Database, 2022, 1–11
DOI: https://doi.org/10.1093/database/baac024
Original article

PCRMS: a database of predicted cis-regulatory modules and constituent transcription factor binding sites in genomes

Pengyu Ni and Zhengchang Su*

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

*Corresponding author: Tel: +01-704-687-7996; Fax: +01-704-687-8667; Email: zcsu@uncc.edu

Citation details: Ni, P. and Su, Z. PCRMS: a database of predicted cis-regulatory modules and constituent transcription factor binding sites in genomes. *Database* (2022) Vol. 2022: article ID baac024; DOI: https://doi.org/10.1093/database/baac024

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

Abstract

More accurate and more complete predictions of cis-regulatory modules (CRMs) and constituent transcription factor (TF) binding sites (TFBSs) in genomes can facilitate characterizing functions of regulatory sequences. Here, we developed a database predicted *cis*-regulatory modules (PCRMS) (https://cci-bioinfo.uncc.edu) that stores highly accurate and unprecedentedly complete maps of predicted CRMs and TFBSs in the human and mouse genomes. The web interface allows the user to browse CRMs and TFBSs in an organism, find the closest CRMs to a gene, search CRMs around a gene and find all TFBSs of a TF. PCRMS can be a useful resource for the research community to characterize regulatory genomes.

Database URL:https://cci-bioinfo.uncc.edu/

Introduction

Cis-regulatory modules (CRMs), such as enhancers, promoters, silencers and insulators, are composed of clusters of short DNA sequences where transcription factors (TFs) can bind to regulate the expression of target genes in many biological processes (1). Recent studies have showed that the vast majority of complex trait-associated single nucleotide variants (SNVs) are located in noncoding sequences (NCSs) and often disrupt TF binding sites (TFBSs) in CRMs (2, 3). Variation of CRMs also plays a crucial role in divergence in closely related species (4–7). In principle, variation in TFBSs in a CRM could affect the affinity of cognate TFs, resulting in changes in chromatin modifications and target gene expression in specific cell types in tissues and ultimately leading to diversity of complex traits, including susceptibility to common complex diseases (8–14). Therefore, more accurate and more complete categorization of CRMs and constituent TFBSs in the human and important model organisms' genomes can greatly facilitate characterizing functions of regulatory sequences and their roles in many important biological processes including disease and evolution.

Recently, a plethora of next-generation sequencing (NGS)-based technologies have been developed to characterize different features of CRMs at a genome scale, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) (15) to locate TFs binding or histone modifications, and DNase I hypersensitive sites sequencing (DNase-seq) (16), assay for transposase-accessible chromatin using sequencing (ATAC-seq) (17), formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) (18) and micrococcal nuclease digestion with deep sequencing (MNase-seq) (19) to identify the chromatin accessibility. Consequently, an exponentially increased number of datasets have been generated using these technologies by consortia such as ENCODE (20, 21), Epigenomics Roadmap (22, 23) and Genotype-Tissue Expression (GTEx) (24). Based on different data types that capture different features of CRMs, many computational strategies have been developed to predict CRMs. For instances based on TF ChIP-seq data, methods such as SpaMo (25), CPModule (26), COPS (27) and INSECT (28) have been developed to identify regions of TF binding as potential CRMs. Based on histone modification marks and chromatin accessibility data, hidden Markov models (29, 30) and dynamics Bayesian models (31) have been developed to predict CRMs and their functional states in different cell types. Based on bidirectionally transcribed pairs of capped RNAs or enhancer RNA (eRNA), the FANTOM 5 project identified active enhancers in various human and mouse tissues (32). By integrating

Received 11 December 2021; Revised 20 February 2022

multiple tracks of epigenetics marks, TF binding, as well as predicted and experimentally validated enhancers, several groups have developed CRM/enhancer databases, such as dbSUPER (33), SEdb (34), DENdb (35), EPDnew promoters (36), UCNEbase (37), CraniofacialAtlas (38), GeneHancer (39), HACER (40), RAEdb (41), HEDD (42), DiseaseEnhancer (43), SEA (44), EnhancerAtlas (45) and SCREEN (46). However, these databases only cover a small portion of enhancers/CRMs encoded in the genomes, and some may have a high false positive rate (FDR) (47). For instance, even the most currently updated SCREEN database that stores candidate cis-regulatory elements (cCREs) predicted by the ENCODE phase 3 consortium contains only 926 535 and 339 815 cCREs in the human and mouse genomes, with a mean length of 273bp and 272bp, respectively (46), which is much shorter than the mean length (~2000bp) of known human and mouse CRMs in the VISTA database (48), indicating that cCREs in both the human and mouse genomes might be underpredicted for both their lengths and numbers. Moreover, none of these databases provide *de novo* predicted TFBSs in enhancers/CRMs, which are critical to understand the mechanisms of transcriptional regulation, and to pinpoint causal variants of phenotype diversity and disease risks.

Using a highly efficient CRM and TFBS prediction pipeline dePCRM2 that we developed recently (47), we have predicted CRMs and constituent TFBSs in the human (Homo sapiens) and mouse (Mus musculus) genomes using a large number of TF ChIP-seq datasets in the organisms. Comparative analysis indicates that our predictions are substantially more accurate and more complete than those in existing databases (47). To facilitate the research community to use these predictions for various purposes, we constructed an online database PCRMS The database currently contains 1 404 973 and 920 068 CRM candidates (CRMCs) and 90 671 016 and 104 251 155 TFBSs for 201 and 210 unique motif (UM) families in the human and mouse genomes, respectively. To our best knowledge, these represent the most complete collections of accurately predicted CRMs and constituent TFBSs in the human and mouse genomes. The web interface to PCRMS allows a quick search, browse and visualization of the contents of the database, and provides three functional analysis modules. Using these modules, a user can find the closest CRMs to a gene; search CRMs that are located in a specified region around a gene; and search TFBSs in CRMs for a given TF. The interface also provides copy, export and download functions of selected CRMs or all the predicted CRMs in BED format. We hope that PCRMS will facilitate the research community's efforts to characterize the regulatory genomes in important organisms.

Materials and methods

Datasets

We downloaded (6/1/2019) 6092 TF ChIP-seq datasets for 779 TFs in 468 cells/tissues/organs of humans, and 4786 TF ChIP-seq datasets for 501 TFs in 162 cells/tissues/organs of mice from the CISTROME database (49). After filtering out called binding peaks with low quality, for each left peak, we extracted 1000bp genome sequence centring on the middle of the binding peaks. As most called binding peaks have a length shorter than 1000bp (47), we extended most of them. We have shown earlier that such extension could greatly increase the power of the datasets without including much noise (47, 50).

Prediction of CRMs and constituent TFBSs

To predict CRMs and TFBSs, we applied our dePCRM2 algorithm (47) to the datasets with extended binding peaks from each organism using the default parameters. Briefly, dePCRM2 first finds over-represented motifs and co-occurring motifs pairs (CPs) in each dataset in an organism. It then identifies UMs by combining highly similar motifs in CPs across all the datasets in the organism. To model the interactions among coperative TFs, dePCRM2 constructs an interaction network, where UMs are the nodes, and two nodes are connected by a edge with a weight being their interaction score, defined as,

$$S_{INTER}(U_i, U_j) = \frac{1}{|D(U_i, U_j)|} \sum d \in D(U_i, U_j)(\frac{1}{|dU_i|}) + (\frac{1}{|d(U_j)|}) \sum s \in S(d(U_i), (d(U_j)) \frac{150}{r(s)'} \quad (1)$$

where $D(U_i, U_j)$ is the dataset 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, andr (s) the shortest distance between any TFBS of U_i and any TFBS of U_j in a sequence $s \in S(d(U_i), (d(U_j))$. Clearly, S_{INTER} allows flexible adjacency and orientation of TFBSs in a CRM and at the same time, it rewards motifs with binding sites cooccurring frequently in a shorter distance in a CRM. Next, dePCRM2 connects two adjacent TFBSs of the UMs if their distance $d \leq 300$bp (about the length of two nucleosomes), and predicts each resulting connected DNA segment as a CRM candidate (CRMC). In this way, dePCRM2 partitions the genome regions covered the extended binding peaks into a CRMC set and a non-CRMC set. Finally, dePCRM2 evaluates each CRMC containing $b_1, b_2 \cdots, b_n$ TFBSs by computing a score defined as,

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

where $U(b_k)$ is the UM of TFBS b_k, $S_{INTER}[U(b_i), U(b_j)]$ the interaction score between $U(b_i)$ and $U(b_j)$, $S(b_k)$ the binding score of b_k based on the position weight matrix (PWM) of $U(b_k)$. Only TFBSs with a positive score are considered. dePCRM2 also computes a *P*-value for each CRMC as follows. For each predicted CRMC, dePCRM2 generates a Null CRMC that has the same length and 4-mer nucleotide frequencies as the CRMC using a third-order Markov chain model (50), and computes the S_{CRM} score for each Null CRMC based on a random interaction network which is generated by randomly rewiring the nodes of the UM interaction network. Then, an empirical *P*-value for a CRMC with a $S_{CRM} =$ sis computed based on the distribution of S_{CRM} score of the Null CRMCs,

$$p = \frac{n(s)}{N}, \quad (3)$$

where $n(s)$ is the number of Null CRMCs with a S_{CRM} score greater than s and N the total number of the CRMCs.

Technical implementation

The current version of PCRMS (v2) was developed using MySQL 5.7.17 (http://www.mysql.com) and it runs on a Linux-based Apache2 server (http://www.apache.org). PHP 7.4 (http://www.php.net) scripts were used for back-end processing. The interactive interface and responsive features were implemented using Bootstrap 4 (https://getbootstrap.com/), JQuery (http://jquery.com) and dataTables (https://datatables.net/). NCBI sequence viewer 3.38.0 (https://www.ncbi.nlm.nih.gov/projects/sviewer/) was used for visualization.

Results and discussion

Predicted CRMs and constituent TFBSs in the human and mouse genomes

Applying dePCRM2 to the TF ChIP-seq datasets available to us (6/1/2019) in each organism, we predicted 1 404 973 and 920 068 CRMCs in the human (47) and mouse genomes, comprising 44.03% and 50.39% of their genomes, respectively. These CRMCs contain 90 671 016 and 104 251 155 TFBSs, comprising 16.71% and 20.34% of the human and mouse genomes, respectively. We compared the numbers and lengths of our CRMCs with those of cCREs in the SCREEN database (46) and those of enhancers in the GeneHancer database (39). cCREs were predicted based on overlaps among hundreds or thousands of DNase-seq, ATAC-seq and histone marks ChIP-seq datasets in various cell/tissue types in an organism (46). Enhancers in GeneHancer were predicted by combing nine sets of earlier predicted and experimentally determined human CRMs using a voting schema (39). As shown in Table 1, the numbers of our predicted CRMCs are much larger than those of cCREs and that of GeneHancer enhancers. Our predicted CRMCs also comprise much larger proportions of the genomes than do those in CREEN or GeneHancer (Table 1). We attribute to two reasons the larger numbers and higher genome coverages of our predicted CRMCs. First, the types of input data used by the three methods were different, which might capture different features of CRMs, thus have different capabilities of predicting CRMs. Specifically, the input data for predicting cCREs were DNase-seq, ATAC-seq and histone marks ChIP-seq data, those for predicting GeneHancer enhancers were earlier predicted and experimentally determined CRM sets by different groups, and those for predicting CRMCs were TF ChIP-seq data. Second, the number of predicted cCREs was limited by the number of called DNase I hypersensitive sites, transposase-accessible sites and epigenetic mark peaks, while the number of GeneHancer enhancers was constrained by the sizes of earlier predicted and experimentally determined enhancer sets. In contrast, by appropriately extending the originally called short TF binding peaks, we could greatly increase the power of available TF ChIP-seq data as we demonstrated earlier (47), since binding sites of co-operative TFs tend to be closely located on a genome segment to form a CRM (1), while a called short binding peak to which the ChIP-ed TF bind can be only a part of a longer CRM. For instance, the extended binding peaks (1000bp) in the 6079 human ChIP-seq datasets cover 77.47% of the mappable genome, and the extended parts of the peaks contribute to almost half (47.10%) of the coverage (47). dePCRM2 predicts 56.84% of the covered genome to be CRMC positions, and 42.13% of them are predicted solely based on the extended parts of originally called binding peaks (47). Importantly, we have shown that CRMC positions predicted by the extended parts of originally called binding peaks are under similarly strong evolutionary constraints as those predicted by the originally called binding peaks, thus, are likely true CRMC positions (47). On the other hand, due to the noisy nature of ChIP-seq data (51–53), 37.82% of genome positions covered by originally called binding peaks are not predicted to be CRMCs, and they are largely selectively neutral (47).

The lengths of our predicted CRMCs in the human (Figure 1A) and mouse (Figure 1B) genomes have similar distributions, ranging from a few hundred bp to a few thousand bp with a mean length of 981bp and 1,439bp, respectively, which are shorter than those of known human (2049bp) and mouse (2432bp) enhancers in the VISTA database (48), indicating that a portion of our CRMCs are only components of longer CRMs as we argued earlier (47). In contrast, the lengths of cCREs in the human and mouse genomes are almost uniform with a mean length of 273bp and 272bp, respectively (46) (Table 1), while the lengths of GeneHancer enhancers show a periodic pattern (47) with a mean length of 1489bp. Such erratic lengths of cCREs and GeneHancer enhancers are likely artifacts of the underlying algorithms as we argued earlier (47). On the other hand, as we pointed out earlier (47), accurate prediction of the lengths of CRMs is a highly challenge task, because a truncated enhancer can still be functional (1), and a super-enhancer may contain multiple discrete short enhancers (54). Thus, the length of a CRM depends on how it is defined. dePCRM2 predicts a CRMC as a cluster of TFBSs with the distance between any two adjacent TFBSs being short than 300bp (Methods and (47)). While cCREs might be shorter discrete units of longer CRMs, we estimated a FDR of 23.12% for the human cCREs positions based on their largely neutrally evolutionary behaviors (47). GeneHancer enhancers have a mean length of 1,489bp, which is shorter than that of known human enhancers (2049bp) in the VISTA database

Table 1. Comparison of the contents of the three databases

		CRMCs			TFBSs	
Databases	Species	Mean length (bp)	Number	Coverage of genome (%)	Number	Coverage of genome (%)
PCRMS	*H. sapiens*	981	1 404 973	44.03	90 671 016	16.71
	M. musculus	1493	920 068	50.39	104 251 155	20.34
GeneHancer	*H. sapiens*	1489	394 086	18.99	X	X
	M. musculus	X	X	X	X	X
SCREEN	*H. sapiens*	273	926 535	8.2	X	X
	M. musculus	272	339 815	3.39	X	X

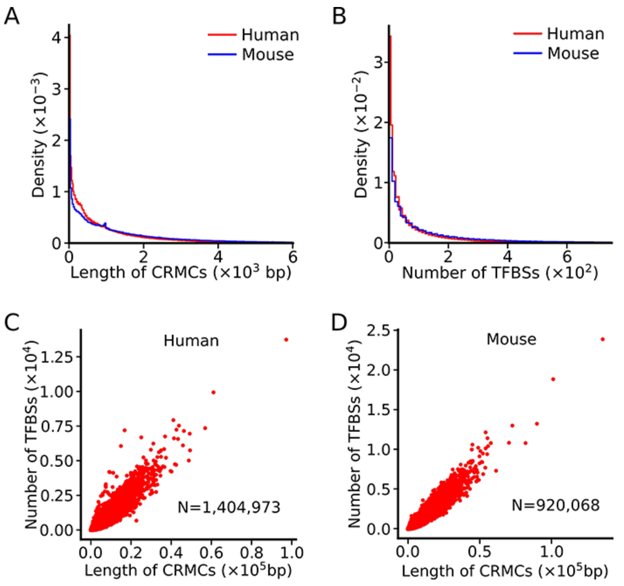


Figure 1. Summary of the lengths of CRMCs and the numbers of TFBSs in a CRMC in the human and mouse genomes. A. Distributions of the lengths of CRMCs in the human and mouse genomes. B. Distributions of the number of TFBSs in a CRMC in the human and mouse genomes. C. Scatter plot of the number of TFBSs in a CRMC vs its length in the human genome. D. Scatter plot of the number of TFBSs in a CRMC vs its length in the mouse genome.

(48), we estimated a FDR of 29.28% for the genome positions of the GeneHancer enhancers based on their largely neutrally evolutionary behaviors (47). We have shown that our predicted CRMCs and TFBSs positions in the human genome are highly accurate based on validations using multiple independent data (47), and the same is true for the predicted CRMCs and TFBSs in the mouse genome (manuscript in preparation, P.N. and Z.S).

The number of TFBSs in a CRMC in either the human genome or the mouse genome varies widely, ranging from a few to a few hundreds, with a mean/median number of 90/34 and 183/67, respectively (Figure 1B). Interestingly, the number of TFBSs in a CRMCs is largely linearly related to the length of the CRMC in both the human (Figure 1C) and the mouse (Figure 1D) genomes, indicating that the density of TFBSs is largely the same in the most of the CRMCs. In contrast, no information of de novo predicted TFBSs in cCREs or enhancers is available in the SCREEN or GeneHancer databases (Table 1).

To evaluate the significance of the CRMCs, dePCRM2 computes a *P*-value for each predicted CRMCs based on its S_{CRM}. We have shown earlier that the longer a CRMC, the higher its S_{CRM} score, the smaller its *P*-value, and the stronger evolutionary constraint it is subject to (47). Therefore, both the S_{CRM} score and its associated *P*-value capture essential features of a true CRM. This result also justifies our assumption that a genome segment containing closely located putative TFBSs is more likely a CRM than a segment without such sequence patterns. It is based on this assumption that dePCRM2 predicts CRMs The assumption is clearly in agreement with the well-known notion that a functional genome segment such as a CRM must contain certain sequence patterns (i.e, clusters of TFBSs) that are unlikely to occur by chance, and that the longer the patterns, the less likely they occur by chance.

However, as dePCRM2 predicts CRMCs based on the predicted TFBSs in the genome, false positive and false negative predictions of TFBSs would result in false positive, false negative and incomplete predictions of CRMs. We estimated the FDR of our motif-finder ProSampler used in the dePCRM2 pipeline to be about 8% (50). Thus, we designed dePCRM2 to further filter out potentially false positive motifs returned by ProSampler in the extended binding peaks in a dataset based on their cooccurring patterns (see Method and (47)). We estimated the FDR of the predicted CRMC positions to be about 0.05%, thus, FDR for TFBSs are likely further reduced (47). However, as we indicated earlier (47), due to the limitation of the available TF ChIP-seq datasets, our predicted TFBSs are still incomplete, and a proportion of our predicted CRMCs might be only components of long CRMs whose full prediction depends on more data available in the future. Nonetheless, the short CRMC components can be effectively filtered out using a higher S_{CRM} score cutoff or a lower *P*-value cutoff (47). To assist the users who might be interested in CRMCs with different lengths, statistical significance or with different evolutionary constraints, in addition to making the entire sets of predicted CRMs available for bulk downloading, we provide four options of *P*-value cutoffs (*P*-value <0.05, 0.01, 5×10^{-6} and 1×10^{-6}) to query the database. Table 2 summarizes the predicted CRMs using these *P*-value cutoffs; they are subsets of the CRMCs with different length distributions and conservation levels (47). Clearly, the smaller a *P*-value cutoff, the longer the predicted CRMs.

Web interface to the database

We provide a user-friendly web interface to the PCRMS database for quickly inquiring and browsing predicted CRMs and TFBSs at different statistically significant levels in each organism as well as three functional analysis modules. Using these modules, the user can (i) search the closest CRM to a given gene, (ii) search all CRMs in the upstream and/or downstream regions of a gene of interest and (iii) search the TFBSs of a TF on one or more chromosomes in an organism (Figure 2).

Browse of database contents

We provide a Browse function by which the user can browse all CRMs predicted at a selected *P*-value cutoff on one or multiple selected chromosomes in a selected organism and inspect each CRMs and constituent TFBSs in detail. The user starts in the search form (Figure 3A) by selecting an organism (e.g. H. sapiens), one or more chromosomes (e.g. chrX) and a *P*-value cutoff (e.g. 1E-06). The search returns all the predicted CRMs ($n = 8762$) on the chromosome (chrX) of the organism (H. sapiens) in the interactive CRM list table (Figure 3B). Clicking on a CRM of interest, e.g. the first CRM hse1000017 in the table pops up the CRM information table (Figure 3C), where some parameter of the CRM are shown in the left panel and the locus of the CRM is displayed in the NCBI sequence viewer (shadowed rectangle) in the right panel, enabling detailed inspections of the genomic context of the CRM, including its neighboring genes and other annotations using the zooming and the translation functions of the viewer. For instance, the viewer reveals that

Table 2. Summary of the predicted CRMs at different *P*-values in the human and mouse genomes

		CRMs			TFBSs	
Species	*P*-value	Mean length (bp)	Number	Coverage of genome (%)	Number	Coverage of genome (%)
H. sapiens	0.05	1162	1 155 151	43.47	89 948 206	16.54
	0.01	1292	1 020 679	42.72	88 912 654	16.32
	5.00E-06	2292	428 628	31.81	71 478 114	12.88
	1.00E-06	2624	327 396	27.82	64 136 635	11.47
M. musculus	0.05	1749	777 409	49.9	103 718 473	20.21
	0.01	1944	688 033	49.06	102 730 265	19.99
	5.00E-06	3182	338 635	39.53	88 579 892	16.96
	1.00E-06	3780	250 606	34.75	80 002 349	15.2

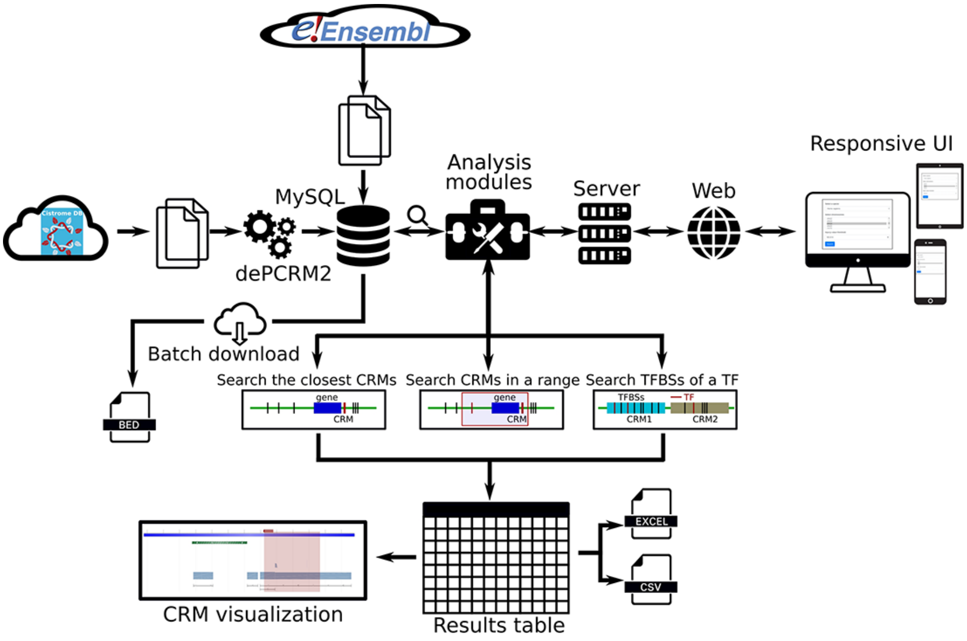


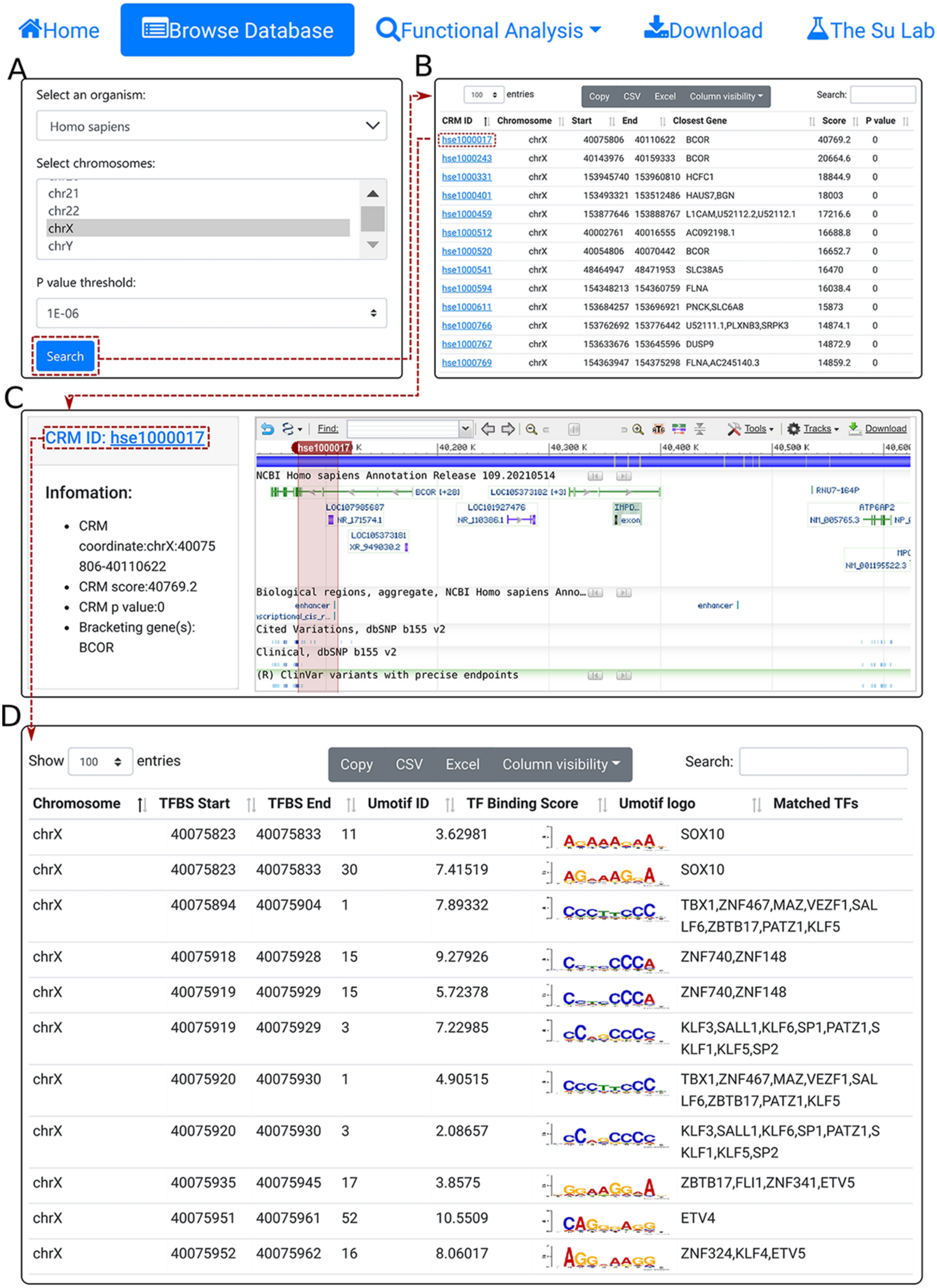
Figure 2. Overview of data integration and analysis modules and features of the PCRMS database.

hse1000017 is located in the second through the fifth introns, and spans the third through fifth exons, of the BCOR gene that codes for a corepressor of a transcription repressor BCL6. Both BCOR and BCL6 are involved in B lymphocytes differentiation (55, 56). Interestingly, hse1000017 overlaps two regulatory sequences annotated as 'enhancer' and 'transcriptional cis-regulation', while many ClinVar variants are located in hse1000017 (Figure 3C). Finally, clicking on the CRM ID (e.g. hse1000017) in the right panel of the CRM information table (Figure 3C) displays the CRM's 5094 constituent TFBSs in the interactive TFBS table (Figure 3D), which includes the coordinates of the TFBSs, their UM IDs, binding scores, UM logos and matched known motifs. The vast majority of these TFBSs match those of known TF families (Figure 3D).

In both the interactive CRM list table (Figure 3B) and the interactive TFBS table (Figure 3C), the user can change the number of entries to display in a page, sort results based on different columns, filter the results using the search box and set visible columns. The user can copy or export the selected items in a file in the CSV or Excel formats or export all records if no item is selected by default (Figure 3).

Functional analyses

To facilitate analyzing potential CRM-gene relationships and TFBSs landscape of specific TFs, we provide three functional analysis modules. First, using the 'select the closest CRMs to a gene' function, the user can search the closest CRMs to a gene (e.g. GL13) in an organism (e.g. H. sapiens) at a *P*-value cutoff (e.g., 1×10^{-6}) (Figure 4A). The search returns the interactive CRM list table containing all CRMs to which the gene is the closest among all other genes in the chromosome (Figure 4B). In the example of the GLI3 gene, a total of 53 CRMs are returned. The user can inspect any of them by clicking on the CRM ID, which pops up the information table of the CRM as we demonstrated earlier (Figure 3C). For instance, clicking on the third CRM hse1003435 in the table displays it in the NCBI sequence viewer, revealing that the CRM is located in the third and fourth introns, and spans the fourth and fifth exons, of the GL13 gene (shadowed rectangle in Figure 4C). Interestingly, hse1003435 overlaps two annotated enhancers and many ClinVar variants (Figure 4C). Finally, clicking on the CRM ID hse1003435 in the right panel of the CRM information table (Figure 4C) displays the CRM's 988 constituent TFBSs in the interactive TFBS table



Select an organism:

Homo sapiens

Select chromosomes:

chr21
chr22
chrX
chrY

P value threshold:

1E-06

Search

100 entries Copy CSV Excel Column visibility Search:

CRM ID	Chromosome	Start	End	Closest Gene	Score	P value
hse1000017	chrX	40075806	40110622	BCOR	40769.2	0
hse1000243	chrX	40143976	40159333	BCOR	20664.6	0
hse1000331	chrX	153945740	153960810	HCFC1	18844.9	0
hse1000401	chrX	153493321	153512486	HAUS7,BGN	18003	0
hse1000459	chrX	153877646	153888767	L1CAM,U52112.2,U52112.1	17216.6	0
hse1000512	chrX	40002761	40016555	AC092198.1	16688.8	0
hse1000520	chrX	40054806	40070442	BCOR	16652.7	0
hse1000541	chrX	48464947	48471953	SLC38A5	16470	0
hse1000594	chrX	154348213	154360759	FLNA	16038.4	0
hse1000611	chrX	153684257	153696921	PNCK,SLC6A8	15873	0
hse1000766	chrX	153762692	153776442	U52111.1,PLXNB3,SRPK3	14874.1	0
hse1000767	chrX	153633676	153645596	DUSP9	14872.9	0
hse1000769	chrX	154363947	154375298	FLNA,AC245140.3	14859.2	0

CRM ID: hse1000017

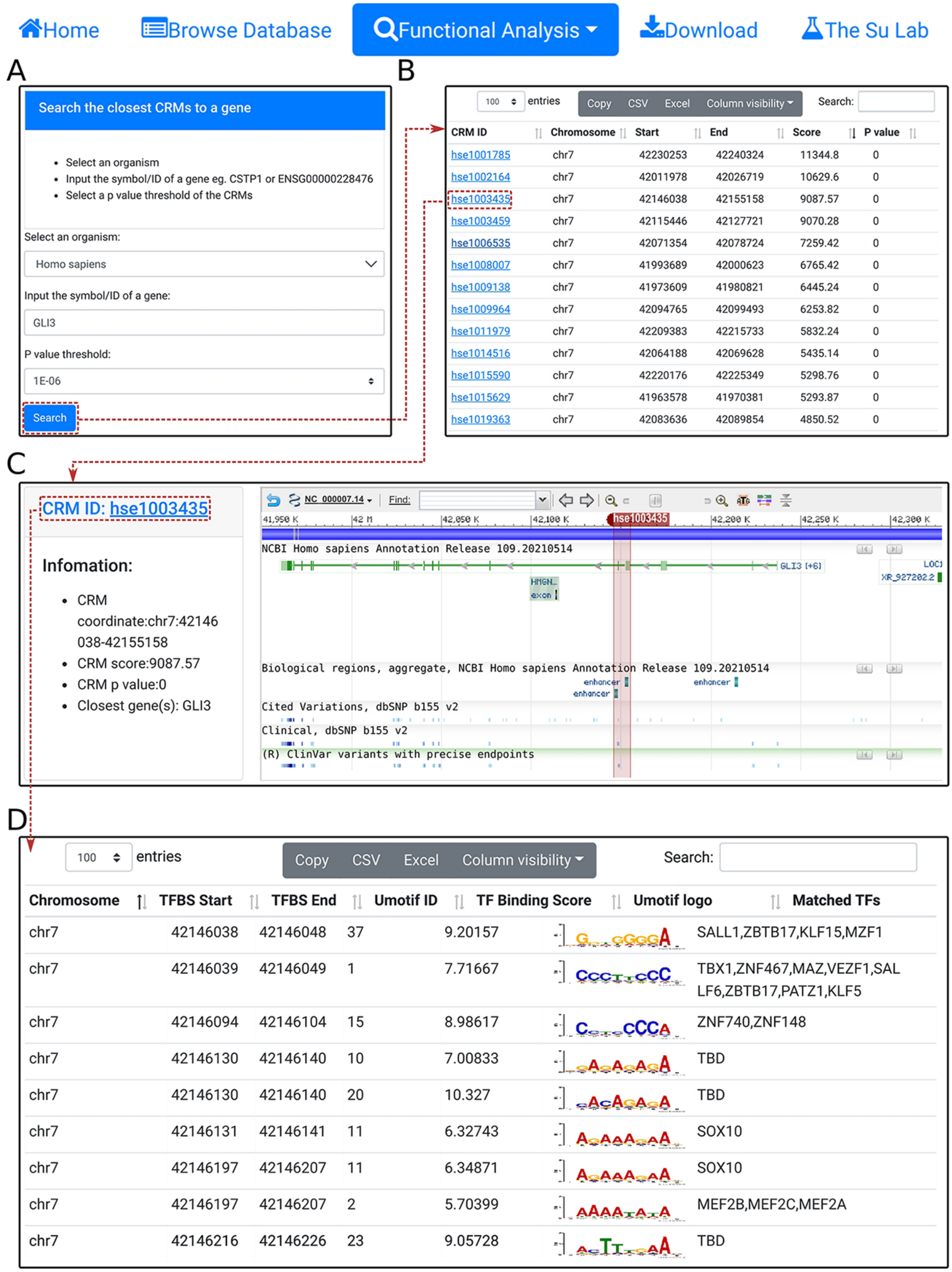
Infomation:

- CRM coordinate:chrX:40075806-40110622
- CRM score:40769.2
- CRM p value:0
- Bracketing gene(s): BCOR

Show 100 entries Copy CSV Excel Column visibility Search:

Chromosome	TFBS Start	TFBS End	Umotif ID	TF Binding Score	Umotif logo	Matched TFs
chrX	40075823	40075833	11	3.62981		SOX10
chrX	40075823	40075833	30	7.41519		SOX10
chrX	40075894	40075904	1	7.89332		TBX1,ZNF467,MAZ,VEZF1,SALLF6,ZBTB17,PATZ1,KLF5
chrX	40075918	40075928	15	9.27926		ZNF740,ZNF148
chrX	40075919	40075929	15	5.72378		ZNF740,ZNF148
chrX	40075919	40075929	3	7.22985		KLF3,SALL1,KLF6,SP1,PATZ1,SKLF1,KLF5,SP2
chrX	40075920	40075930	1	4.90515		TBX1,ZNF467,MAZ,VEZF1,SALLF6,ZBTB17,PATZ1,KLF5
chrX	40075920	40075930	3	2.08657		KLF3,SALL1,KLF6,SP1,PATZ1,SKLF1,KLF5,SP2
chrX	40075935	40075945	17	3.8575		ZBTB17,FLI1,ZNF341,ETV5
chrX	40075951	40075961	52	10.5509		ETV4
chrX	40075952	40075962	16	8.06017		ZNF324,KLF4,ETV5

Figure 3. The browse functions. A. In the search form, the use selects an organism (e.g. H. sapiens), one or more chromosomes (e.g. chrX) and a *P*-value cutoff (e.g. 1E-06). B. The searching results are displayed in the CRM list table. Shown is a snapshot of the resulting CRM list table containing 8762 predicted CRMs on chrX of H. sapiens. The first CRM hse1000017 in the list table is selected for further visualization. C. In the CRM information table, some parameters of the selected CRM hse1000017 is shown in the right panel, and the locus is displayed in the NCBI sequence viewer for further inspection. Clicking on 'hse1000017' in the right panel of the CRM information table displays its constituent TFBSs. D. A snapshot of the TFBS table of hse1000017 containing its 5094 constituent TFBSs.



Search the closest CRMs to a gene

- Select an organism
- Input the symbol/ID of a gene eg. CSTP1 or ENSG00000228476
- Select a p value threshold of the CRMs

Select an organism: Homo sapiens

Input the symbol/ID of a gene: GLI3

P value threshold: 1E-06

Search

100 entries Copy CSV Excel Column visibility Search:

CRM ID	Chromosome	Start	End	Score	P value
hse1001785	chr7	42230253	42240324	11344.8	0
hse1002164	chr7	42011978	42026719	10629.6	0
hse1003435	chr7	42146038	42155158	9087.57	0
hse1003459	chr7	42115446	42127721	9070.28	0
hse1006535	chr7	42071354	42078724	7259.42	0
hse1008007	chr7	41993689	42000623	6765.42	0
hse1009138	chr7	41973609	41980821	6445.24	0
hse1009964	chr7	42094765	42099493	6253.82	0
hse1011979	chr7	42209383	42215733	5832.24	0
hse1014516	chr7	42064188	42069628	5435.14	0
hse1015590	chr7	42220176	42225349	5298.76	0
hse1015629	chr7	41963578	41970381	5293.87	0
hse1019363	chr7	42083636	42089854	4850.52	0

CRM ID: hse1003435

Infomation:

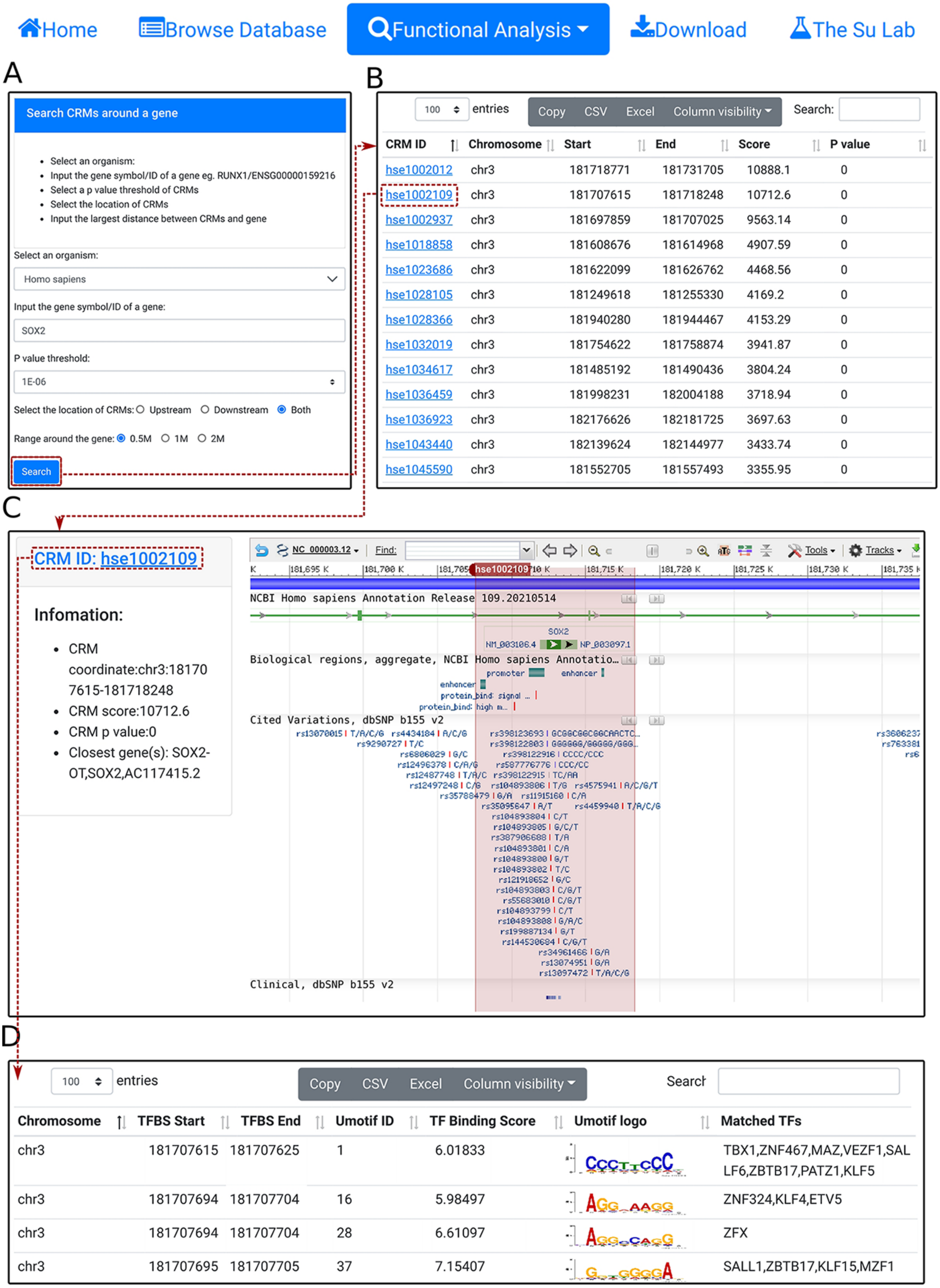
- CRM coordinate:chr7:42146038-42155158
- CRM score:9087.57
- CRM p value:0
- Closest gene(s): GLI3

NC_000007.14
Find:
41,950 K
42 M
42,050 K
42,100 K
hse1003435
42,200 K
42,250 K
42,300 K
NCBI Homo sapiens Annotation Release 109.20210514
GLI3 [+6]
LOC1
XR_927202.2
HMGN...
exon
Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109.20210514
enhancer
enhancer
enhancer
Cited Variations, dbSNP b155 v2
Clinical, dbSNP b155 v2
(R) ClinVar variants with precise endpoints

100 entries Copy CSV Excel Column visibility Search:

Chromosome	TFBS Start	TFBS End	Umotif ID	TF Binding Score	Umotif logo	Matched TFs
chr7	42146038	42146048	37	9.20157	G GGGGA	SALL1,ZBTB17,KLF15,MZF1
chr7	42146039	42146049	1	7.71667	CCCT CCC	TBX1,ZNF467,MAZ,VEZF1,SALLF6,ZBTB17,PATZ1,KLF5
chr7	42146094	42146104	15	8.98617	C CCCA	ZNF740,ZNF148
chr7	42146130	42146140	10	7.00833	AAAAA	TBD
chr7	42146130	42146140	20	10.327	ACAGAGA	TBD
chr7	42146131	42146141	11	6.32743	AAAAAA	SOX10
chr7	42146197	42146207	11	6.34871	AAAAAA	SOX10
chr7	42146197	42146207	2	5.70399	AAAATAA	MEF2B,MEF2C,MEF2A
chr7	42146216	42146226	23	9.05728	A TT AA	TBD

Figure 4. Search the closest CRM(s) to a gene. A. In the search form, the use selects an organism (e.g. H. sapiens) and a *P*-value cutoff (e.g. 1E-06), and inputs a gene name (e.g. GL13). B. The searching results are displayed in the CRM list table. Shown is a snapshot of the returned CRM list table containing 53 predicted CRMs. The third CRM hse1003435 in the list table is selected for further inspection. C. In the CRM information table, some parameters of the selected CRM hse10003435 is displayed in the right panel, and the locus is displayed in the NCBI sequence viewer for further inspections. Clicking on ‘hse1003435’ in the right panel of the CRM information table displays its constituent TFBSs. D. A snapshot of the TFBS table of hse1003435 containing its 988 constituent TFBSs.



Search CRMs around a gene

- Select an organism:
- Input the gene symbol/ID of a gene eg. RUNX1/ENSG00000159216
- Select a p value threshold of CRMs
- Select the location of CRMs
- Input the largest distance between CRMs and gene

Select an organism: Homo sapiens

Input the gene symbol/ID of a gene: SOX2

P value threshold: 1E-06

Select the location of CRMs: ○ Upstream ○ Downstream ● Both

Range around the gene: ● 0.5M ○ 1M ○ 2M

Search

100 entries Copy CSV Excel Column visibility Search:

CRM ID	Chromosome	Start	End	Score	P value
hse1002012	chr3	181718771	181731705	10888.1	0
hse1002109	chr3	181707615	181718248	10712.6	0
hse1002937	chr3	181697859	181707025	9563.14	0
hse1018858	chr3	181608676	181614968	4907.59	0
hse1023686	chr3	181622099	181626762	4468.56	0
hse1028105	chr3	181249618	181255330	4169.2	0
hse1028366	chr3	181940280	181944467	4153.29	0
hse1032019	chr3	181754622	181758874	3941.87	0
hse1034617	chr3	181485192	181490436	3804.24	0
hse1036459	chr3	181998231	182004188	3718.94	0
hse1036923	chr3	182176626	182181725	3697.63	0
hse1043440	chr3	182139624	182144977	3433.74	0
hse1045590	chr3	181552705	181557493	3355.95	0

CRM ID: hse1002109

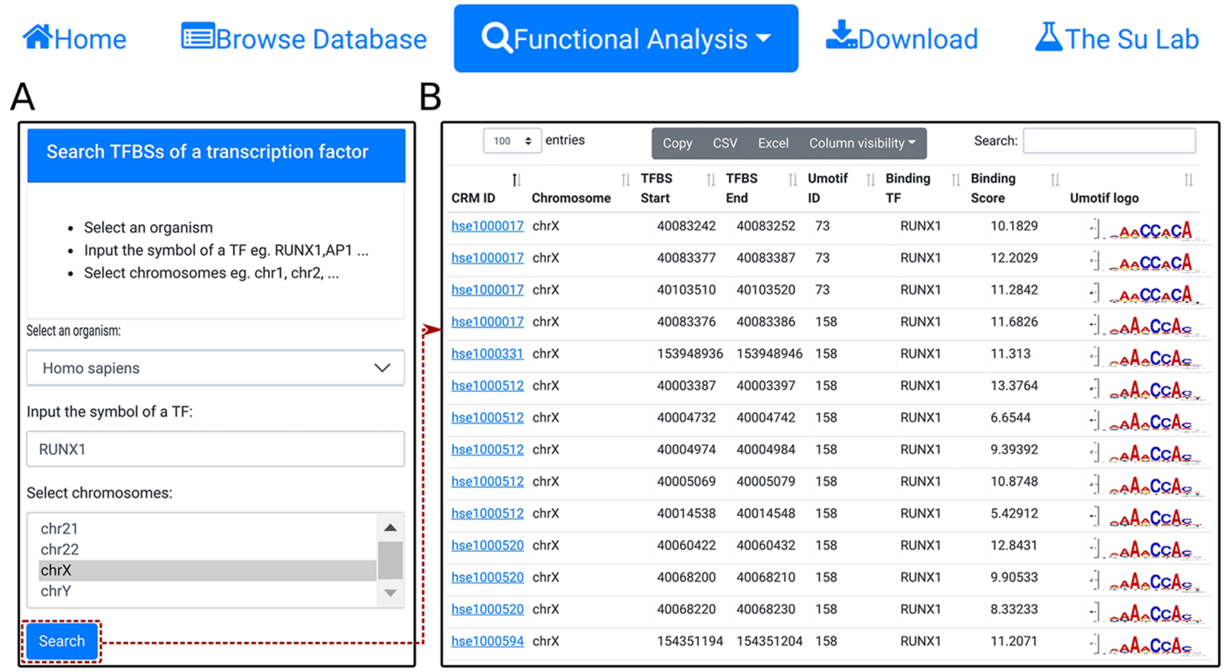
Infomation:

- CRM coordinate:chr3:181707615-181718248
- CRM score:10712.6
- CRM p value:0
- Closest gene(s): SOX2-OT,SOX2,AC117415.2

100 entries Copy CSV Excel Column visibility Searcl

Chromosome	TFBS Start	TFBS End	Umotif ID	TF Binding Score	Umotif logo	Matched TFs
chr3	181707615	181707625	1	6.01833		TBX1,ZNF467,MAZ,VEZF1,SALLF6,ZBTB17,PATZ1,KLF5
chr3	181707694	181707704	16	5.98497		ZNF324,KLF4,ETV5
chr3	181707694	181707704	28	6.61097		ZFX
chr3	181707695	181707705	37	7.15407		SALL1,ZBTB17,KLF15,MZF1

Figure 5. Search CRM(s) in a region around a gene. A. In the search form, the use selects an organism (e.g. H. sapiens) and a *P*-value cutoff (e.g. 1E-06), and inputs a gene name (e.g.SOX2). B. The searching results are displayed in the CRM list table. Shown is a snapshot of the 102 returned CRMs in the table. The second CRM hse1002109 in the list table is selected for further inspection. C. In the CRM information table, parameters of the selected CRM hse1002109 is shown in the right panel, and the locus is displayed in the NCBI sequence viewer for further inspections. Clicking on 'hse1002109' in the right panel of the CRM information table displays its constituent TFBSs. D. A snapshot of the TFBS table of hse1002109 containing its 1344 constituent TFBSs.



Search TFBSs of a transcription factor

- Select an organism
- Input the symbol of a TF eg. RUNX1,AP1 ...
- Select chromosomes eg. chr1, chr2, ...

Select an organism:

Homo sapiens

Input the symbol of a TF:

RUNX1

Select chromosomes:

chr21
chr22
chrX
chrY

Search

100 entries

Copy CSV Excel Column visibility

Search:

CRM ID	Chromosome	TFBS Start	TFBS End	Umotif ID	Binding TF	Binding Score	Umotif logo
hse1000017	chrX	40083242	40083252	73	RUNX1	10.1829	
hse1000017	chrX	40083377	40083387	73	RUNX1	12.2029	
hse1000017	chrX	40103510	40103520	73	RUNX1	11.2842	
hse1000017	chrX	40083376	40083386	158	RUNX1	11.6826	
hse1000331	chrX	153948936	153948946	158	RUNX1	11.313	
hse1000512	chrX	40003387	40003397	158	RUNX1	13.3764	
hse1000512	chrX	40004732	40004742	158	RUNX1	6.6544	
hse1000512	chrX	40004974	40004984	158	RUNX1	9.39392	
hse1000512	chrX	40005069	40005079	158	RUNX1	10.8748	
hse1000512	chrX	40014538	40014548	158	RUNX1	5.42912	
hse1000520	chrX	40060422	40060432	158	RUNX1	12.8431	
hse1000520	chrX	40068200	40068210	158	RUNX1	9.90533	
hse1000520	chrX	40068220	40068230	158	RUNX1	8.33233	
hse1000594	chrX	154351194	154351204	158	RUNX1	11.2071	

Figure 6. Search TFBSs of a TF. A. In the search form, the use selects an organism (e.g. H. sapiens), input the name of a TF (e.g. RUNX1), and select a chromosome (e.g. chrX). B. A snapshot of the resulting TFBS table containing 2678 TFBSs of RUNX1 in chrX of H. sapiens.

(Figure 4D). Most of these TFBSs match those of known TF families, while a few need to be determined (TBD) for their cognate TFs (Figure 4D).

Second, using the 'select CRMs around a gene' function, the user can search in an organism (e.g. H. sapiens) all CRMs in the upstream and/or downstream regions (e.g. 0.5Mbp) of a given gene (SOX2) (Figure 5A). The search returns all CRMs in the interactive CRM list table (Figure 5B). As before, each CRM can be inspected in its information table by clicking on the CRM ID. In the example of the SOX2 gene of H. sapiens, a total of 102 CRMs on chr3 are returned in the CRM list table (Figure 5B). Inspection of the second CRM hse1002109 in the NCBI sequence viewer reveals that the CRM is located in the sixth and seventh intron, and spans the seventh exon, of the SOX2 gene. Interestingly, it overlaps two annotated enhancer sequences, as well as many ClinVar variants (Figure 5C). Clicking on the CRM ID hse1002109 in the right panel of the CRM information table (Figure 5C) displays the CRM's 1344 constituent TFBSs in the interactive TFBS table (Figure 5D). Some of these TFBSs match those of known TF families, while others need to be determined (TBD) for their cognate TFs.

Using the 'search TFBSs of a transcription factor' function, the user can retrieve all TFBSs of a given TF (e.g. RUNX1) in one or more selected chromosomes (e.g. chrX) in an organism (e.g. H. sapiens) (Figure 6A). The results are returned in the interactive TFBS table (Figure 6B). In the example of the TF RUNX1, a total of 2678 binding sites are found in chrX of H. sapiens.

Batch download

Using the Download function from the home page, the user can download all predicted CRMCs and constituent TFBSs in an organism in a file in BED format.

Future development

In the future, we will add predicted CRMCs and TFBSs in other important model organisms such as the worm (C. elegans) and the fly (D. melanogaster). We will also update the predictions in each organism when more data are available. We will add more information about the CRMCs, including their predicted functional states (active or non-active) of the CRMCs in various cell/tissue types, predicted target genes and causal variants of complex traits and diseases by integrating more data sources.

Conclusions

We have developed the PCRMS database that contains the most comprehensive collections of accurately predicted CRMs and constituent TFBSs in the human and mouse genomes. The web interface to PCRMS allows the user to browse, search and visualize the CRMs and constituent TFBSs. It also provides three functional analysis modules to search the closest CRM(s) to a gene, CRM(s) in a region around a gene and TFBSs landscape of a specific TF. The results can be inspected in interactive ways and exported in files in different formats. All the predicted CRMCs and TFBSs in an organism can be download in BED format. PCRMS will facilitate the research community's efforts to characterize the regulatory genomes in important organisms

Acknowledgements

The authors would like to acknowledge members from Office of Technical Service of College of Computing and Informatics at UNC Charlotte for security reviewing and deploying the databases.

Funding

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.

Conflict of interest

The authors declare that they have no competing interests.

Declarations

Ethics approval and consent to participate. Ethics approval is not applicable to this study.

Data availability

All predicted CRMCs of human and mouse can be freely downloaded at https://cci-bioinfo.uncc.edu.

Consent for publication

Not applicable.

Author contributions

Z.S. and P.N. conceived and designed the project. P.N. carried out the computational analysis and built the database and the web interface. Z.S. and P.N. wrote the manuscript. All the authors read and approved the final manuscript.

References

1. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, Amsterdam.
2. Hindorff,L.A., Sethupathy,P., Junkins,H.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
3. Ramos,E.M., Hoffman,D., Junkins,H.A. *et al.* (2014) Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147.
4. Wittkopp,P.J. and Kalay,G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, **13**, 59–69.
5. Rubinstein,M. and de Souza,F.S. (2013) Evolution of transcriptional enhancers and animal diversity. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **368**, 20130017.
6. Siepel,A. and Arbiza,L. (2014) Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.*, **29**, 81–89.
7. King,M. and Wilson,A. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
8. Maurano,M.T., Humbert,R., Rynes,E. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
9. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
10. Kilpinen,H., Waszak,S.M., Gschwind,A.R. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.
11. McVicker,G., van de Geijn,B., Degner,J.F. *et al.* (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–749.
12. Huang,D. and Ovcharenko,I. (2015) Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res.*, **43**, 225–236.
13. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
14. Pai,A.A., Pritchard,J.K. and Gilad,Y. (2015) The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.*, **11**, e1004857.
15. Schmidt,D., Wilson,M.D., Spyrou,C. *et al.* (2009) ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods*, **48**, 240–248.
16. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, **2010**, pdb–prot5384.
17. Buenrostro,J.D., Wu,B., Chang,H.Y. *et al.* (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **109**, 21–29.
18. Simon,J.M., Giresi,P.G., Davis,I.J. *et al.* (2012) Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat. Protoc.*, **7**, 256.
19. Schones,D.E., Cui,K., Cuddapah,S. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
20. Consortium EP. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
21. Consortium EP. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
22. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
23. Kundaje,A., Meuleman,W., Ernst,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
24. Consortium,G.T. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
25. Whitington,T., Frith,M.C., Johnson,J. *et al.* (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39**, e98.
26. Sun,H., Guns,T., Fierro,A.C. *et al.* (2012) Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res.*, **40**, e90.
27. Ha,N., Polychronidou,M. and Lohmann,I. (2012) COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PLoS One*, **7**, e52055.
28. Rohr,C.O., Parra,R.G., Yankilevich,P. *et al.* (2013) INSECT: IN-silico SEarch for Co-occurring Transcription factors. *Bioinformatics*, **29**, 2852–2858.
29. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
30. Ernst,J., Kheradpour,P., Mikkelsen,T.S. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
31. Hoffman,M.M., Buske,O.J., Wang,J. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
32. Andersson,R., Gebhard,C., Miguel-Escalada,I. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
33. Khan,A. and Zhang,X. (2016) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, **44**, D164–D171.
34. Jiang,Y., Qian,F., Bai,X. *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.*, **47**, D235–D243.

35. Ashoor,H., Kleftogiannis,D., Radovanovic,A. *et al.* (2015) DENdb: database of integrated human enhancers. *Database : j. biol. databases curation*, **2015**, bav085.
36. Dreos,R., Ambrosini,G., Cavin Perier,R. *et al.* (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.*, **41**, D157–D164.
37. Dimitrieva,S. and Bucher,P. (2013) UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, **41**, D101–D109.
38. Visel,A., Taher,L., Girgis,H. *et al.* (2013) A high-resolution enhancer atlas of the developing telencephalon. *Cell*, **152**, 895–908.
39. Fishilevich,S., Nudel,R., Rappaport,N. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : j. biol. databases curation*, **2017**, bax028.
40. Wang,J., Dai,X., Berry,L.D. *et al.* (2019) HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.*, **47**, D106–D112.
41. Cai,Z., Cui,Y., Tan,Z. *et al.* (2019) RAEdb: a database of enhancers identified by high-throughput reporter assays. *Database: j. biol. databases curation*, **2019**, bay140.
42. Wang,Z., Zhang,Q., Zhang,W. *et al.* (2018) HEDD: human enhancer disease database. *Nucleic Acids Res.*, **46**, D113–D120.
43. Zhang,G., Shi,J., Zhu,S. *et al.* (2018) DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res.*, **46**, D78–D84.
44. Wei,Y., Zhang,S., Shang,S. *et al.* (2016) SEA: a super-enhancer archive. *Nucleic Acids Res.*, **44**, D172–D179.
45. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
46. Moore,J.E., Purcaro,M.J., Pratt,H.E. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
47. Ni,P. and Su,Z. (2021) Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genom. Bioinform.*, **3**, lqab052.
48. Visel,A., Minovitsky,S., Dubchak,I. *et al.* (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
49. Mei,S., Qin,Q., Wu,Q. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
50. Li,Y., Ni,P., Zhang,S. *et al.* (2019) ProSampler: an ultra-fast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics*, **35**, 4632–4639.
51. Mendoza-Parra,M.A., Van Gool,W., Mohamed Saleem,M.A. *et al.* (2013) A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res.*, **41**, e196.
52. Marinov,G.K., Kundaje,A., Park,P.J. *et al.* (2014) Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, **4**, 209–223.
53. Devailly,G., Mantsoki,A., Michoel,T. *et al.* (2015) Variable reproducibility in genome-scale public data: a case study using ENCODE ChIP sequencing resource. *FEBS Lett.*, **589**, 3866–3870.
54. Hnisz,D., Abraham,B.J., Lee,T.I. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
55. Basso,K. and Dalla-Favera,R. (2012) Roles of BCL6 in normal and transformed germinal center B cells. *Immunol. Rev.*, **247**, 172–183.
56. Damm,F., Chesnais,V., Nagata,Y. *et al.* (2013) BCOR and BCORL1 mutations in myelodysplastic syndromes and related disorders. *Blood*, **122**, 3169–3177.