Quantitative Biology
https://doi.org/10.1007/s40484-020-0204-7

PROTOCOL AND TUTORIAL

Cistrome Data Browser and Toolkit: analyzing human and mouse genomic data using compendia of ChIP-seq and chromatin accessibility data

Rongbin Zheng[1,2,†], Xin Dong[1,2,†], Changxin Wan[1,2], Xiaoying Shi[1,2], Xiaoyan Zhang[2,*], Clifford A. Meyer[3,4,*]

[1] Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Science and Technology, Tongji University, Shanghai 200433, China
[2] Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai 200092, China
[3] Department of Data Science, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
[4] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA
* Correspondence: xyzhang@tongji.edu.cn, cliff_meyer@mail.dfci.harvard.edu

Received December 12, 2019; Revised January 14, 2020; Accepted January 21, 2020

The Cistrome Data Browser (DB) at the website (cistrome.org/db) provides about 56,000 published human and mouse ChIP-seq, DNase-seq, and ATAC-seq chromatin profiles, which we have processed using uniform analysis and quality control pipelines. The Cistrome DB Toolkit at the website (dbtoolkit.cistrome.org) was developed to allow users to investigate fundamental questions using this data collection. In this tutorial, we describe how to use the Cistrome DB to search for publicly available chromatin profiles, to assess sample quality, to access peak results, to visualize signal intensities, to explore DNA sequence motifs, and to identify putative target genes. We also describe the use of the Toolkit module to seek the factors most likely to regulate a gene of interest, the factors that bind to a given genomic interval (enhancer, SNP, etc.), and samples that have significant peak overlaps with user-defined peak sets. This tutorial guides biomedical researchers in the use of Cistrome DB resources to rapidly obtain valuable insights into gene regulatory questions

Keywords: ChIP-seq; chromatin accessibility; gene regulatory analysis; transcription factor

INTRODUCTION

Chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) is a widely used technique for studying genome-wide DNA-protein interactions and histone [1–3]. The DNase I hypersensitivity (DNase-seq) [4] and transposase-accessible chromatin (ATAC-seq) [5] assays facilitate genome-wide mapping of accessible chromatin, which reflects potential cis-regulatory elements bound by trans-acting factors [6]. ChIP-seq, DNase-seq, and ATAC-seq experiments are being carried out to acquire information about the complex biology of gene regulation. The Encyclopedia of DNA Elements (ENCODE) Consortium [7] and NIH Roadmap Epigenomics Project [8] have generated many high-quality ChIP-seq samples, targeting various transcription factors (TF) and histone marks, as well as DNase-seq samples in many cell and tissue types. Besides these projects, a large quantity of ChIP-seq, DNase-seq, and ATAC-seq data has been deposited in the NCBI Gene Expression Omnibus (GEO) [9], another invaluable resource for gene regulation research. Such data generated by consortiums or individual groups is hard to integrate due to the non-unified metadata annotation, processing, quality control,

[†] These authors contributed equally to this work.

and downstream analyses. The Cistrome Data Browser [10] was developed to overcome these challenges and is the most comprehensive database that provides re-annotated, uniformly processed, and quality-controlled human and mouse ChIP-seq, DNase-seq, and ATAC-seq datasets [8]. Cistrome DB is a user-friendly online platform for the query and visualization of published ChIP-seq, DNase-seq, and ATAC-seq samples.

Since TFs play a crucial role in gene regulation [11,12], public ChIP-seq data downloaded from databases, such as Cistrome DB, has been successfully re-utilized to explain differential expression [13] and identify the potential regulators that bind to a genomic interval or target a gene [14]. With the integration of public ChIP-seq data, researchers have successfully constructed gene regulatory networks [15] and predicted the functional regulatory elements genome-wide [16]. However, it is still difficult for many experimental biologists who may not have bioinformatics expertise to re-use the public ChIP-seq, DNase-seq, and ATAC-seq data for gene regulatory analysis. Cistrome DB Toolkit is a module that was developed to find regulators that potentially target a gene or bind to intervals based on the collection of cistrome in the database. The cistrome is defined as a set of cis-acting targets of a trans-acting factor on a genome-wide scale.

To assist users, here we describe a tutorial for the Cistrome DB and Toolkit. We show how to quickly find the publicly available ChIP-seq / DNase-seq / ATAC-seq samples in human and mouse using the Cistrome DB. We also describe how to use the Cistrome DB Toolkit to prioritize the factors that are likely to regulate a gene of interest, explain how to find the factors or histone marks that have binding on a given genomic interval, and how to search for ChIP-seq / DNase-seq / ATAC-seq profiles that significantly overlap to a given peak set.

MATERIALS AND METHODS

Data collection and processing of Cistrome DB

To efficiently collect the rapidly accumulating public ChIP-seq, DNase-seq, and ATAC-seq data, an automated parser script was written in Python to identify and collect such samples from NCBI GEO and other public repositories. The collected metadata was saved in the MySQL database, which allows convenient and fast data retrieval. Next, the raw data (FASTQ file) were downloaded and processed through unified pipelines, "ChiLin" [17] and its updated version "Chips". The outputs of each processed sample include a mapping result, quality control metrics, peaks, signal intensities, a conservation plot, motif scanning results (for TFs ChIP-seq), and putative targets.

Data preparation for Cistrome DB Toolkit

There are three functions in Cistrome DB Toolkit for re-utilizing the collected data. The first function, "what factors regulate your gene of interest," allows users to rank ChIP-seq samples by the likelihood of regulation on a given gene. Regulatory potential (*RP*) score has been demonstrated as an efficient method to find putative targets based on ChIP-seq peaks and reflect the regulation likelihood [18]. *RP* score is a weighted sum of peak contributions, where the weights decay exponentially with distance to the TSS [19]. The calculation formula is:

$$RP_g = \sum_{i=1}^{k} 2^{-\frac{d_i}{d_0}}$$

where the k means all the binding sites around TSS of a gene g in a range of $15 \times d_0$, the d_0 is a parameter that controls the relative distance to TSS where the weight is decreased to half of the highest value, the d_i means the distance between i-th binding site and the TSS. It has been known that there are factors binding on the proximal promoter region and also other factors binding on distal enhancers. We, therefore, set the range d_0 to be 1 Kb, 10 Kb, and 100 Kb to calculate the *RP* score for each sample and each gene. Since the total peak number of the ChIP-seq sample varies, we scaled the genome-wide *RP* score to the same range from 0 to 1 using the minimum-maximum normalization method. These calculations produce an *RP* score matrix in which rows are transcripts, and columns are ChIP-seq samples of Cistrome DB. With this data matrix, Cistrome DB Toolkit can conveniently query the gene or transcript of interest and rank all available ChIP-seq samples based on the normalized *RP* score vector.

In the second and third functions of Cistrome DB Toolkit, we are answering what regulators bind to a genomic interval or enrich on a set of intervals. To compare the user-defined intervals with Cistrome DB sample peaks, we use an efficient search engine for a large scale genomic loci, GIGGLE [20]. Firstly, we build the GIGGLE index on peaks of Cistrome DB ChIP-seq samples. When users input a genomic interval in the second function, "what factors bind in your interval", or the third function, "what factors have a significant binding overlap with your peak set", Cistrome DB Toolkit handles the searching or comparing jobs by "giggle search" function. To include high-confident peaks, we only include peaks whose fold-enrichment to the background is greater than 5. Since it would take a longer time to run GIGGLE on all Cistrome DB peaks, we built the GIGGLE index using 1,000 and 10,000 peaks with the highest fold-enrichment to provide options that would reduce the running time but with comparable results.

TUTORIAL

Design of Cistrome DB and Toolkit

Cistrome DB provides over 56,000 published human and mouse ChIP-seq, DNase-seq, and ATAC-seq data that were collected from NCBI Gene Expression Omnibus (GEO) [9], Encyclopedia of DNA Elements (ENCODE) [7], and Roadmap Epigenetics Project [8]. The main page of Cistrome DB contains four panels, including searching, results, inspector, and tools. The searching panel allows users to query samples through keywords, species, factor names, and biological sources. The result panel shows a list of samples that meet the queries of users. Furthermore, we designed each sample an inspector panel to show meta-annotation and quality control (QC) reports. The tools panel provides downstream analysis results, including motif scanning and putative targets. Besides, Cistrome DB allows users to conveniently visualize peaks and signal intensities through both UCSC Genome Browser [21] and Wash U Epigenome Browser [22]. Cistrome DB also provides download functions for peak files and putative targets. To better utilize Cistrome DB resource, Cistrome DB Toolkit was developed to address the following three questions: 1) "What factors regulate your gene of interest?" 2) "What regulator binds in your interval?" and 3) "What factors have a significant binding overlap to your peak set?"

Quality control (QC) of Cistrome DB samples

To describe sample quality, Cistrome DB uses three metrics related to DNA sequence read analysis and three metrics that relate more to peaks. The read-based metrics include the FastQC score to assess raw sequence quality, the mapping rate to indicate the mapping ability to the genome, and the PCR bottleneck coefficient (PBC) to elucidate the diversity of the sequencing library. Lower PBC values can indicate technique problems, such as PCR bias. The peak-based metrics include FRiP score, the number of high-quality peaks with 10-fold enrichment over background, and overlap ratio to union DNase-I hypersensitive sites (uDHS). The FRiP score assesses the ChIP-seq signal to noise ratio, and is defined as the fraction of mapped or usable reads that locate in the called peaks. Besides, it has been accepted that transcriptional regulator bindings generally happen on chromatin accessible regions [5,23,24]. Therefore, we calculate the overlap rate of ChIP-seq peaks to the uDHS as a peak-based quality control metric. It is expected that the high-quality ChIP-seq samples of transcriptional regulators and active histone marks have high overlap rates to uDHS.

To help users easily distinguish high quality samples from poor, we set thresholds for all six metrics (Table 1) based on the distribution of the QC metric of collected samples (detailed description at the website (cistrome.org/db/#/about)).). Each sample is labeled by six dots that corresponding to the six QC metrics. The green dots denote the sample that has passed the threshold, while red dots label samples that are failed to pass the quality control threshold (Fig. 1A-e, B-b; Table 1). Cistrome DB also shows the complete quality report in the "QC Reports" panel describing the actual value of the six quality control metrics.

It has been reported that transcription factor binding sites are highly conserved [28,29]. To help describe sample quality in this aspect, a conservation plot is provided for each sample, which we draw by summarizing the PhastCons value [29] around summits of most significant 5,000 peaks (Fig. 1C-d). Moreover, since the distribution of transcriptional factor bindings and histone modifications on genome can be different, we calculate the proportion of ChIP-seq peaks in promoter, exon, intron, and intergenic regions (Fig. 1C-c) for each sample. Users can further determine the sample quality based on the reasonability of peak distribution on genome.

Sample queries in Cistrome DB

Cistrome DB provides two ways to search for sample based on metadata. First, users can enter keywords related to biological sources, factor names, or GEO accessions to query ChIP-seq/DNase-seq/ATAC-seq samples of interest (Fig.1A-a). Second, users can search for samples by clicking the selectors with a combination of species, biological sources, and factor names before or after the keyword search (Fig. 1A-b). To view only samples with QC metrics indicative of high quality, users can use the drop down "Options" tab and click "Samples passing peak quality control" to filter out samples that are failed to pass peak quality control (FRiP score, 10-fold peak number, and overlap rate to uDHS) or click "Sample passing all quality control" to get samples that all quality control metrics (both read-based and peak-based) are passed (Fig. 1A-c). It is not rare that many redundant samples are available for the same factor in same biological source. In such a scenario, it is suggested to prioritize samples by all six QC metrics. However, there are factors that only have limited number of samples in our database, users can more focus on peak-based quality control if most samples are poor in read-based quality control. As the searched result can be a long list, Cistrome DB demonstrates at most 20 samples on each page. Users can turn to other pages by entering the page number or clicking the "Next" or "Prev" button at the end of the result panel (Fig. 1A).

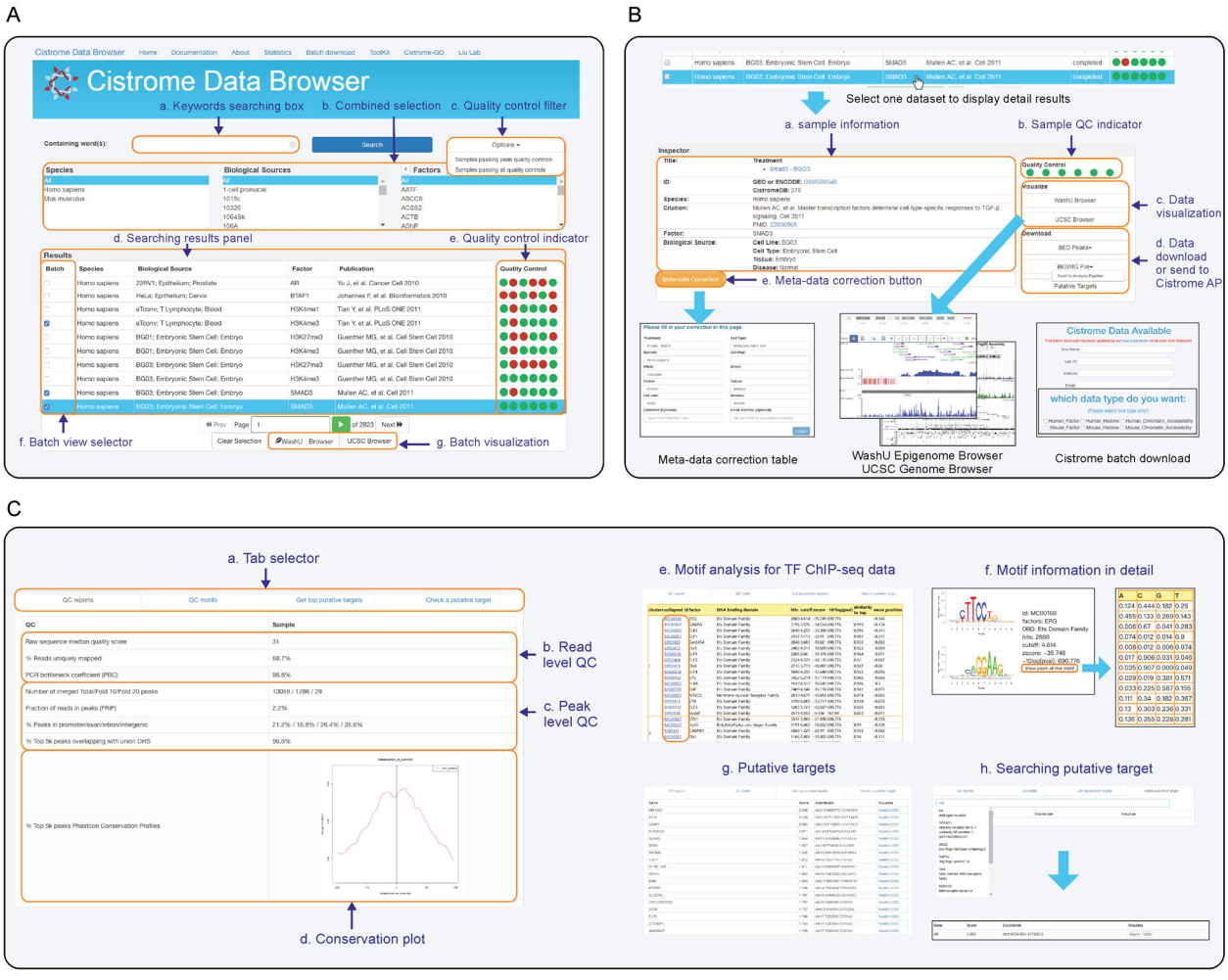

Figure 1. Introduction to the primary function of the Cistrome Data Browser. (A) The main page of Cistrome Data Browser: (a) A keyword searching box that allows to input factors, biological source, Cistrome ID, or GSM accession. (b) A filter for refining searched results from three aspects, including species, biological sources, and target factors. (c) A quality control filter for refining searched results by peak level quality control or all quality control metrics. (d) The results that matched users' queries. (e) Six quality control indicators for each sample. From left to right, each dot means raw sequence quality, mapping rate, library complexity (PBC), ChIP enrichment (10-fold peaks), signal to noise ratio (FRiP), and peak overlap ratio of union DNase-I sensitive sites (DHS). Green dots represent samples that are passed the quality control, while the red dots represent failed samples. (f, g) Batch view operator. Select multiple samples by batch view selector and view the peak intensity on browsers by clicking "WashU Browser" or "UCSC Browser." (B) Sample inspector page: (a) The inspector panel shows samples of metadata information. (b) Sample quality control indicator as the same as Figure A-e. (c) Signal intensity visualization buttons designed for each sample. (d) Buttons for downloading peak BED file and putative targets TXT file, and a button for sending BIGWIG file to Cistrome Analysis Pipeline (AP). (e) Metadata correction button for reporting incorrect metadata. (C) Quality control report and downstream analysis result: (a) Tab selector to switch between QC reports, QC motifs, and putative targets. (b) Read level quality control report containing raw sequence median quality score, the percentage of reads uniquely mapped, and PCR bottleneck coefficient. (c) The report of peaks level quality control includes the number of merged total/10/20-fold enriched peaks, the fraction of reads in peaks, percentage of peaks in promoter/exon/intron/intergenic, and percentage of top 5,000 peaks overlapped with union DHS. (d) Conservation plot for showing the average Phastcon conservation score of top 5,000 peaks. (e) Motif scanning results on peaks of transcription factor ChIP-seq samples. (f) Motif sequence logo and position-weight matrix. (g) Putative targets ranked by the regulatory potential score from high to low. (h) Searching box for putative targets.

Table 1 Explanation and default cutoff of six quality control metrics used in Cistrome DB

Indicators	Description	Thresholds indicate good quality
Sequence quality	Raw sequence median quality score was calculated by FastQC	>= 25
Mapping quality	The uniquely mapped read number is the number of reads with BWA [25] mapping quality above 1	>= 60%
Library complexity	PBC (PCR Bottleneck Coefficient) score [26], which is the number of locations with exactly one uniquely mapped reads divided by the number of unique locations	>= 80%
Signal to noise ratio	FRiP (Fraction of reads in peaks) score, which is the percentage of uniquely mapped tags from autosomal chromosomes that fall in MACS2 peaks	>= 1%
ChIP enrichment	The number of 10 fold confident peaks called by MACS2 [27], where the peaks fold change is greater than 10	>= 500
Regulatory region	The proportion of the most significant 5,000 peaks (ordered by MACS2-log(q.value)) that overlap with a union of DNase-seq peaks	>= 70%

Sample annotation in Cistrome DB

Besides the species, biological source, factor name, publication, and quality controls, the full description of each sample can be found by clicking a particular row in the result panel. The additional annotation includes the cell line, cell type, tissue type, disease, etc. (Fig. 1B-a). Moreover, users can click the hyperlink of the GEO accession (or ENCODE ID) to direct to the original sample web-page and click the hyperlink of PMID to the literature page (Fig. 1B-a). In this way users can obtain more details about particular samples including experimental conditions and protocol details that could influence the interpretation of results. Since the Cistrome DB meta-data collection relies on automated parsers, mistakes may exist in the thousands of samples, although our team has manually corrected most of them. To improve sample annotations, users can help to manually correct our annotation through the "meta-correction" button that can be found at the bottom of the inspector panel for each sample (Fig. 1B-e). The corrected information will be displayed once maintainers have reviewed the submission.

Motif and putative targets for factor ChIP-seq in Cistrome DB

TFs have been known to have high affinity for a set of specific DNA sequences, namely, motifs. It is meaningful to check whether the ChIP targeted TF occupies a strong binding motif, and whether there are other known TF motifs which could be potential cooperative factors enriching in the ChIP-seq peaks. Therefore, the Cistrome DB scans DNA sequence motifs in all samples except histone mark and histone variants ChIP-seq, chromatin accessibility, and ChIP-seq samples with a total peak number less than 200 (Fig. 1C-e). The result is shown at the "QC Motifs" panel, where users are allowed to check the significantly enriched motifs and visualize the sequence logo and position-specific scoring matrix (PSSM) by clicking the collapsed motif ID (Fig. 1C-f). The motif result includes the enrichment of previously known motifs and *de novo* motifs that are newly identified from DNA sequence around the peak summits of a particular ChIP-seq. The *de novo* motif indicates the novel sequence affinity of the factor. The higher motif enrichment the more frequent that the motif sequence appears in the high-confident ChIP-seq peaks [30]. The smaller z-score of motifs indicates the stronger motif enrichment. Since certain groups of TFs tend to cooperatively enable gene regulation, the scanned motifs are further clustered into groups based on their sequence logo similarity [31]. The motifs of TFs in same cluster indicates cooperativity. With such motif enrichment results, users can understand the binding pattern of the factor.

Identifying target genes is a typical task for ChIP-seq. Cistrome DB samples were analyzed by *RP* model [18,19] which calculates each gene a score based on the ChIP-seq peaks. The formula of *RP* calculation can be found at method and materials. Users can also calculate their own *RP* score on Cistrome-GO webserver by clicking the "Cistrome-GO" tab at the top of Cistrome DB main page or directly visiting http://go.cistrome.org. The higher *RP* score indicates the higher likelihood to be a target for the particular factor. The list of putative targets that ordered by *RP* score from high to low was provided in the "Get Top Putative Targets tab" (Fig. 1C-g). To help check a particular gene of interest for the sample, a search box was designed in the "Check a Putative Targets" tab, which allows users to further search putative target *RP* score by gene symbols (Fig. 1C-h).

Data visualization and download in Cistrome DB

Cistrome DB supports to visualize signal intensity in batch or sample by sample on both Wash U Epigenome Browser and UCSC genome browser. To visualize

samples in batch mode, users can select 20 samples in maximum each time through the batch view list selector, as shown in Fig. 1A-fg. Cistrome DB also designed a visualization button for each sample in the inspector panel, as shown in Fig. 1B-c. Since users may want to download the analysis result of their interesting ChIP-seq data, Cistrome DB provides downloadable peak information in BED format files, and putative targets in tab-delimited TXT files (Fig. 1B-d). Users can either download the peak BED file sample by sample in the "Inspector" panel or use the batch download page to down files by data types (Fig. 1B). For users who want to further analyze the data for their research but do not have computing resource, we designed buttons for sending peak file and signal intensity file of a particular sample to Cistrome Analysis Pipeline (AP). Cistrome AP is a galaxy platform that incorporates various analysis tools regarding gene regulatory analysis (Fig. 1B-d).

Prioritizing factors through Cistrome DB Toolkit

We developed a Toolkit module provides three functions that allow users to query the Cistrome DB on the basis of genomic coordinates, gene names or SNP identifiers instead of sample metadata (Figs. 2 and 3). The first function in the Cistrome DB Toolkit was designed to prioritize transcriptional regulators that may regulate a particular gene of interest (Fig. 2). A query gene symbol or RefSeq transcript accession ID produces an ordered list of TF ChIP-seq, histone modification or DNase-seq sample associated with the query. The user can adjust three parameters [1]: the species (human or mouse) [2]; a choice between TRs and chromatin state data such as histone mark ChIP-seq and chromatin accessibility [3]; the decay rate used in the *RP* score calculation (1,000, 10,000, 100,000 base pairs) can be specified, which we calculate *RP* score through summarizing and weighting peaks surrounding the TSS of given genes or transcripts in the range determined by the user-selected decay rate (Fig. 2A). Next, users are asked to choose a particular transcript if there are multiple isoforms of the query gene. The relative genomic position is illustrated on the web page to help users select the transcript they are interested in (Fig. 2B). The Cistrome DB Toolkit returns 200 ChIP-seq samples ordered by *RP* scores in a table (Fig. 2C). Here, the *RP* score has been normalized into same range from 0 to 1 in sample-wise through minimum-maximum normalization method. The highly ranked ChIP-seq samples are more likely to regulate the query gene. The Cistrome DB Toolkit displays the results in three ways [1]. The tabular display allows users to select samples and visualize the signal intensity on the genome browser in batch (Fig. 2C, D) [2]. The interactive dot plot provides functions to highlight samples of a particular factor or biological resource, and to filter samples by sliding an *RP* score cutoff (Fig. 2E) [3]. The static plot result is downloadable and includes the 20 regulators with the highest *RP* score (Fig. 2F).

The second function of the Cistrome DB Toolkit prioritizes the regulators binding on a genomic interval. Such genomic intervals can be an enhancer, a SNP, or a promoter of a long non-coding RNA. Users can select species and data types in a similar way described as before. For any intervals less than 2 Mb, users can query genomic coordinate range in the format: "chrom:start-end" (*e.g.*, chr6:151690496-152103274). Notably, we have incorporated SNPs from dbSNP; and users can search ChIP-seq samples that have peaks overlapping with the SNP regions through directly using the SNP identifiers (*e.g.*, rs2222162) (Fig. 3A).

The third function of the Toolkit module is to find ChIP-seq samples with significant overlap with a set of query intervals uploaded by the user. This function can help investigators answer a variety of research questions. Researchers might want to know, for example, which other TFs bind to similar regions as their TF of interest; which TFs bind to a set of differentially accessible regions identified by ATAC-seq; or which histone modifications are associated with the binding of a TF. Users upload the interval set in BED or tab-delimited file formats, in which the first three columns represent chromosome, start position, and end position, respectively. The parameter "Peak number of Cistrome data to use" is designed to use the most confident peaks of samples in Cistrome DB (*e.g.*, top 1k peaks).

The second and third functions present results in three ways, similar to the first function. In the table format, ChIP-seq samples of the second function are ordered by overlapping peak number on the given interval over total peak number. The third function ranks the ChIP-seq samples based on GIGGLE score [20] which represents the overlap significance between the given peak set and Cistrome DB sample peaks (Fig. 3A-bc, B-b). Alternatively, we also provide a boxplot in which all the samples of the same factor are aggregated (Fig. 3A-d, B-c) to show the ranking of factors. The Cistrome DB Toolkit also provides a downloadable plot showing the top 20 regulators in the "Static plot" tab (Fig. 3A-e, B-d).

DISCUSSION

We describe a comprehensive tutorial for the Cistrome DB and Toolkit, which is a resource of published ChIP-seq /DNase-seq /ATAC-seq samples in human and mouse and a platform for finding regulators that are associated with users' input. Users can query Cistrome DB samples by keyword, including factor name or biological source. For each sample, the Cistrome DB provides six quality

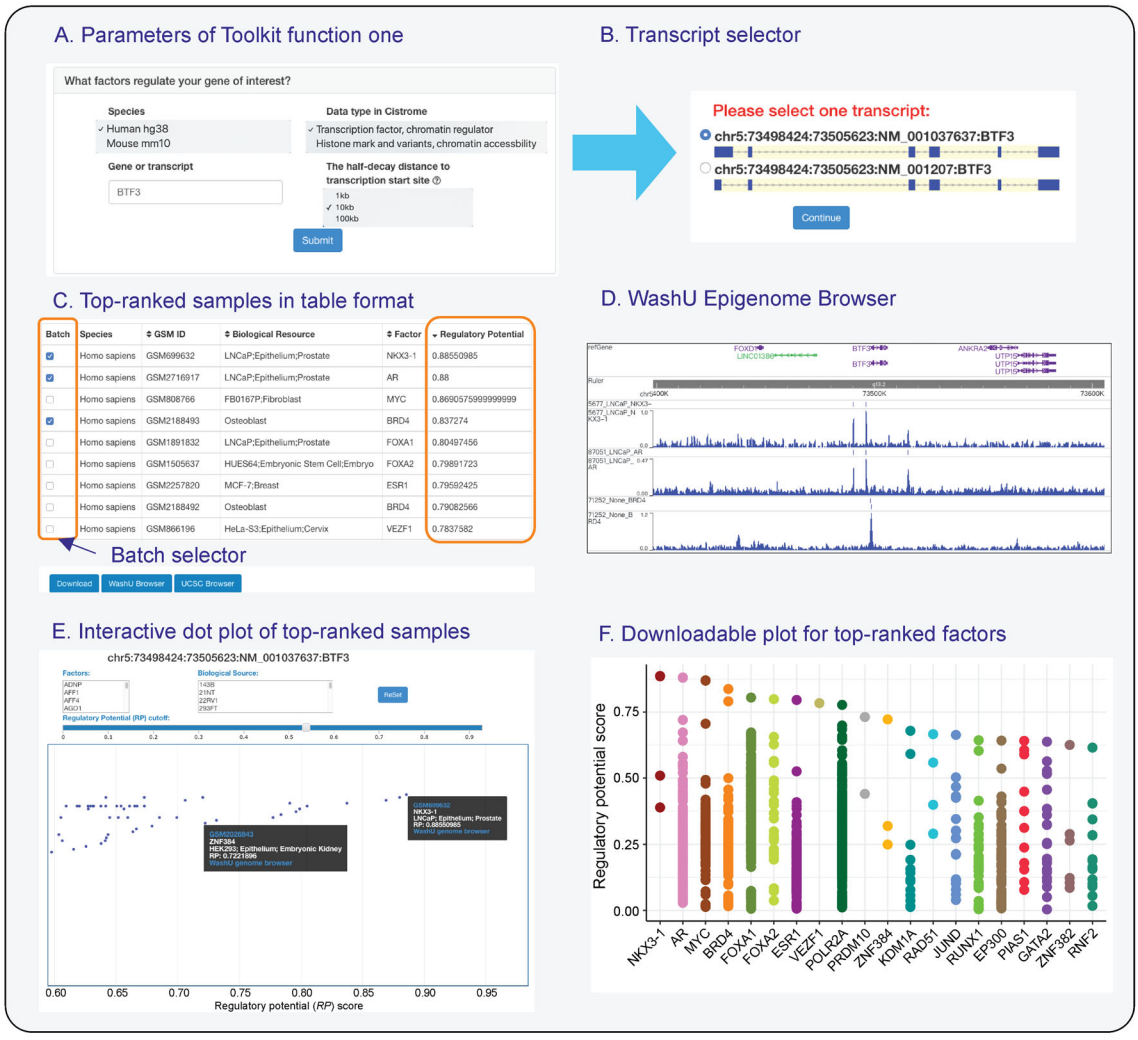

Figure 2. Introduction to the first function of Cistrome DB Toolkit, "what factors regulate your gene of interest?" (A) An input box for gene name or RefSeq transcript ID, and parameter selections for species of human (hg38) and mouse (mm10), data type of "transcription factors, chromatin regulator" and "Histone mark and variants, chromatin accessibility", and half decay rate of 1 kb, 10 kb, and 100 kb to weight peaks up- and down-stream of TSS. (B) Transcript selection only when a gene symbol matches multiple transcripts. (C) Results of top-ranked samples by the regulatory potential score. Select samples in the batch selector and click the bottom button to view tracks in Wash U or UCSC Browser. (D) A track example in the Wash U Epigenome Browser. (E) An interactive dot plot for showing ChIP-seq samples that RP score of the user-given gene greater than zero. Users filter samples by selecting factor or biological resources or moving the RP cutoff slider above the plot. (F) A downloadable plot for showing 20 regulators with the highest score.

control metrics covering library complexity, raw sequencing, and peak enrichment. Cistrome DB also provides downstream analysis results, including motif enrichment and putative targets for TF ChIP-seq samples. With the Cistrome DB Toolkit, users can find the potential transcriptional regulators of a given gene and find TFs that bind to a given interval or a peak set. The Cistrome DB and Toolkit is a widely used resource that has received over 123,000 visits and 1,500 bulk data downloads. To better support gene regulatory analysis, the Cistrome DB and Toolkit will be improved in three ways in the future. Firstly, the regulator lists provided by the Cistrome DB Toolkit can be further summarized into protein complexes to provide more biological insights. Secondly, single-cell genomics technologies, including single-cell ATAC-seq (scATAC-seq) have been recently developed [32], raising opportunities for biomedical researchers to understand gene regulation rules at a single cell resolution [33]. We will consider incorporating scATAC-seq data into the Cistrome DB and allow users to explore the scATAC-seq on the same platform with bulk cistrome data. Thirdly, aberrant gene regulation is related to the development of

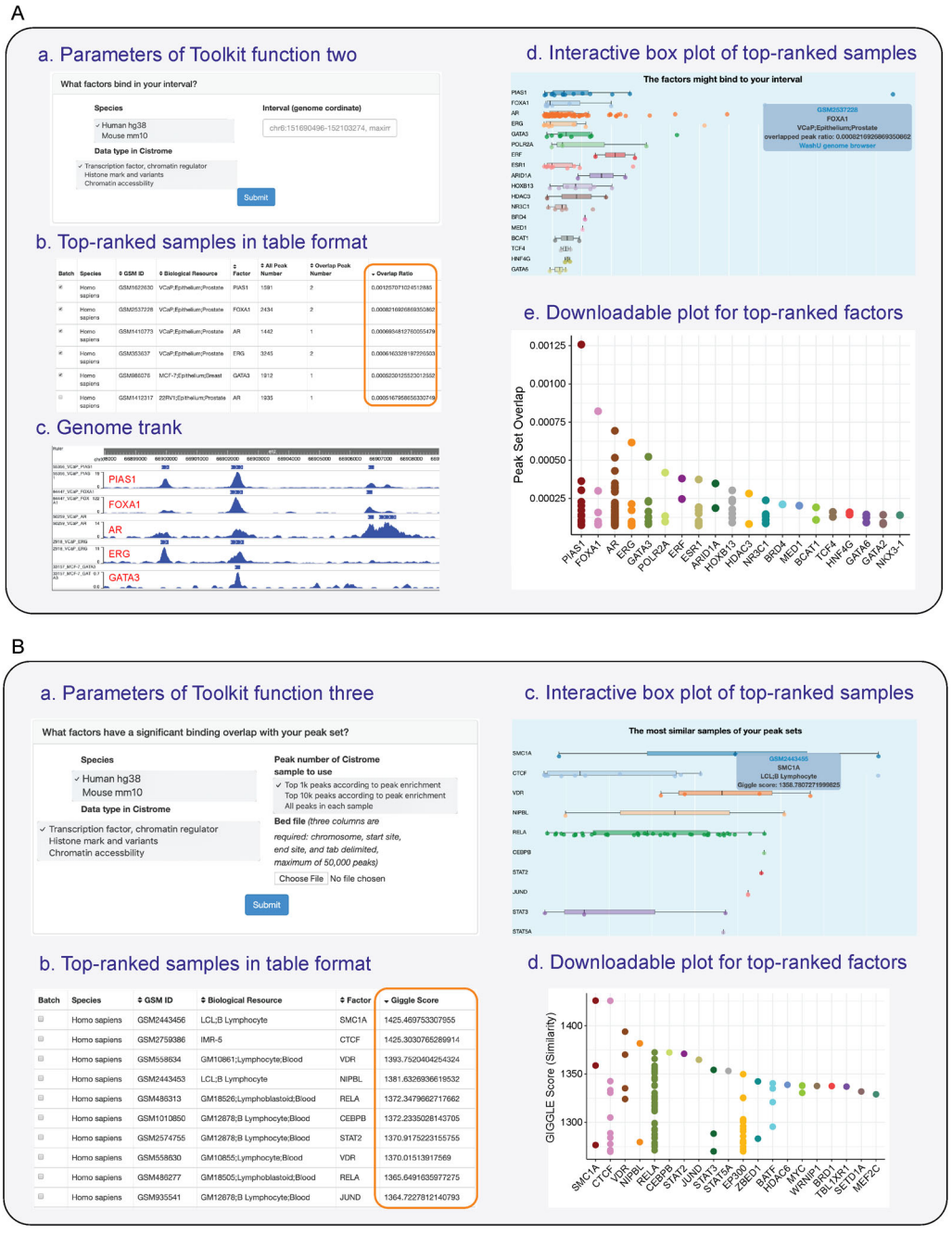

Figure 3. Introduction to the second and third functions of Cistrome DB Toolkit: "what factors binding in your interval?" "what factors have a significant binding overlap with your peak set?" (A) The primary usage of the second function: (a) An input box for a genomic interval, and parameter selections for species of human (hg38) and mouse (mm10), the data type of "transcription factors, chromatin regulator" and "Histone mark and variants, chromatin accessibility." (b) Results of top-ranked samples by overlap ratio to total peaks in the sample. (c) Batch view of signal intensity on a user-given interval using top-ranked samples. (d) An interactive box plot for showing top-ranked samples. Users could check each samples' details by moving the mouse over the dot. (e) A downloadable plot for showing 20 regulators with the highest peak overlap ratio. (B) The primary usage of function three: (a) BED file uploading and parameter selections for species and data type. (b) Results of top-ranked samples by giggle score which represents the similarity between a sample and user-defined peak sets. (c) An interactive box plot for showing top-ranked samples. Users can check each samples' details by moving the mouse over the dot. (d) A downloadable plot for showing 20 regulators with the highest giggle score.

disease, such as cancer [34,35]. Many cistrome data sets pertain to cancer related transcription factors and are carried out on cancer cell lines, mouse tumor models, and cancer patient samples. To enhance the use of disease information in Cistrome DB, we will improve the disease annotation for samples to allow users to find ChIP-seq of a particular disease type. The Cistrome DB and related resources are easily accessible, highly informative, resources for gene regulatory analysis that benefit the biomedical community.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Zhiping Weng for providing the backup of the Cistrome DB and Dr. Ting Wang for the Wash U Epigenome Gateway Browser. This work is supported by National Institutes of Health of US (U24 CA237617).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Rongbin Zheng, Xin Dong, Changxin Wan, Xiaoying Shi, Xiaoyan Zhang and Clifford A. Meyer declare that they have no conflict of interests.

The article does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. 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
2. Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680
3. Furey, T. S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet., 13, 840–852
4. 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
5. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods, 10, 1213–1218
6. 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
7. The ENCODE Project Consortium (2013) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74
8. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
9. Barrett, T., Wilhite, S. E.,Ledoux, P., Evangelista, C., Kim, I.T., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res., 41, D991–D995
10. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C. H., Brown, M., Zhang, X., Meyer, C. A., *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res., 47, D729–D735
11. Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., *et al.* (2018) Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat. Genet., 50, 238–249
12. Ballaré, C., Castellano, G., Gaveglia, L., Althammer, S., González-Vallinas, J., Eyras, E., Le Dily, F., Zaurin, R., Soronellas, D., Vicent, G. P., *et al.* (2013) Nucleosome-driven transcription factor binding and gene regulation. Mol. Cell, 49, 67–79
13. Ouyang, Z., Zhou, Q. and Wong, W. H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. USA, 106, 21521–21526
14. Jiang, S. and Mortazavi, A. (2018) Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics, 17, 104–115
15. Guan, D., Shao, J., Deng, Y., Wang, P., Zhao, Z., Liang, Y., Wang, J. and Yan, B. (2014) CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. Bioinformatics, 30, 1190–1192
16. Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet., 5, 276–287
17. 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
18. 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
19. Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C. A. and Liu, X. S. (2019) Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. Nucleic Acids Res., 47, W206–W211
20. 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
21. Karolchik, D. and Kent, W. J. (2003) The UCSC Genome Browser. Curr. Protoc. in Bioinforma., 00, 1.4.1–1.4.23
22. Li, D., Hsu, S., Purushotham, D., Sears, R. L. and Wang, T. (2019) WashU Epigenome Browser update 2019. Nucleic Acids Res., 47, W158–W165
23. Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., *et al.* (2018) The chromatin accessibility landscape of primary human cancers. Science, 362, eaav1898
24. Bell, O., Tiwari, V. K., Thomä, N. H. and Schübeler, D. (2011)

Determinants and dynamics of genome accessibility. Nat. Rev. Genet., 12, 554–564
25. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics, 25, 1754–1760
26. Marinov, G. K., Kundaje, A., Park, P. J. and Wold, B. J. (2014) Large-scale quality analysis of published ChIP-seq data. G3: Genes, Genomes. Genetics, 4, 209–223
27. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137
28. Siepel, A. and Haussler, D. (2005) Phylogenetic Hidden Markov Models. In: Statistical Methods in Molecular Evolution. Statistics for Biology and Health. New York: Springer
29. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. D. W., Richards, S., *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034–1050
30. Meyer, C. A., He, H. H., Brown, M. and Liu, X. S. (2011) BINOCh: Binding inference from nucleosome occupancy changes. Bioinformatics, 27, 1867–1868
31. Jiang, P. and Singh, M. (2014) CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. Nucleic Acids Res., 42, 2833–2847
32. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. and Greenleaf, W. J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523, 486–490
33. Jia, G., Preussner, J., Chen, X., Guenther, S., Yuan, X., Yekelchyk, M., Kuenne, C., Looso, M., Zhou, Y., Teichmann, S., *et al.* (2018) Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. Nat. Commun., 9, 4877
34. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. and Garraway, L. A. (2013) Highly recurrent TERT promoter mutations in human melanoma. Science, 339, 957–959
35. Demichelis, F., Setlur, S. R., Banerjee, S., Chakravarty, D., Chen, J. Y. H., Chen, C. X., Huang, J., Beltran, H., Oldridge, D. A., Kitabayashi, N., *et al.* (2012) Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc. Natl. Acad. Sci. USA, 109, 6686–6691

