



DATABASE

hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets



Qiong Zhang^{1, #}, Wei Liu^{1, 2, #}, Hong-Mei Zhang¹, Gui-Yan Xie¹, Ya-Ru Miao¹, Mengxuan Xia¹, An-Yuan Guo^{1, *}

¹ Department of Bioinformatics and Systems Biology, Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

² Clinical Center of Human Genomic Research, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China

Received 23 April 2019; revised 17 August 2019; accepted 23 October 2019
 Available online 26 August 2020

Handled by Jiang Qian

KEYWORDS

Transcription factor;
 ChIP-seq;
 Transcriptional regulation;
 Human;
 Database

Abstract Transcription factors (TFs) as key regulators play crucial roles in biological processes. The identification of TF–target regulatory relationships is a key step for revealing functions of TFs and their regulations on gene expression. The accumulated data of chromatin immunoprecipitation sequencing (**ChIP-seq**) provide great opportunities to discover the TF–target regulations across different conditions. In this study, we constructed a **database** named hTFtarget, which integrated huge **human** TF target resources (7190 ChIP-seq samples of 659 TFs and high-confidence binding sites of 699 TFs) and epigenetic modification information to predict accurate TF–target regulations. hTFtarget offers the following functions for users to explore TF–target regulations: (1) browse or search general targets of a query TF across datasets; (2) browse TF–target regulations for a query TF in a specific dataset or tissue; (3) search potential TFs for a given target gene or non-coding RNA; (4) investigate co-association between TFs in cell lines; (5) explore potential co-regulations for given target genes or TFs; (6) predict candidate TF binding sites on given DNA sequences; (7) visualize ChIP-seq peaks for different TFs and conditions in a genome browser. hTFtarget provides a comprehensive, reliable and user-friendly resource for exploring

* Corresponding author.

E-mail: guoay@hust.edu.cn (Guo AY).

Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2019.09.006>

1672-0229 © 2020 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

human TF–target regulations, which will be very useful for a wide range of users in the TF and gene expression regulation community. hTFtarget is available at <http://bioinfo.life.hust.edu.cn/hTFtarget>.

Introduction

Transcriptional regulation is a fundamental and vital process for general and condition-specific gene expression [1]. Transcription factors (TFs) are the key regulators involved in transcriptional regulation [2]. Most TFs recognize and bind to specific DNA sequences named as transcription factor binding sites (TFBSs), leading to specific spatiotemporal expression patterns of target genes [3]. The disorder of TF–target regulation can disrupt normal biological processes, which may result in severe damage and diseases [4]. Thus, the identification of TF–target regulation is an important issue for understanding transcriptional regulation underlying complex biological processes [5]. Moreover, investigating spatiotemporal TF–target regulations can provide comprehensive insights into the regulatory mechanisms of gene expression in different cell status and diseases [6].

Chromatin immunoprecipitation sequencing (ChIP-seq) technology provides convenience to systematically investigate target genes of a TF at the genome-wide level [7]. The accumulation of ChIP-seq data offers great opportunities to characterize the interactions between TFs and their targets in different conditions [8]. Meanwhile, comprehensive utilization and visualization of TF–target data can offer systematic views for TF–target regulations. Although several resources, such as ReMap [9], CistromeDB [10], Factorbook [11], ChIPBase [12], and TRRUST [13], were conducted to display the relationships between mammalian TFs and their targets, most of them indirectly provided evidence for TF–target regulations. For example, ReMap and Factorbook present experimental designs of ChIP-seq datasets and peak information instead of putative TF–target regulatory relationships, while TRRUST deposits TF–target regulatory networks using a sentence-based text mining approach. Therefore, integrating large-scale omics datasets for TFs (including TFBSs, target prediction, mRNA profiling, and epigenetic status of chromatin) can provide a comprehensive resource of TF–target regulations and benefit researchers in transcriptional regulation studies [14].

In this study, we developed a database named hTFtarget by creating a comprehensive repertoire of TF–target relationships for humans, which offers an almost one-stop solution for studies involved in TF–target regulation. The hTFtarget has integrated thousands of ChIP-seq datasets and epigenetic modification information to predict reliable TF–target regulations, and also provides online tools to predict potential co-association and co-regulation between TFs. In a word, hTFtarget can serve as a useful resource for researchers in the community of TF regulation and gene expression.

Implementation

hTFtarget was implemented with HTML, JavaScript (<https://www.javascript.com/>), Python (<https://www.python.org/>), Flask (<http://flask.pocoo.org/>), Bootstrap (<https://getbootstrap.com/>), AngularJS (a model-view-controller frame for

Javascript web service, <https://angularjs.org>), and WashU Epi-Genome Browser [15]. MongoDB (a cross-platform document-oriented database engine, <https://www.mongodb.com/>) was used to store metadata information. hTFtarget was hosted on the Ubuntu Linux system (version 16.04) with the Apache HTTP Server to provide a stable and open service.

Database content and methods

Data collection and quality control

Non-redundant ChIP-seq datasets of human TFs were curated from public databases, including NCBI Gene Expression Omnibus (GEO), NCBI Sequence Read Archive (SRA), and ENCODE [8]. In GEO and SRA databases, ChIP-seq datasets were collected through the E-Utilities toolkit with filter criteria “gds or sra, human or *Homo sapiens*, ChIP-seq or ChipSeq or ChIP sequencing, transcription factor or transcriptional factor or TF”. ChIP-seq datasets from ENCODE database were enrolled using parameters “assay_term_name = ChIP-seq, assembly = hg19/hg38, type = experiment, status = released, organism = *Homo sapiens*, target.investigated_as = transcription factor”. All datasets were manually curated to discard the non-TF and abnormal datasets, such as artificial TFs (mutated or fused), transcriptional co-factors, general TFII family members, and unclear descriptions for experimental designs. After completing the aforementioned procedures, FastQC (v0.11.5) was used for data quality control (QC) to obtain clean reads. Bowtie (v1.2.1) was employed to align clean reads to human reference genome GRCh38. Datasets with <5 million reads after the QC procedure or an alignment ratio <50% were discarded. Finally, 7190 samples of 659 TFs from 569 conditions (399 cell lines, 129 tissues or cells, and 141 treatments) were kept for further analyses.

Peak detection and motif discovery

Bam files of technically replicated samples from the same dataset were merged together, and the input or IgG samples were used as controls according to the experiment design (IgG samples were used as controls only at the absence of input samples). The peak calling procedure followed the protocol of MACS2 (v2.1.0) pipeline with the following parameters (q value ≤ 0.01 , fix-bimodal) [16]. Putative motifs of the TF in a dataset were identified using a similar method proposed by Wang and colleagues [17]. The detailed procedures are described as follows. (1) All of the peaks were ranked by the value of enrichment signal, and then the top 500 peaks (training set) were used for motif discovery using MEME-ChIP suite (v4.10.0) [18]. (2) The top five motifs (ranked by E-value, E-value $\leq 10^{-5}$ and the “match sites” ≥ 100 in the top 500 peaks of step (1) were considered as confident ones, and the top 501–1000 peaks from the step 1 served as the testing set to measure the power of the top five motifs. The same number of random genomic regions of GRCh38 reference genome

(non-peaks) with similar GC contents and length were selected as the control set. The FIMO (v4.10.0) [19] was used to scan both the testing and control sets, and then the numbers of recurrent motifs within the two sets were used to evaluate the significance of motifs for the TF (*t*-test with Bonferroni corrected *P* value < 0.01). Motifs with significant reoccurrence only in the testing sets were considered as high-confidence ones and used for peak filtration. (3) Peaks with a *P* value $\leq 10^{-5}$ and containing the aforementioned high-confidence motif(s) were considered as putative functional peaks and then used for further analyses.

Identification of TF–target regulation

Based on the putative functional peaks, we implemented the beta-model to identify candidate TF–target regulation [20], in which the beta-model score was used to measure the potential power of peaks for the TF–target regulation.

$$S_g = \sum_{i=1}^k e^{-(0.5+4\Delta i)}$$

where S_g is the beta-model score represented as the sum of the weighted scores of peaks nearby the TSS of gene *g*, and the parameter *k* is the number of TFBSs within the 50 kb upstream from the TSS, while Δi is the distance between the summit of peak *i* and the TSS (which was normalized to 50 kb; e.g., the value 1 represents a 50 kb distance, while 0.04 indicates 2 kb). For an extreme case, only one peak was detected within 50 kb, and the summit of the peak was exactly at 2 kb upstream the TSS, the beta-model score was 0.517. Thus, the value 0.517 was used as the cutoff for regulatory capacity. If the putative functional peak with a beta-model score ≥ 0.517 locates within 50 kb upstream of the TSS, we consider the TF–target regulation is reliable. Further analyses and functional modules in hTFtarget were based on reliable TF–target regulation.

Moreover, we integrated TF–target regulations from multiple datasets of each TF and the chromatin status of the upstream region of a target gene, to survey whether the TF–target relationship is general or condition-specific. First, for each TF, we combined all of the targets of the TF from multiple datasets together, and the upstream 50 kb region from the TSS of each target was divided to subunits with a bin length of 200 bp. Then we examined whether any of the subunits was labeled with activated epigenetic modifications (data source curated from Roadmap project, <http://www.roadmapepigenomics.org/>). Finally, if the TF–target regulation is supported by one peak with a beta-model score ≥ 0.517 in more than 30% datasets of the TF, and any of the subunits of the region 50 kb upstream of the target gene is labeled with an activated epigenetic modification status within at least 30% of the Roadmap epigenomic samples, we consider the TF–target regulation is a general case; otherwise it is a condition-specific one.

Detection of TF co-association

The co-association of TFs, which was predicted according to the method proposed by Mark and colleagues [21], indicates the probability of TFs co-binding to common genomic regions. Briefly, to investigate the co-associations between a given TF (focus-factor) and other TFs (partner-factors) in a specific con-

dition (a cell line or experiment), we first collected the overlapping regions of peaks between the focus-factor and each partner-factor to obtain a systematic co-binding map. We then implemented a machine learning method, which combined the RuleFit3 algorithm and mutual information of the overlapping regions between the focus-factor and partner-factors, to calculate the quantitative relationship as the relative importance score between the focus-factor and each partner-factor. A higher relative importance score of two given TFs indicates a stronger and more confident co-association relationship between them.

Co-regulation of TFs for the same target gene

Target gene co-regulated by more than two TFs was predicted based on the reliable TF–target regulations detected from curated ChIP-seq datasets in hTFtarget. Briefly, for a query gene, the region 50 kb upstream of the transcription start site (TSS) served as a core area to predict the candidate co-regulation of TFs. TFs with high-confidence peaks within the region 50 kb upstream of TSS of target gene and the beta-model score of each peak ≥ 0.517 were considered as putative co-regulators to the query gene.

Sequence-based TFBS prediction

For the prediction of candidate TFBS(s) on given sequence(s), we integrated the motif matrices of humans from the hTFtarget database and other resources, including HOCOMOCO [22], JASPAR [23], and TRANSFAC v2018 databases [24]. In total, we collected TFBSs for 864 human TFs, and the FIMO tool was employed for the TFBS discovery on given sequences.

Data resource and web-interface features

hTFtarget deposited 3.2 million records of TF–target regulations for 659 TFs in 569 experimental conditions and integrated 2737 high-confidence motifs of 699 TFs from other databases to predict 3.5 million records of candidate TF–target regulations. Moreover, 20,320 motifs were curated in the hTFtarget database from ChIP-seq data. Meanwhile, 408 TFs were found to possess co-association ability in 10 cell lines. Furthermore, hTFtarget provides a user-friendly web interface to facilitate search, browse, and download of comprehensive TF–target relationships in multiple experimental datasets, including putative TF–target regulations, ChIP-seq peaks, epigenetic modification status of TFBSs and targets, co-regulation of TFs of the same targets, and co-association of TFs. All of the data sources and functional modules in hTFtarget are shown in Figure 1. We also compared hTFtarget with other resources related to TF–target regulation (Table 1). We found that hTFtarget may be the most comprehensive database providing various functions for the exploration of TF–target regulations in humans.

Usage

First, hTFtarget provides a “Quick Search” function on the top right of each page for users to conveniently survey the

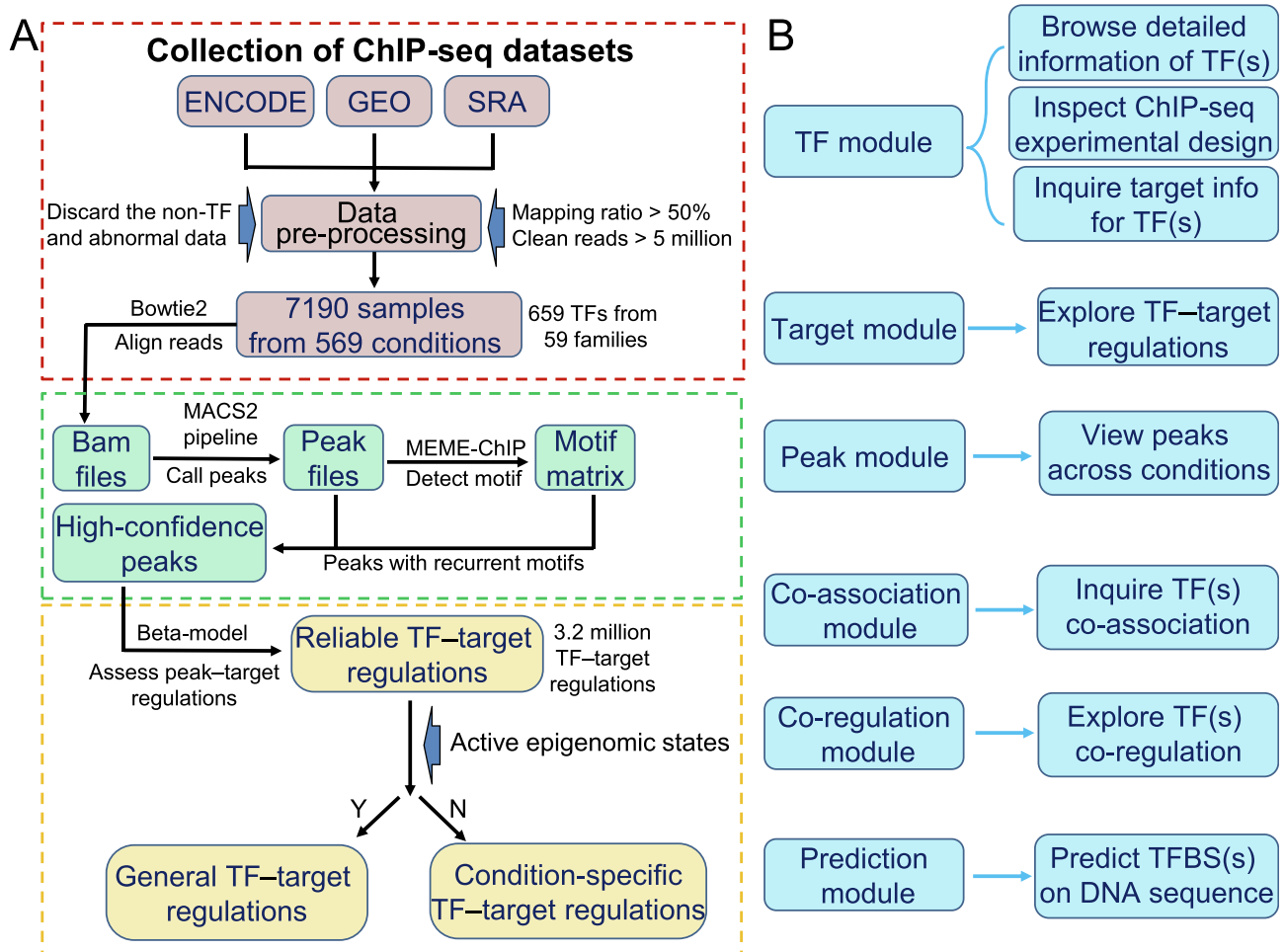


Figure 1 Overview of data resources and functional modules of hTFtarget

A. The resource summary and workflow for the detection of TF–target regulations in hTFtarget. **B.** The main functional modules of hTFtarget. TF, transcription factor; ChIP-seq, Chromatin immunoprecipitation sequencing.

TF–target regulations. Various types of inputs are acceptable, such as the gene name, the Ensembl gene ID, gene symbol, and alias. Users can obtain basic information for the query TF or gene, related TF–target regulations, epigenetic modification status in the promoter region of target gene, and detailed evidence for the TF–target regulation in different conditions. For example, when inputting gene *ASCL2* on the “Quick Search” function (Figure 2), all records of *ASCL2* as a TF gene or a target gene are shown in Figure 2A.

On one hand, when the query gene *ASCL2* is a TF gene, the “Details” icon below the term “Targets” (red box in Figure 2A) links to a page displaying its targets with three levels of evidence (Figure 2B). Additionally, users can filter target genes by support levels (“Epigenetic Evidence”, “ChIP-seq Evidence”, and “Motifs Evidence”) and browse the peak information through the “view” icon (Figure 2C). On the other hand, when the query gene *ASCL2* is a target gene, the “Details” label underlying the term “TFs” (blue box in Figure 2A) shows all records about which TFs could regulate the target gene *ASCL2* (Figure 2D). Users can browse the peak information

for the TF–target regulation by clicking the “Details” icon (Figure 2E).

Other than the quick search function, hTFtarget offers the following six modules for users to conveniently explore TF–target regulations.

TF module

The TF module provides details of TF–target relationships and ChIP-seq experimental designs (Figure 3A and B). Users can explore the general (Figure 3A) or condition-specific TF–target regulations (Figure 3B), as well as browse the TF annotation (the bottom of Figure 3A) and the experimental design of the interested dataset (the bottom of Figure 3B). The “details” icons in Figure 3A enable users to browse the detailed information of TF–target regulations (Figure 3C), including the epigenetic status of the upstream region for the target gene and the visualized peak information in the genome browser (Figure 3D). Additionally, beyond surveying TF–target regulation in hTFtarget online, users can download the records as well.

Table 1 Summary of well-known databases related to TF–target regulation

Parameter	hTFtarget	Current study	ReMap [9]	CistromeDB [10]	Factorbook [11]	TRRUST [13]	ChIPBase [12]
Data	Experiment	Experiment	Experiment	Experiment	Experiment	Literature	Experiment
Source	ChIP-seq	ChIP-seq	ChIP-seq	ChIP-seq, DNase-seq, ATAC	ChIP-seq	Text mining	ChIP-seq, DNase-seq, ATAC
Technology	ChIP-seq	ChIP-seq	ChIP-seq	ChIP-seq, DNase-seq, ATAC	ChIP-seq	Text mining	ChIP-seq, DNase-seq, ATAC
No. of TFs	659	659	485	1700	167	800	480
No. of datasets	7190	7190	2829	13,976	837	ND	2498
No. of cells/conditions	399/170	399/170	346/0	ND	ND	ND	ND
Species	Human	Human	Human	Human, mouse	Human	Human, others	Human, others
TFBS prediction	Y	Y	N	N	N	N	N
TFs co-association	Y	Y	N	N	N	N	N
TFs co-regulation	Y	Y	N	N	N	N	N
Target search for TF	Y	Y	N	Y	N	Y	Y
TF search for target	Y	Y	N	Y	N	N	N
Peak view	Y	Y	Y	Y	N	N	Y
Peak comparison	Y	Y	N	Y	N	N	N
Epigenetic status	Y	Y	Y	Y	Y	N	Y

Note: CistromeDB (<http://cistrome.org/>) collects a large number of transcriptional co-factors, RNA polymerases, as well as TFII family members and their components, resulting in an extremely high number of TFs and datasets. TF, transcription factor; TFBS, transcription factor binding site; ChIP-seq, chromatin immunoprecipitation sequencing; ATAC, assay for transposase-accessible chromatin. ReMap: <http://pedagogix-tage.univ-mrs.fr/remap/>; Factorbook: <https://factorbook.org/>; TRRUST: <http://www.grnpedia.org/trrust/>; ChIPBase: <http://rna.sysu.edu.cn/chipbase/>. ND, not declared (the corresponding databases do not declare the numbers of cells/conditions for humans in papers and websites).

Target module

The target module helps users to investigate the TF(s) that regulate(s) the query gene. The function and the interface of this module are similar to what are shown in Figure 2D and 2E.

Peak module

The peak module provides comprehensive information of peaks in a user-defined manner (Figure 4A). Users can browse the peaks of different TFs in the selected condition (e.g., the same cell line or tissue) or investigate TF–target regulations for the same TF in different conditions (e.g., the distinct spatiotemporal TF–target regulatory relationships in different tissues).

Co-regulation module

The co-regulation module helps users to search the common targets that can be regulated by the given TFs, or search common TFs that can regulate the given genes (Figure 4B).

Co-association module

The co-association module predicts and visualizes potential co-association TFs in the collected cell lines. The co-association of two TFs indicates the probability of their combinatorial occupancies on the same genomic regions. The human TFs have always shown distinct co-association relationships in combinatorial and context-specific patterns for gene regulation [21,25]. The relative importance score indicates the probability of co-association between two TFs, and the higher score means stronger co-association in the corresponding condition.

Prediction module

The prediction module employs motif matrices to predict potential TFBSs on the user-provided sequence(s) (Figure 4C). The motif matrices were curated from TRANSFAC, JASPAR, HOCOMOCO, and hTFtarget databases.

Discussion

Understanding TF–target regulatory relationships is a critical issue when investigating complex molecular mechanisms underlying biological processes and diseases [26,27]. Up to date, several resources have been dedicated to identify TFs, such as PlantTFDB and AnimalTFDB [28,29], and large-scale of ChIP-seq data have been available for the detection of TF–target regulatory relationships as well. Integrative resources would offer significant insights into the dynamic TF–targets regulatory mechanisms [30]. In this study, we presented the hTFtarget, a comprehensive resource focusing on the regulation between human TFs and their targets by integrating ChIP-seq data and TFBS prediction.

Although several resources could indirectly investigate potential regulatory interaction between TF(s) and target(s) based on ChIP-seq peak signals, a database concisely indicat-

A

Result(s) of query gene <i>ASCL2</i> as a TF						
#	TF	No. of datasets	Full Name	TF Family	Location	Targets
1	ASCL2	3	Achaete-scute family bHLH transcription factor 2	bHLH	11p15.5	Details

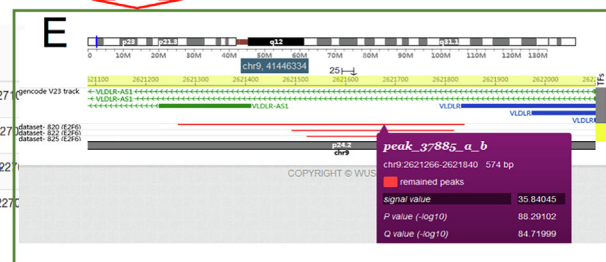
Query gene: *ASCL2* may be regulated by the following TF(s):

#	Ensembl ID	Gene Name	Chromosome	Start	End	Biotype	TFs
1	ENSG00000183734	ASCL2	Chr11	2268495	2270952	Protein_coding	Details

Chromosome	Start	End	Gene id	Gene name	Gene type	Epigenomic evidence	ChIP-seq evidence	Motif evidence	Detailed information
Chr1	924880	944581	ENSG00000187634	SAMD11	Protein_coding	⊖	⊕	⊕	View
Chr1	1020123	1056118	ENSG00000188157	AGRN	Protein_coding	⊖	⊖	⊕	View
Chr1	1070966	1074307	ENSG00000237330	RNF223	Protein_coding	⊖	⊕	⊖	View

#	Cell line	Cell type	Tissue	ES	No. of targets	Browser
173	LoVo	Colorectal carcinoma cell	Colon	⊖	1,619	Browser
174	VCaP	Epithelial	Prostate	⊖	4,395	Browser
175	VCaP	Epithelial	Prostate	⊕	4,773	Browser

Browse peaks	No. of datasets	TF	Tissue	No. of peaks (total/average)	No. of peaks in gene body (total/average)	No. of peaks around TSS (total/average)	
Details	31	SPI1	Muscle	76 / 2	57 / 1	19 / 0	Chr11,227
Details	27	CTCF	Blood	26 / 0	26 / 0	0 / 0	Chr11,227
Details	24	CTCF	Breast	44 / 1	20 / 1	12 / 0	Chr11,227
Details	24	ESR1	Breast	76 / 3	45 / 1	31 / 1	Chr11,227

**Figure 2** Snapshots of the quick search function in hTFtarget

A. Results of a quick search function provide comprehensive views for TF–target regulations. **B.** A partial screenshot of the results for the query gene as a TF gene after clicking the “details” icon. Each record represents a target gene regulated by the query TF. **C.** Browsing of the condition-specific TF–target regulation in different experimental conditions or cell lines. **D.** A part of the results for the query gene as a target gene. Each record indicates a TF–target regulation in a certain tissue. **E.** The peak information of the query gene as a target gene.

ing target genes for specific TFs and revealing comprehensive regulations of TF–targets is still lacking. For example, although a large number of datasets and TFs has been collected in CistromeDB, it does not distinguish TFs, transcriptional co-factors, RNA polymerases, TFII family members, or their compounds. Additionally, although CistromeDB provides a search function for TF–target regulations of a single query gene and browse peak information of a given TF, it ignores the co-regulation information for TF–targets, and does not refer to the TF–target regulations at the genome-wide level. The ReMap database just provides peak information for TFs in ChIP-seq datasets. The HOCOMOCO [22], JASPAR [23], TRANSFAC v2018 [24], and TFBSbank [31] databases focus on the collection and discovery of the motif profiles for TFBSs, while TRRUST elucidates TF–target interactions using text mining [13]. hTFtarget is the only currently available database integrating various resources to provide an almost one-stop solution in investigating TF–target regulations in humans.

Further development of hTFtarget will be focused on spatiotemporal classification of TF–target regulations in different

gene families and diseases, such as integrating with GSCALite cancer analysis [32]. Additionally, we will continue to upgrade hTFtarget with new datasets and more epigenetic modifications, keep improving algorithms for better accuracy, and offer more functions and tools. We will maintain and regularly update hTFtarget as we do for our AnimalTFDB database, which have been maintained for more than eight years. We believe that hTFtarget will be a valuable resource for the research community.

Conclusion

The identification of TF–target regulations is an important issue for revealing the molecular mechanisms underlying complex biological processes. Huge amount of ChIP-seq resources and epigenetic modification information have been integrated into hTFtarget to identify putative TF–target regulations in humans. hTFtarget provides a comprehensive, reliable, and user-friendly resource, and an almost one-stop solution to explore TF–target regulations in humans.

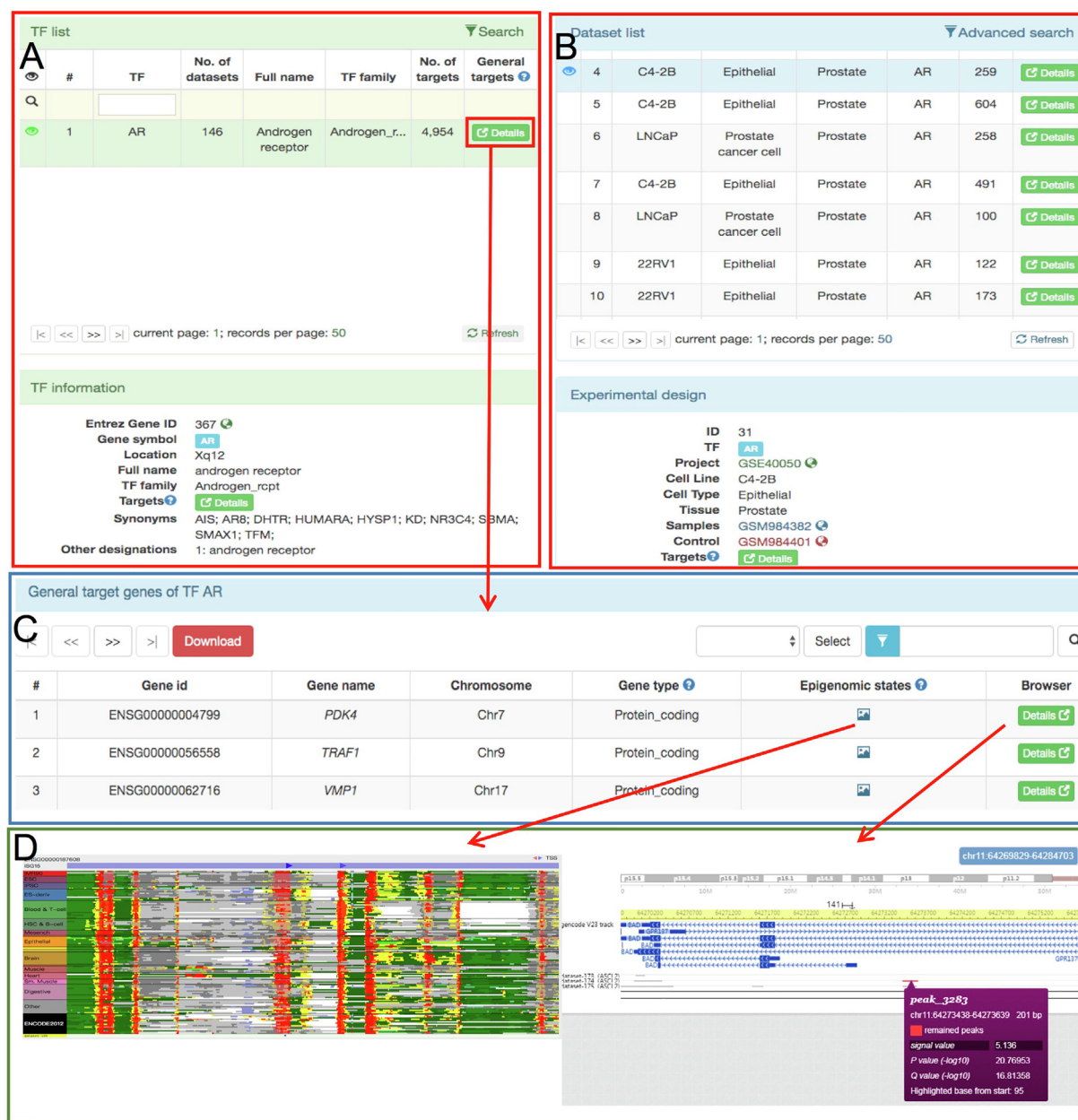


Figure 3 Views for the TF and target related modules in hTFtarget

A. The basic information of the collected TFs and datasets in hTFtarget. **B.** The experimental design and data information of datasets for the given TF. **C.** General target genes of the selected TF. **D.** Epigenetic status and peak details within the flanking region of the target gene.

Data availability

hTFtarget is freely accessible at <http://bioinfo.life.hust.edu.cn/hTFtarget/>.

CRedit author statement

Qiong Zhang: Methodology, Software, Data curation, Resources, Formal analysis, Visualization, Writing - original draft, Writing - review & editing, Funding acquisition. **Wei Liu:** Methodology, Software, Data curation, Formal analysis,

Visualization. **Ya-Ru Miao:** Formal analysis. **Mengxuan Xia:** Formal analysis. **Hong-Mei Zhang:** Methodology, Software, Data curation, Formal analysis. **Gui-Yan Xie:** Formal analysis. **Ya-Ru Miao:** Formal analysis. **Mengxuan Xia:** Formal analysis. **An-Yuan Guo:** Conceptualization, Project administration, Funding acquisition, Writing - review & editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

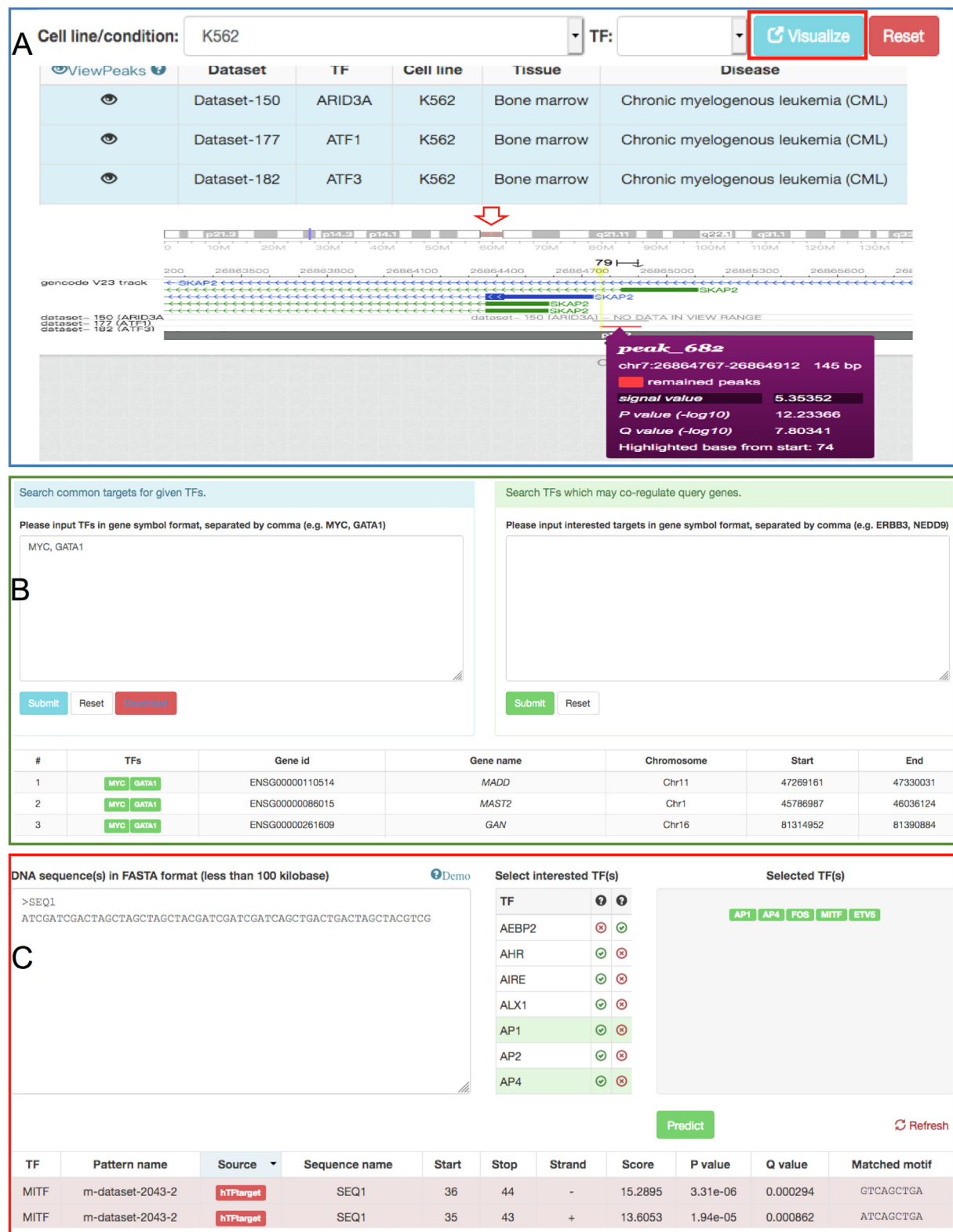


Figure 4 Other important modules in hTFtarget

A. Peak visualization for TFs in user-customized cell lines. **B.** Potential targets and TF co-regulation analysis for input gene sets. **C.** TFBS prediction for given sequences. TFBS, transcription factor binding site.

Acknowledgments

We would like to thank colleagues in groups of Ensembl, GEO, SRA, ENCODE, TRANSFAC, JASPAR, HOCOMOCO, Roadmap Epigenomics, and AnimalTFDB. We also acknowledge the funding from the National Natural Science Foundation of China (Grant Nos. 31822030, 31801113, and 31771458), National Key R&D Program of China (Grant No. 2017YFA0700403), and China Postdoctoral Science Foundation (Grant No. 2018M632830).

ORCID

0000-0001-5001-8434 (Zhang, Q)
 0000-0001-9881-785X (Liu, W)
 0000-0001-9835-9684 (Zhang, HM)
 0000-0002-2924-5306 (Xie, GY)
 0000-0002-1982-3141 (Miao, YR)
 0000-0002-8270-4444 (Xia, M)
 0000-0002-5099-7465 (Guo, AY)

References

- [1] Lin Y, Zhang Q, Zhang HM, Liu W, Liu CJ, Li Q, et al. Transcription factor and miRNA co-regulatory network reveals shared and specific regulators in the development of B cell and T cell. *Sci Rep* 2015;5:15215.
- [2] Zhang Q, Hu H, Chen SY, Liu CJ, Hu FF, Yu J, et al. Transcriptome and regulatory network analyses of CD19-CAR-T immunotherapy for B-ALL. *Genomics Proteomics Bioinformatics* 2019;17:190–200.
- [3] Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 2004;116:247–57.
- [4] Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell* 2013;152:1237–51.
- [5] Tang Q, Zhang Q, Lv Y, Miao YR, Guo AY. SEGReg: a database for human specifically expressed genes and their regulations in cancer and normal tissue. *Brief Bioinform* 2019;20:1322–8.
- [6] Zhang Q, Liu W, Liu C, Lin SY, Guo AY. SEGtool: a specifically expressed gene detection tool and applications in human tissue and single-cell sequencing data. *Brief Bioinform* 2018;19:1325–36.
- [7] Kaufmann K, Muiño JM, Østerås M, Farinelli L, Krajewski P, Angenent GC. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nat Protoc* 2010;5:457–72.
- [8] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
- [9] Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res* 2018;46:D267–75.
- [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 Res* 2017;45:D658–62.
- [11] Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013;41:D171–6.
- [12] Zhou KR, Liu S, Sun WJ, Zheng LL, Zhou H, Yang JH, et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res* 2017;45:D43–50.
- [13] Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* 2018;46:D380–6.
- [14] Thompson PJ, Macfarlan TS, Lorincz MC. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol Cell* 2016;62:766–76.
- [15] Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The Human Epigenome Browser at Washington University. *Nat Methods* 2011;8:989–90.
- [16] Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9:R137.
- [17] Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012;22:1798–812.
- [18] Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc* 2014;9:1428–50.
- [19] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;27:1017–8.
- [20] Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, Wang Q, et al. A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res* 2011;71:6940–7.
- [21] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489:91–100.
- [22] Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* 2018;46:D252–9.
- [23] Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016;44:D110–5.
- [24] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;34:D108–10.
- [25] Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* 2014;512:453–6.
- [26] Lim WA, Lee CM, Tang C. Design principles of regulatory networks: searching for the molecular algorithms of the cell. *Mol Cell* 2013;49:202–12.
- [27] Zhang HM, Kuang S, Xiong X, Gao T, Liu C, Guo AY. Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief Bioinform* 2015;16:45–58.
- [28] Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 2017;45:D1040–5.
- [29] Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res* 2019;47:D33–8.
- [30] MacNeil LT, Walhout AJM. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* 2011;21:645–57.
- [31] Chen D, Jiang S, Ma X, Li F. TFBSbank: a platform to dissect the big data of protein–DNA interaction in human and model species. *Nucleic Acids Res* 2017;45:D151–7.
- [32] Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. *Bioinformatics* 2018;34:3771–2.