Published online 20 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D729–D735***
doi: 10.1093/nar/gky1094

Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis

Rongbin Zheng[1,†], Changxin Wan[1,†], Shenglin Mei[1], Qian Qin[1], Qiu Wu[1], Hanfei Sun[1], Chen-Hao Chen[2,3,4], Myles Brown[3,5], Xiaoyan Zhang[1,*], Clifford A. Meyer[2,3,*] and X. Shirley Liu[2,3,1,*]

[1]Shanghai Key Laboratory of Tuberculosis, Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA, [3]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [4]Biological and Biomedical Science Program, Harvard Medical School, Boston, MA 02115, USA and [5]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02215, USA

Received September 11, 2018; Revised October 18, 2018; Editorial Decision October 19, 2018; Accepted November 05, 2018

ABSTRACT

The Cistrome Data Browser (DB) is a resource of human and mouse cis-regulatory information derived from ChIP-seq, DNase-seq and ATAC-seq chromatin profiling assays, which map the genome-wide locations of transcription factor binding sites, histone post-translational modifications and regions of chromatin accessible to endonuclease activity. Currently, the Cistrome DB contains approximately 47,000 human and mouse samples with about 24,000 newly collected datasets compared to the previous release two years ago. Furthermore, the Cistrome DB has a new Toolkit module with several features that allow users to better utilize the large-scale ChIP-seq, DNase-seq, and ATAC-seq data. First, users can query the factors which are likely to regulate a specific gene of interest. Second, the Cistrome DB Toolkit facilitates searches for factor binding, histone modifications, and chromatin accessibility in any given genomic interval shorter than 2Mb. Third, the Toolkit can determine the most similar ChIP-seq, DNase-seq, and ATAC-seq samples in terms of genomic interval overlaps with user-provided genomic interval sets. The Cistrome DB is a user-friendly, up-to-date, and well maintained resource, and the new tools will greatly benefit the biomedical research community. The database is freely available at http://cistrome.org/db, and the Toolkit is at http://dbtoolkit.cistrome.org.

INTRODUCTION

Transcription factors (TFs) bind to cis-regulatory elements and regulate the transcription rates of genes through complex mechanisms, which involve the disruption of nucleosomes, the alteration of histone post-translational modifications, the recruitment or eviction of protein complexes, etc. (1). Cistromes, defined as genome-wide maps of the cis-regulatory binding sites of trans-acting factors, are invaluable for understanding the complex biology of gene regulation (2,3). Chromatin immunoprecipitation and DNA sequencing (ChIP-seq) experiments (4–7) targeting histones in particular post-translational modification states have revealed that histone marks can be used to identify promoters and enhancers (8,9), discriminate between repressive and activating regulatory states (8,10), and distinguish genes that are actively transcribed from silent ones (11). It has been estimated that over 1,600 TFs exist in human and mouse (12,13). As these are expressed in different combinations according to cell type and state, comprehensive mapping of these cistromes by ChIP-seq is an enormous challenge. DNase-seq (14) and ATAC-seq (15) are technologies developed to comprehensively map most of the TF binding sites in a biological sample through the characterization of regions that are accessible to DNase I or Tn5 transposase enzymatic activity. The raw sequencing data from tens of thousands of ChIP-seq, DNase-seq and ATAC-seq experiments, carried out by consortia such as ENCODE (16) and the Epigenomics Roadmap Project (17), as well as by individual research groups are publicly available in repositories such as GEO (18). The Cistrome Data Browser (DB) is a platform that extracts useful cis-regulatory informa-

*To whom correspondence should be addressed. Tel: +1 617 632 3012/3498; Fax: +1 617 632 2444; Email: cliff@jimmy.harvard.edu
Correspondence may also be addressed to Xiaoyan Zhang. Tel: +86 21 65980233; Fax: +86 21 65981041; Email: xyzhang@tongji.edu.cn
Correspondence may also be addressed to X. Shirley Liu. Tel: +1 617 632 3012/3498; Fax: +1 617 632 2444; Email: xsliu@jimmy.harvard.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

tion from these datasets and provides features that allow the biomedical research community to readily find and re-use this information (19).

Although there are other ChIP-seq databases, including ChIP-Atlas (BioRxiv: https://doi.org/10.1101/262899), ChIPBase (20) and ReMap (21), the Cistrome DB differs from these in terms of sample coverage, comprehensive quality control metrics, data browsing and querying capabilities, and downstream analysis functions. We reported the first version of the Cistrome DB in the 2017 Nucleic Acids Research database issue (19). Here, we present an updated version which doubles the original datasets (before 1 February 2018), including ∼25,000 human and 22,000 mouse samples. To increase the utility of these resources we have also implemented several Toolkit features for querying the Cistrome DB data. These new features allow users to find the predicted regulators of a specific gene, determine factors that bind to a specific genomic interval, and identify factors with similar cistromes to a user provided cistrome.

MATERIALS AND METHODS

Data collection

ChIP-seq, DNase-seq, and ATAC-seq samples were identified in the public databases: NCBI Gene Expression Omnibus (GEO), Encyclopedia of DNA Elements (ENCODE), and Roadmap Epigenetics Project. In the case of GEO, all sample identifiers (GSM ID) were obtained from the SRA database using the query '(homo sapiens[Organism] OR mus musculus[Organism])'. Sample XML files were downloaded from GEO and parsed to determine the species ('Organism'), and data type ('Library Strategy') based on 'ChIP-Seq' and 'DNase-seq' labels. Since ATAC-Seq data is usually labeled as 'OTHER' in library strategy, the Cistrome DB parser identified ATAC-seq data by matching the keywords in the GEO sample description text. Single-cell ATAC-seq data were excluded if they match terms such as 'scATAC-seq', 'single cell ATAC' etc, in the sample description.

Data processing and quality control

The data in these public databases were produced by numerous laboratories, and the processed results were derived using a variety of algorithms. To improve the consistency of Cistrome DB data, raw DNA sequence data for each sample was downloaded and uniformly processed by the ChiLin pipeline (22), which uses BWA (23) to map reads to the hg38 or mm10 genomes and MACS2 (24) to identify statistically significant peaks. The raw data of SRA file was downloaded from NCBI at ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/. We obtained FASTQ files from SRA files using the fastq-dump software (https://ncbi.github.io/sra-tools/fastq-dump.html). Motif scanning was also performed on transcription factor or chromatin regulator ChIP-seq samples based on enrichment of the motif sequence relative to the center of the peaks (25). Target genes were predicted from ChIP-seq peaks using the regulatory potential model which weighs the impact of each peak by exponential decay of distance to gene transcription start site (TSS) (26). Additional information about these data can be found on the Cistrome DB document page at http://cistrome.org/db/#/documents.

Cistrome DB data quality controls include six metrics, representing DNA sequencing quality, ChIP quality, and genomic distribution characteristics. Read quality is based on the median FASTQ read quality, mapping quality is measured by the percentage of reads that each map to a unique genomic locus, and the PCR bottleneck coefficient (PBC) is used to estimate the rate of read duplication through PCR amplification (27,28). The fraction of non-mitochondrial reads in peak regions (FRiP) and the number of peaks with 10-fold enrichment are used to reflect the quality of the ChIP experiment (27,28). A union of DNase hypersensitive sites (Union DHS) was summarized using a large collection of DNase-seq samples from the Cistrome DB (19,29). The percentage of peaks that overlap with the union of DHS sites is used to characterize the data quality based on the genomic distribution of the peaks. Although most TFs and chromatin associated factors tend to bind at DHS sites, some histone marks and factors do not follow this trend. Cutoffs were determined based on the distribution of these quality control metrics in the Cistrome DB (22), and a red dot indicates data with lower quality on a metric while a green dot indicates higher quality of a sample (Figure 1). These QC measures are meant to guide users in their appraisal of data, instead of being used strictly to categorize samples as pass or fail. Although the Cistrome DB includes some samples which appear to be of poor quality by several metrics, these samples may nevertheless hold valuable clues to some aspect of regulatory biology not represented by other samples in the database.

Toolkit development

To enhance the usage of Cistrome DB data, three new 'Toolkit' functionalities have been developed. These can be accessed through a link on the Cistrome DB webpage or at the URL: http://dbtoolkit.cistrome.org. The first function addresses the question: 'What factors regulate your gene of interest?' The assignment of TFs to genes is based on regulatory potential scores that reflect the collective influence of the binding sites of a given TF on genes nearby these sites (30), and assume that TF binding sites near the TSS are more likely to regulate the gene than those further away. As different TFs may regulate genes over different ranges of genomic influence, short (1 kb), mid-range (10 kb) and long-range (100 kb) influence scores are calculated for each TF. These distances represent the exponential decay parameter to estimate the impact of each TF binding site by its distance to TSS. To focus on high quality and high confidence peaks, only peaks with 5-fold enrichment over background were used in these RP score calculations. As the total number of peaks varies between samples and this number influences the RP scores, the RP scores for each sample were standardized to fit into a range between 0 and 1 to enable cross-sample comparison. Through the interactive web interface, users can input a coding gene name and select the required parameters (species, distance). The Cistrome DB Toolkit queries RP scores across all the samples and returns samples, ranked based on the RP score for this gene.

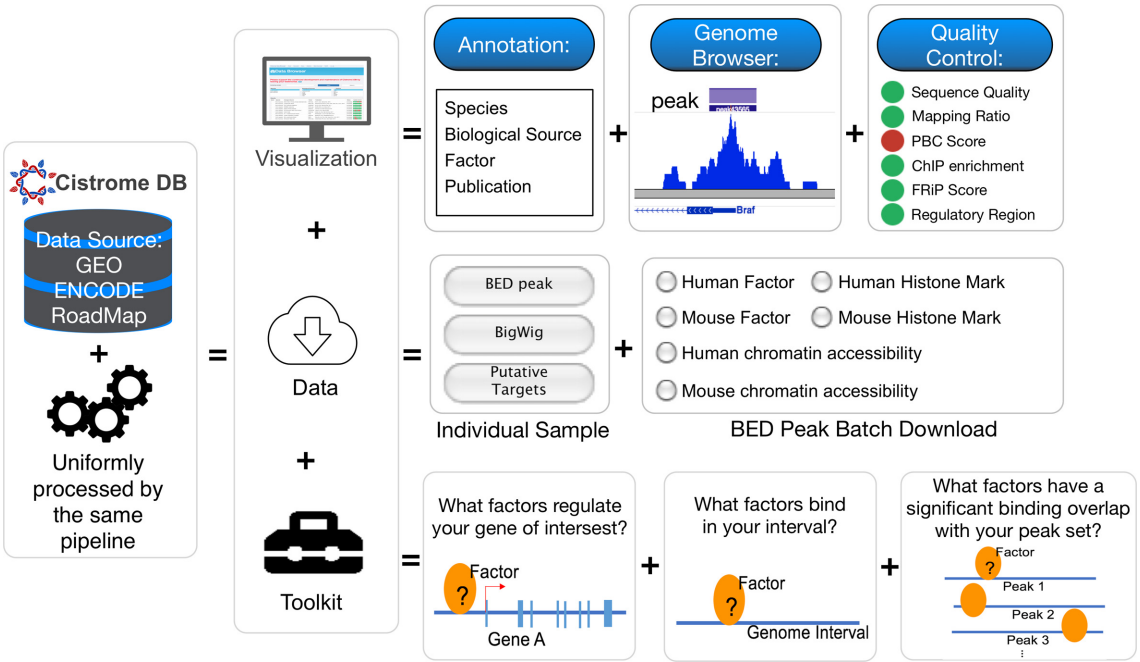

Figure 1. Overall design of Cistrome Data Browser and Toolkit. Cistrome DB incorporates publicly available ChIP-seq, DNase-seq, and ATAC-seq data collected from Gene Expression Omnibus (GEO), Encyclopedia of DNA Elements (ENCODE), and RoadMap Epigenomics. Cistrome DB provides sample annotations and uniformly processed results that allow for comparisons of peaks, signal files, quality control metrics, motifs and imputed target genes. To easily access Cistrome DB data, users can conveniently visualize BigWig files in genome browsers and download peak BED files and putative target gene results. The new Toolkit module includes functionalities that answer three questions: 'What factors regulates your gene of interest?', 'What regulator bind in your interval?' and 'What factors have a significant binding overlap to your peak set?'

Two additional Cistrome DB Toolkit functions were developed to address the questions: 'What factors bind in your interval?' and 'What factors have a significant binding overlap with your peak set?' The GIGGLE algorithm (31), with high speed and accuracy, is used to search and compare Cistrome samples with the user defined intervals or peak sets. Only samples which have >1000 five-fold enriched peaks were used to build the GIGGLE search index. Further details about the Cistrome DB toolkit can be found at http://dbtoolkit.cistrome.org/document.

RESULTS

Design of the Cistrome DB

The Cistrome DB concentrates on collecting publicly available ChIP-seq, DNase-seq and ATAC-seq data in human and mouse and providing functionalities to yield useful insights from the collected data (Figure 1). Cistrome DB users can search published ChIP-seq or chromatin accessibility data by factor, biological source (cell line, cell type and tissue type), and species. Sample quality control reports are available and the quality of multiple samples can be assessed simultaneously by green and red dots which indicate high and low quality control metrics, respectively. Visualization of multiple samples is provided through the UCSC Genome Browser (32,33) and the WashU Epigenome Browser (34). In addition, users can conveniently download peaks from one particular sample or from a bulk collection. In terms of downstream analysis, Cistrome DB predicts target genes and evaluates motif enrichments for transcription factor ChIP-seq data. The Cistrome DB Toolkit is a new module which enables better re-use of the data collection.

Integration of data sources

The total number of human and mouse samples in the Cistrome DB has grown steadily since 2008 (Figure 2A). In the current collection (February in 2018), the Cistrome DB incorporates ~25,000 human and ~22,000 mouse samples, which doubles the number of samples in the last release (19). This collection not only increases the sample size in the trans-factor/histone mark ChIP-seq, and chromatin accessibility in human and mouse, but also increases the types of factors and histone marks (Figure 2B and C). The current Cistrome DB contains ~1,700 factors and 132 histone marks/variants in human, and 965 factors and 120 histone marks/variants in mouse (Figure 2B). Examples of new factors include ZBTB48 (35,36) and ZMYM3 (37) in human, and SPEN and TERF2IP (38) in mouse; and examples of new histone modifications / variants include H3F3A (37) and H2AFZ (37) in human, and H3K9BHB (39) and H2BK5me1 (40) in mouse (Figure 2D). The new data in the Cistrome DB is of a similar high quality as the previous collection (Figure 2E), as evident from the number of highly enriched peaks and the overlap with the union of DHS sites (Figure 2E).

Query, visualization, and download

The Cistrome DB provides a drop-down menu to find samples with certain annotations, such as TF name, histone

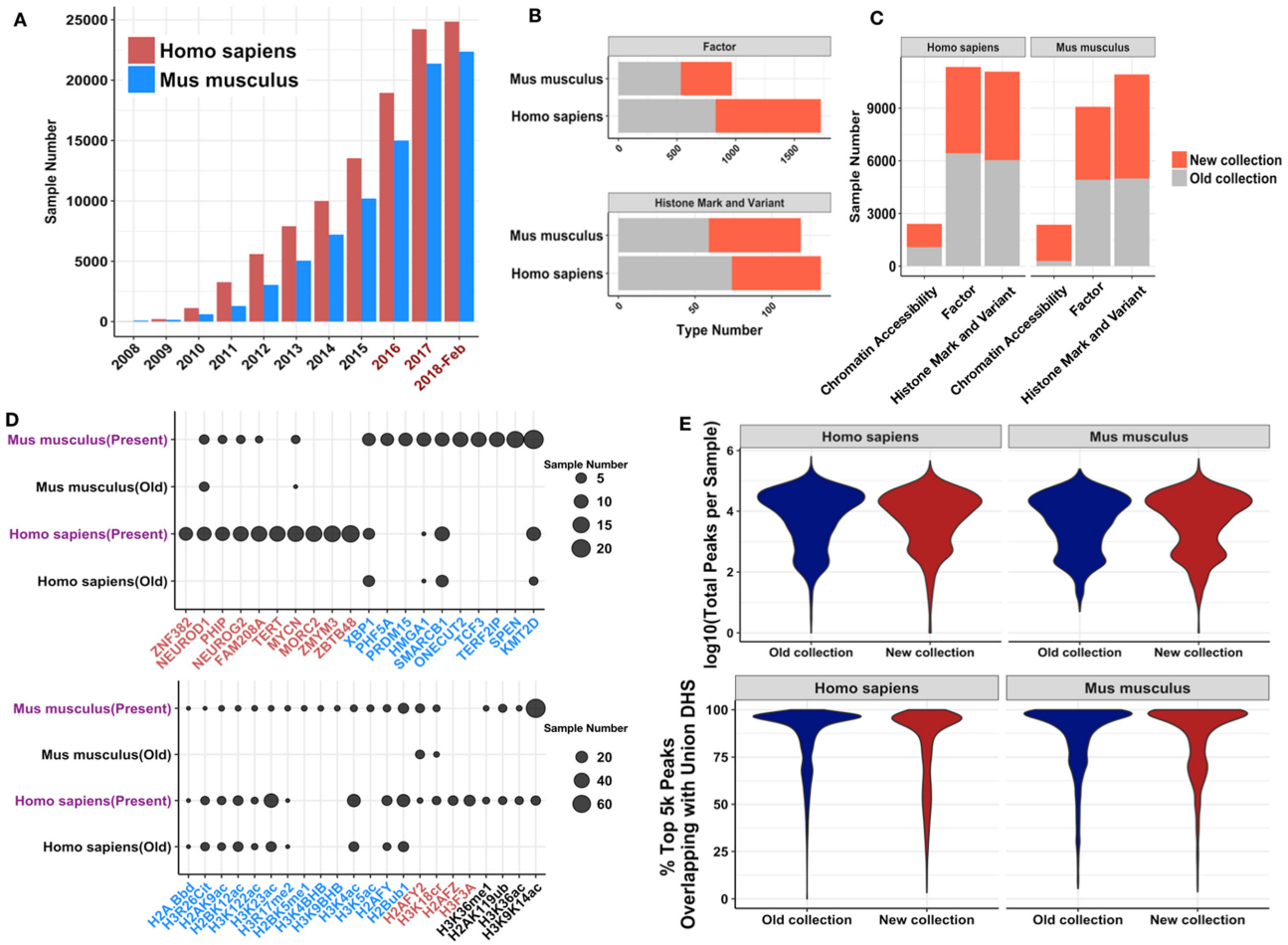

Figure 2. Quantity and quality of new Cistrome DB data. (**A**) Cumulative size of human and mouse data collection by year. Collection years before the last Cistrome DB release are shown in black, while new collection years are red. (**B**, **C**) In the new collection, Cistrome DB increased not only the sample number of each data type, but also the types of factor and histone marks and variants. (**D**) The TFs (upper) and histone marks or variants (lower) with the most new samples. Blue labels on the x-axis indicate new factors for mouse; red labels indicate new factors for human, and black labels for factors that are novel in Cistrome DB for both human and mouse. (**E**) Violin plots showing an overview of data quality for old and new collections. Total peak numbers on the $\log_{10}$ scale and the percentage of peaks overlapping with a union of DNase hypersensitive sites (DHS) were calculated.

modification, cell line, cell type, and tissue type. Alternatively, users can directly search for Cistrome DB data by typing keywords. After finding relevant samples and filtering using quality control metrics, users can visualize sample batches on the WashU Epigenome Browser and on the UCSC Genome Browser. The Cistrome DB also displays the enrichment levels of known and *de novo* motifs with a sequence logo for each transcription factor and chromatin regulator ChIP-seq sample in the collection. A list of genes that are predicted to be directly regulated by the factor is provided for ChIP-seq samples, and users can further search by gene name to check whether a given gene can be targeted by the factor. Bulk download of peak files of many samples is supported, which could be a useful resource for computational groups.

Cistrome DB Toolkit

The Cistrome DB Toolkit was designed to help users easily extract useful *cis*-regulatory information from the large collection of Cistrome DB data. In this module, we provide tools to address three questions that are likely to be of interest to many users. The first tool addresses the question: 'What factors regulate your gene of interest?' This function returns a list of the transcription factors in the Cistrome DB that are the most likely regulators of a query gene based on the positions of transcription factor ChIP-seq peaks relative to the transcription start site. As an example, we asked what regulators target the human Androgen Receptor (AR) gene. To include long-range enhancer effects in this case, we set the distance influence parameter to 100 kb. The top factors returned by the Toolkit function are GATA2, AR, ERG, FOXA1, PIAS1, consistent with the known regulators of AR (41–43) (Figure 3A).

The second tool answers the question: 'What factors bind in your interval?' This function identifies TF binding, histone modifications, and chromatin accessibility in any query genomic interval shorter than 2Mb. As an example, we queried an interval with known distal enhancers of the *AR* gene (chrX:66,897,958–66,908,958 hg38) in human prostate cancer cells (44). Since the number of peaks varies between different ChIP-seq samples, the number of peaks in this in-

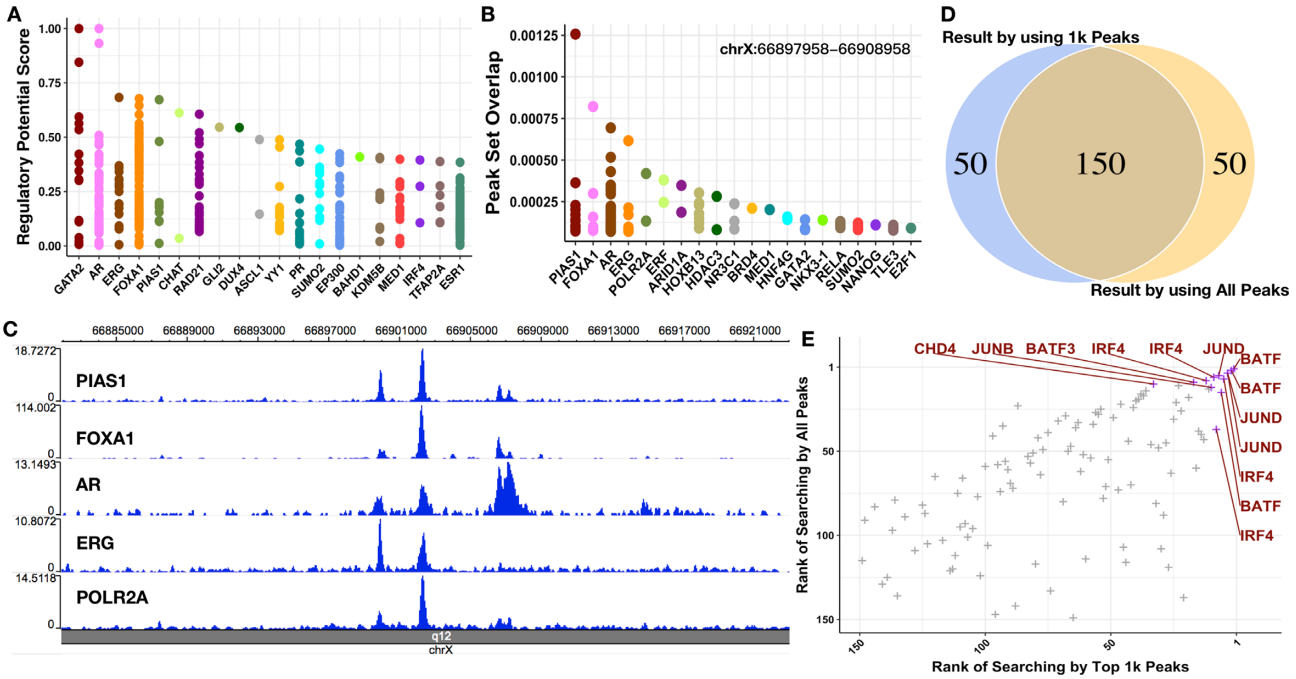

Figure 3. Cistrome DB Toolkit. (**A**) An example of the first Cistrome DB Toolkit function, showing putative regulators of the human androgen receptor (*AR*) gene. A parameter of 100kb regulatory potential decay rate was selected to include long-range enhancers of AR. Each dot in this figure represents a ChIP-seq sample. The x-axis includes the top 20 factors, ranked by the maximum regulatory potential score over all ChIP-seq samples representing each factor. (**B**) The second Toolkit function was used to discover the TFs binding to a known *AR* enhancer (chrX:66,897,958-66,908,958, Hg38) in prostate cancer. For each sample, the number of peaks overlapping with the interval divided by the total number of peaks in the sample was calculated, and shown on the x axis. The top 200 samples were plotted, categorized by factor on the x-axis. (**C**) WashU Epigenome Browser tracks of the 5 top-most ranked samples from panel B show the peaks within the examined genomic region. (**D**, **E**) Cistromes in the Cistrome DB similar to peaks of an input BATF peak set as determined by the third Toolkit function. The top-most 200 samples detected using two parameter choices (Cistrome DB top 1000 peaks or all peaks) are compared by Venn Diagram in D and by scatter plot in E. The Venn Diagram in D shows that 150 samples out of the top 200 samples are common to both parameter choices. The scatter plot in E depicts the rank comparison of the overlapping top 150 samples, and the TFs represented by the top ten samples are labeled with the TF name.

terval divided by the total number of peaks for the factor is used to rank the result. The top factors returned by the Toolkit function are PIAS1, FOXA1, AR, ERG, POLR2A, etc (Figure 3B). The WashU Epigenome Browser view (45,46) (Figure 3C) shows the binding peaks within this enhancer, which can help determine the functional sequence and the factors bound to this sequence.

The third tool answers the question: 'What factors have a significant binding overlap with your peak set?'. This function compares the strongest peaks in each cistrome with the peak set provided by the user. Users can upload their own set of genomic intervals, such as a ChIP-seq peak set in a BED file format. The function then identifies the samples in the Cistrome DB that have the most significant peak overlaps with the input, which might be cofactors, histone marks, or chromatin accessibility profiles associated with the input sample. We tested this function using ChIP-seq peaks of BATF (GSM1370277) (47), and compared the results using either the top one thousand peaks or all the peaks in each Cistrome DB sample. The top 200 hits in the results using the two options share 150 common samples (Figure 3D), including ChIP-seq samples of BATF, JUND, IRF4, JUNB, BATF3 and other factors that are known to co-bind with BATF (48,49) (Figure 3E).

DISCUSSION

We report an update of the Cistrome DB which includes an expanded data collection and new functionalities. Users can search by keyword or by drop-down menu for any factor they are interested in, and evaluate the quality of the data and the characteristics of the resulting cistromes. In addition, users can find informative data using the new Toolkit functions which are based on genomic binding patterns rather than metadata annotations. This way of finding data can lead to new hypotheses regarding cis-elements or trans-factors that might be functionally associated with the user input on gene regulation. The Cistrome DB is currently the most comprehensive resource for searching, visualizing, and exploring publicly available ChIP-seq and chromatin accessibility data of human and mouse. Because it is based on the collection of public data and relies on the automatic parsing of sample metadata from data source, occasional mis-annotation, incompleteness or ambiguity in the system is unavoidable. Correction of these types of error will require involvement from the community, especially the data contributors, and we are working on developing the web interface for users to conveniently correct meta-data errors. In the future, the Cistrome DB team will continue to collect all newly produced ChIP-seq and chromatin accessibility data, but will prioritize factors and histone modifications

that are less well represented in the existing collection. In addition, we will explore the use of long-range chromatin interaction data, such as those available at The 3D Genome Browser (50) to improve TF target predictions. We hope that an awareness of the available data in the Cistrome DB will lead data producers to explore factors and cell types that are not well represented and thereby enrich the diversity and utility of cistromes. We will continue to maintain the database, incorporate new data, and develop new features into the Cistrome DB, to help accelerate the investigations and understanding of gene regulatory mechanisms in biological processes and diseases.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr Zhiping Weng for providing backup of the Cistrome DB and Dr Ting Wang for the WashU Epigenome Gateway Browser.

FUNDING

National Key Research and Development Program of China [2016YFC1303200 to X.Z., 2017YFC0908500 to X.S.L.]; National Natural Science Foundation of China [31801110 to S.M., 81573023 to X.Z.]; National Institutes of Health of US [U24 HG009446 and U01 CA180980 to X.S.L.]. Funding for open access charge: National Institutes of Health of US [U24 HG009446 and U01 CA180980 to X.S.L.].
Conflict of interest statement. None declared.

REFERENCES

1. Lelli,K.M., Slattery,M. and Mann,R.S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, **46**, 43–68.
2. Liu,T., Ortiz,J.A., Taing,L., Meyer,C.A., Lee,B., Zhang,Y., Shin,H., Wong,S.S., Ma,J., Lei,Y. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.
3. Mei,S., Meyer,C.A., Zheng,R., Qin,Q., Wu,Q., Jiang,P., Li,B., Shi,X., Wang,B., Fan,J. *et al.* (2017) Cistrome cancer: a web resource for integrative gene regulation modeling in cancer. *Cancer Res.*, **77**, e19–e22.
4. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
5. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
6. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.-K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
7. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
8. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
9. Wang,S., Zang,C., Xiao,T., Fan,J., Mei,S., Qin,Q., Wu,Q., Li,X., Xu,K., He,H.H. *et al.* (2016) Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Res.*, **26**, 1417–1429.
10. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
11. Karlić,R., Chung,H.-R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2926–2931.
12. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
13. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
14. 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**, doi:10.1101/pdb.prot5384.
15. Buenrostro,J.D., Wu,B., Chang,H.Y. and Greenleaf,W.J. (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **109**, doi:10.1002/0471142727.mb2129s109.
16. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
17. Chadwick,L.H. (2012) The NIH Roadmap Epigenomics Program data resource. *Epigenomics*, **4**, 317–324.
18. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
19. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
20. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
21. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
22. Qin,Q., Mei,S., Wu,Q., Sun,H., Li,L., Taing,L., Chen,S., Li,F., Liu,T., Zang,C. *et al.* (2016) ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics*, **17**, 404.
23. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
24. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
25. Meyer,C.A., He,H.H., Brown,M. and Liu,X.S. (2011) BINOCh: binding inference from nucleosome occupancy changes. *Bioinformatics*, **27**, 1867–1868.
26. Tang,Q., Chen,Y., Meyer,C., Geistlinger,T., Lupien,M., Wang,Q., Liu,T., Zhang,Y., Brown,M. and Liu,X.S. (2011) A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res.*, **71**, 6940–6947.
27. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
28. Bailey,T., Krajewski,P., Ladunga,I., Lefebvre,C., Li,Q., Liu,T., Madrigal,P., Taslim,C. and Zhang,J. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.
29. He,H.H., Meyer,C.A., Hu,S.S., Chen,M.-W., Zang,C., Liu,Y., Rao,P.K., Fei,T., Xu,H., Long,H. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.
30. Wang,S., Sun,H., Ma,J., Zang,C., Wang,C., Wang,J., Tang,Q., Meyer,C.A., Zhang,Y. and Liu,X.S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.*, **8**, 2502–2515.

31. Layer,R.M., Pedersen,B.S., DiSera,T., Marth,G.T., Gertz,J. and Quinlan,A.R. (2018) GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods*, **15**, 123–126.
32. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
33. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
34. Zhou,X. and Wang,T. (2012) Using the Wash U Epigenome Browser to examine genome-wide sequencing data. *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi1010s40.
35. Jahn,A., Rane,G., Paszkowski-Rogacz,M., Sayols,S., Bluhm,A., Han,C., Draškovič,I., Londoño-Vallejo,J.A., Kumar,A.P., Buchholz,F. *et al.* (2017) ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.*, **18**, 929–946.
36. Schmitges,F.W., Radovani,E., Najafabadi,H.S., Barazandeh,M., Campitelli,L.F., Yin,Y., Jolma,A., Zhong,G., Guo,H., Kanagalingam,T. *et al.* (2016) Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.*, **26**, 1742–1752.
37. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic. Acids. Res.*, **46**, D794–D801.
38. Martínez,P., Gómez-López,G., Pisano,D.G., Flores,J.M. and Blasco,M.A. (2016) A genetic interaction between RAP1 and telomerase reveals an unanticipated role for RAP1 in telomere maintenance. *Aging Cell*, **15**, 1113–1125.
39. Xie,Z., Zhang,D., Chung,D., Tang,Z., Huang,H., Dai,L., Qi,S., Li,J., Colak,G., Chen,Y. *et al.* (2016) Metabolic regulation of gene expression by histone lysine β-Hydroxybutyrylation. *Mol. Cell*, **62**, 194–206.
40. Vian,L., Pękowska,A., Rao,S.S.P., Kieffer-Kwon,K.-R., Jung,S., Baranello,L., Huang,S.-C., El Khattabi,L., Dose,M., Pruett,N. *et al.* (2018) The energetics and physiological impact of cohesin extrusion. *Cell*, **173**, 1165–1178.
41. He,B., Lanz,R.B., Fiskus,W., Geng,C., Yi,P., Hartig,S.M., Rajapakshe,K., Shou,J., Wei,L., Shah,S.S. *et al.* (2014) GATA2 facilitates steroid receptor coactivator recruitment to the androgen receptor complex. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 18261–18266.
42. Grad,J.M., Lyons,L.S., Robins,D.M. and Burnstein,K.L. (2001) The androgen receptor (AR) amino-terminus imposes androgen-specific regulation of AR gene expression via an exonic enhancer. *Endocrinology*, **142**, 1107–1116.
43. Toropainen,S., Malinen,M., Kaikkonen,S., Rytinki,M., Jääskeläinen,T., Sahu,B., Jänne,O.A. and Palvimo,J.J. (2015) SUMO ligase PIAS1 functions as a target gene selective androgen receptor coregulator on prostate cancer cell chromatin. *Nucleic Acids Res.*, **43**, 848–861.
44. Takeda,D.Y., Spisák,S., Seo,J.-H., Bell,C., O'Connor,E., Korthauer,K., Ribli,D., Csabai,I., Solymosi,N., Szállási,Z. *et al.* (2018) A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell*, **174**, 422–432
45. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
46. Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.*, **1418**, 335–351.
47. Care,M.A., Cocco,M., Laye,J.P., Barnes,N., Huang,Y., Wang,M., Barrans,S., Du,M., Jack,A., Westhead,D.R. *et al.* (2014) SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. *Nucleic Acids Res.*, **42**, 7591–7610.
48. Man,K., Gabriel,S.S., Liao,Y., Gloury,R., Preston,S., Henstridge,D.C., Pellegrini,M., Zehn,D., Berberich-Siebelt,F., Febbraio,M.A. *et al.* (2017) Transcription factor IRF4 promotes CD8+ T cell exhaustion and limits the development of Memory-like T Cells during chronic infection. *Immunity*, **47**, 1129–1141.
49. Kurachi,M., Barnitz,R.A., Yosef,N., Odorizzi,P.M., DiIorio,M.A., Lemieux,M.E., Yates,K., Godec,J., Klatt,M.G., Regev,A. *et al.* (2014) The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. *Nat. Immunol.*, **15**, 373–383.
50. Wang,Y., Song,F., Zhang,B., Zhang,L., Xu,J., Kuang,D., Li,D., Choudhary,M.N.K., Li,Y., Hu,M. *et al.* (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.*, **19**, 151.