Recall rates for recalling VISTA enhancers, ClinVar SNPs and GWAS SNPs, by the predicted enhancers/CRMs that at least partially overlap the covered genome regions. C. Upset plot showing numbers of nucleotide positions shared among the predicted CRMs, GeneHancer enhancers, EnhancerAtlas enhancers and cCREs. D. Distributions of GERP

scores of positions of the CRMs predicted at p-value $\leq$ 0.05 and p-value $\leq$ 2.2X10$^{-308}$, and the non-CRMCs, as well as of positions that GeneHancer enhancers, EnhancerAtlas enhancers and cCREs share or do not share with the predicted CRMs at p-value $\leq$ 0.05. E. Distributions of lengths of the four sets of predicted enhancers/CRMs in comparison the that of the VISTA enhancers. The inset is a blow-up view of the indicated region on the Density axis.

**At least half of the human genome might code for CRMs**

What is the proportion of the human genome coding for CRMs and TFBSs? The high accuracy of our predicted CRMs and constituent TFBSs might well position us to more accurately address this interesting and important, yet unanswered question(Kellis et al., 2014; Pennacchio, Bickmore, Dean, Nobrega, & Bejerano, 2013). To this end, we took a semi-theoretic approach. Specifically, we calculated the expected number of true positives and false positives in the CRMCs in each non-overlapping $S_{CRM}$ score interval based on the predicted number of CRMCs and the density of $S_{CRM}$ scores of Null CRMCs in the interval (Figure 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 a low $S_{CRM}$ score < 4 (inset in Figure 9A) with a mean length of 28 pb (Figure 4D), making up 0.02% (21,821x28/3,088,269,832) of the genome and 0.05% (0.02%/0.4403) of the total length of the CRMCs, i.e., a FDR of 0.05% for nucleotide positions (Figure 9C). On the other hand, as the CRMCs miss 2.49% of ClinVar SNPs in the covered genome regions (Figure 6A), the FNR of partitioning the genome in CRMCs and non-CRMCs would be < 2.49%(1-0.40)=1.49%, given the proportion of neutrality of 0.4 for the unrecalled ClinVar SNPs (Figure 6B). We therefore estimate false negative CRMCs make up 0.67% of the genome and 1.99% of the total length of the non-CRMCs, i.e., a false omission rate (FOR) of 1.99% for nucleotide positions (Figure 9C). Hence, the true CRM positions in the covered regions make up 44.68% (44.03%-0.02%+0.67%) of the genome (Figure 9C). In addition, as we argued earlier, the uncovered 22.53% genome regions have 79.4% CRMC density as in the covered regions, thus, CRMCs in the uncovered regions would be about 10.32% (0.2253 x 0.4468x0.7940/0.7747) of the genome (Figure 9C). Taken together, we estimate that about 55.00% (44.68%+10.32%) of the genome might code for CRMs,

for which we have predicted 80.02% [(44.03-0.02)/55.00]. Moreover, as we predict that about 40% of CRCs are made up of TFBSs (Figure 4C), we estimate that about 22.00% of the genome might encode TFBSs. Furthermore, assuming the mean length of CRMs is 1,162pb as our predicted CRMs at p-value <0.05, we estimate that the human genome encodes about 1,461,746 CRMs (3,088,269,832x0.55/1,162) and 67,941,936 TFBSs.

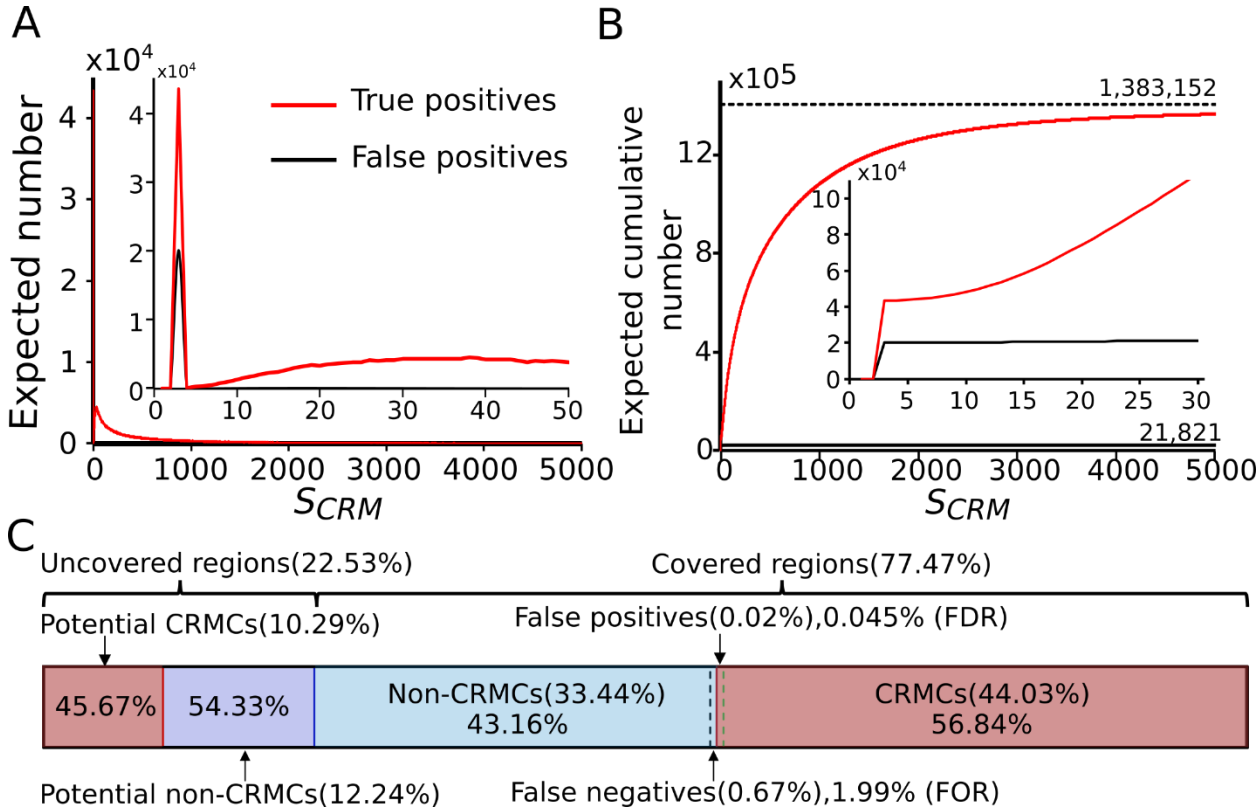


Figure 9. Estimation of the portion of the human genome encoding CRMs. A. Expected number of true positive and false positive CRMCs in the predicted CRMCs in each one-unit interval of the $S_{CRM}$ score. The inset is a blow-up view of the axes defined region. B. Expected cumulative number of true positives 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, and enormous CRM function-related data have been produced(Gasperini et

al., 2020; Snyder et al., 2020). Although great progresses have been made to predict CRMs in the genomes (Fishilevich et al., 2017; Gao & Qian, 2020; Moore et al., 2020; X. Wang et al., 2018; Zerbino et al., 2015) using these data[26, 53, 55, 59, 11], most existing methods attempt to predict cell/tissue specific CRMs using CA and multiple tracks of histone marks collected in the same cell /tissue types(Ernst & Kellis, 2012; Fishilevich et al., 2017; Gao & Qian, 2020; Hoffman et al., 2012; Moore et al., 2020). These methods are limited by the scope of applications(Ernst & Kellis, 2012; Hoffman et al., 2012; Moore et al., 2020), low resolution of predicted CRMs(Gao & Qian, 2020; Moore et al., 2020), lack of constituent TFBS information(Gao & Qian, 2020; Moore et al., 2020), and high FDRs(Dogan et al., 2015)(Figure 8D). To overcome these limitations, we proposed a different approach to first predict a cell type agnostic or static map of CRMs and constituent TFBBs in the genome(Niu et al., 2018; Niu et al., 2014) by identifying repeatedly cooccurring patterns of motifs found in appropriately expanded binding peaks in a large number of TF ChIP-seq datasets for different TFs in various cell/tissue types. As it is mainly TFBSs in a CRM that define its structure and function, it not surprising that TF ChIP-seq data are more accurate predictor of CRMs than CA and histone mark data(Arbel et al., 2019; Dogan et al., 2015; Kleftogiannis et al., 2015). Therefore, our approach might hold promise for more accurate predictions of CRMs and constituent TFBSs, notwithstanding computational challenges. Once a map of CRMs and constituent TFBBs in the gnome is available, functions of CRMs and constituent TFBSs in cell/tissue types could be studied in a more focused and cost-effective ways. Another advantage of our approach is that we do not need to exhaust all TFs and all cell/tissue types of the organism in order to predict most, if not all, of CRMs and constituent TFBBSs in the genome(Niu et al., 2018; Niu et al., 2014), because CRMs are often repeatedly used in different cell/tissue types, developmental stages and physiological homeostasis(Davidson, 2006). Moreover, by appropriately extending the binding peaks in each dataset, we can largely increase the chance of including and identifying cooperative motifs(Y. Li et al., 2019) and full-length CRMs(Niu et al., 2018; Niu et al., 2014), thereby increasing the power of existing data and

further reducing the number of datasets needed as we show in this study. In other word, we might only need a large but limited number of datasets to predict most, if not all, CRMs and TFBSs in the genome(Niu et al., 2018; Niu et al., 2014). Our earlier application of the approach resulted in very promising results in the fly(Niu et al., 2014) and human(Niu et al., 2018) genomes even using a relatively small number of strongly biased datasets available then. However, the earlier implementations were limited by computational inadequacies of underlying algorithms to find and integrate motifs in ever increasing number of large TF ChIP-seq datasets in mammalian cell/tissues(Niu et al., 2018; Niu et al., 2014). In this study, we circumvent the limitations by developing the new pipeline dePCRM2 based on an ultrafast and accurate motif finder ProSampler, an efficient motif pattern integration method, and a novel CRM scoring function that capture essential features of full-length CRMs.

Remarkably, dePCRM2 enables us to partition the 77.47% genome regions covered the extended binding peaks in ~6,000 TF ChIP-seq datasets into two exclusive sets, i.e., the CRMCs, which are a part of the *cis*-regulatory genome; and the non-CRMCs, which are a part of the non-*cis*-regulatory genome. Multiple pieces of evidence strongly suggest that the partition might be highly accurate. First, the vast majority of the CRMCs have a small p-values(Figure 4B), and thus, are unlikely predicted by chance. Second, even the subset of the CRMCs with the lowest $S_{CRM}$ scores ((0,1]) are under stronger evolutionary constraints than the non-CRMCs (Figures 5B and S3B), indicating that even these low-scoring CRMCs are more likely to be functional than non-CRMCs, not to mention CRMCs with higher $S_{CRM}$ scores that are under stronger evolutionary constraints. Third, the vast majority of the CRMCs are under similarly strong evolutionary constraints, and those with a high $S_{CRM}$ score evolve like the ultra-conserved, development-related VISTA enhancers (Figures 5A, 5C, S3A and S3C). Fourth, all experimentally validated VISTA enhancers and almost all (97.51%) of well-documented ClinVar SNPs in the covered regions are recovered by the CRMCs (Figure 6), indicating that the CRMCs have a very low

FDR. Finally, consistent with these findings, our simulation study also indicates that the CRMCs have a very low FDR of 0.045%, and the non-CRMCs have a low FOR of 1.99% (Figure 9C), strongly suggesting that both sensitivity and specificity of the predictions are very high. To the best of our knowledge, we are the first to accurately partition large regions of the genome (77.47**%)** into a set (CRMCs) with *cis*-regulatory function, and a set (non-CRMCs) without *cis*-regulatory function.

Accurate prediction of the length of CRMs is critical, but it appeared to be a difficult problem as evidenced by the peculiar distributions of the lengths of GeneHancer enhancers and cCREs (Figure 8E). Nevertheless, dePCRM2 also is able to largely correctly predict the lengths of most CRMCs, even though the TF ChIP-seq data used are still insufficient. First, the 783,132 CRMCs (55.74%) with a total length of 92.58% of the total length of the 1,404,973 CRMCs are longer than the shortest (428bp) known enhancer (428bp) in the VISTA database, suggesting that at least most of them might be full-length CRMs. The remaining 621,841 (44.26%) CRMCs with a total length of only 7.42% of the total length of the CRMCs are shorter than 428bp, suggesting that they might be either short CRMs or component of long CRMs. Therefore, it is highly likely that the vast majority of the CRMC positions might be in a predicted full-length CRMs. Second, with the increase in $S_{CRM}$ score cutoff α (or decrease in p-value), lengths of the predicted CRMs increase (Figure 4E). In particular, at α=676 (p-value ≤ $5x10^{-6}$), the resulting 428,628 predicted CRMs have an almost identical length distribution to that of the VISTA enhancers (Figure 4E), while these CRMs recall 99.10% of the VISTA enhancers in the coverage regions. 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 (mean length 981bp) predicted using the much larger and more diverse datasets albeit still strongly biased to a

882 few TFs and cell/tissue types (Figures 2A and 2B). We anticipate that full-length CRMs of these short
883 CRM components can be predicted using even larger and more diverse TF ChIP-seq data. Thus, efforts
884 should be made in the future to increase the genome coverage and reduce data biases by including
885 more untested TFs and untested cell types in the TF ChIP-seq data generation.
886
887 Interestingly, our predicted CRMs (at p-value < 0.05) achieve perfect (100.00%) and very high
888 (97.43) sensitivity for recalling VISTA enhancers (Visel et al., 2007) and ClinVAR SNPs (Landrum et al.,
889 2018), respectively, but varying intermediate sensitivity ranging from 64.50% (for GWAS SNPs) to
890 88.77% (for FPs) for recalling other CRM function-related elements in the remaining eight datasets
891 (Figure 6A). It appears that such varying sensitivity is due to varying FDRs ranging from 0% (for VISTA
892 enhancers) to 35.5% (for GWAS SNPs) of the methods used to characterize the elements as CRMs or
893 parts of CRMs (Figure 6B). Our finding that DHSs, TASs, and histone mark (H3K4m1, H3K4m3 and
894 H3K27ac) peaks have high FDRs for predicting CRMs are consistent with an earlier study showing that
895 histone marks or CA were less accurate predictor of active enhancers than TF binding data(Dogan et al.,
896 2015). In this sense, it is not surprising that our predicted CRMs substantially outperforms the three
897 state-of-the-art sets of predicted enhancers/cCREs, i.e., GeneHancer 4.14 (Fishilevich et al., 2017),
898 EnhancerAtals2.0 (Gao & Qian, 2020) and cCREs(Moore et al., 2020), both for recalling VISTA enhancers
899 and ClinVar SNPs (Figure 8B) and for predicting the lengths of CRMs (Figure 8E), because these three
900 sets were mainly predicted based on overlaps between multiple tracks of CA and histone marks in
901 various cell/tissue type. Although great efforts have been made to improve the accuracy of
902 EnhancerAtlas 2.0 enhancers(Gao & Qian, 2020), GeneHancer 4.14 enhancers(Fishilevich et al., 2017)
903 and cCREs(Moore et al., 2020), they still suffer quite high FDRs (49.14%, 29.28% and 23.12%,
904 respectively).
905 Although dePCRM2 can predict the functional states of CRMs in a cell/tissue type that have

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 available binding peaks of all TFs tested in the cell/tissue type. However, the functional state of each CRM in the map in any cell/tissue type could be predicted based on overlap between the CRM and a single or few epigenetic mark datasets collected from the very cell/tissue type, such as CA, H3K27ac and/or H3K4m3 data. Anchored by correctly predicted CRMs, these epigenetic marks could accurately predict the functional states of the CRMs(Dogan et al., 2015). 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.

Although originally called binding peaks in the strongly biased 6,090 TF ChIP-seq datasets cover only 40.96% of the genome, moderately extended peaks increase genome coverage to 77.47%, an 89.14 % increase. Remarkably, positions of the extended parts of the peaks consisting of 43.17% of positions of the extended peaks contribute 42.12% positions of the predicted CRMCs. Therefore, appropriate extension of called binding peaks in the datasets indeed can substantially increase the power of available data. On the other hand, we abandoned 38.60% of positions covered by the original binding peaks, which might be nonfunctional as they evolve like the non-CRMCs (Figures 5A and S3A). Therefore, called binding peaks cannot be equivalent to CRMs or parts of CRMs as has been indicated earlier(Devailly et al., 2015; Marinov et al., 2014; Mendoza-Parra et al., 2013).

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 (Gasperini et al., 2020; D. C. King et al., 2005; Pennacchio et al., 2013; Stamatoyannopoulos, 2012; Thurman et al., 2012). The major disagreement is for the proportion of functional NCSs in the genome, mainly CRMs, which has been

930 coarsely estimated to be from 8% to 40% of the genome (Bernstein et al., 2012; Stamatoyannopoulos,
931 2012). Moreover, a wide range of CRM numbers from 400,000 (Bernstein et al., 2012) to more than a
932 few million (Gao & Qian, 2020; Gasperini et al., 2020) encoded in the human genome have been
933 suggested. However, to our knowledge, no estimate has been made on substantial evidence. Our
934 predicted CRMCs cover a lower proportion (44.03%) of the genome than EnhancerAtlas 2.0 enhancers
935 (58.99%)(Gao & Qian, 2020) that might have a FDR of 49.14%, but a higher proportion than cCREs
936 (7.9%)(Moore et al., 2020) and GeneHancer v.4.14 enhancers (18.99%) (Fishilevich et al., 2017), even
937 though both sets have high FDRs (Figure 7D). The much higher accuracy of our predicted CRMs suggests
938 that cCREs (7.9%)(Moore et al., 2020) and GeneHancer enhancer might underpredict, whereas
939 EnhancerAtlas 2.0 might overpredict CRMs. Based on the estimated FDR and FNR of the predicted
940 CRMCs and non-CRMCs as well as the estimated density of CRMs in the uncovered regions relative to
941 the covered regions (Figure 9C), we estimate that about 55.00% and 22.00% of the genome might code
942 for CRMs and TFBSs, respectively, which encode about 1.5 million CRMs and 68 million TFBSs.
943 Therefore, the number of our predicted CRMs is almost four times more than an earlier estimate of
944 400,000 (Bernstein et al., 2012), and they are encoded by a higher (55.00%) proportion of the genome
945 than earlier thought 40%(Bernstein et al., 2012; Stamatoyannopoulos, 2012). We estimate our true
946 positive CRMs cover 44.01% (44.03-0.02) of the genome, therefore, we might have predicted 80.02 %
947 (44.01/55.00) CRM positions encoded in the genome. In summary, it appears that the cis-regulatory
948 genome is more prevalent than originally thought.

949 **Conclusions**

950 We have developed a new highly accurate and scalable pipeline dePCRM2 for predicting CRMs
951 and constituent TFBSs in large genomes by integrating a large number of TF ChIP-seq datasets for
952 various TFs in a variety of cell/tissue types of the organisms. Applying dePCRM2 to all available ~6,000
953 TF ChIP-seq datasets, we predicted an unprecedentedly complete, high resolution map of CRMs and

constituent TFBSs in 77.47% of the human genome covered by the extended binding peaks of the datasets. Evolutionary and experimental data suggest that dePCRM2 achieves very high prediction sensitivity and specificity. Based on the predictions, we estimate that about 55.00% and 22.00% of the genome might code for CRMs and TFBSs, encoding about 1.5 million CRMs and 68 million TFBSs, respectively; for both of which we predicted about 80%. Therefore, the *cis*-regulatory genome is more prevalent than originally thought. With the availability in the future of more diverse and balanced data covering more regions of the genome, it is possible to predict a more complete map of CRMs and constituent TFBSs in the genome.

## 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 genomesequence 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), 424,622 ClinVar SNPs from the ClinVar database(Landrum et al., 2018), 32,689 enhancers(Andersson et al., 2014) and 184,424 promoters(Forrest et al., 2014) from the FANTOM5 project website, 91,369 GWAS SNPs from GWAS Catalog(Buniello et al., 2019), and 122,468,173 DHSs in 1,353 datasets (Table S2), 29,520,736 transposase-accessible sites (TASs) in 1,059 datasets (Table S3), 99,974,447 H3K27ac peaks in 2,539 datasets (Table S4), 77,500,232 H3K4me1 peaks in 1,210 datasets (Table S5), and 70,591,888 H3K4me3 peaks in 2,317 datasets (Table S6) from the Cistrome database(Mei et al., 2017).

### 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,

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

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

**Parameters for accuracy evaluation**

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

**The dePCRM2 pipeline**

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

**Step 2.** Compute pairwise motif co-occurring scores and find co-occurring motif pairs (CPs): 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 CPs in each dataset (Figure 1C). Specifically, for each pair of motifs $M_d(i)$ and $M_d(j)$ in each data set *d*, we compute their co-occurring scores $S_c$ defined as,

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

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 $o(M_d(i), M_d(j))$ the number of binding peaks containing TFBSs of both the motifs in $d$. We identify CPs with an $S_c \geq \beta$. We choose $\beta$ such that the component with the highest scores in the trimodal distribution $S_c$ is kept (Figures 1C and 3B) (by default $\beta = 0.7$).

**Step 3.** Construct a motif similarity graph and find unique motifs (UMs): We combine highly similar motifs in the CPs from different datasets to form a 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 $M_a(i)$ and $M_b(j)$ from different datasets $a$ and $b$, respectively, we compute their similarity score $S_s$ using our SPIC(S. Zhang, Jiang, Du, & Su, 2013) 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, $\beta$ =0.8, Figure 1D). We apply the Markov cluster (MCL) algorithm (van Dongen & 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(Y. Li et al., 2019)(Figure 1D).

**Step 4.** Construct the integration networks of the UMs/TFs: 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 & Kulkarni, 2005; Vockley, Barrera, & Reddy, 2017; Yanez-Cuna, Kvon, & Stark, 2013), 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, $U_i$ and $U_j$, defined as,

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

where $D(U_i, U_j)$ is the datasets in which TFBSs of both $U_i$ and $U_j$ occur, $d(U_k)$ the subset of dataset $d$, containing at least one TFBS of $U_k$, $S(d(U_i), (d(U_j))$ the subset of $d$ containing TFBSs of both $U_i$ and $U_j$, and $r(s)$ the shortest distance between any TFBS of $U_i$ and any TFBS of $U_j$ in a sequence $s$. We construct UM/TF interaction networks using the UMs as nodes and connecting two nodes with their $S_{\mathrm{INTER}}$ being the weight on the edge (Figure 1E). Therefore, $S_{\mathrm{INTER}}$ 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, particularly within a nucleosome with a length of about 150bp(Arnosti & Kulkarni, 2005; Vockley, McDowell, D'Ippolito, & Reddy, 2017; Yanez-Cuna et al., 2013)**.**

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

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

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

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 7**. 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(Y. 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 an empirical p-value for each CRMC, and predict those with a p-value smaller than a preset cutoff as CRMs in the genome (Figure 1G).

**Step 8.** Prediction of the functional states of CRMs in a given cell type: For each predicted CRM at p-value <0.05, 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).

**Generation of control sequences for validation**

To create a set of matched control sequences for validating the predicted CRMs using experimentally determined elements (Figure 6A), for each predicted CRMC, we produced a control sequence by randomly selecting a sequence segment with the same length as the CRMC from the genome regions covered by the extended binding peaks. To calculate the $S_{CRM}$ score of a control sequence, we assigned it the TFBS positions and their UMs according to those in the counterpart CRMC. Thus the control set contains the same number and length of sequences as in the CRMCs, but with arbitrarily assigned TFBSs and UMs.

**Authors’ contributions**

ZS conceived the project. ZS and PN developed the algorithms and PN carried out all computational experiments and analysis. ZS and PN wrote the manuscripts. All authors read and approved the final manuscript.

**Funding**

The work was supported by US National Science Foundation (DBI- 1661332). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Competing interests**

The authors declare no competing financial interests.

**REFERENCES**

Aday, A. W., Zhu, L. J., Lakshmanan, A., Wang, J., & Lawson, N. D. (2011). Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. *Dev Biol, 357*(2), 450-462. doi:10.1016/j.ydbio.2011.03.007

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat Rev Genet, 16*(4), 197-212. doi:10.1038/nrg3891

Allen, J. E., Pertea, M., & Salzberg, S. L. (2004). Computational gene prediction using multiple sources of evidence. *Genome Res, 14*(1), 142-148. doi:10.1101/gr.1562804 [doi]

14/1/142 [pii]

Ambrosini, G., Vorontsov, I., Penzar, D., Groux, R., Fornes, O., Nikolaeva, D. D., . . . Bucher, P. (2020). Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol, 21*(1), 114. doi:10.1186/s13059-020-01996-3

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., . . . Makeev, V. J. (2014). An atlas of active enhancers across human cell types and tissues. *Nature, 507*(7493), 455-461. doi:nature12787 [pii]

10.1038/nature12787 [doi]

Arbel, H., Basu, S., Fisher, W. W., Hammonds, A. S., Wan, K. H., Park, S., . . . Brown, J. B. (2019). Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proc Natl Acad Sci U S A, 116*(3), 900-908. doi:10.1073/pnas.1808833115

Arnosti, D. N., & Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem., 94*(5), 890-898. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/jcb.20352/abstract

Ashoor, H., Kleftogiannis, D., Radovanovic, A., & Bajic, V. B. (2015). DENdb: database of integrated human enhancers. *Database (Oxford), 2015*. doi:10.1093/database/bav085

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics, 27*(12), 1653-1659. doi:btr261 [pii]

10.1093/bioinformatics/btr261 [doi]

Bailey, T. L., & 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, T. L., & Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res*. doi:gks433 [pii] 10.1093/nar/gks433 [doi]

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi:S0092-8674(07)00600-9 [pii]

10.1016/j.cell.2007.05.009 [doi]

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science, 304*(5675), 1321-1325. doi:10.1126/science.1098119

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:nature11247 [pii]

10.1038/nature11247 [doi]

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi:10.1038/nmeth.2688

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res, 47*(D1), D1005-D1012. doi:10.1093/nar/gky1120

Catarino, R. R., & Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev, 32*(3-4), 202-223. doi:10.1101/gad.310367.117

Chen, C., Zhou, D., Gu, Y., Wang, C., Zhang, M., Lin, X., . . . Zhang, Y. (2020). SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. *Nucleic Acids Res, 48*(D1), D198-d203. doi:10.1093/nar/gkz1028

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*(D1), D267-d275. doi:10.1093/nar/gkx1092

Chereji, R. V., Eriksson, P. R., Ocampo, J., Prajapati, H. K., & Clark, D. J. (2019). Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome Res, 29*(12), 1985-1995. doi:10.1101/gr.249326.119

Consortium, G. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science, 348*(6235), 648-660. doi:10.1126/science.1262110

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., & Nickerson, D. A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods., 7*(4), 250-251. doi: 210.1038/nmeth0410-1250.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A, 107*(50), 21931-21936. doi:1016071107 [pii]

10.1073/pnas.1016071107 [doi]

Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*: Academic Press.

Devailly, G., Mantsoki, A., Michoel, T., & Joshi, A. (2015). Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. *FEBS Lett, 589*(24 Pt B), 3866-3870. doi:10.1016/j.febslet.2015.11.027

Dimitrieva, S., & Bucher, P. (2013). UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res, 41*(Database issue), D101-109. doi:10.1093/nar/gks1092

Dogan, N., Wu, W., Morrissey, C. S., Chen, K. B., Stonestrom, A., Long, M., . . . Hardison, R. C. (2015). Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin, 8*, 16. doi:10.1186/s13072-015-0009-5

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*(Database issue), D157-164. doi:10.1093/nar/gks1233

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods, 9*(3), 215-216. doi:10.1038/nmeth.1906

Firpi, H. A., Ucar, D., & Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics, 26*(13), 1579-1586. doi:btq248 [pii]

10.1093/bioinformatics/btq248 [doi]

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., . . . Cohen, D. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford), 2017*. doi:10.1093/database/bax028

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V., . . . Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature, 507*(7493), 462-470. doi:10.1038/nature13182

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*(23), 3543-3551. doi:10.1093/bioinformatics/btw495

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*(D1), D58-d64. doi:10.1093/nar/gkz980

Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet, 21*(5), 292-310. doi:10.1038/s41576-019-0209-0

Goi, C., Little, P., & Xie, C. (2013). Cell-type and transcription factor specific enrichment of transcriptional cofactor motifs in ENCODE ChIP-seq data. *BMC Genomics, 14 Suppl 5*, S2. doi:1471-2164-14-S5-S2 [pii]

10.1186/1471-2164-14-S5-S2 [doi]

Hartmann, H., Guthohrlein, E. W., Siebert, M., Luehr, S., & Soding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res, 23*(1), 181-194. doi:gr.139881.112 [pii]

10.1101/gr.139881.112 [doi]

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell, 38*(4), 576-589. doi:10.1016/j.molcel.2010.05.004

Herz, H. M., Hu, D., & Shilatifard, A. (2014). Enhancer malfunction in cancer. *Mol Cell, 53*(6), 859-866. doi:10.1016/j.molcel.2014.02.033

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367. doi:0903103106 [pii]

10.1073/pnas.0903103106 [doi]

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods, 9*(5), 473-476. doi:nmeth.1937 [pii]

10.1038/nmeth.1937 [doi]

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., . . . Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res, 41*(2), 827-841. doi:gks1284 [pii]

10.1093/nar/gks1284 [doi]

Jiang, P., & Singh, M. (2014). CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Res, 42*(5), 2833-2847. doi:10.1093/nar/gkt1302

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502.

Kang, R., Zhang, Y., Huang, Q., Meng, J., Ding, R., Chang, Y., . . . Guo, Z. (2019). EnhancerDB: a resource of transcriptional regulation in the context of enhancers. *Database (Oxford), 2019*. doi:10.1093/database/bay141

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., . . . Snyder, M. (2013). Extensive variation in chromatin states across humans. *Science, 342*(6159), 750-752. doi:10.1126/science.1242510

Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., . . . Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science, 317*(5840), 915. doi:10.1126/science.1142430

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., . . . Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A., 111*(17), 6131-6138. doi: 6110.1073/pnas.1318948111. Epub 1318942014 Apr 1318948121.

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*(5), 2976-2987. doi:10.1093/nar/gkt1249

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat Rev Genet, 17*(2), 93-108. doi:10.1038/nrg.2015.17

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., . . . Dermitzakis, E. T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science, 342*(6159), 744-747. doi:science.1242463 [pii]

10.1126/science.1242463 [doi]

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., . . . Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell, 128*(6), 1231-1245. doi:S0092-8674(07)00205-X [pii]

10.1016/j.cell.2006.12.048 [doi]

King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., & Hardison, R. C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res, 15*(8), 1051-1060. doi:10.1101/gr.3642605

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

Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2015). DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res, 43*(1), e6. doi:10.1093/nar/gku1058

Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2016). Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform, 17*(6), 967-979. doi:10.1093/bib/bbv101

Koyabu, Y., Nakata, K., Mizugishi, K., Aruga, J., & Mikoshiba, K. (2001). Physical and functional interactions between Zic and Gli proteins. *J Biol Chem, 276*(10), 6889-6892. doi:10.1074/jbc.C000773200

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., . . . Makeev, V. J. (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*(D1), D252-d259. doi:10.1093/nar/gkx1106

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature., 518*(7539), 317-330. doi: 310.1038/nature14248.

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Res, 24*(10), 1595-1602. doi:10.1101/gr.173518.114

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-

beta and wnt pathways. *Proc Natl Acad Sci U S A, 97*(15), 8358-8363. doi:10.1073/pnas.150152697

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., . . . Weirauch, M. T. (2018). The Human Transcription Factors. *Cell, 175*(2), 598-599. doi:10.1016/j.cell.2018.09.045

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res, 46*(D1), D1062-d1067. doi:10.1093/nar/gkx1153

Li, L., & Wunderlich, Z. (2017). An Enhancer's Length and Composition Are Shaped by Its Regulatory Task. *Front Genet, 8*, 63. doi:10.3389/fgene.2017.00063

Li, X., Wang, W., Wang, J., Malovannaya, A., Xi, Y., Li, W., . . . Chen, J. (2015). Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Mol Syst Biol, 11*(1), 775. doi:10.15252/msb.20145504

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, D. L., & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-138.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res, 45*(D1), D896-D901. doi:10.1093/nar/gkw1133

Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics, 27*(12), 1696-1697. doi:btr189 [pii]

10.1093/bioinformatics/btr189 [doi]

Marinov, G. K., Kundaje, A., Park, P. J., & Wold, B. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda). 4*(2), 209-223. doi: 210.1534/g1533.1113.008680. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3931556/pdf/209.pdf

Martin, N., Benhamed, M., Nacerddine, K., Demarque, M. D., van Lohuizen, M., Dejean, A., & Bischof, O. (2012). Physical and functional interaction between PML and TBX2 in the establishment of cellular senescence. *EMBO J, 31*(1), 95-109. doi:10.1038/emboj.2011.370

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., . . . Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res, 44*(D1), D110-115. doi:10.1093/nar/gkv1176

Mathelier, A., Shi, W., & Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends Genet, 31*(2), 67-76. doi:10.1016/j.tig.2014.12.003

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science, 337*(6099), 1190-1195. doi:10.1126/science.1222794

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., . . . Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science, 342*(6159), 747-749. doi:science.1242429 [pii]

10.1126/science.1242429 [doi]

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., . . . Liu, X. S. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res, 45*(D1), D658-d662. doi:10.1093/nar/gkw983

Mendoza-Parra, M. A., Van Gool, W., Mohamed Saleem, M. A., Ceschin, D. G., & Gronemeyer, H. (2013). A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res, 41*(21), e196. doi:10.1093/nar/gkt829

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., . . . Consortium, E. P. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature, 583*(7818), 699-710. doi:10.1038/s41586-020-2493-4

Niu, M., Tabari, E., Ni, P., & Su, Z. (2018). Towards a map of cis-regulatory sequences in the human genome. *Nucleic Acids Res, 46*(11), 5395-5409. doi:10.1093/nar/gky338

Niu, M., Tabari, E. S., & 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*(1), 1047. doi: 1010.1186/1471-2164-1015-1047.

Ongen, H., Andersen, C. L., Bramsen, J. B., Oster, B., Rasmussen, M. H., Ferreira, P. G., . . . Dermitzakis, E. T. (2014). Putative cis-regulatory drivers in colorectal cancer. *Nature., 512*(7512), 87-90. doi: 10.1038/nature13602. Epub 12014 Jul 13623. Retrieved from http://www.nature.com/nature/journal/v512/n7512/pdf/nature13602.pdf

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., . . . Hermjakob, H. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res, 42*(Database issue), D358-363. doi:10.1093/nar/gkt1115

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., . . . Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res, 47*(D1), D529-D541. doi:10.1093/nar/gky1079

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet., 11*(1), e1004857. doi: 1004810.1001371/journal.pgen.1004857. eCollection 1002015 Jan.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nat Rev Genet, 14*(4), 288-295. doi:nrg3458 [pii]

10.1038/nrg3458 [doi]

Perrot, C. Y., Gilbert, C., Marsaud, V., Postigo, A., Javelaud, D., & Mauviel, A. (2013). GLI2 cooperates with ZEB1 for transcriptional repression of CDH1 expression in human melanoma cells. *Pigment Cell Melanoma Res, 26*(6), 861-873. doi:10.1111/pcmr.12149

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res, 20*(1), 110-121. doi:10.1101/gr.097857.109

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., . . . Ren, B. (2013). RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol, 9*(3), e1002968. doi:10.1371/journal.pcbi.1002968 [doi]

PCOMPBIOL-D-12-01314 [pii]

Ramos, E. M., Hoffman, D., Junkins, H. A., Maglott, D., Phan, L., Sherry, S. T., . . . Hindorff, L. A. (2014). Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet, 22*(1), 144-147. doi:ejhg201396 [pii]

10.1038/ejhg.2013.96 [doi]

Rubinstein, M., & de Souza, F. S. (2013). Evolution of transcriptional enhancers and animal diversity. *Philos Trans R Soc Lond B Biol Sci, 368*(1632), 20130017. doi:rstb.2013.0017 [pii]

10.1098/rstb.2013.0017 [doi]

Sánchez-Tilló, E., de Barrios, O., Valls, E., Darling, D. S., Castells, A., & Postigo, A. (2015). ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene, 34*(46), 5760-5770. doi:10.1038/onc.2015.352

Siepel, A., & Arbiza, L. (2014). Cis-regulatory elements and human evolution. *Curr Opin Genet Dev, 29*, 81-89. doi:10.1016/j.gde.2014.08.011

Sinha, S. (2003). Discriminative motifs. *J Comput Biol, 10*(3-4), 599-615. doi:10.1089/10665270360688219 [doi]

Smith, E., & Shilatifard, A. (2014). Enhancer biology and enhanceropathies. *Nat Struct Mol Biol., 21*(3), 210-219. doi: 210.1038/nsmb.2784. Retrieved from http://www.nature.com/nsmb/journal/v21/n3/pdf/nsmb.2784.pdf

Snetkova, V., & Skok, J. A. (2018). Enhancer talk. *Epigenomics, 10*(4), 483-498. doi:10.2217/epi-2017-0157

Snyder, M. P., Gingeras, T. R., Moore, J. E., Weng, Z., Gerstein, M. B., Ren, B., . . . Consortium, E. P. (2020). Perspectives on ENCODE. *Nature, 583*(7818), 693-698. doi:10.1038/s41586-020-2449-8

Stamatoyannopoulos, J. A. (2012). What does our genome encode? *Genome Res, 22*(9), 1602-1611. doi:22/9/1602 [pii]

10.1101/gr.146506.112 [doi]

Stergachis, A. B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S. L., . . . Stamatoyannopoulos, J. A. (2013). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell, 154*(4), 888-903. doi:S0092-8674(13)00891-X [pii]

10.1016/j.cell.2013.07.020 [doi]

Sun, H., Guns, T., Fierro, A. C., 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*(12), e90. doi:gks237 [pii]

10.1093/nar/gks237 [doi]

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., . . . Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75-82. doi:nature11232 [pii]

10.1038/nature11232 [doi]

van Dongen, S., & Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol Biol, 804*, 281-295. doi:10.1007/978-1-61779-361-5_15 [doi]

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., . . . Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell, 160*(3), 554-566. doi:10.1016/j.cell.2015.01.006

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res, 35*(Database issue), D88-92. doi:10.1093/nar/gkl822

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., . . . Pennacchio, L. A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet, 40*(2), 158-160.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., 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*(Database issue), D95-97.

Vockley, C. M., Barrera, A., & Reddy, T. E. (2017). Decoding the role of regulatory element polymorphisms in complex disease. *Curr Opin Genet Dev, 43*, 38-45. doi:10.1016/j.gde.2016.10.007

Vockley, C. M., McDowell, I. C., D'Ippolito, A. M., & Reddy, T. E. (2017). A long-range flexible billboard model of gene activation. *Transcription, 8*(4), 261-267. doi:10.1080/21541264.2017.1317694

Wang, M., Zhang, K., Ngo, V., Liu, C., Fan, S., Whitaker, J. W., . . . Wang, W. (2019). Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res, 47*(13), 6753-6768. doi:10.1093/nar/gkz483

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., . . . Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun, 9*(1), 5380. doi:10.1038/s41467-018-07746-1

Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol, 30*(11), 1095-1106. doi:nbt.2422 [pii]

10.1038/nbt.2422 [doi]

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., . . . Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell., 158*(6), 1431-1443. doi: 1410.1016/j.cell.2014.1408.1009. Retrieved from http://ac.els-cdn.com/S0092867414010368/1-s2.0-S0092867414010368-main.pdf?_tid=3da24598-fcf0-11e4-a3ed-00000aab0f01&acdnat=1431907128_20fe1f58badebd42bca9b4e3254ca437

Whitaker, J. W., Chen, Z., & Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nat Methods, 2015 Mar;12*(3), 265-272.

Whitington, T., Frith, M. C., Johnson, J., & Bailey, T. L. (2011). Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res, 39*(15), e98. doi:gkr341 [pii]

10.1093/nar/gkr341 [doi]

Wilczynski, B., & Furlong, E. E. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol, 6*, 383. doi:msb201035 [pii]

10.1038/msb.2010.35 [doi]

Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J. P., & Cotney, J. (2018). High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell Rep, 23*(5), 1581-1597. doi:10.1016/j.celrep.2018.03.129

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., . . . Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res., 29*, 281-283.

Won, K. J., Chepelev, I., Ren, B., & Wang, W. (2008). Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics, 9*, 547. doi:1471-2105-9-547 [pii]

10.1186/1471-2105-9-547 [doi]

Yanez-Cuna, J. O., Kvon, E. Z., & Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. *Trends Genet., 29*(1), 11-22. doi: 10.1016/j.tig.2012.1009.1007. Epub 2012 Oct 1023. Retrieved from http://ac.els-cdn.com/S0168952512001473/1-s2.0-S0168952512001473-main.pdf?_tid=26fa13e4-f8e9-11e4-a50a-00000aacb360&acdnat=1431464279_d73a9bcd8006d26385fd518eb70211b8

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., . . . Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol, 13*(9), R48. doi:gb-2012-13-9-r48 [pii]

10.1186/gb-2012-13-9-r48 [doi]

Young, R. S., Kumar, Y., Bickmore, W. A., & Taylor, M. S. (2017). Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol, 18*(1), 242. doi:10.1186/s13059-017-1379-8

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The ensembl regulatory build. *Genome Biol., 16:56.*(doi), 10.1186/s13059-13015-10621-13055.

Zhang, G., Shi, J., Zhu, S., Lan, Y., Xu, L., Yuan, H., . . . Li, X. (2018). DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res, 46*(D1), D78-d84. doi:10.1093/nar/gkx920

Zhang, S., Jiang, L., Du, C., & Su, Z. (2013). SPIC: A novel information contents based similarity metric for comparing transcription factor binding site motifs. *BMC Syst Biol, 7*(Suppl 2), S14. doi:doi:10.1186/1752-0509-7-S2-S14

Zhou, S., Treloar, A. E., & Lupien, M. (2016). Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. *Cancer Discov, 6*(11), 1215-1229. doi:10.1158/2159-8290.cd-16-0745

Zuin, J., Dixon, J. R., van der Reijden, M. I., Ye, Z., Kolovos, P., Brouwer, R. W., . . . Wendt, K. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A, 111*(3), 996-1001. doi:10.1073/pnas.1317788111