IVYSPRING INTERNATIONAL PUBLISHER

International Journal of Biological Sciences

2018; 14(12): 1724-1731. doi: 10.7150/ijbs.28850

Research Paper

TFmapper: A Tool for Searching Putative Factors Regulating Gene Expression Using ChIP-seq Data

Jianming Zeng, Gang Li✉

Faculty of Health Sciences, University of Macau, Taipa, Macau, China

✉ Corresponding author: Gang Li, Faculty of Health Sciences, University of Macau, Taipa, Macau, China. Email address: gangli@umac.mo

© Ivyspring International Publisher. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/). See http://ivyspring.com/terms for full terms and conditions.

Received: 2018.07.29; Accepted: 2018.07.30; Published: 2018.09.07

Abstract

Background: Next-generation sequencing coupled to chromatin immunoprecipitation (ChIP-seq), DNase I hypersensitivity (DNase-seq) and the transposase-accessible chromatin assay (ATAC-seq) has generated enormous amounts of data, markedly improved our understanding of the transcriptional and epigenetic control of gene expression. To take advantage of the availability of such datasets and provide clues on what factors, including transcription factors, epigenetic regulators and histone modifications, potentially regulates the expression of a gene of interest, a tool for simultaneous queries of multiple datasets using symbols or genomic coordinates as search terms is needed.

Results: In this study, we annotated the peaks of thousands of ChIP-seq datasets generated by ENCODE project, or ChIP-seq/DNase-seq/ATAC-seq datasets deposited in Gene Expression Omnibus (GEO) and curated by Cistrome project; We built a MySQL database called TFmapper containing the annotations and associated metadata, allowing users without bioinformatics expertise to search across thousands of datasets to identify factors targeting a genomic region/gene of interest in a specified sample through a web interface. Users can also visualize multiple peaks in genome browsers and download the corresponding sequences.

Conclusion: TFmapper will help users explore the vast amount of publicly available ChIP-seq/DNase-seq/ATAC-seq data and perform integrative analyses to understand the regulation of a gene of interest. The web server is freely accessible at http://www.tfmapper.org/.

Key words: Chromatin immunoprecipitation; Next-generation sequencing; Regulation of gene expression

Introduction

Next-generation sequencing (NGS) coupled to chromatin immunoprecipitation (ChIP-seq), DNase I hypersensitivity (DNase-seq) and the transposase-accessible chromatin assay (ATAC-seq) provide powerful tools for studying gene regulation by factors in cis and trans, which includes components of the basal transcriptional machinery, transcription factors, chromatin regulators, histone variants, histone modifications and others. A great amount of data of ChIP-seq, DNase-seq and ATAC-seq has been accumulated by individual laboratories and large-scale collaborative projects, including the ENCODE (Encyclopedia of DNA Elements) Consortium [1], the Roadmap Epigenomics Mapping Consortium (Roadmap) [2] and the International Human Epigenome Consortium (IHEC) projects [3]. The datasets are usually publicly accessible through the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI) [4], and the data portals of large-scale projects.

Despite the easy access, mining and interpreting the ChIP-seq/DNase-seq/ATAC-seq data is challenging for regular users, especially for bench biologists with limited bioinformatic expertise. Adding to the complexity, the data qualities and data analysis pipelines are remarkably varied, which hinder their direct use in further analyses [5]. To address the challenges, multiple algorithms/metrics have been developed to evaluate the data quality bioinformatically, such as ENCODE quality metrics, NGS-QC, ChiLin and others [6-9]. NGS-QC developed by Gronemeyer and colleagues built a quality control (QC) indicator database of the largest collection of publicly available NGS datasets [6, 8], which provides a solid start point for further analysis. CistromeDB developed by Liu and colleagues curated and processed a huge collection of human and mouse ChIP-seq and chromatin accessibility datasets from GEO with a standard analysis pipeline ChiLin [9], and further evaluated individual data quality under several scoring metrics [10]. High-quality processed ChIP-seq data generated by ENCODE consortium, including histone modification, chromatin regulator and transcription factor binding data in a selected set of biological samples, are also available through its data portal [11]. ENCODE and CistromeDB provide access to the processed data, and the corresponding metadata including the sources and properties of biological samples, experimental protocols, the antibody used and others, which offer opportunities for users to re-analyze the data and identify the genome-wide targets of a transcription regulator in different cell lines and tissues. Nonetheless, different questions are often asked, such as: Which transcription factors are responsible for the regulation of a gene of interest, and what is the epigenetic landscape of a gene of interest in a particular tissue or cell type? To address these questions, a tool for simultaneous queries of multiple datasets using symbols or genomic coordinates of target genes as search terms is needed.

In this study, we collected and annotated a large number of ChIP-seq, DNase-seq and ATAC-seq datasets including the ENCODE datasets and the GEO datasets curated by CistromeDB [10]. We built a Structured Query Language (SQL) database containing the annotations and the associated metadata, allowing users to search across multiple datasets and identify the putative factors which target a specified gene or genomic locus based on actual NGS data. We also provide links for users to visualize the peaks and download the corresponding sequences. In addition, we included an example to demonstrate the utility of *TFmapper*.

Results

The primary purpose of *TFmapper* is to search all experimental ChIP-seq datasets and identify the trans-acting factors or histone modifications which show peaks at a gene of interest or a specified genomic region in a user-defined biological sample. To achieve this objective, we downloaded ENCODE and GEO/CistromeDB data, and annotated the peaks using HOMER software [12] based on the peaks called by ENCODE and CistromeDB. We then built a MySQL database called TFmapper containing the annotations and associated metadata with a web interface (Fig. 1), which allows searching of thousands of GEO/ENCODE datasets simultaneously to find what TF/histone marks target a gene of interest in a user-specified biological sample. Currently, TFmapper includes 26445 human and mouse datasets, including 11978 ChIP-seq datasets for trans-acting factors, 11073 ChIP-seq datasets for histone marks/variants, and 1371 DNase-Seq/ATAC-seq datasets (Fig. 1).

Website interface

The 'home' page of *TFmapper* consists of three sections (Fig. 2): *Parameters section,* search section, and results section.

Parameters Section. In this section, users can define the searching criteria, including species (human or mouse), types of experiments (ChIP-seq for trans-acting factors, ChIP-seq for histone marks or Chromatin accessibility assays), sources of datasets (GEO or ENCODE) and biological sources. Each biological source is defined by three properties. Based on the original description, the properties could be 'cell line', 'tissue' and 'organ'. All three properties are searchable.

Search Section. Users can query the database by 1) gene symbols or 2) genomic coordinates of GRCh38/hg38 (Homo Sapiens) or GRCm38/mm10 (Mus Musculus) respectively. Users can further define which portion of the gene to query (Fig. 2).

Results Section. The search results are displayed in the form of a table including the following fields: Sample identifiers (GEO sample (GSM) IDs or ENCODE IDs); SampleID defined by CistromeDB; name of factors (trans-acting factors, histone modifications); links for peak visualization in the WashU Epigenome Browser [13] or the UCSC Genome Browser [14]; the coordinates of the peak; distance between the center of a peak and the transcription start site (TSS) of a gene; the enrichment scores, the p- and q- values for ChIP-seq peaks called by Model-based Analysis for ChIP-Seq (MACS2) [15, 16]; the genomic attribute (promoters, intergenic or intragenic peaks); title and source name of the dataset, etc. At the bottom of this section, a graph is automatically generated to display the number of experiments being performed for each factor. The search results are sorted by default per the distance between the feature site and the TSS. All fields can be further sorted by clicking the arrow at the left of the title of each field, or narrowed by specifying the desired value for each field. Under the 'distance' field,

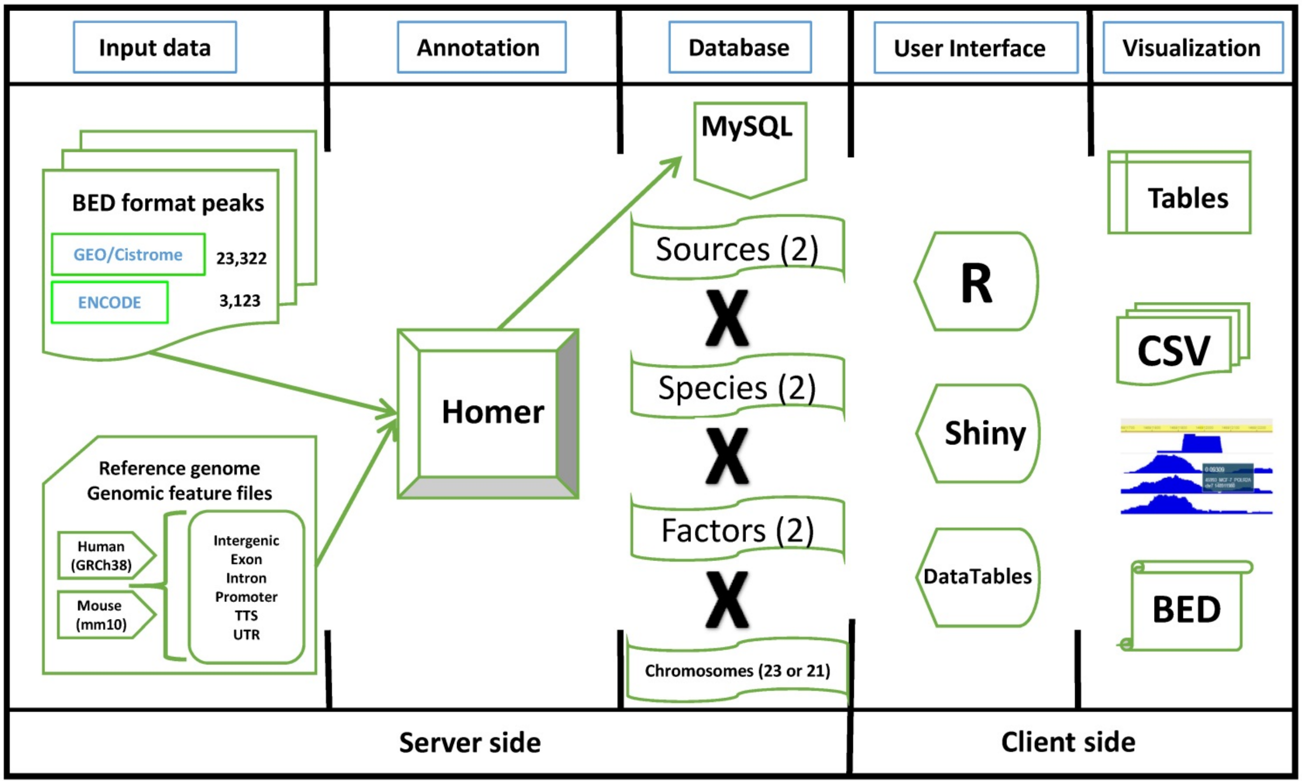


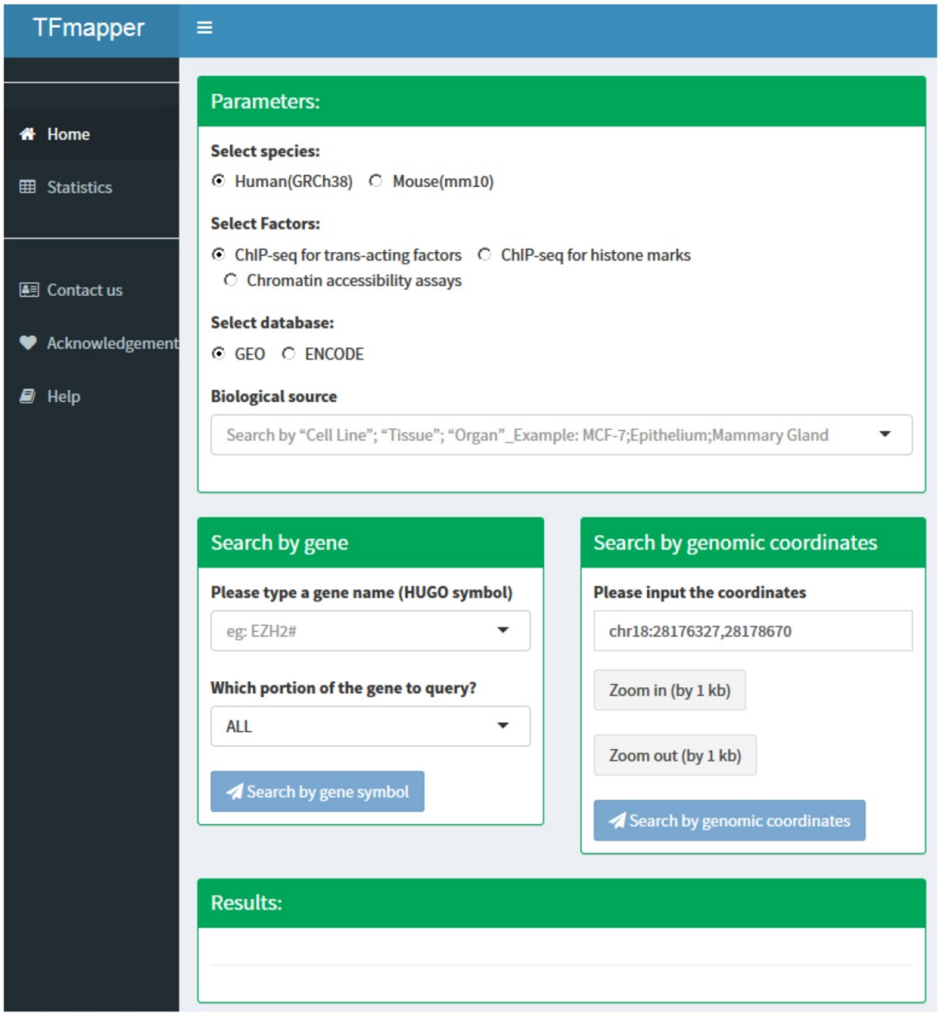
Figure 1. The workflow diagram of *TFmapper*. Peak files in the BED (Browser Extensible Data) format are downloaded from the GEO/Cistrome and the ENCODE data portals. Peaks are annotated to genomic features (Promoter, TTS, 5'UTR, 3'UTR, Intron, Exon, Intergenic) using the software HOMER with GRCh38/hg38 (Human) and GRCm38/mm10 (Mouse) as the reference genomes. The annotation results are stored in a MySQL database. To increase the speed of query processing, the peaks are split by species, sources, factors, and chromosomes. For the client side, all the elements on the HTML page are built by R (Shiny), and result tables are created with the DataTables JavaScript library. Results can be downloaded in the CSV or BED format, and peaks can be directly visualized in the WashU Epigenome Browser or the UCSC Genome Browser.

a slider will show up when clicked; the user can move the slider to narrow down the region. Several other fields are clickable; users will be redirected to the specific GSM or ENCODE information page by clicking the sample identifier field, and the user can retrieve the actual sequence of a peak by clicking the peak coordinates. The query results are stored in a CSV (Comma Separated Values) format file, or a BED (Browser Extensible Data) format file. The CSV format file contains all the fields in the search results, and the BED format file contains the names of factors, coordinates and enrichment scores for all peaks. Both files are downloadable for downstream analysis. For example, the user can load the BED format file into the Integrative Genomics Viewer (IGV) to visualize each peak at a specific genomic region [17]. More importantly, a link will appear on the top of the table when multiple peaks are highlighted in blue, which allows users to visualize multiple peaks in the WashU Epigenome Browser (Fig. 3).

Case studies

To demonstrate the utility of *TFmapper*, we examined the transcriptional regulation of chromobox homologs (CBXs) in the H1 line of the human embryonic stem cells (H1-hESC). CBX proteins (including CBX2, CBX4, CBX6, CBX7, and CBX8) are mammalian homologs of Drosophila Polycomb protein, bind to trimethylated lysine 27 of histone H3 (H3K27me3) through the chromodomain. Orthologs of Cbxs are differentially expressed in mouse embryonic stem cells, in which Cbx7 is highly expressed. Cbx7 represses the expression of other CBXs and contributes to the maintenance of stem cell pluripotency [18, 19]. Nonetheless, analysis of RNA-Seq data of ENCODE clearly demonstrated that the expression profiles of CBXs are strikingly different in human ESCs than in mouse ESCs (Fig. 4A). In H1 cells, CBX2 are highly expressed, CBX6 and CBX7 are moderately expressed, whereas CBX4 and CBX8 are lowly or not expressed. Human CBX2, CBX8 and CBX4 are sequentially localized on chromatin 17. To understand why CBX2 and its neighbor CBX8 express differently in H1 cells, we used *TFmapper* to retrieve information on chromatin regulators that target CBX2 and CBX8 (Supplementary Table S1, S2), and the statuses of histone modifications on CBX2 and CBX8 gene in H1 cells.

To our surprise, although CBX2 and CBX8 are expressed at significantly different levels, the factors targeting CBX2 and CBX8 are largely overlapped, and include those correlated with either gene activation or



Parameters:

Select species:
● Human(GRCh38) ○ Mouse(mm10)

Select Factors:
● ChIP-seq for trans-acting factors ○ ChIP-seq for histone marks
○ Chromatin accessibility assays

Select database:
● GEO ○ ENCODE

Biological source
Search by "Cell Line"; "Tissue"; "Organ"_Example: MCF-7;Epithelium;Mammary Gland ▾

Search by gene

Please type a gene name (HUGO symbol)
eg: EZH2# ▾

Which portion of the gene to query?
ALL ▾

Search by genomic coordinates

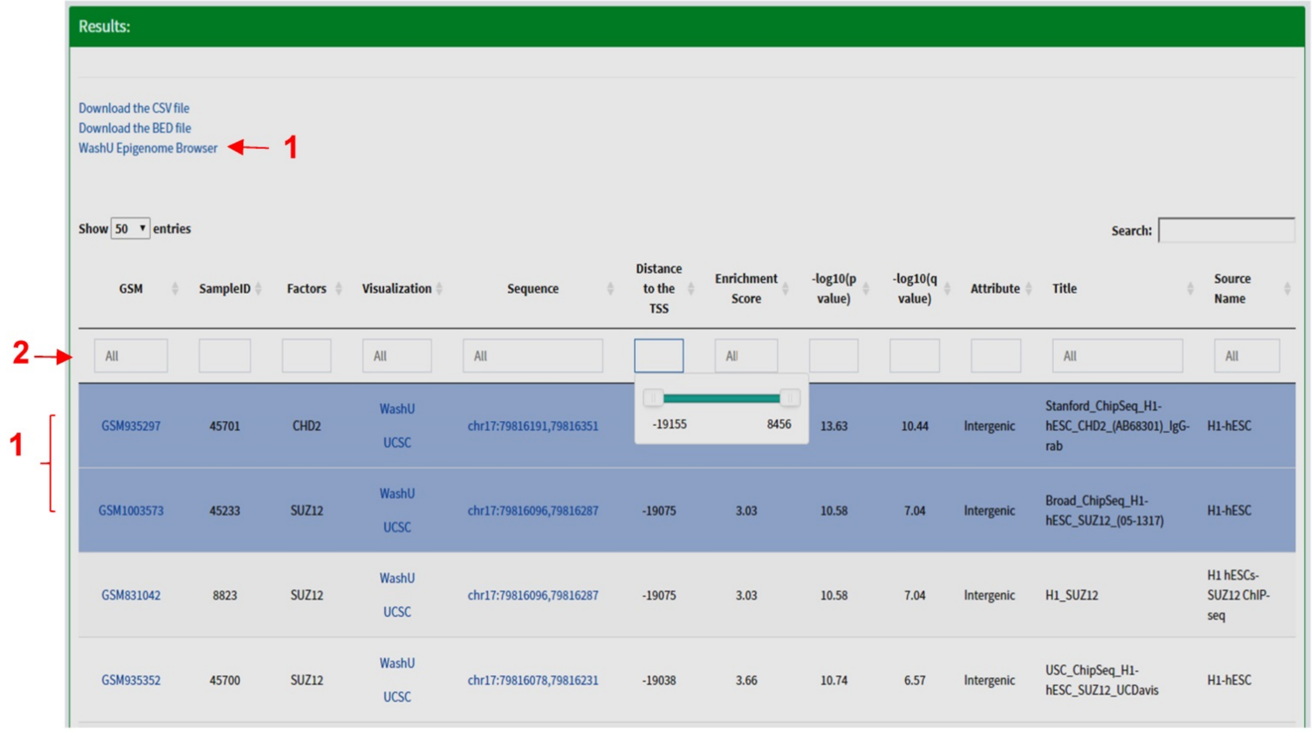
Please input the coordinates
chr18:28176327,28178670

Results:

Figure 2. The homepage of *TFmapper* consists of 'Parameters', 'Search', and 'Results' section.

silencing (Fig. 4B). However, there are several factors which target CBX8 only, such as BRCA1, CHD7, CTBP2, MAFK, MXI1, POU5F1, SOX2, SP2, SRF and USF2 (Fig. 4B, Supplementary Fig. S1). Only GTF2F1, which encodes the alpha subunit of the general transcription factor TFIIF and promotes transcription elongation, targets CBX2 exclusively (Supplementary Fig. S1). As shown in Figure 4C, a CTCF peak sits proximally upstream (-500bp) of CBX2 transcription start site, a double peak of CTCF present in the intergenic region between CBX2 and CBX8 in H1 cells, suggesting CBX2 and CBX8 sit in two neighborhoods, possibly insulated by the intergenic CTCF double peak. A wide binding of EZH2 and enrichment of repressive mark H3K27me3 were found on CBX8, extending almost continuously from upstream 19 Kb of CBX8 transcription start site (TSS) to downstream 2 Kb of CBX8 transcription termination site (TTS), suggesting CBX8 is silenced by Polycomb Repressive Complex 2 (PRC2).

Further detailed examination of the trans-acting factors targeting CBX8 reveals several other intriguing features which are worth to be further investigated: 1) a putative poised enhancer region centered at the serum response factor (SRF) peak, 13Kb upstream of the TSS of CBX8. Around the SRF peak, there are significant POLR2A binding, EP300 binding, H3K27me3 and H3K4me1 enrichment, strongly suggests the existence of a poised enhancer, and SRF might regulate CBX8 expression. 2) a BRCA1 peak at the TSS of CBX8. A significant BRCA1 peak was found at the TSS of CBX8, suggesting the potential involvement of BRCA1 in the regulation of CBX8 expression. Recently, Oza J et al. reported that CBX8 is quickly recruited to the sites of DNA damage [20]. The targeting of CBX8 by BRCA1 further supports the participation of CBX8 in DNA damage response. 3) POU5F1 (OCT4), SOX2 and CHD7 target CBX8. Two significant CHD7 peaks were found 3.5 Kb and 7.5 Kb upstream of CBX8 TSS site; the later peak overlaps with POU5F1 and SOX2 peaks. The transcriptional circuitry mediated by OCT-4 and SOX2 supports the self-renewal and pluripotency of ESC. OCT-4 and SOX2 work cooperatively to regulate genes primarily



Results:

Download the CSV file
Download the BED file
WashU Epigenome Browser

Show 50 entries

Search:

GSM	SampleID	Factors	Visualization	Sequence	Distance to the TSS	Enrichment Score	-log10(p value)	-log10(q value)	Attribute	Title	Source Name
All			All	All		All				All	All
GSM935297	45701	CHD2	WashU UCSC	chr17:79816191,79816351			13.63	10.44	Intergenic	Stanford_ChipSeq_H1-hESC_CHD2_(AB68301)_IgG-rab	H1-hESC
GSM1003573	45233	SUZ12	WashU UCSC	chr17:79816096,79816287	-19075	3.03	10.58	7.04	Intergenic	Broad_ChipSeq_H1-hESC_SUZ12_(05-1317)	H1-hESC
GSM831042	8823	SUZ12	WashU UCSC	chr17:79816096,79816287	-19075	3.03	10.58	7.04	Intergenic	H1_SUZ12	H1 hESCs-SUZ12 ChIP-seq
GSM935352	45700	SUZ12	WashU UCSC	chr17:79816078,79816231	-19038	3.66	10.74	6.57	Intergenic	USC_ChipSeq_H1-hESC_SUZ12_UCDavis	H1-hESC

Figure 3. The 'Results' section of *TFmapper*. When multiple peaks are selected, a link for visualization in the WashU Epigenome Browser will appear (red arrow 1). The results can be further filtered by typing search terms in the boxes (red arrow 2); a slider will appear when the box under 'Distance' being clicked, which can be moved to narrow down the region.

as transcriptional activators [21]. The chromodomain helicase DNA-binding protein 7 (Chd7) physically interacts Sox2 [22], targets active gene enhancer [23], and maintains open chromatin configuration [24, 25]. The peaks of POU5F1, SOX2, and CHD7 on the distal promoter region of CBX8 partly overlap, or reside in between the broad and strong H3K27me3 peaks. We hypothesize these factors might work together to prime CBX8 to be expressed. Whether the hypothesis holds true requires further investigation.

Discussion

A large amount of ChIP-seq, DNase-seq and ATAC-seq data is freely available in the public domain. Unfortunately, the majority of the datasets deposited in GEO of NCBI are not processed, which poses a significant barrier to datamining by non-bioinformaticians. CistromeDB reprocessed a huge collection of human and mouse ChIP-seq and chromatin accessibility datasets from GEO with a standard analysis pipeline, and calculated metrics reflecting data quality, in turn, CistromeDB provides a very useful resource for datamining. However, Cistrome only allows factor-oriented (in trans) queries: to identify the target sites of a specific transcription factor or epigenetic mark on the genome in a specified sample, but not gene-oriented (in cis) query: to identify all of the factors targeting a gene of interest in a specific biological sample by searching multiple datasets. To address this problem, we developed TFmapper allowing for searches of multiple datasets using gene symbols or genomic coordinates as queries. TFmapper contains datasets generated by ChIP-seq/DNase-seq/ATAC-seq experiments, including 23,322 GEO datasets re-processed by Cistrome [10] and 3123 datasets from ENCODE [1].

Gene-oriented (in cis) query of multiple datasets can be achieved by using SCREEN (http://screen.encodeproject.org/) of ENCODE [1], or the Track Hub Registry (http://www.trackhubregistry.org) designed by Ensembl [26]. However, both tools do not allow straightforward identification of transcription factors targeting a gene of interest in a user-specified biological sample. In addition, both tools have not incorporated most of the public available NGS data deposited in GEO. Other tools including BindDB [27], GTRD [28] and Remap [29] are capable of performing similar functions as TFmapper. However, BindDB includes far fewer datasets than TFmapper, only 455 datasets generated from mouse and human ESCs and iPSCs. The query result from BindDB does not include the exact peak coordinates, making it difficult to analyze the relationships between factors. It does not provide adequate descriptive information on the samples and clickable links to GEO. In addition, BindDB does not support sequence retrieving and peak visualization. Remap includes 1085 human

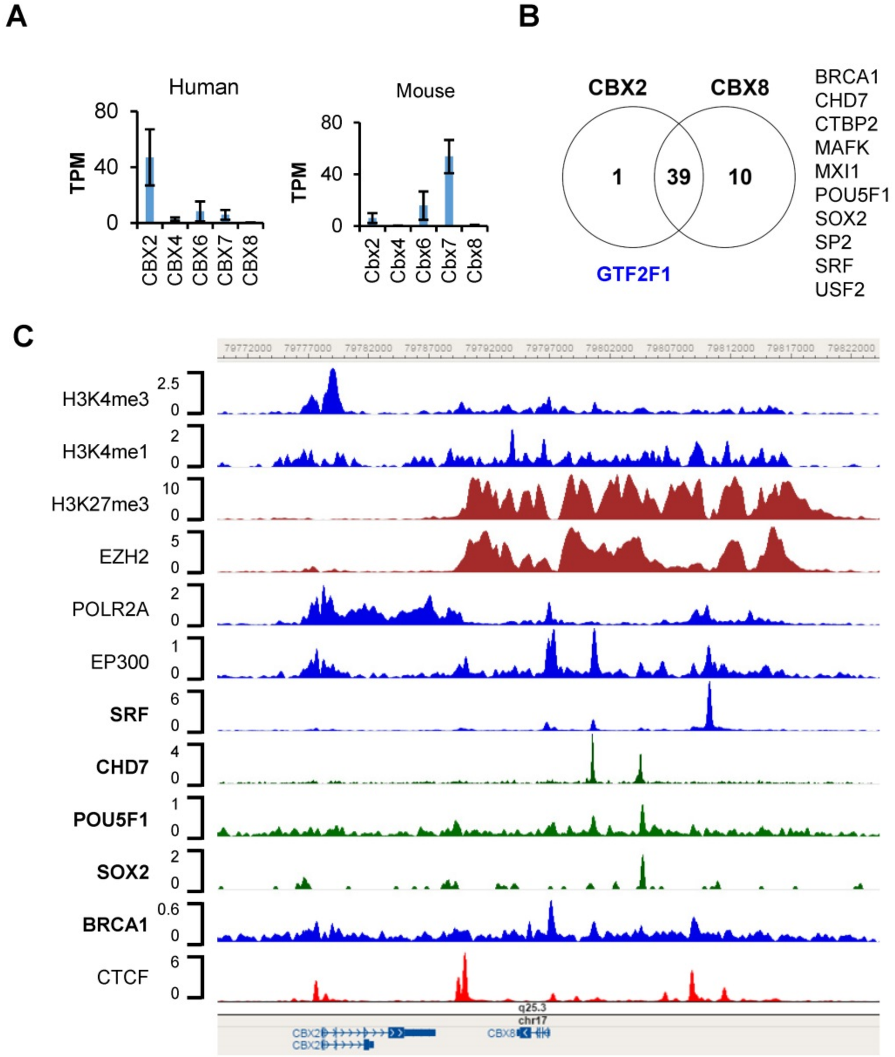


Figure 4. Transcriptional regulation of CBX2 and CBX8 in hESC-H1 cells. (**a**) mRNA expression of CBX2, CBX4, CBX6, CBX7 and CBX8 in human and mouse embryonic stem cells (hESCs and mESCs). The data are retrieved from the RNA-seq datasets of ENCODE. The expression values are presented as transcripts per million (TPM). (**b**) Venn diagram of trans-acting factors targeting CBX2 and CBX8 in hESC-H1 cells. The detailed information on the factors can be found in the Supplementary Table S1 and S2. The factors targeting CBX2 or CBX8 exclusively are indicated in blue and black respectively. (**c**) Binding patterns of selected trans-acting factors and histone modification marks at the CBX2-CBX8 locus in hESC-H1 cells. The peaks are visualized with the WashU Epigenome Browser. The GEO accession number for each dataset can be found in the Supplementary Table S3.

datasets of transcription factors, while GTRD includes 8828 datasets of transcription factors for human and mouse, both tools do not contain datasets for histone marks and chromatin accessibility assays. Both tools merged peaks from different cell lines with distinct gene expression profiles to facilitate the identification of putative transcription factor binding sites (TFBS) or cis-acting regulatory sequences in the genome. Different types of cells have distinct gene expression profiles: correspondingly, different types of cells might use different enhancers to regulate the same gene and have different binding patterns of transcription factors [30, 31]. Although the merging of peaks serves its intended purpose well, it prevents the users from predicting if a targeting event happens in the biological sample they are studying. In addition, both tools don't allow users to trace back to the original GSM dataset easily to check in detail the culture/treatment condition and quality of the data.

In conclusion, *TFmapper* enables users to search across thousands of ChIP-seq/DNase-seq/ATAC-seq datasets, including the ENCODE datasets and the GEO datasets curated by Cistrome, to identify the putative factors which target a genomic region/gene of interest. We believe this tool will help eliminate bioinformatic barriers, allow the scientists to take advantage of the enormous amount of existing ChIP-seq data and perform integrative analysis to understand mechanisms of gene regulation.

Methods

Data collection

We downloaded the GEO datasets processed by Liu and colleagues from CistromeDB as BED (Browser Extensible Data) files, which includes 6092 ChIP-seq datasets for trans-acting factors and 6068 ChIP-seq datasets for histone modifications generated from human samples; 4786 ChIP-seq datasets for trans-acting factors and 5002 ChIP-seq datasets for histone modifications generated from mouse samples; and 1371 DNase-seq/ATAC-seq datasets. For ENCODE datasets, we downloaded the conservative IDR (Irreproducible Discovery Rate)- thresholded peaks, which are called after combinational analysis of two replicates for each ChIP-seq experiment. The ENCODE datasets selected in this study include 955 transcription factor and 771 histone modification datasets for Homo sapiens, and 145 transcription factor and 1252 histone modification datasets for Mus musculus.

Data processing

We used HOMER [12] to annotate the peaks in BED files per the newest reference genome, GRCh38/hg38 for human and GRCm38/mm10 for mouse respectively. HOMER assigns each peak to the nearest gene by calculating the distance between the middle of a peak and the transcription start site (TSS) of a gene. Meanwhile, seven genomic features based on RefSeq annotations were assigned to each peak, which are promoter (by default defined from -1kb to +100bp of TSS), TTS region (transcription termination site, by default defined from -100 bp to +1kb of TTS), Exon, 5' UTR Exon, 3' UTR Exon, Intron and Intergenic.

Database design

To enhance the searching speed in identifying all the peaks targeting a selected gene or specific region, we split the table into subsets based on species, sources, factors, and chromosomes, resulting in 176 tables containing the peaks information. We organized the gene location information according to the longest annotation of the transcript in the Consensus CDS (CCDS) database [32], and only the genes which have peaks in any of the ChIP-seq datasets will be kept for searching. The metadata for GEO and ENCODE datasets were re-organized and uploaded to a MySQL database. The R package GEOmetadb [33] is also used to retrieve further annotation details of the samples in GEO.

Web application

The web application was generated by the R package Shiny and hosted on a Linux server (Ubuntu 16.04), can be accessed by a web browser at http://www.tfmapper.org/. The page layout was made using the R package Shinydashboard, and the interactive tables were displayed using the R package DT.

Abbreviations

NGS: Next-generation sequencing; GEO: the Gene Expression Omnibus; NCBI: the National Center for Biotechnology Information; IHEC: the International Human Epigenome Consortium; ENCODE: the Encyclopedia of DNA Elements) Consortium; ChIP-Seq: Next-generation sequencing coupled to chromatin immunoprecipitation; DNase-seq: Next-generation sequencing coupled to DNase I hypersensitivity; ATAC-seq: Next-generation sequencing coupled to the transposase-accessible chromatin assay.

Supplementary Material

Supplementary figure and tables.
http://www.ijbs.com/v14p1724s1.pdf

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau [137/2014/A3, 095/2015/A3] and the Research & Development Administration Office of the University of Macau [SRG201400015, MYRG201500232, MYRG201700099].

Authors' contributions

J.Z. and G.L. conceived the study. J.Z. and G.L. collected and analyzed the data. J.Z. built the database and the Web server. All authors wrote and approved the manuscript.

Availability of data and material

TFmapper is freely accessible at http://www.tfmapper.org/. The complete source code for TFmapper is freely available at GITHUB (https://github.com/jmzeng1314/TF_map) under a GPLv3 license.

Competing Interests

The authors have declared that no competing interest exists.

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489: 57-74.
2. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology. 2010; 28: 1045-8.
3. Stunnenberg HG, Hirst M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016; 167: 1145-9.
4. Clough E, Barrett T. The Gene Expression Omnibus database. Methods in molecular biology (Clifton, NJ). 2016; 1418: 93-110.
5. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics. 2014; 15: 709.
6. Mendoza-Parra MA, Van Gool W, Mohamed Saleem MA, Ceschin DG, Gronemeyer H. A quality control system for profiles obtained by ChIP sequencing. Nucleic acids research. 2013; 41: e196.
7. Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. G3: Genes, Genomes, Genetics. 2014; 4: 209-23.
8. Mendoza-Parra MA, Gronemeyer H. Assessing quality standards for ChIP-seq and related massive parallel sequencing-generated datasets: When rating goes beyond avoiding the crisis. Genomics data. 2014; 2: 268-73.
9. Qin Q, Mei S, Wu Q, Sun H, Li L, Taing L, et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC bioinformatics. 2016; 17: 404.
10. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic acids research. 2017; 45: D658-d62.
11. Hitz BC, Rowe LD, Podduturi NR, Glick DI, Baymuradov UK, Malladi VS, et al. SnoVault and encodeD: A novel object-based storage system and applications to ENCODE metadata. PloS one. 2017; 12: e0175310.
12. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell. 2010; 38: 576-89.
13. Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The Human Epigenome Browser at Washington University. Nature methods. 2011; 8: 989-90.
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002; 12: 996-1006.
15. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008; 9: R137.
16. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature protocols. 2012; 7: 1728-40.
17. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011; 29: 24-6.
18. Morey L, Pascual G, Cozzuto L, Roma G, Wutz A, Benitah SA, et al. Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. Cell stem cell. 2012; 10: 47-62.
19. O'Loghlen A, Munoz-Cabello AM, Gaspar-Maia A, Wu HA, Banito A, Kunowska N, et al. MicroRNA regulation of Cbx7 mediates a switch of Polycomb orthologs during ESC differentiation. Cell stem cell. 2012; 10: 33-46.

20. Oza J, Ganguly B, Kulkarni A, Ginjala V, Yao M, Ganesan S. A Novel Role of Chromodomain Protein CBX8 in DNA Damage Response. The Journal of biological chemistry. 2016; 291: 22881-93.
21. Schmidt R, Plath K. The roles of the reprogramming factors Oct4, Sox2 and Klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation. Genome biology. 2012; 13: 251.
22. Engelen E, Akinci U, Bryne JC, Hou J, Gontan C, Moen M, et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. Nature genetics. 2011; 43: 607-11.
23. Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF, Pereira CF, Fisher AG, et al. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. PLoS genetics. 2010; 6: e1001023.
24. Feng W, Kawauchi D, Korkel-Qu H, Deng H, Serger E, Sieber L, et al. Chd7 is indispensable for mammalian brain development through activation of a neuronal differentiation programme. Nature communications. 2017; 8: 14758.
25. Whittaker DE, Riegman KL, Kasah S, Mohan C, Yu T, Sala BP, et al. The chromatin remodeling factor CHD7 controls cerebellar development by regulating reelin expression. The Journal of clinical investigation. 2017; 127: 874-87.
26. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic acids research. 2017; 45: D635-d42.
27. Livyatan I, Aaronson Y, Gokhman D, Ashkenazi R, Meshorer E. BindDB: An Integrated Database and Webtool Platform for "Reverse-ChIP" Epigenomic Analysis. Cell stem cell. 2015; 17: 647-8.
28. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic acids research. 2017; 45: D61-d7.
29. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic acids research. 2015; 43: e27.
30. Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell. 2016; 167: 1170-87.
31. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Molecular cell. 2013; 49: 825-37.
32. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research. 2009; 19: 1316-23.
33. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics (Oxford, England). 2008; 24: 2798-800.