Computational and Structural Biotechnology Journal 23 (2024) 1877–1885

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database article

TFTG: A comprehensive database for human transcription factors and their targets

Xinyuan Zhou [a,j,1], Liwei Zhou [h,1], Fengcui Qian [a,c,e,1], Jiaxin Chen [i,1], Yuexin Zhang [e,1], Zhengmin Yu [d], Jian Zhang [i], Yongsan Yang [i], Yanyu Li [i], Chao Song [e], Yuezhu Wang [i], Desi Shang [e], Longlong Dong [i], Jiang Zhu [i,*], Chunquan Li [a,b,c,d,e,f,g,**], Qiuyu Wang [e,*]

[a] The First Affiliated Hospital & Hunan Provincial Key Laboratory of Multi-omics And Artificial Intelligence of Cardiovascular Diseases, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China
[b] Hunan Provincial Maternal and Child Health Care Hospital, National Health Commission Key Laboratory of Birth Defect Research and Prevention, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China
[c] Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences & MOE Key Lab of Rare Pediatric Diseases, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China
[d] School of Computer, University of South China, Hengyang, Hunan 421001, China
[e] The First Affiliated Hospital, Cardiovascular Lab of Big Data and Imaging Artificial Intelligence, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China
[f] The First Affiliated Hospital, Institute of Cardiovascular Disease, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China
[g] The First Affiliated Hospital, Department of Cardiology, Hengyang Medical School, University of South China, Hengyang, China
[h] State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China
[i] School of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing 163319, China
[j] College of Artificial Intelligence and Big Data For Medical Sciences, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, China

ARTICLE INFO

Keywords:
Transcription factor
Target gene
Epigenetic annotation
Functional annotation
Enrichment analysis

ABSTRACT

Transcription factors (TFs) are major contributors to gene transcription, especially in controlling cell-specific gene expression and disease occurrence and development. Uncovering the relationship between TFs and their target genes is critical to understanding the mechanism of action of TFs. With the development of high-throughput sequencing techniques, a large amount of TF-related data has accumulated, which can be used to identify their target genes. In this study, we developed TFTG (Transcription Factor and Target Genes) database (http://tf.liclab.net/TFTG), which aimed to provide a large number of available human TF-target gene resources by multiple strategies, besides performing a comprehensive functional and epigenetic annotations and regulatory analyses of TFs. We identified extensive available TF-target genes by collecting and processing TF-associated ChIP-seq datasets, perturbation RNA-seq datasets and motifs. We also obtained experimentally confirmed relationships between TF and target genes from available resources. Overall, the target genes of TFs were obtained through integrating the relevant data of various TFs as well as fourteen identification strategies. Meanwhile, TFTG was embedded with user-friendly search, analysis, browsing, downloading and visualization functions. TFTG is designed to be a convenient resource for exploring human TF-target gene regulations, which will be useful for most users in the TF and gene expression regulation research.

1. Introduction

Transcriptional regulation serves as a pivotal mechanism in dictating the gene expression within organisms [1,2]. Serving as cellular markers, transcription factors (TFs) directly interpret the genome and interact with transcription co-factors (TcoFs) to bind DNA sequences regulate gene expression [3–5]. With the development of high-throughput sequencing technology, a large amount of TFs and TF-related data

* Corresponding authors.
** Corresponding author at: The First Affiliated Hospital & Hunan Provincial Key Laboratory of Multi-omics And Artificial Intelligence of Cardiovascular Diseases, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China.
E-mail addresses: hydzhujiang@126.com (J. Zhu), lcqbio@163.com (C. Li), wangqiuyu900490@163.com (Q. Wang).
[1] The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

https://doi.org/10.1016/j.csbj.2024.04.036
Received 27 January 2024; Received in revised form 15 April 2024; Accepted 15 April 2024
Available online 16 April 2024

(such as ChIP-seq data) has been accumulated, which can be used for identifying the TF binding sites (TFBS) on the whole genome and constructing transcriptional regulatory network. These networks are crucial in synchronizing the spatial and temporal expression patterns of genes [6,7]. The existing methods for identifying TF-target genes are classified into four categories: (I) The ChIP-seq for TFs in special cellular contexts can be used for identifying cell-type-specific TF-target genes using software such as BETA [8]. (II) Motif scanning can predict TF binding sites on the whole genome using software such as FIMO [9]. (III) The differential expressed genes obtained from RNA-seq data before and after perturbing TF are usually considered as TF-target genes at the expression level [10]. (IV) The accurate TF-target genes can be confirmed using low-throughput experiments. Furthermore, the previous studies focused on TF binding information in promoter regions [11,12]. Increasing evidence showed that TFs could also regulate genes by binding to distal regulatory elements, including enhancers, super-enhancers (SEs) and chromatin accessibility regions [13,14]. For example, typical enhancer and SE could cause the tissue-specific transcription of genes in liver metastatic colorectal cancer (CRC) tumors. FOXA2 and HNF1A, as the affirmed liver-specific TFs, could mediate unique enhancer changes and activate a set of liver-specific genes in hepatic metastatic CRC cells which leads to CRC liver metastasis [15]. Overall, the identification of TF-target genes based on distal regulatory element has received an increasing attention.

Some TF-related resources have been developed to provide TF-target genes based on different strategies, Among these, the CistromeDB [16] and hTFtarget [17] databases store human TF-target genes obtained by the ChIP-seq datasets using the BETA method. The KnockTF [18] database identifies TF-target genes using perturbation RNA-seq data of TFs. TRRUST [19] provides validated TF-target genes interactions through manual curation from the literature. MotifMap [20] predicts TF-target genes based on TF motif profiles. These databases can help researchers obtain information on TF-target genes of interest. However, none of them combine multiple methods to balance comprehensiveness and accuracy. More importantly, most resources focus on the proximal regulation of TFs by binding to promoter regions, while ignoring distal regulatory mechanisms through DNA regulatory elements such as SEs and enhancers. Hence, many potential target genes are missed. Until now, a relatively comprehensive database to describe TF-target genes has not been available. As a large amount of data associated with TFs and DNA regulatory elements has accumulated, the comprehensive function annotations of TFs and target genes to facilitate further dissection of the regulatory mechanisms of TFs has become an urgent need. Therefore, it is highly necessary to construct a fully integrated TF-target genes resource and provide the associated regulatory analyses and annotations of TFs.

To address these needs, we developed the TFTG database platform (http://tf.liclab.net/TFTG), which aimed to document broadly TF-target gene resources combining multiple strategies and provide extensive annotations and analysis. The current version of TFTG integrates 11,056 human TF ChIP-seq datasets with more than 700 tissues/cell types, 414 TF perturbation RNA-seq datasets involving about 200 tissues/cell types, 7966 TF-target regulations from more than 5000 published studies and more than 3000 DNA binding motifs for 805 TFs. Furthermore, we collected the most comprehensive DNA regulatory regions, including promoters, super enhancers, typical enhancers, silencers and chromatin accessibility regions, to identify TF-target genes in distal regulation. In particular, TFTG also focused on TF-related annotation information, including TF-associated pathways, Gene Ontology (GO) terms, cancer hallmarks, expression and disease information. Besides the three query modes, TFTG embedded four analyses for TF regulation, including TF gene set enrichment, TF downstream regulatory analysis,

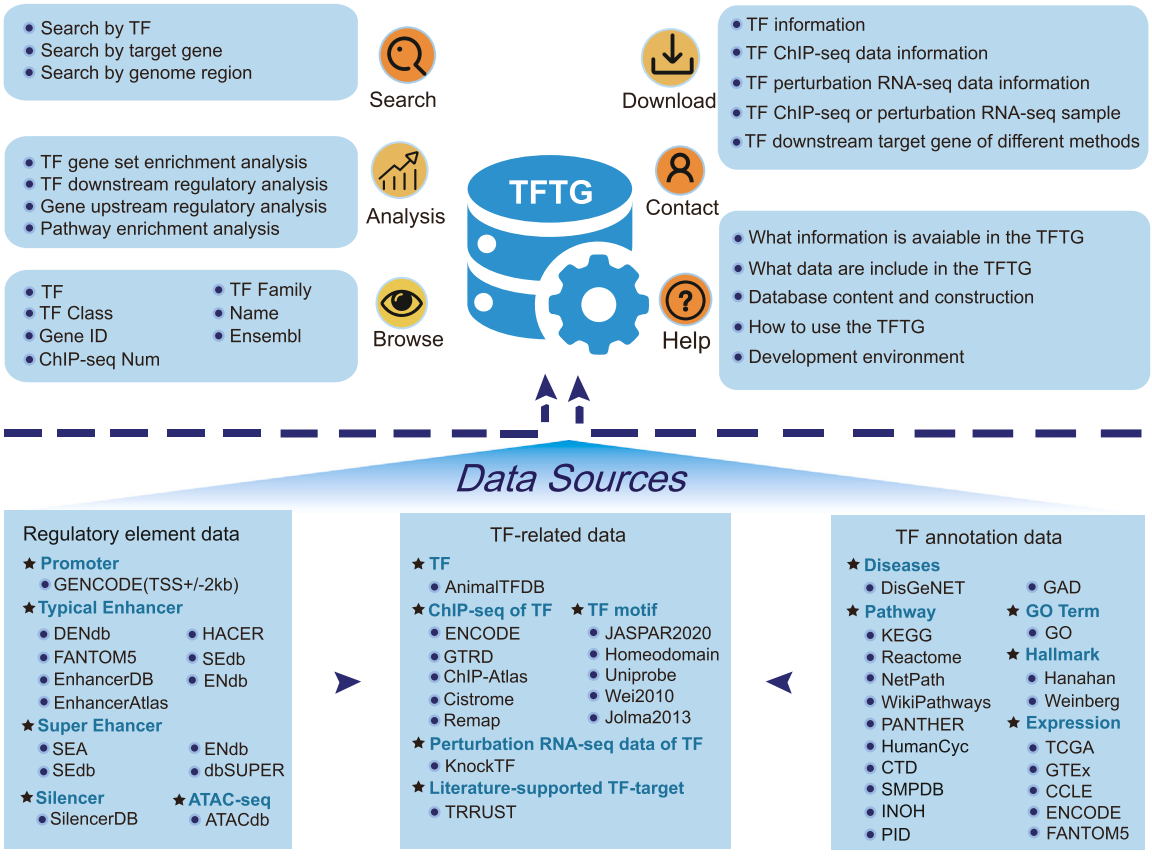

Fig. 1. Database content and construction. TFTG has plenty of TF-related data, functional and epigenetic annotations for TFs and multiple functions including browse, search, analysis, and download.

gene upstream regulatory analysis and pathway enrichment analysis. TFTG was a relatively comprehensive human TF-target gene database that integrated multiple functions of annotation, storage, browsing, search and analysis (Fig. 1). Overall, TFTG will be helpful for elucidating TF regulation and gene expression and exploring potential biological mechanisms.

2. Materials and methods

2.1. TF-related data

The list for human TFs was obtained from AnimalTFDB 3.0 [21] (Fig. 1 middle-bottom panel). To comprehensively annotate TF-target genes, we curated a large number of TF associated data that can be used for identification of TF-target genes.

TF ChIP-seq data. The TF ChIP-seq datasets involving a large number of human tissue/cell types were collected from five public sources, including ENCODE [22], Cistrome [16], Remap [23], ChIP-Atlas [24] and GTRD [25] (Fig. 1 middle-bottom panel). All ChIP-seq datasets were further deduplicated by manual screening according to the unique GEO/SRA ID to avoid duplication (Supplementary Materials). For the uniformity of genome version, peaks from ChIP-seq were converted to the hg38 genome using the liftOver tool of UCSC [26] (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Finally, 11,056 datasets were collected and processed, involving 1043 TFs and 743 tissue/cell types.

TF perturbation RNA-seq data. All TF perturbation RNA-seq data were collected from the KnockTF database developed by our group (Fig. 1 middle-bottom panel). KnockTF curated these datasets from NCBI GEO [27] and ENCODE [22] using a list of keywords, such as 'knockout', 'knockdown', 'siRNA', 'shRNA' and 'CRISPR'. We further traversed the title, summary and protocol of preliminary screening results to ensure data quality. Overall, we collected 414 perturbation RNA-seq datasets of 219 TFs involving 187 tissues/cell types.

TF Motif profile data. The motif of TF was also used to identify TF-target genes. As a result, we obtained >3000 DNA binding motifs from the TRANSFAC [28] and MEME suite [29,30] deriving from the following five sources: JASPAR CORE 2020 vertebrates [31], Jolma2013 [32], Homeodomain [33], UniPROBE [34] and Wei2010 [35] (Fig. 1 middle-bottom panel). Finally, we collected 805 TFs.

Literature-supported TF-target pairs. Literature-supported TF-target genes are generally considered the gold standard dataset. Thus, we also collected literature-supported TF-target pairs from TRRUST [19], which contained 7966 TF-target relationships from 5256 published studies (Fig. 1 middle-bottom panel). Some of TF-target relationships also indicated the activation or repression.

2.2. Regulatory element data

TFTG not only focused on the promoter regions but also contained the distal regulatory regions to explore TF-target genes comprehensively, thus providing a better understanding for the research on the function and regulatory mechanism of TFs. The DNA regulatory elements used in TFTG included promoters, super-enhancers (SEs), enhancers, silencers and accessible chromatin regions.

Promoter regions. Promoters are DNA sequences located upstream of the 5′ end of structural genes that activate RNA polymerase to bind to template DNA accurately and have specificity for transcription initiation. We defined promoter regions as 2 kb upstream and 2 kb downstream of the transcription start sites (TSSs) of genes, which were collected from GENCODE [36] (Fig. 1 bottom left panel).

Enhancers. Enhancers can be occupied by a large number of TFs to enhance the transcription of genes, which play important role in biological processes. We comprehensively curated the resources provided in the existing enhancer database, including EnhancerAtlas [37], FAMTOM5 [38], HANCER [39], EnhancerDB [40], DENdb [41], SEdb [42] and ENdb [43] (Fig. 1 bottom left panel). Finally, 27,468,231 enhancer regions from 2459 tissues/cell types were collected.

Super-enhancers. Super-enhancers are a large cluster of transcriptionally active enhancers that are richer in enhancer-associated chromatin features and have a greater ability to control and define cell-specific gene expression than enhancers. We obtained SE regions from SEdb [42] database developed by our group. Briefly, we curated H3K27ac ChIP-seq raw data from Roadmap [44], ENCODE [22], NCBI GEO/SRA [27,45] and Genomics of Gene Regulation Project (GGR) [22]. We identified 331,551 SE regions involving 542 tissues/cell types using the streamlined pipeline of Bowtie-MACS14-ROSE [46–48]. In addition, we also downloaded SEs from other projects including SEA [49], ENdb [43] and dbSuper [50] (Fig. 1 bottom left panel). At last, we collected 4335,093 SEs.

Silencers. Silencer is a special sequence in a eukaryotic gene that remotely regulates the promoter to slow down transcription. We collected 3558,081 silencers for 201 tissues/cell types from SilencerDB [51] (Fig. 1 bottom left panel).

Accessible chromatin regions. Accessible chromatin regions are a highly informative structural feature for identifying regulatory elements, which provides a large amount of information about transcriptional activity and gene regulatory mechanisms. More than 2200 publicly available human ATAC-seq samples were manually curated from NCBI GEO/SRA [27,45] by ATACdb [52] developed by us (Fig. 1 bottom left panel). After filtering and running Bowtie2-MACS2 [46,53] 52,078,883 accessible chromatin regions were identified covering 1400 tissues/cell types.

2.3. Identification of TF-target relationships

TFTG employed a variety of strategies to identify TF-target genes so as to provide more evidence of the regulation between TFs and target genes and their regulatory mechanisms. We divided all the TF-target genes resources into 4 categories (ChIP-seq_class, Perturbation_class, Motif_class and Curate_class) and 14 sub-categories based on the different identification strategies (Supplementary Materials and Fig. S1).

TF-targets based on TF ChIP-seq data (ChIP-seq_class). The ChIP-seq category was divided into seven sub-categories as follows: (I) For TF peaks, we measured the potential power for the TF-target regulations using the BETA [8] method. The beta-model score was calculated based on the number of peaks within a certain range and the distance between the peak and TSS; (II) We used a Python script geneMapper.py from ROSE [47] to annotate downstream target genes for TFs. This script provided the closest, proximal and overlapping genes for TFs according to the genomic distance; (III) A gene was considered as TF-target when the peaks of this TF overlapped with the gene promoter region. The overlapping information was calculated using BEDTools [54] software; (IV-VII) Numerous studies showed that TFs could regulate target genes through corresponding regulatory elements. Thus, we first used ROSE geneMapper.py [47] to annotate downstream target genes of these regulatory elements. Then, we identified TF-target genes when the gene-associated SEs, enhancers, silencers or accessible chromatin regions were occupied by TF binding sites in matched tissues/cell types.

TF-targets based on motif scanning (Motif_class). The motif category was divided into five sub-categories according to the types of DNA regulatory elements with TF motif occurrence. We used FIMO [9] software of the MEME suite [29] to perform motif scanning. We took the TF-targets as these genes associated with promoters, SEs, enhancers, silencers or accessible chromatin regions with TF motif occurrence.

Perturbation (Perturbation_class). These genes with expression changes observed after interfering with TF are generally considered to be TF-target genes. We mapped Ensembl IDs to gene symbols with regard to each gene expression profile and deleted genes with zero values in all control or case samples. Then the raw expression values of gene expression profiles were processed by Log2 transformation. We further computed fold change (FC) for each gene. For datasets with more than three samples, The R package limma-voom was used to compute the

statistical significance of differential expression. We extracted differentially expressed genes under the threshold of FC $\geq$ 3/2 or FC $\leq$ 2/3 as TF-target genes. In the end, we obtained 902,693 TF-target pairs.

Curation (Curate_class). We collected literature-supported TF-target regulations from more than 5000 published studies from TRRUST [19]. TRRUST first manually collected Medline abstracts. Then, it used a method named sentence-based text_mining to extract text sentences that might be related to transcriptional regulation, which were finally subjected to manual curation to gain TF-target genes.

2.4. Annotations of TFs

To better understand biological functions of TFs, TFTG not only provides the comprehensive TF-target genes but also more functional annotation information on TFs, including TF-associated GO terms, pathways, cancer hallmarks, expression and disease information from multiple-sources. Specifically, we collected the expression profiles of TFs from ENCODE [22], TCGA [55], CCLE [56] and GTEX [57], including 31 cell types, 33 cancers types, 41 sample types and 30 tissue types, respectively (Fig. 1 bottom right panel). Meanwhile, the experimentally supported human TF-related diseases were derived from GAD [58] and DisGeNET [59] involving 7353 TF-disease pairs (Fig. 1 bottom right panel). We integrated 2881 pathways and their components from 10 databases, including KEGG, Reactome, NetPath, WikiPathways, PANTHER, PID, CTD, SMPDB, HumanCyc and INOH [60–62] (Fig. 1 bottom right panel). In addition, we obtained 33 cancer hallmarks from Hanahan and Weinberg (2011) [63] and 31 GO terms from the GO database [64] (Fig. 1 bottom right panel).

In addition to the functional annotations of TFs, the abundant epigenetic annotations for TF binding regions were also displayed in TFTG, which aimed to mine the deeper functions of TFs, including promoters, SEs, enhancers, silencers and accessible chromatin. The BEDTools software was used to annotate the corresponding information for TFs and displayed details of the epigenetic annotation using interactive tables.

3. Database use and access

3.1. A search interface for retrieving TFs

TFTG provides multi-type query modes to facilitate the users to query TF-target genes flexibly. The users can determine the scope of TF-target genes through 'Search by TF', 'Search by target gene' and 'Search by genomic region' (Fig. 2A and B). In 'Search by TF', the users can search TF-target genes by inputting the TF name. After submitting, the users can obtain the detailed information of on the input TF, including TF overview, TF target gene network, table of TF-target genes, detailed regulatory information of TF-targets, TF expression, disease information and annotation of TF (Fig. 2C). In the table of TF-target genes, TFTG provides 14 identification methods (C_SE, C_TE, C_ATAC, C_Silencer, C_Promoter, C_BETA, C_Genemapper, Perturbation, Curate, M_SE, M_TE, M_ATAC, M_Silencer and M_Promoter) and a weighted score among TF-target genes. In addition, the users can click gene names to view details about this gene (Fig. 2D). TFTG displays gene information and TFs that regulate this gene in different tissues/cell lines by different methods. The users can also click Sample ID in 'detailed regulatory information of TF-targets' module to view the details of the TF ChIP-seq sample (Fig. 2E). TFTG displays sample overview, detailed target genes information in this TF ChIP-seq sample and peak annotation visualization using ChIPseeker [65].

In the 'Search by target gene' mode, the users can select one of 14 methods, set a weight threshold and input a gene name of interest to search for TFs that regulate this gene. The weights are calculated by counting the number of methods that determine the regulatory relationship of TFs to target genes. Then, TFTG returns the detailed information of these TFs in the result table. In the 'Search by genomic region' mode, with the input of a genomic region and selection of a method, TFTG returns the details of all TF-target genes associated with the input region in the result table.

3.2. A user-friendly browsing interface

The 'Browse' page is an interactive and alphanumerically sorted table that allows the users to quickly search for TFs with custom filters for 'TF Family', 'TF Class' and 'TF Name' (Fig. 2F). The users can use the 'Show entries' drop-down menu to change the number of records per page. They can click the 'TF' button to further view the detailed information for TFs.

3.3. Online analysis tools

TFTG designs four types of analyses for the TF-target network to elucidate the regulatory mechanisms and function of TF (Fig. 2G). (I) *TF gene set enrichment analysis*. With the input of the gene set of interest, TFTG returns the list of TFs that are significantly enriched based on target genes using the hypergeometric test. TFTG also provides various parameters to facilitate the users to filter the results, including TF-target genes identification method, the P-value/FDR, number of gene and sample of interest. (II) *TF downstream regulatory analysis*. By inputting TFs of interest and selecting specificity or non-specificity network, TFTG provides the regulatory network visualizations of these TFs and targets. The node properties in this network are also displayed in a table for insight into interactions. (III) *Gene upstream regulatory analysis*. By inputting genes of interest, the users can visualize the regulatory network formed by these genes and associated TFs. (IV) *Pathway enrichment analysis*. The users can select at least one pathway source and input a gene set of interest, TFTG returns the significantly enriched pathways. The significance P-values were calculated using the hypergeometric test. In the result table, the terminal TFs of the pathway and their target genes are displayed.

3.4. Data download

TFTG provides a convenient and flexible download function for TF-related files, including 'TF information', 'TF ChIP-seq data information', 'TF perturbation RNA-seq data information', 'TF ChIP-seq or perturbation RNA-seq sample' and 'TF downstream target gene of different methods' (Fig. 2H). Meanwhile, the database provides batch download for all TF-target genes identified by different methods. In addition, TFTG supports to export all result tables in search and analysis results.

3.5. Case study

Case study of ATF3. As a key TF in atherosclerosis, the expression level of ATF3 is correlated with the stability of atherosclerotic plaques [66]. We took TF ATF3 as the input of 'Search by TF' to illustrate the function of TFTG. In the returned results page, the user was first be presented with an overview of ATF3 and a visualization of the target gene regulatory network (Fig. 3A).

TFTG also provides a table for downstream target genes of ATF3. This table includes 14 columns representing 14 methods for identifying TF-target genes. The last column of the table is the weight for each TF-target gene based on the number of methods, with higher weights indicating the relationship identified by more methods ('$\surd$' means identified by the current method) (Fig. 3B). In the results of ATF3 target genes, the highest weight is 8. Furthermore, after the removal of ATF3 target genes confirmed by Curate_class (Literature-supported TF-target genes) from TRRUST. We compared our results with TRRUST and hTFtarget, respectively. The Venn diagram shows that all 16 ATF3 target genes from TRRUST completely overlap with TFTG and 9 are intersect with hTFtarget.(Fig. 3C) We further verified the weight distribution of

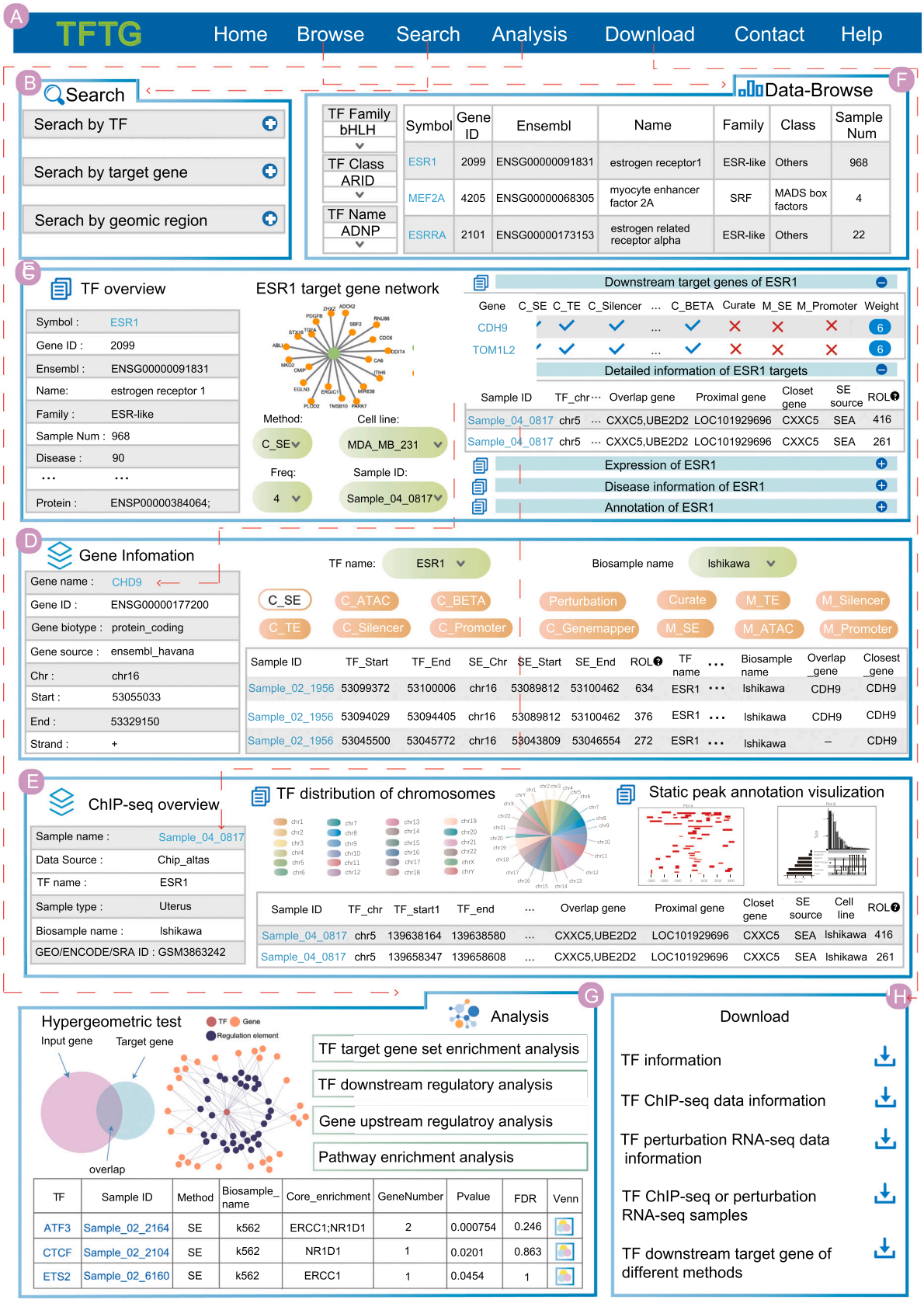

Fig. 2. Main functions and usage of TFTG. (A) Navigation bar of TFTG. (B) Three available inquiry modes. (C) Search results including TF overview, ESR1 target gene network, downstream target genes of ESR1, detailed information of ESR1 targets, expression of ESR1, disease information of ESR1 and annotation of ESR1. (D) Gene information and detailed regulation information by TFs under different methods and different tissues/cell lines. (E) Interactive table with detailed information about ChIP-seq samples of interest and visualization of peak annotation. (F) Browsing TFs. (G) Four online analysis tools for TF targets. (H) Data download.

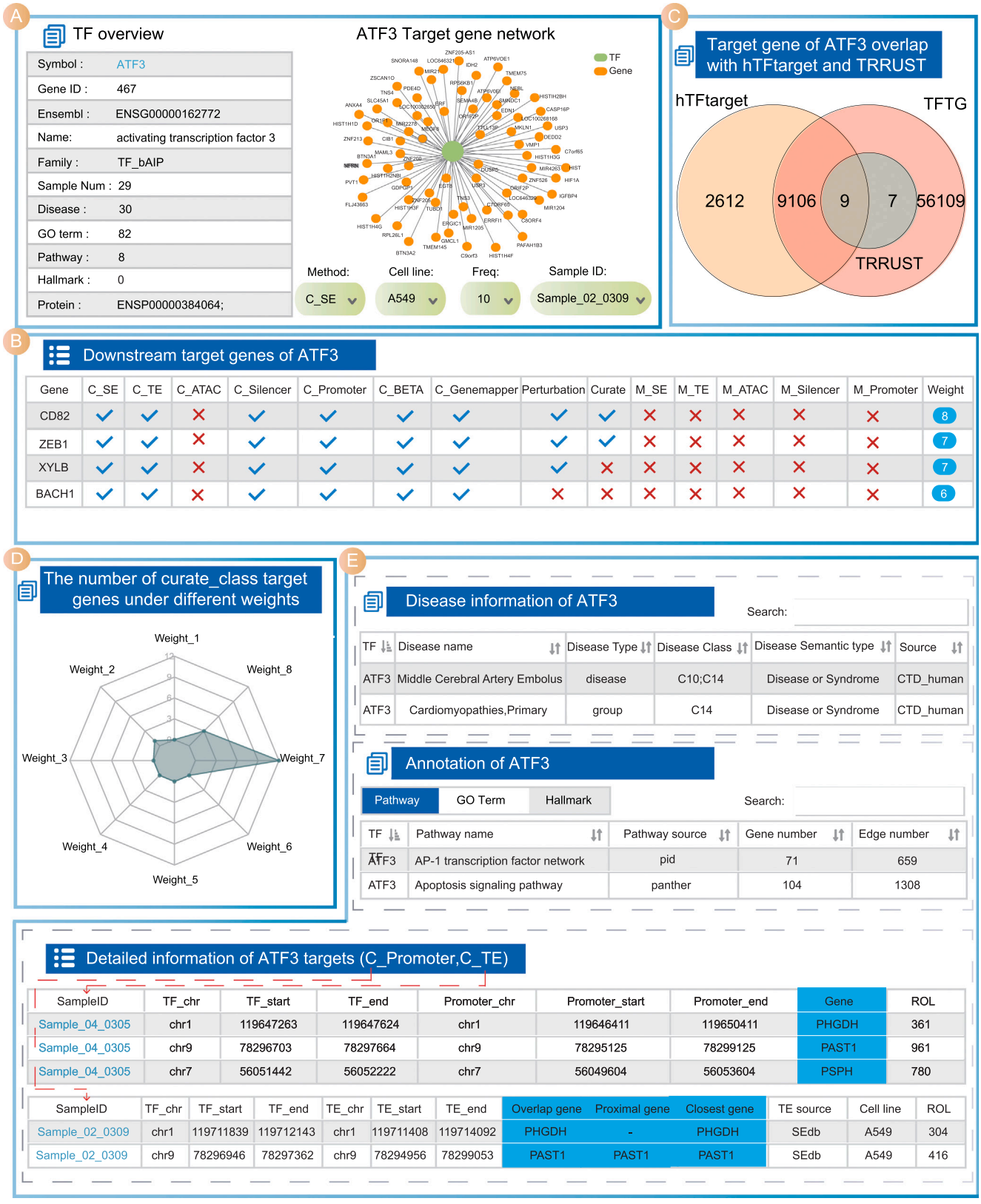

Fig. 3. Verification results of ATF3. (A) ATF3 overview and ATF3 target gene network. (B) Downstream target genes of ATF3. (C) Proportion of Curate_class target genes under different weights. (D) AUROC curves of ATF3, ESR1, AR, MYC and TP53. (E) Detailed information of ATF3 targets, disease information of ATF3 and annotation of ATF3.

the 16 ATF3 target genes from TRRUST in TFTG and the results indicated that 15 out of the 16 ATF3 target genes have high weights in TFTG. (Fig. 3D) Importantly, all target genes with a weight of 8 were confirmed by the Curate_class method, including CD82 which is associated with blood pressure proved by GWAS and GEO analyses [67]. Next, we further checked the target genes with the high weight that were not confirmed by Curate_class. Among these, we found a large number of atherosclerosis-related genes, such as ZEB1, XYLB, BACH1 and so forth. Type H vessel formation is reduced with endothelial ZEB1 deletion and XYLB plays a significant role in anti-hypertension [68,69]. While the transcriptional network formed by BACH1 and YAP is crucial for vascular inflammation and atherosclerosis [70].

TFTG also provide upstream regulatory information, function annotation and disease information for TFs to help researchers explore their regulatory mechanisms. In 'Detailed regulatory information of ATF3 targets', we further verified that ATF3 could promote the expression of PHGDH, PSAT1 and PSPH by binding to their enhancers/promoters [71] (Fig. 3E). The results suggested that the TFTG could assign TF to distal target genes. The aforementioned results indicated that TFTG could be used to find comprehensive TF-target genes and further understand their regulatory mechanisms.

Case study of breast cancer. We also provided a case study of online analysis to better display the analysis ability of TFTG. First, we obtained 743 differential expressed genes of breast cancer from GEPIA [72] with $\log_2FC \geq 2$ or $\log_2FC \leq -2$. Then, we put these genes into 'TF gene set enrichment' analysis. The parameters were set as follows: 'Method: All, Threshold: P-value < 0.05, GeneNumber: Min: 10 and Max: 300, Biosample name: MCF7′. The analysis result page sequentially displayed the TFs that are significantly enriched in 14 TF-target identification methods (Fig. S2A and Table S1). In the result of 'C_Promoter' method, FOXM1 was identified as the top enriched TF (Fig. S2B), which played a significant role in the proliferation, invasion, metastasis and chemoresistance of breast cancer [73]. After clicking 'FOXM1′, the detailed information and function annotations of FOXM1 were shown in the detail page (Fig. S2C). It is worth noting that as the target gene of FOXM1, the transcriptional activity of YAP1 was increased by FOXM1 which leading to cell proliferation, clonal formation and migration capacity in triple-negative breast cancer [74]. In addition, several evidence confirmed that FOXM1 was associated with breast cancer in the 'Disease Information of FOXM1′ module. Meanwhile, TF ESR1 which could encode protein ERα was also significantly enriched in method 'C_TE' (Fig. S2B), indicating that ERα played an important role in the occurrence and development of breast cancer through the distal regulation of enhancers. After clicking the 'ESR1′, the detailed page showed that the targets of ERα included the gene ESR1. This result was consistent with previous findings [75]. Furthermore, as the most significantly enriched TF in 'M_TE' (Fig. S2B), PAX5 was found to be a marker for breast cancer diagnosis and treatment strategy design [76]. Given the importance of FOXM1, ESR1 and PAX5 in breast cancer, the value of TFTG was indicated.

4. System design and implementation

The current version of TFTG was developed using MySQL 5.7.17 (http://www.mysql.com) and runs on a Linux-based Apache Web server (http://www.apache.org). SpringBoot (https://spring.io/projects/spring-boot) was used for server-side scripting. The interactive interface was designed and built using Bootstrap v3.3.7 (https://v3.bootcss.com) and JQuery v2.1.1 (http://jquery. com). ECharts (https://www.echartsjs.com/) and Highcharts (https://www.highcharts.com.cn/) were used as a graphical visualization framework. We recommend to use a modern web browser that supports the HTML5 standard, such as Firefox, Google Chrome, Safari, Opera or IE 9.0 + for the best display.

5. Discussion

The study of TFs has made rapid progress and is one of the most in-depth research fields [77]. The identification of TF-target regulatory relationships is pivotal for understanding the mechanisms of disease development and biological processes [78]. It is critical to unravel the complexity of various biological processes by understanding the gene regulatory networks and the identification of the TF-target regulatory relationships is the first step in constructing a gene regulatory network [79]. Some databases have investigated the potential regulatory interaction between TFs and their targets. For instance, CistromeDB [16], hTFtarget [17], KnockTF [18], TRRUST [19] and MotifMap [20] databases have already counted and stored TF-target genes. However, the integration of multiple TF-related data and methods to provide comprehensive TF-target genes resources is still lacking. Meanwhile, the epigenetic and functional annotations of TFs are just as important for researchers. Based on the above need and the continuous accumulation of a large amount of available data, we developed TFTG to offer relatively extensive available human TF-target genes, abundant annotations and useful analyses for TFs. In brief, TFTG is an available database integrating various resources to provide relatively comprehensive in investigating TF-target regulations in humans.

As we can see, TFTG is a resource focused on providing target genes for TFs and user-friendly interface to search, browse, analyze and visualize information about TF-target genes. Meanwhile, TFTG has abundant functional and epigenetic annotations for TFs and useful analysis tools. Compared TFTG with other TF-target gene databases, Table S2 displays multiple advantages, including (i) The target genes of TFs were identified by fourteen methods through integrating the relevant data of various TFs (TF ChIP-seq data, TF perturbation RNA-seq data and TF Motif profile data); (ii) TFs binding to promoters, enhancers, super-enhancers (SEs), silencers and accessible chromatin regions were simultaneously considered, with strong tissue/cell specificity; (iii) We provided comprehensive functional and epigenetic annotations in TFBS; (iv) We provided useful online analysis tools, (v) three search methods to access TF-target genes and (vi) user-friendly browsing of TFs.

The current version of TFTG not only processed a large amount of TF-related data (ChIP-seq data, perturbation RNA-seq data, TF motif profiles and literature-supported TF-target genes) and offered the most comprehensive TF-target resources, but also provided the functional annotations (pathway, GO terms, cancer hallmarks, expression and disease information) of TFs and epigenetic annotations (promoters, SEs, enhancers, silencers and accessible chromatin) for TF binding regions from multiple-sources. In future version, TFTG will cover more comprehensive TFs. Taking into account the impact of ChIP-Seq peak strength, perturbation response fold-change and motif binding strength on the recognition of target genes by TFs. We will assign different weights to the methods of TF-target gene identification and employ better TF-target gene identification strategies. We will also expand the range of species and further add annotation information and practical analysis functions. Overall, the TFTG database aims to provide a valuable resource for the scientific community to explore gene expression and transcriptional regulation in human diseases and biological processes.

Funding

This work was supported by National Natural Science Foundation of China [62272212, 62171166, 62301246]; Natural Science Foundation of Hunan Province [2023JJ30536]; Natural Science Foundation of Heilongjiang Province [LH2021F044]; Research Foundation of the First Affiliated Hospital of University of South China for Advanced Talents [20210002–1005 USCAT-2021–01]; Clinical Research 4310 Program of the University of South China [20224310NHYCG05]; Research Foundation of Education Bureau of Hunan Province [22C0210].

CRediT authorship contribution statement

Xinyuan Zhou: The first author, whose main work is the analysis and processing of overall data, the writing of manuscripts and the analysis and design of databases, etc. Liwei Zhou: Mainly responsible for building the database background. Fengcui Qian: Helped to check the English. Jiaxin Chen: Mainly responsible for the analysis of some data in the revision process. Yuexin Zhang: Mainly responsible for checking the website page in the revision process. Other Author: The above authors are responsible for a series of related work, such as guidance in data processing, server operation and management, etc.

Declaration of Competing Interest

The manuscript is not submitted to print and electronic manuscripts elsewhere, and there is no economic benefit (except for the author's basic academic career) that may lead to the appearance of a conflict of interest. We are glad to take this opportunity to submit our work to show our platform. We are very grateful for your editorial attention and suggestions for this manuscript.

Data Availability

The research community can access information freely in the TFTG without registration or logging in. The URL of TFTG is http://tf.liclab.net/TFTG.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.036.

References

[1] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet 2009;10:252–63.
[2] Thurlings I, de Bruin A. E2F transcription factors control the roller coaster ride of cell cycle gene expression. Methods Mol Biol 2016;1342:71–88.
[3] Soto LF, Li Z, Santoso CS, Berenson A, Ho I, Shen VX, Yuan S, Fuxman Bass JI. Compendium of human transcription factor effector domains. Mol Cell 2022;82:514–26.
[4] Schmeier S, Alam T, Essack M, Bajic VB. TcoF-DB v2: update of the database of human and mouse transcription co-factors and transcription factor interactions. Nucleic Acids Res 2017;45:D145–50.
[5] Xu M, Bai X, Ai B, Zhang G, Song C, Zhao J, Wang Y, Wei L, Qian F, Li Y, et al. TF-marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. Nucleic Acids Res 2022;50:D402–12.
[6] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The human transcription factors. Cell 2018;172:650–65.
[7] Muley VY, König R. Human transcriptional gene regulatory network compiled from 14 data resources. Biochimie 2022;193:115–25.
[8] Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, Tang Q, Meyer CA, Zhang Y, Liu XS. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc 2013;8:2502–15.
[9] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8.
[10] Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, et al. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res 2009;19:2172–84.
[11] Qian X, Zhao FQ. Regulatory roles of Oct proteins in the mammary gland. Biochim Biophys Acta 2016;1859:812–9.
[12] Thorne JL, Campbell MJ, Turner BM. Transcription factors, chromatin and cancer. Int J Biochem Cell Biol 2009;41:164–75.
[13] Esposito C, Miccadei S, Saiardi A, Civitareale D. PAX 8 activates the enhancer of the human thyroperoxidase gene. Biochem J 1998;331(Pt 1):37–40.
[14] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. Cell 2013;155:934–47.
[15] Teng S, Li YE, Yang M, Qi R, Huang Y, Wang Q, Zhang Y, Chen S, Li S, Lin K, et al. Tissue-specific transcription reprogramming promotes liver metastasis of colorectal cancer. Cell Res 2020;30:34–49.
[16] Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen CH, Brown M, Zhang X, Meyer CA, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 2019;47:D729–35.
[17] Zhang Q, Liu W, Zhang HM, Xie GY, Miao YR, Xia M, Guo AY. hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. Genom Proteom Bioinforma 2020;18:120–8.
[18] Feng C, Song C, Liu Y, Qian F, Gao Y, Ning Z, Wang Q, Jiang Y, Li Y, Li M, et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. Nucleic Acids Res 2020;48:D93–d100.
[19] Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 2018;46:D380–6.
[20] Daily K, Patel VR, Rigor P, Xie X, Baldi P. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. BMC Bioinforma 2011;12:495.
[21] 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.
[22] An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[23] Chèneby J, Ménétrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, Bergon A, Lopez F, Ballester B. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. Nucleic Acids Res 2020;48:D180–8.
[24] Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, Kawaji H, Nakaki R, Sese J, Meno C. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep 2018;19.
[25] Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation-2019 update. Nucleic Acids Res 2019;47:D100–5.
[26] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC genome browser database: 2019 update. Nucleic Acids Res 2019;47:D853–8.
[27] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5.
[28] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006;34:D108–10.
[29] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009;37:W202–8.
[30] Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young RA. Models of human core transcriptional regulatory circuitries. Genome Res 2016;26:385–96.
[31] Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2020;48:D87–92.
[32] Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors. Cell 2013;152:327–39.
[33] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 2008;133:1266–76.
[34] Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res 2011;39:D124–8.
[35] Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. Embo J 2010;29:2147–60.
[36] Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res 2023;51:D942–9.
[37] Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res 2020;48:D58–64.
[38] de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, Åström G, Babina M, Bertin N, Burroughs AM, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol 2017;35:872–8.
[39] Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res 2019;47:D106–12.
[40] Kang R, Zhang Y, Huang Q, Meng J, Ding R, Chang Y, Xiong L, Guo Z. EnhancerDB: a resource of transcriptional regulation in the context of enhancers. Database (Oxf) 2019;2019.
[41] Ashoor H, Kleftogiannis D, Radovanovic A, Bajic VB. DENdb: database of integrated human enhancers. Database (Oxf) 2015;2015.
[42] Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, Han X, Shi S, Zhang J, Li X, et al. SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res 2019;47:D235–43.
[43] Bai X, Shi S, Ai B, Jiang Y, Liu Y, Han X, Xu M, Pan Q, Wang F, Wang Q, et al. ENdb: a manually curated database of experimentally supported enhancers for human and mouse. Nucleic Acids Res 2020;48:D51–7.
[44] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 2010;28:1045–8.

[45] Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 2012;40:D54–6.
[46] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
[47] Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 2013;153:320–34.
[48] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
[49] Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H, et al. SEA: a super-enhancer archive. Nucleic Acids Res 2016;44:D172–9.
[50] Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res 2016;44:D164–71.
[51] Zeng W, Chen S, Cui X, Chen X, Gao Z, Jiang R. SilencerDB: a comprehensive database of silencers. Nucleic Acids Res 2021;49:D221–8.
[52] Wang F, Bai X, Wang Y, Jiang Y, Ai B, Zhang Y, Liu Y, Xu M, Wang Q, Han X, et al. ATACdb: a comprehensive human chromatin accessibility database. Nucleic Acids Res 2021;49:D55–64.
[53] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.
[54] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[55] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362.
[56] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
[57] The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
[58] Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004;36:431–2.
[59] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48:D845–55.
[60] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62.
[61] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway commons, a web resource for biological pathway data. Nucleic Acids Res 2011;39:D685–90.
[62] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503.
[63] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
[64] The Gene Ontology project in 2008. Nucleic Acids Res 2008;36:D440–4.
[65] Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31:2382–3.
[66] Qin W, Yang H, Liu G, Bai R, Bian Y, Yang Z, Xiao C. Activating transcription factor 3 is a potential target and a new biomarker for the prognosis of atherosclerosis. Hum Cell 2021;34:49–59.
[67] Sun G, Chen J, Ding Y, Wren JD, Xu F, Lu L, Wang Y, Wang DW, Zhang XA. A bioinformatics perspective on the links between tetraspanin-enriched microdomains and cardiovascular pathophysiology. Front Cardiovasc Med 2021;8:630471.
[68] Fu R, Lv WC, Xu Y, Gong MY, Chen XJ, Jiang N, Xu Y, Yao QQ, Di L, Lu T, et al. Endothelial ZEB1 promotes angiogenesis-dependent bone formation and reverses osteoporosis. Nat Commun 2020;11:460.
[69] Zhao LY, Li J, Huang XQ, Wang GH, Lv XF, Meng WF, Chen WL, Pang JY, Lin YC, Sun HS, et al. Xyloketal B exerts antihypertensive effect in renovascular hypertensive rats via the NO-sGC-cGMP pathway and calcium signaling. Acta Pharm Sin 2018;39:875–84.
[70] Jia M, Li Q, Guo J, Shi W, Zhu L, Huang Y, Li Y, Wang L, Ma S, Zhuang T, et al. Deletion of BACH1 attenuates atherosclerosis by reducing endothelial inflammation. Circ Res 2022;130:1038–55.
[71] Li X, Gracilla D, Cai L, Zhang M, Yu X, Chen X, Zhang J, Long X, Ding HF, Yan C. ATF3 promotes the serine synthesis pathway and tumor growth under dietary serine restriction. Cell Rep 2021;36:109706.
[72] Li C, Tang Z, Zhang W, Ye Z, Liu F. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. Nucleic Acids Res 2021;49:W242–6.
[73] Zhang YL, Ma Y, Zeng YQ, Liu Y, He EP, Liu YT, Qiao FL, Yu R, Wang YS, Wu XY, et al. A narrative review of research progress on FoxM1 in breast cancer carcinogenesis and therapeutics. Ann Transl Med 2021;9:1704.
[74] Sun HL, Men JR, Liu HY, Liu MY, Zhang HS. FOXM1 facilitates breast cancer cell stemness and migration in YAP1-dependent manner. Arch Biochem Biophys 2020;685:108349.
[75] Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell 2015;58:362–70.
[76] Li X, Huang J, Luo X, Yang D, Yin X, Peng W, Bi C, Ren G, Xiang T. Paired box 5 is a novel marker of breast cancers that is frequently downregulated by methylation. Int J Biol Sci 2018;14:1686–95.
[77] Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell 2013;152:1237–51.
[78] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337:1190–5.
[79] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12:56–68.