D402–D412 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 19 November 2021*
https://doi.org/10.1093/nar/gkab1114

TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human

Mingcong Xu[1,2,†], Xuefeng Bai[1,3,†], Bo Ai[1,†], Guorui Zhang[1], Chao Song[1], Jun Zhao[1], Yuezhu Wang[1], Ling Wei[1], Fengcui Qian[1], Yanyu Li[1], Xinyuan Zhou[1], Liwei Zhou[1], Yongsan Yang[1], Jiaxin Chen[1], Jiaqi Liu[2,4,5,6], Desi Shang[2,4,5,6], Xuan Wang[1], Yu Zhao[2,4,5,6], Xuemei Huang[2,4,5,6], Yan Zheng[3], Jian Zhang[1,*], Qiuyu Wang[2,1,4,5,6,*] and Chunquan Li[1,2,4,5,6,7,8,*]

[1]School of Medical Informatics, Daqing Campus, Harbin Medical University. Daqing 163319, China, [2]The First Affiliated Hospital, Institute of Cardiovascular Disease, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China, [3]State Key Laboratory of Genetic Engineering, Human Phenome Institute and School of Life Sciences, Fudan University, Shanghai 200438, China, [4]School of Computer, University of South China, Hengyang, Hunan 421001, China, [5]The First Affiliated Hospital, Cardiovascular Lab of Big Data and Imaging Artificial Intelligence, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China, [6]Hunan Provincial Base for Scientific and Technological Innovation Cooperation, University of South China, Hengyang, Hunan 421001, China, [7]General Surgery Department, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China and [8]Guangxi Key Laboratory of Diabetic Systems Medicine, Guilin Medical University, Guilin, Guangxi 541199, China

Received August 20, 2021; Revised October 21, 2021; Editorial Decision October 21, 2021; Accepted October 25, 2021

ABSTRACT

Transcription factors (TFs) play key roles in biological processes and are usually used as cell markers. The emerging importance of TFs and related markers in identifying specific cell types in human diseases increases the need for a comprehensive collection of human TFs and related markers sets. Here, we developed the TF-Marker database (TF-Marker, http://bio.liclab.net/TF-Marker/), aiming to provide cell/tissue-specific TFs and related markers for human. By manually curating thousands of published literature, 5905 entries including information about TFs and related markers were classified into five types according to their functions: (i) TF: TFs which regulate expression of the markers; (ii) T Marker: markers which are regulated by the TF; (iii) I Marker: markers which influence the activity of TFs; (iv) TFMarker: TFs which play roles as markers and (v) TF Pmarker: TFs which play roles as potential markers. The 5905 entries of TF-Marker include 1316 TFs, 1092 T Markers, 473 I Markers, 1600 TFMarkers and 1424 TF Pmarkers, involving 383 cell types and 95 tissue types in human. TF-Marker further provides a user-friendly interface to browse, query and visualize the detailed information about TFs and related markers. We believe TF-Marker will become a valuable resource to understand the regulation patterns of different tissues and cells.

INTRODUCTION

Marker genes are signatures in specific cell and tissues. And marker genes can be also used as biomarkers in certain diseases, while cell/tissue-specific marker genes can also help automatically annotate cell types in single-cell sequencing technology. Transcription factors (TFs) can recognize and bind to specific DNA sequences to guide expression of cell marker genes and maintain cell identity. Marker genes can enhance the ability to characterize cell types (1–4). Cell/tissue-specific TFs have multiple relationships with markers. For example, TFs can regulate the expression of cell markers (5–7). FOXA1 (HNF3A) is a TF involved in

*To whom correspondence should be addressed. Tel: +86 1327 2311691; Fax: +86 0734 8279018; Email: lcqbio@163.com
Correspondence may also be addressed to Qiuyu Wang. Tel: +86 1335 1294769; Fax: +86 0734 8279018; Email: wangqiuyu900490@163.com
Correspondence may also be addressed to Jian Zhang. Tel: +86 1383 6963949; Fax: +86 0459 2796867; Email: hmudqzj@163.com
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

embryonic development which plays an important role in cancer. Studies have shown that FOXA1 could regulate the expression of cell marker PLOD2 by binding to promoters, thereby affecting the occurrence and development of lung cancer (8). TF POU5F1 controls the expression of a number of cell markers (e.g. YES1, FGF4, UTF1 and ZFP206) involved in embryonic development, which is critical for early embryogenesis and embryonic stem cell pluripotency (9). Furthermore, the expression of cell markers can influence the activity of TFs. Dang *et al*. evaluated the role of CD27 in inducing the expression of TFs (PRDM1 and XBP1) involved in plasma cell differentiation. They demonstrated that CD27 could activate PRDM1 and XBP1 by binding to CD70 on B cells (10). Moreover, a number of TFs play crucial roles as verified cell markers or potential markers in biological processes. TF GATA3 is a definitive cell marker of breast cancer. Visvader *et al*. identified GATA3, which promoted the differentiation of progenitor cells, as an important marker of tumor initiation (11). Becker *et al*. (12) found that TF LGR5 was a potential marker of intestinal stem cells in human. Overall, a variety of relationships between TFs and related markers have been confirmed by low-throughput biological experiments such as quantitative reverse transcription-polymerase chain reaction (qRT-PCR), western blot, knock down and luciferase reporter assays (13,14).

At present, some databases have been published for TFs or markers. For example, CellMarker (16) has been established to collect cell/tissue-specific cell markers for human and mouse. The data of CellMarker were collected from the PubMed database, handbooks and instructional websites from eight companies (Bio-Rad, Labome, BD Biosciences, R&D Systems, BioLegend, Abcam, Miltenyi Biotec and Thermo Fisher Scientific). MarkerDB (17) consolidates information on clinical and a selected set of pre-clinical molecular biomarkers for human disease into a single resource. However, these existing cell marker databases do not fully focus on cell/tissue-specific TFs and TF-related markers backed by experimental evidence. Other databases and algorithms such as TRANSFAC (18), JASPAR (19), TFCat (20), AnimalTFDB (21), TcoF-DB (22), KnockTF (23), WSMD (24) and TFBSImpute (25) have been developed for TFs and provide a resource for the expression, interactions and functions of TFs. These databases have also become valuable resources for TF research. For example, ReMap (26) is a database which provides the largest catalog of high-quality regulatory regions from an integrative analysis. ReMap helps researchers analyze the regulatory relationships between TFs and marker genes. Expression Atlas (27) is an added-value database that provides information about gene and protein expression in different species and contexts, such as tissue, developmental stage, disease or cell type. However, these databases do not fully explore the links betwteen TFs and cell markers. A large number of studies have shown that human TFs play crucial roles as cell markers in specific cells and tissues. More importantly, TFs and related markers have multiple relationships in specific cells and tissues. Therefore, it is highly desirable to construct a comprehensive resource of manually curated human TFs and related markers which provides comprehensive experimental evidence.

Here, we developed the TF-Marker database (TF-Marker, http://bio.liclab.net/TF-Marker/) which is committed to a comprehensive manual curation of TFs and related markers with experimental evidence in specific cell and tissue types in human. Currently, through reviewing 2,091 published literature, we have manually classified TFs and related markers into five types according to their functions: (i) *TF*: TFs, which regulate the expression of markers. For example, FOXA1 (HNF3A) is a TF involved in embryonic development which plays an important role in cancer. Studies showed that FOXA1 could regulate the expression of cell marker PLOD2 by binding to its promoters, thereby affecting the occurrence and development of lung cancer (Figure 1 A 1); (ii) *T Marker*: markers, which are regulated by TFs. For example, TF POU5F1 controls the expression of marker genes (e.g. YES1, FGF4, UTF1 and ZFP206) involved in embryonic development. These genes are defined as Tmarker (Figure 1 A 2); (iii) *I Marker*: markers, which influence the activity of TFs. For example, Marker gene CD27 could activate PRDM1 and XBP1 by binding to TF CD70 on B cells. CD27 plays a role as I Marker (Figure 1A3); (iv) *TFMarker*: TFs, which play roles as markers. TFMarkers are usually cell/tissue-specific TFs and used as cell markers. For example, TF GATA3 is a definitive cell marker of breast cancer (Figure 1A4); and (v) *TF Pmarker*: TFs, which play roles as potential markers. For example, Becker *et al*. (12) found that TF LGR5 was a potential marker of intestinal stem cells in human (Figure 1A5). By curating thousands of published literature, 5905 entries including 1316 TFs, 1092 T Markers, 473 I Markers, 1600 TFMarkers and 1424 TF Pmarkers, were annotated in 383 cell types and 95 tissue types in human. Moreover, TF-Marker divided markers into disease markers and tissue/cell-specific markers. TF-Marker is an elaborate database, which provides TFs and related markers supported by experimental evidence. We believe that TF-Marker will provide strong support for research into cell/tissue-specific TFs and related markers.

DATA COLLECTION AND DATABASE CONTENT

To ensure high quality data collection, we referred to the steps involved in the manual collection of other databases, such as CellMarker (14), ENdb (28) and EVLncRNAs (29). The literature with ‘transcription factor(s)’ was initially retrieved from the PubMed database. We found >70 000 literature with TFs and related markers mentioned in the titles or abstracts. Then, we carefully read these abstracts and retained >10 000 literature that included the relationships between TFs and markers. We applied the following workflow to obtain specific information about the TFs and related markers. First, we screened the literature based on two standards: (i) experimental evidence that could confirm the relationships between TFs and markers was mentioned in each article (e.g. qRT-PCR, western blot); and (ii) names of the cells or tissues were mentioned in the corresponding experiments. As a result, a total of 2091 literature were obtained according to our extraction requirements. Second, we carefully scrutinized the full text of the 2091 literature and obtained detailed information about the TFs and related markers. This detailed information included the PubMed ID of the literature, gene name, gene

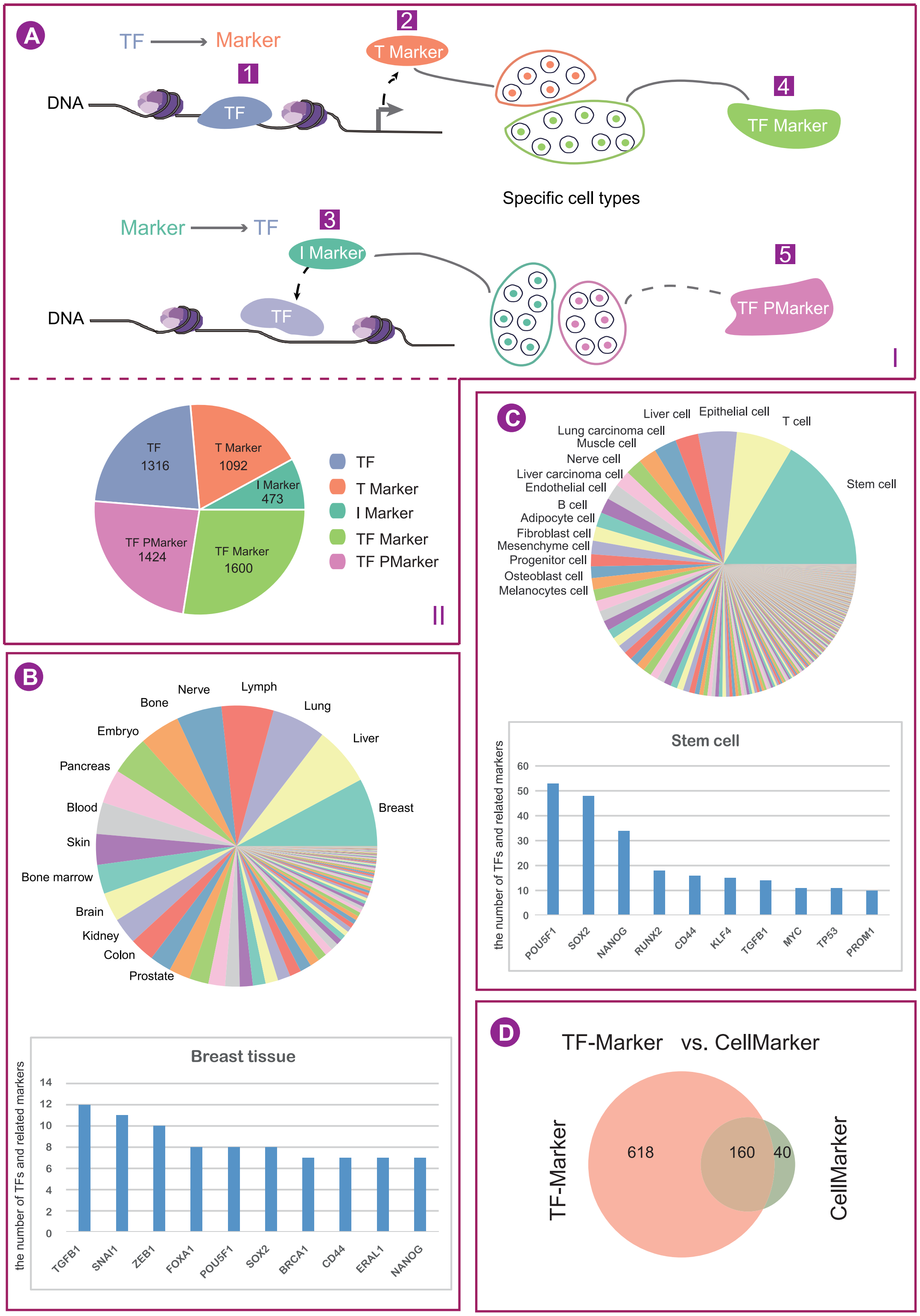


Figure 1. Statistics of TFs and related markers in TF-Marker. (**AI**) TFs and related markers were classified into five types according to their functions: one TF: TFs which regulate expression of the markers; two T Marker: markers which are regulated by the TF; three I Marker: markers which influence the activity of TFs; four TFMarker: TFs which play roles as markers and five TF Pmarker: TFs which play roles as potential markers. (**A II**) Number of TFs and related marker entries in TF-Marker. (**B**) The top 15 tissue types ranked by the number of entries in TF-Marker, and the top 10 TFs and related markers in breast tissue. (**C**) The top 15 cell types ranked by the number of entries in TF-Marker. The top 10 TFs and related markers in stem cells. (**D**) TF-Marker includes 80% of TFs in listed in CellMarker, which were collected from experiments and reviews.

type (TF, T Marker, I Marker, TFMarker or TF Pmarker), detailed description of the TFs and related markers, cell name, cell type, tissue type, experimental technique (e.g. qRT-PCR, western blot, knock down, luciferase reporter assay) and experiment type ('low-throughput' or 'high-throughput'). Third, the information about TFs and related markers was further expanded and standardized. We normalized the official names of TFs and related markers from the Gene (http://www.ncbi.nlm.nih.gov/gene) and Ensembl databases (http://ensemblgenomes.org/), and provided Entrez and Ensembl gene ID. Furthermore, names of tissues were normalized into the standard tissue list from UniProt (9) and cell names were normalized into an integrated reference list based on the Human Cell Atlas (2) and CELLPEDIA (30). TF family information was provided by TFClass (31). We also obtained gene expression atlases from GTEx (32), CCLE (https://sites.broadinstitute.org/ccle/), TCGA (https://cancergenome.nih.gov/) and ENCODE. Finally, we obtained 5905 TFs and related markers involved in 383 cell types and 95 tissue types in human. For each literature, two biological researchers carefully read the full text and examined the information in the literature twice.

The relationships between TFs, super enhancers (SEs) and marker genes

TF-Marker collected experimentally confirmed TFs and their related markers. In order to better understand the relationship between TFs and marker genes, we constructed TF-SE-Marker gene transcriptional regulatory relationships using SEanalysis (34), which was developed by our group. SEanalysis is an SE upstream and downstream transcription regulation analysis tool. The TF-SE pair is predicted based on two methods: (i) TF ChIP-seq data obtained from databases such as ReMap (26) and Cistrome (35); and (ii) Motif scanning based on FIMO. The SE-Marker gene pair is predicted based on four strategies: closest active genes (36), overlapping genes, proximal genes and the closest genes (37). In order to further understand the regulation of TF-SE-Marker gene, we have checked the interaction in SEanalysis. We provide the TF-SE-Marker gene pair and other information on the detail page.

The core TFs in core transcriptional regulatory circuit (CRC)

The CRC is comprised of a group of interconnected auto-regulating TFs forming loops (38–40). The core TFs in CRCs have been shown to be important for cell type-specific transcriptional regulation in normal cells and disease cells (41,42). The core TFs in CRCs are expected to be a reference for markers used to identify specific cell types (38,43). Therefore, we determined the core TFs in CRCs by integrating human H3K27ac ChIP-seq data from SEdb (44). Specifically, we first collected H3K27ac ChIP-seq data from NCBI GEO (15), ENCODE (33), Roadmap (33,45) and GGR (33). Next, Bowtie (v0.12.9) (46,47) and MACS were used for sequence alignment and peak calling of the ChIP-seq data, respectively (48). Finally, all CRC and core TFs were predicted by ROSE (37), CRCmapper (36) and the Coltron (39) program using default parameters and were added to TF-Marker. Users can browse the information of the core TFs in CRCs in TF-Marker.

DATABASE STATISTICS

The current version of TF-Marker includes 1316 TFs, 1092 T Markers, 473 I Markers, 1600 TFMarkers and 1424 TF Pmarkers, involving 95 tissue types and 383 cell types in human (Figure 1AII). The top 15 tissue types ranked by the number of entries in TF-Marker included breast, liver, lung, lymph, nerve and other tissues. (Figure 1B). TGFB1, SNAI1 and ZEB1 were the top three genes collected in breast tissue (Figure 1B). Inflammatory breast cancer cells are usually characterized by the expression of these three genes (49–51). Furthermore, the statistical results for cell types showed that stem cells were the top cell type ranked by the number of all entries in TF-Marker. The top ten TFs and related markers such as POU5F1, SOX2, NANOG, RUNX2 and CD44 have been extensively researched in stem cells (Figure 1C). The expression of these markers controls the cell phenotype (52). TF-Marker contains 778 TFs with experimental evidence, and also includes 80% of the TFs in CellMarker, which were collected from experiments and reviews (Figure 1D).

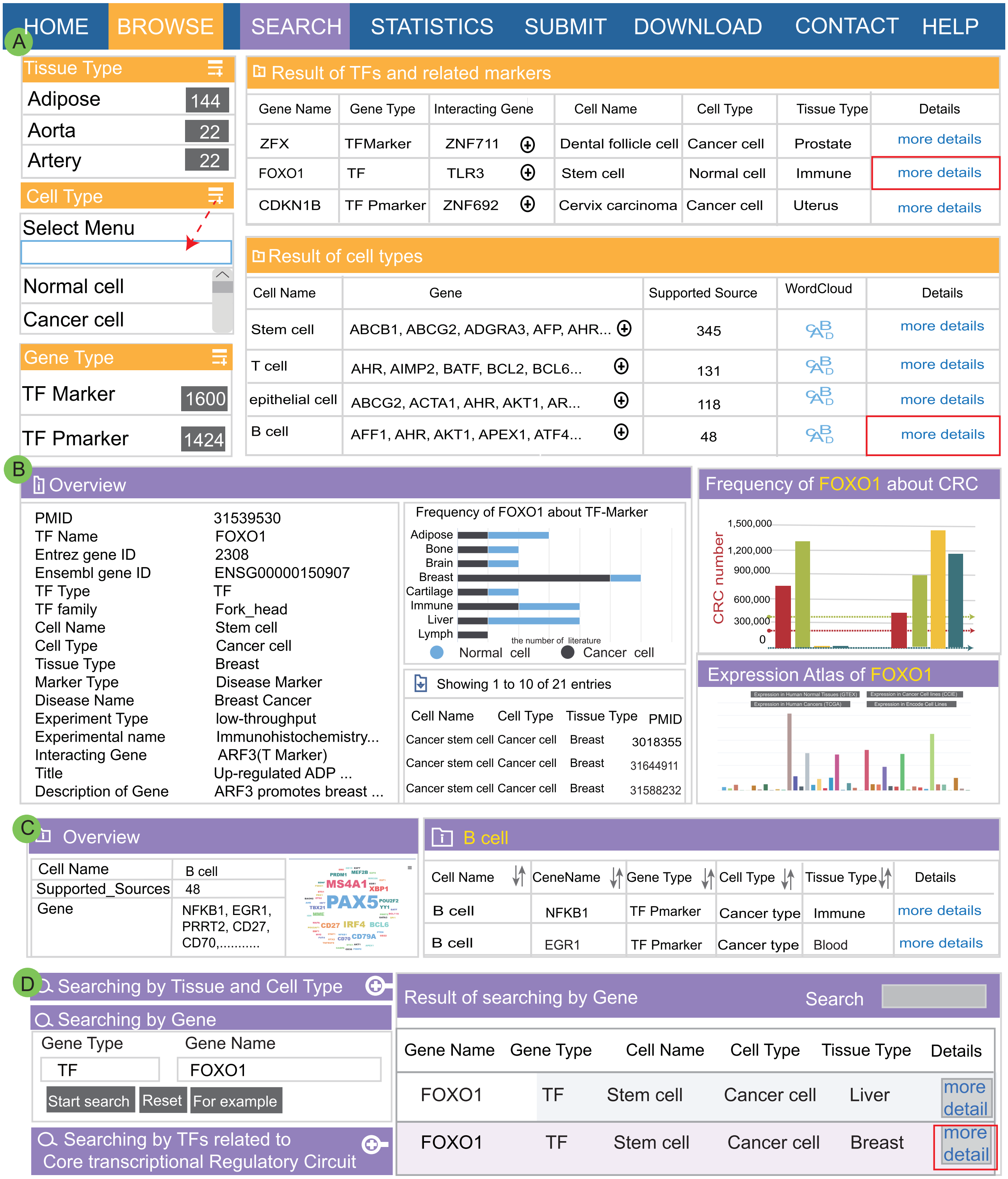
USER INTERFACE

User-friendly interface for browsing TFs and markers

The 'Browse' page is organized as an interactive and alphanumerically sortable table that allows users to quickly browse through 'Tissue Type', 'Gene Type' and 'Cell Type'. Users can browse TFs and related markers of interest via fuzzy search functions. TF-Marker presents two visual tables for users to view the information about TFs and related markers. One table is designed to list the information about multiple TFs and related markers via gene names. The other table is designed to display information about cell-specific TFs and related markers via cell types (Figure 2A). For browsing the details of TFs and related markers, users can click on 'more details'. TF-Marker will return an overview of genes of interest (Figure 2B). Users can also obtain the list of TFs and related markers by selecting their corresponding cell types. Users can select tissue types and cell types to access TFs and related marker entries quickly. TF-Marker also adds a drop-down menu of 'Show entries' to change record numbers per page. Furthermore, users can click 'more details' to view details of TFs and related markers of interest (Figure 2C).

Search interface for conveniently retrieving TFs and related markers

TF-Marker provides a convenient interface for retrieving genes on the 'Search' page. Users can search for information about TFs and related markers through three paths, including 'Searching by Tissue and Cell Type', 'Searching by Gene' and 'Searching by TFs related to CRC' (Figure 2D, left). Through 'Searching by Gene', the information about TFs and related markers can be obtained by inputting a single gene name (or gene alias) or gene lists. Brief information from the search results is then displayed in a table on the results page (Figure 2D, right). Clicking on 'more details', TF-Marker will display 'PMID', 'Gene Name', 'Gene Type', 'Cell Name', 'Gene ID', 'Ensembl ID', 'TF Family',



Tissue Type	
Adipose	144
Aorta	22
Artery	22

Cell Type: Normal cell, Cancer cell

Gene Type	
TF Marker	1600
TF Pmarker	1424

Result of TFs and related markers

Gene Name	Gene Type	Interacting Gene	Cell Name	Cell Type	Tissue Type	Details
ZFX	TFMarker	ZNF711 ⊕	Dental follicle cell	Cancer cell	Prostate	more details
FOXO1	TF	TLR3 ⊕	Stem cell	Normal cell	Immune	more details
CDKN1B	TF Pmarker	ZNF692 ⊕	Cervix carcinoma	Cancer cell	Uterus	more details

Result of cell types

Cell Name	Gene	Supported Source	WordCloud	Details
Stem cell	ABCB1, ABCG2, ADGRA3, AFP, AHR... ⊕	345		more details
T cell	AHR, AIMP2, BATF, BCL2, BCL6... ⊕	131		more details
epithelial cell	ABCG2, ACTA1, AHR, AKT1, AR... ⊕	118		more details
B cell	AFF1, AHR, AKT1, APEX1, ATF4... ⊕	48		more details

Overview

PMID	31539530
TF Name	FOXO1
Entrez gene ID	2308
Ensembl gene ID	ENSG00000150907
TF Type	TF
TF family	Fork_head
Cell Name	Stem cell
Cell Type	Cancer cell
Tissue Type	Breast
Marker Type	Disease Marker
Disease Name	Breast Cancer
Experiment Type	low-throughput
Experimental name	Immunohistochemistry...
Interacting Gene	ARF3(T Marker)
Title	Up-regulated ADP ...
Description of Gene	ARF3 promotes breast ...

Frequency of FOXO1 about TF-Marker
Adipose
Bone
Brain
Breast
Cartilage
Immune
Liver
Lymph
the number of literature
● Normal cell
● Cancer cell
Frequency of FOXO1 about CRC
CRC number
1,500,000
1,200,000
900,000
600,000
300,000
0
Expression Atlas of FOXO1
Expression in Human Normal Tissues (GTEX)
Expression in Cancer Cell lines (CCIE)
Expression in Human Cancers (TCGA)
Expression in Encode Cell Lines

Showing 1 to 10 of 21 entries

Cell Name	Cell Type	Tissue Type	PMID
Cancer stem cell	Cancer cell	Breast	3018355
Cancer stem cell	Cancer cell	Breast	31644911
Cancer stem cell	Cancer cell	Breast	31588232

Overview

Cell Name	B cell
Supported_Sources	48
Gene	NFKB1, EGR1, PRRT2, CD27, CD70,...........

B cell

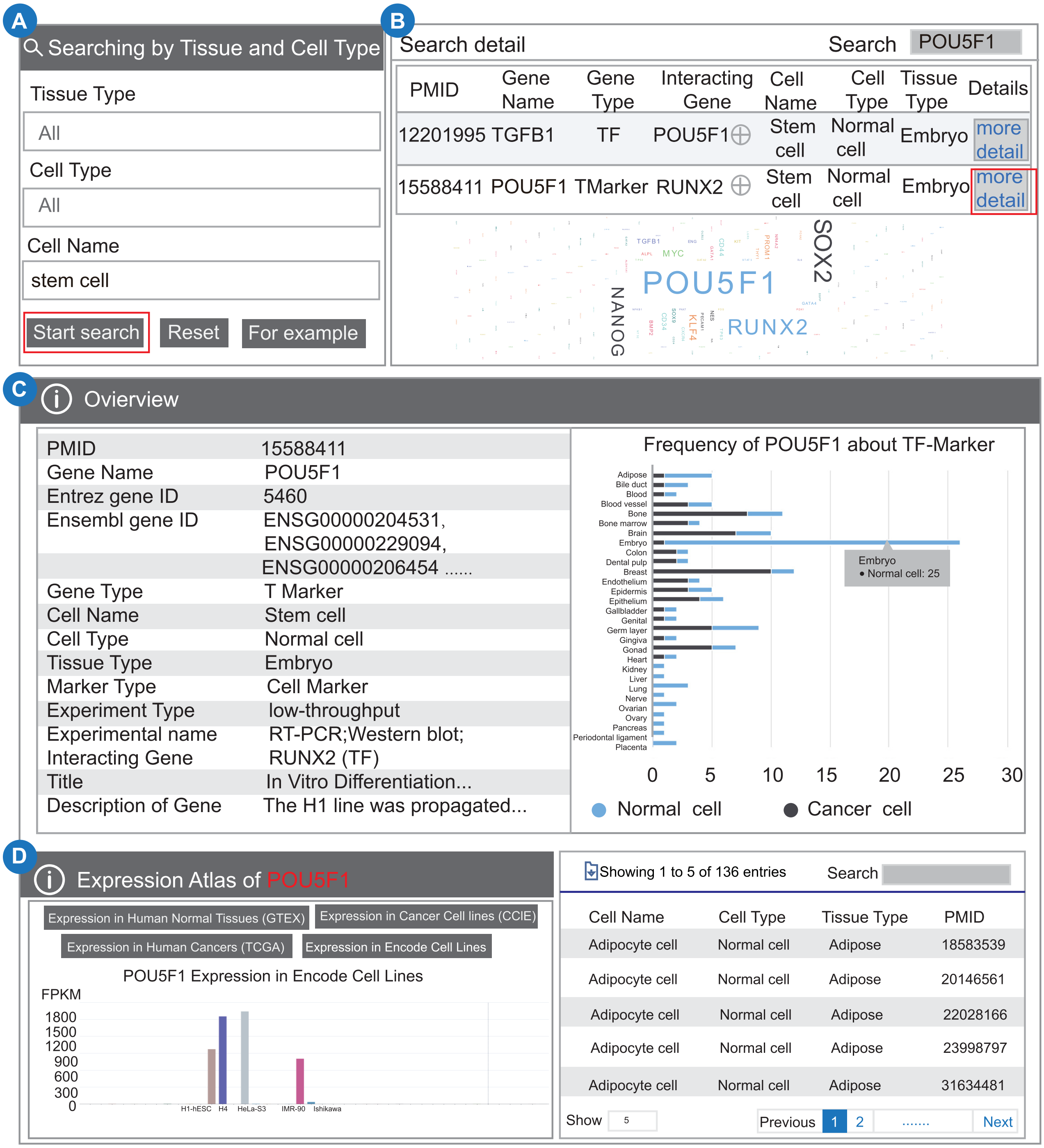
Cell Name	CeneName	Gene Type	Cell Type	Tissue Type	Details
B cell	NFKB1	TF Pmarker	Cancer type	Immune	more details
B cell	EGR1	TF Pmarker	Cancer type	Blood	more details

Gene Type	Gene Name
TF	FOXO1

Result of searching by Gene

Gene Name	Gene Type	Cell Name	Cell Type	Tissue Type	Details
FOXO1	TF	Stem cell	Cancer cell	Liver	more detail
FOXO1	TF	Stem cell	Cancer cell	Breast	more detail

Figure 2. Main functions and usage of TF-Marker. (**A**) User-friendly interface for browsing TFs and markers. (**B**) The details of TFs and related markers. (**C**) Overview of details of TFs and related markers based on specific cell types. (**D**) Three paths for searching cell/tissue-specific TFs and related markers.



Searching by Tissue and Cell Type

Tissue Type: All

Cell Type: All

Cell Name: stem cell

Start search | Reset | For example

Search detail — Search: POU5F1

PMID	Gene Name	Gene Type	Interacting Gene	Cell Name	Cell Type	Tissue Type	Details
12201995	TGFB1	TF	POU5F1 ⊕	Stem cell	Normal cell	Embryo	more detail
15588411	POU5F1	TMarker	RUNX2 ⊕	Stem cell	Normal cell	Embryo	more detail

Overview

PMID	15588411
Gene Name	POU5F1
Entrez gene ID	5460
Ensembl gene ID	ENSG00000204531, ENSG00000229094, ENSG00000206454 ......
Gene Type	T Marker
Cell Name	Stem cell
Cell Type	Normal cell
Tissue Type	Embryo
Marker Type	Cell Marker
Experiment Type	low-throughput
Experimental name	RT-PCR;Western blot;
Interacting Gene	RUNX2 (TF)
Title	In Vitro Differentiation...
Description of Gene	The H1 line was propagated...

Frequency of POU5F1 about TF-Marker

Expression Atlas of POU5F1

Expression in Human Normal Tissues (GTEX) | Expression in Cancer Cell lines (CCIE) | Expression in Human Cancers (TCGA) | Expression in Encode Cell Lines

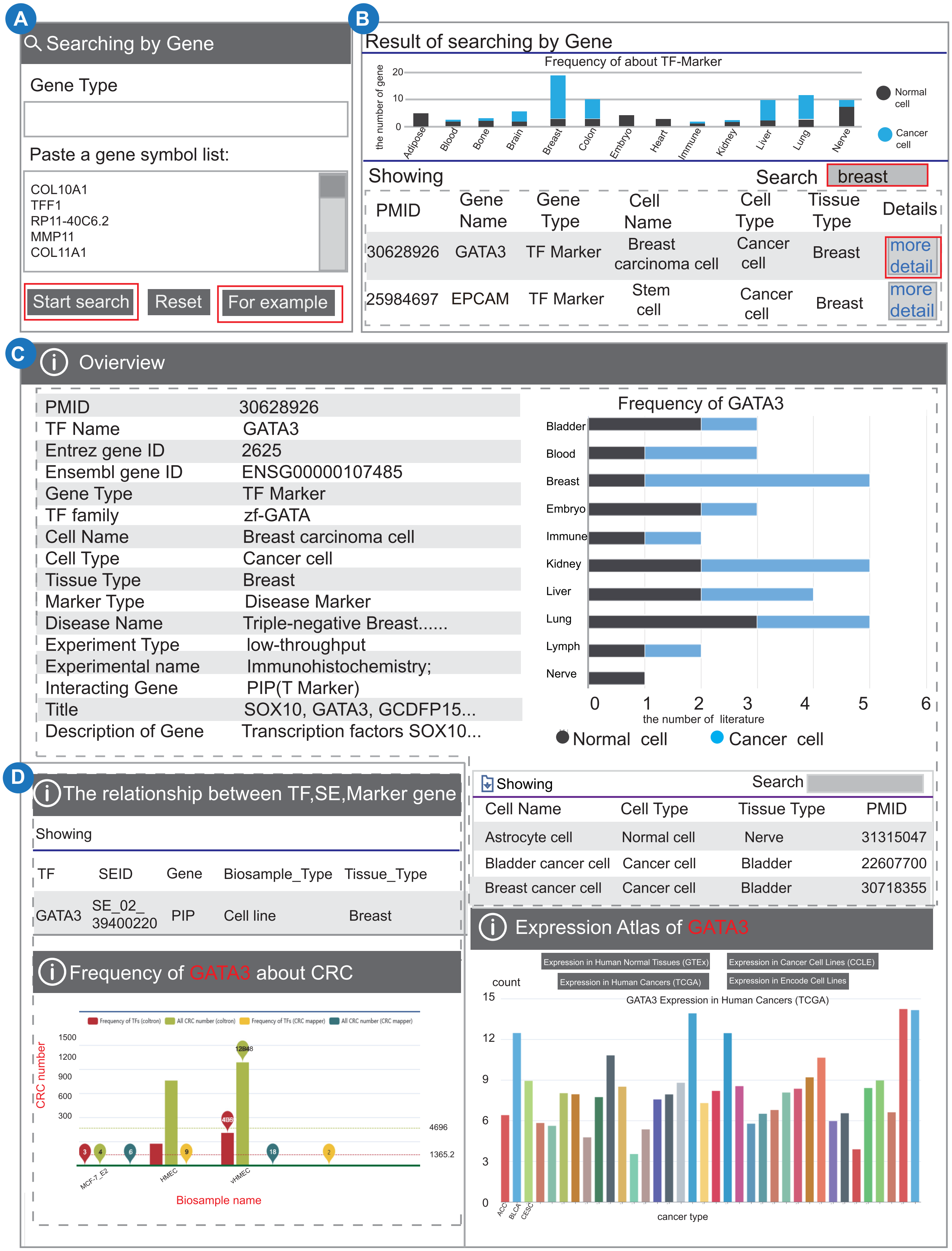
POU5F1 Expression in Encode Cell Lines

Showing 1 to 5 of 136 entries — Search

Cell Name	Cell Type	Tissue Type	PMID
Adipocyte cell	Normal cell	Adipose	18583539
Adipocyte cell	Normal cell	Adipose	20146561
Adipocyte cell	Normal cell	Adipose	22028166
Adipocyte cell	Normal cell	Adipose	23998797
Adipocyte cell	Normal cell	Adipose	31634481

Show 5 — Previous 1 2 .... Next

Figure 3. TFs and related markers in stem cells. (**A**) Searching for TFs and related markers in stem cells. (**B**) A summarized results table for stem cells. (**C**) The distribution graph is displayed based on the number of entries occupied by POU5F1 in the TF-Marker total results. The distribution shows 25 records of POU5F1 studies in embryo research. The list of the literature recorded in TF-Marker for POU5F1 is shown below. (**D**) The expression of POU5F1 in GTEx, CCLE, TCGA and ENCODE.



Searching by Gene

Gene Type

Paste a gene symbol list:

COL10A1
TFF1
RP11-40C6.2
MMP11
COL11A1

Start search | Reset | For example

Result of searching by Gene

Showing

Search breast

PMID	Gene Name	Gene Type	Cell Name	Cell Type	Tissue Type	Details
30628926	GATA3	TF Marker	Breast carcinoma cell	Cancer cell	Breast	more detail
25984697	EPCAM	TF Marker	Stem cell	Cancer cell	Breast	more detail

Overview

PMID	30628926
TF Name	GATA3
Entrez gene ID	2625
Ensembl gene ID	ENSG00000107485
Gene Type	TF Marker
TF family	zf-GATA
Cell Name	Breast carcinoma cell
Cell Type	Cancer cell
Tissue Type	Breast
Marker Type	Disease Marker
Disease Name	Triple-negative Breast......
Experiment Type	low-throughput
Experimental name	Immunohistochemistry;
Interacting Gene	PIP(T Marker)
Title	SOX10, GATA3, GCDFP15...
Description of Gene	Transcription factors SOX10...

Showing

Search

Cell Name	Cell Type	Tissue Type	PMID
Astrocyte cell	Normal cell	Nerve	31315047
Bladder cancer cell	Cancer cell	Bladder	22607700
Breast cancer cell	Cancer cell	Bladder	30718355

The relationship between TF,SE,Marker gene

Showing

TF	SEID	Gene	Biosample_Type	Tissue_Type
GATA3	SE_02_39400220	PIP	Cell line	Breast

Frequency of GATA3 about CRC

Expression Atlas of GATA3

Figure 4. Differential expressed TFs and related markers in breast tissue. (**A**) Searching for differential expressed genes using the TF-Marker function ‘Searching by Gene’. (**B**) TF-Marker provides the distribution of the differential expressed genes in different tissues. (**C**) The detailed information of GATA3 is provided. (**D**) The information of GATA3 is displayed.

'Tissue Type', 'Marker Type', 'Disease Name', 'Experimental name', and a description of TFs and related markers in the literature. The detailed information can also be obtained by the visualization method. Additionally, TF-Marker display a graph to show entries of genes in different cell types of multiple tissues. The annotation information of TFs in CRCs is also presented on this page (Figure 2B). In addition, TF-Marker integrates the gene expression data of GTEx, CCLE, TCGA and ENCODE to display the expression levels of the gene. TF-Marker also provides users with more reference information such as TFs related to CRCs (Figure 2B).

Download, submit and help interface

TF-Marker supports user downloads of all data and the submittal of new data via the 'Download' and 'Submit' page, respectively. The 'Help' page provides a detailed tutorial for users.

Applications of TF-Marker

Case 1. Obtaining TFs and related markers in stem cells. The TF-Marker database provides users with a function for obtaining TFs and related markers of interest in specific cell types of different tissues. Stem cells have major implications for research into developmental biology. As part of the process of cellular turnover and regeneration, stem cells are essential in the development of human tissues. Here, we applied TF-Marker to investigate the relevant TFs and related markers in stem cells. Users can select the cell name of 'Stem cell' in the 'Searching by Tissue and Cell Type' section (Figure 3A). TF-Marker will return a table and an intuitive summarized statistical graph of the prevalence of TFs and related markers, which shows all TFs and related markers of the stem cell. Users can download the list of TFs and related markers in stem cells by clicking the download button. Among these TFs and related markers, we found that POU5F1, SOX2 and RUNX2 were the top three genes investigated in stem cells. Users can extract the POU5F1-related contents by inputting POU5F1 in the search box, based on the total results. TF-Marker will display all the entries about POU5F1 based on the filter (Figure 3B). Users can click 'more details' to obtain the information about POU5F1. According to the search results, as a T Marker, POU5F1 is regulated by TF RUNX2 in embryo stem cells. The distribution is displayed based on the number of entries occupied by POU5F1 in the TF-Marker total results. It aims to show in which tissues and cells POU5F1 has been widely studied. The results indicate that POU5F1 is a widely studied marker in embryo tissue. The distribution shows 25 records of POU5F1 studies in embryo research (Figure 3C upper right panel). The list of the literature recorded in TF-Marker for POU5F1 is shown below (Figure 3C, bottom right panel). The expression of POU5F1 in GTEx, CCLE, TCGA and ENCODE is also displayed (Figure 3D).

Case 2. Searching for differential expressed TFs and related markers in breast tissue. To further verify the function of TF-Marker, we downloaded gene expression data for breast cancer from TCGA and obtained 248 differential expressed genes with $|\log FC|>2$ and $P < 0.001$. The 248 differential expressed genes were used as input for 'Searching by Gene' in TF-Marker (Figure 4A, Supplementary Table S1). Firstly, TF-Marker converted the gene alias to the official names when click the 'Start search' button. Then, the results of output will provide the distribution of the differential expressed genes in different tissues. As shown in the bar plot, we found that these genes were studied in multiple different tissues. Notably, the records of breast ranked first among all tissues. We extracted the breast-related contents by inputting 'Breast' in the search box. We could learn more about the function of GATA3 in breast cancer when clicked on 'more detail' (Figure 4B). As shown, GATA3 is a TF-Marker which can regulate the T Marker PIP in breast cancer. GATA3 is regarded as a biomarker in Triple-negative breast cancer (Figure 4C, left). The distribution indicates that GATA3 is a marker which has been widely studied in breast tissue (Figure 4C, upper right panel). The list of the literature recorded in TF-Marker for GATA3 is shown below (Figure 4C, bottom right panel). We also found that GATA3 can regulate PIP gene by SE (Figure 4D, upper left panel). Moreover, GATA3 was further identified as a core TF (Figure 4D, bottom left panel) and has high expression in breast cancer (Figure 4D, right). Taken together, TF-Marker can help better understand regulatory relationship in biological research.

Table 1. Comparison of information in TF-Marker with other databases

Attribution	TF-Marker	CellMarker	MarkerDB
TF number[a]	778	200	Unknown
Gene type	√	–	–
Interacting gene[b]	√	–	–
Experiment type	√	√	–
Experiment name	√	–	–
Biomarkers[c]	√	–	√
Description of the literature	√	–	–
Cell name	√	√	–
Cell type	√	√	–
Tissue Type	√	√	–
Expression atlas[d]	√	–	–
TF-SE-Marker gene[e]	√	–	–
The core TFs in CRCs[f]	√	–	–

[a] TF Number was the experimentally verified TFs for human.
[b] Interacting Gene was the genes that have some relationship in the biology process with the TFs or markers.
[c] In biological research, marker genes can be used as biomarkers in certain diseases.
[d] TF-Marker provides users with more TF reference information like expression atlas from GTEx, CCLE, TCGA (https://cancergenome.nih.gov/) and ENCODE.
[e] TF-SE-Marker gene regulation was constructed by SEanalysis.
[f] The core TFs in CRCs were determined by integrating human H3K27ac ChIP-seq data from SEdb.

SYSTEM DESIGN AND IMPLEMENTATION

We developed the current version of TF-Marker using MySQL 5.7.27. TF-Marker runs on a Linux-based Apache web server. We utilized PHP 5.6.40 for sever-side scripting, Bootstrap v3.37 and JQuerry v2.1.1 for interactive interface building, and Echats for visualization. For better display, we recommend using a comprehensive web server that supports HTML5 standard, for example, Firefox, Google

Chrome and Safari. The research community can freely access information in the TF-Marker database without registering or logging in. The web link for TF-Marker is http://bio.liclab.net/TF-Marker/. PHP program is provided in Github website (https://github.com/LicLab-bio/PHP/tree/master/TF-Marker). The underlying data has been released in (https://doi.org/10.5281/zenodo.5574651).

DISCUSSION

TFs play crucial roles in biological processes and are usually used as cell markers. The emerging importance of TFs and related markers in identifying specific cell types in human diseases increases the need for a comprehensive collection of human TFs and related markers sets (1). Therefore, we established a comprehensive database called TF-Marker which manually curates TFs and related markers in specific cell and tissue types in human. In the field of transcriptional regulation, ReMap (26) provides the largest catalog of high-quality regulatory regions resulting from a large-scale integrative analysis of TFs and regulators from DNA-binding experiments. MarkerDB and CellMarker have been constructed for marker research. Compared with these databases, TF-Marker focuses on cell/tissue-specific TFs and TF-related markers with experimental evidence (Table 1). MarkerDB consolidates information on clinical and a selected set of pre-clinical molecular biomarkers into a single resource, but it does not contain markers identified by histological, flow cytometry or other experiments. TF-Marker provides this information and further includes markers identified by single-cell RNA sequencing for different cell types. Compared with MarkerDB and CellMarker, TF-Marker finely divided TFs and related markers into five types according to the descriptions of research conclusions. In addition, we also provide information for core TFs in CRCs and their expression atlas in different cell lines/tissues.

The main advantages of the database are illustrated below: (i) TF-Marker provides comprehensive TF and related marker reference sets with classifications of TFs and related markers. We divided the TFs and related markers into five types according to their functions (TF, T Marker, I Marker, TFMarker and TF Pmarker). (ii) TF-marker is dedicated to collecting cell/tissue-specific TFs and related markers backed by experimental evidence. (iii) TF-Marker supports CRC TFs which have been proven highly valuable for understanding cell type-specific transcriptional regulation in normal and disease cells; (iv) TF-Marker supports the use of specific experimental names and descriptions of TFs and related markers; (v) TF-Marker provides relevant descriptions of TFs and markers in each published article; (vi) Users can obtain their TFs of interest and related markers by different searching interfaces; (vii) TF-Marker provides users with more TF reference information from expression atlases such as GTEx, CCLE, TCGA (https://cancergenome.nih.gov/) and ENCODE.

The current version of TF-Marker curates 5905 TFs and related markers. Through building a comprehensive database for TFs and related markers in various cells/tissues, TF-Marker contributes by advancing objective research into cell markers, while classification of markers through reading of the literature is subjective. We focus on the experimentally verified TFs and related markers, and the genes related to human diseases usually experimented on model organisms. We also collected them and normalized official names of TFs and related markers from Gene (http://www.ncbi.nlm.nih.gov/gene) and Ensembl database (http://ensemblgenomes.org/). With the development of single-cell sequencing technology and accumulation of experimental data, extensive literature regarding TFs and related markers will become available. We will manually curate these data in order to update this database in a timely manner. Furthermore, TF-Marker will be supplemented with additional functional information for TFs and related markers. TF-Marker will strive to expand upon the number of species and collections, including additional experimental methods to extend our data sources, constructing networks of TFs and disease-related markers, and providing users with powerful analysis tools in future versions. We believe this first version and future updates of TF-Marker will provide biologists with accessible information for research into different diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [62171166]; Research Foundation of the First Affiliated Hosptial of University of South China for Advanced Talents; Natural Science Foundation of Heilongjiang Province [JQ2020C004]; Open project of the Key Laboratory of Myocardial Ischemia of the Ministry of Education [KF202005]; China Postdoctoral Science Foundation [2019M661311]; Heilongjiang Provincial Postdoctoral Science Foundation; Open Project of Centre of Diabetic Systems Medicine, Guangxi Key Laboratory of Excellence, Guilin Medical University; Shanghai Municipal Science and Technology Major Project [2017SHZDZX01].

Conflict of interest statement. None declared.

REFERENCES

1. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
2. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The Human Cell Atlas. *Elife*, **6**, e27041.
3. Cochain,C., Vafadarnejad,E., Arampatzi,P., Pelisek,J., Winkels,H., Ley,K., Wolf,D., Saliba,A.E. and Zernecke,A. (2018) Single-cell RNA-Seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. *Circ. Res.*, **122**, 1661–1674.
4. Han,X., Chen,H., Huang,D., Chen,H., Fei,L., Cheng,C., Huang,H., Yuan,G.C. and Guo,G. (2018) Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.*, **19**, 47.
5. Deng,X.S., Meng,X., Li,F., Venardos,N., Fullerton,D. and Jaggers,J. (2019) MMP-12-induced pro-osteogenic responses in human aortic valve interstitial cells. *J. Surg. Res.*, **235**, 44–51.
6. Horton,C., Davies,T.J., Lahiri,P., Sachamitr,P. and Fairchild,P.J. (2020) Induced pluripotent stem cells reprogrammed from primary

dendritic cells provide an abundant source of immunostimulatory dendritic cells for use in immunotherapy. *Stem Cells*, **38**, 67–79.

7. Bhuria,V., Xing,J., Scholta,T., Bui,K.C., Nguyen,M.L.T., Malek,N.P., Bozko,P. and Plentz,R.R. (2019) Hypoxia induced Sonic Hedgehog signaling regulates cancer stemness, epithelial-to-mesenchymal transition and invasion in cholangiocarcinoma. *Exp. Cell Res.*, **385**, 111671.
8. Du,H., Chen,Y., Hou,X., Huang,Y., Wei,X., Yu,X., Feng,S., Wu,Y., Zhan,M., Shi,X. *et al.* (2017) PLOD2 regulated by transcription factor FOXA1 promotes metastasis in NSCLC. *Cell Death. Dis.*, **8**, e3143.
9. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
10. Dang,L.V., Nilsson,A., Ingelman-Sundberg,H., Cagigi,A., Gelinck,L.B., Titanji,K., De Milito,A., Grutzmeier,S., Hedlund,J., Kroon,F.P. *et al.* (2012) Soluble CD27 induces IgG production through activation of antigen-primed B cells. *J. Intern. Med.*, **271**, 282–293.
11. Asselin-Labat,M.L., Sutherland,K.D., Vaillant,F., Gyorki,D.E., Wu,D., Holroyd,S., Breslin,K., Ward,T., Shi,W., Bath,M.L. *et al.* (2011) Gata-3 negatively regulates the tumor-initiating capacity of mammary luminal progenitor cells and targets the putative tumor suppressor caspase-14. *Mol. Cell. Biol.*, **31**, 4609–4622.
12. Becker,L., Huang,Q. and Mashimo,H. (2008) Immunostaining of Lgr5, an intestinal stem cell marker, in normal and premalignant human gastrointestinal tissue. *Scientific World J.*, **8**, 1168–1176.
13. Habashy,H.O., Powe,D.G., Rakha,E.A., Ball,G., Paish,C., Gee,J., Nicholson,R.I. and Ellis,I.O. (2008) Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *Eur. J. Cancer*, **44**, 1541–1551.
14. Luo,P., Peng,S., Yan,Y., Ji,P. and Xu,J. (2020) IL-37 inhibits M1-like macrophage activation to ameliorate temporomandibular joint inflammation through the NLRP3 pathway. *Rheumatology (Oxford)*, **59**, 3070–3080.
15. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
16. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
17. Wishart,D.S., Bartok,B., Oler,E., Liang,K.Y.H., Budinski,Z., Berjanskii,M., Guo,A., Cao,X. and Wilson,M. (2021) MarkerDB: an online database of molecular biomarkers. *Nucleic Acids Res.*, **49**, D1259–D1267.
18. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
19. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranasic,D. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.
20. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
21. Hu,H., Miao,Y.R., Jia,L.H., Yu,Q.Y., Zhang,Q. and Guo,A.Y. (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, **47**, D33–D38.
22. Schmeier,S., Alam,T., Essack,M. and Bajic,V.B. (2017) TcoF-DB v2: update of the database of human and mouse transcription co-factors and transcription factor interactions. *Nucleic. Acids. Res.*, **45**, D145–D150.
23. Feng,C., Song,C., Liu,Y., Qian,F., Gao,Y., Ning,Z., Wang,Q., Jiang,Y., Li,Y., Li,M. *et al.* (2020) KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res.*, **48**, D93–D100.
24. Zhang,H., Zhu,L. and Huang,D.S. (2017) WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data. *Sci. Rep.*, **7**, 3217.
25. Guo,W.L. and Huang,D.S. (2017) An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency. *Mol. Biosyst.*, **13**, 1827–1837.
26. Cheneby,J., Menetrier,Z., Mestdagh,M., Rosnet,T., Douida,A., Rhalloussi,W., Bergon,A., Lopez,F. and Ballester,B. (2020) ReMap 2020: a database of regulatory regions from an integrative analysis of human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.*, **48**, D180–D188.
27. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Fullgrabe,A., Fuentes,A.M., George,N. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
28. Bai,X., Shi,S., Ai,B., Jiang,Y., Liu,Y., Han,X., Xu,M., Pan,Q., Wang,F., Wang,Q. *et al.* (2020) ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res.*, **48**, D51–D57.
29. Zhou,B., Zhao,H., Yu,J., Guo,C., Dou,X., Song,F., Hu,G., Cao,Z., Qu,Y., Yang,Y. *et al.* (2018) EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.*, **46**, D100–D105.
30. Hatano,A., Chiba,H., Moesa,H.A., Taniguchi,T., Nagaie,S., Yamanegi,K., Takai-Igarashi,T., Tanaka,H. and Fujibuchi,W. (2011) CELLPEDIA: a repository for human cell information for cell studies and differentiation analyses. *Database (Oxford)*, **2011**, bar046.
31. Wingender,E., Schoeps,T., Haubrock,M., Krull,M. and Donitz,J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**, D343–D347.
32. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
33. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
34. Qian,F.C., Li,X.C., Guo,J.C., Zhao,J.M., Li,Y.Y., Tang,Z.D., Zhou,L.W., Zhang,J., Bai,X.F., Jiang,Y. *et al.* (2019) SEanalysis: a web tool for super-enhancer associated regulatory analysis. *Nucleic Acids Res.*, **47**, W248–W255.
35. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
36. Saint-Andre,V., Federation,A.J., Lin,C.Y., Abraham,B.J., Reddy,J., Lee,T.I., Bradner,J.E. and Young,R.A. (2016) Models of human core transcriptional regulatory circuitries. *Genome Res.*, **26**, 385–396.
37. Loven,J., Hoke,H.A., Lin,C.Y., Lau,A., Orlando,D.A., Vakoc,C.R., Bradner,J.E., Lee,T.I. and Young,R.A. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320–334.
38. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
39. Ott,C.J., Federation,A.J., Schwartz,L.S., Kasar,S., Klitgaard,J.L., Lenci,R., Li,Q., Lawlor,M., Fernandes,S.M., Souza,A. *et al.* (2018) Enhancer architecture and essential core regulatory circuitry of chronic lymphocytic leukemia. *Cancer Cell*, **34**, 982–995.
40. Chew,J.L., Loh,Y.H., Zhang,W., Chen,X., Tam,W.L., Yeap,L.S., Li,P., Ang,Y.S., Lim,B., Robson,P. *et al.* (2005) Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.*, **25**, 6031–6046.
41. Kim,J., Chu,J., Shen,X., Wang,J. and Orkin,S.H. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.
42. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
43. Wang,Z., Oron,E., Nelson,B., Razis,S. and Ivanova,N. (2012) Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell*, **10**, 440–454.

44. Jiang,Y., Qian,F., Bai,X., Liu,Y., Wang,Q., Ai,B., Han,X., Shi,S., Zhang,J., Li,X. *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.*, **47**, D235–D243.
45. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
46. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
47. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
48. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
49. Dhasarathy,A., Phadke,D., Mav,D., Shah,R.R. and Wade,P.A. (2011) The transcription factors Snail and Slug activate the transforming growth factor-beta signaling pathway in breast cancer. *PLoS One*, **6**, e26514.
50. Rypens,C., Marsan,M., Van Berckelaer,C., Billiet,C., Melis,K., Lopez,S.P., van Dam,P., Devi,G.R., Finetti,P., Ueno,N.T. *et al.* (2020) Inflammatory breast cancer cells are characterized by abrogated TGFbeta1-dependent cell motility and SMAD3 activity. *Breast Cancer Res. Treat.*, **180**, 385–395.
51. Nakshatri,H., Kumar,B., Burney,H.N., Cox,M.L., Jacobsen,M., Sandusky,G.E., D'Souza-Schorey,C. and Storniolo,A.M.V. (2019) Genetic ancestry-dependent differences in breast cancer-induced field defects in the tumor-adjacent normal breast. *Clin. Cancer Res.*, **25**, 2848–2859.
52. Lee,S., Wottrich,S. and Bonavida,B. (2017) Crosstalks between Raf-kinase inhibitor protein and cancer stem cell transcription factors (Oct4, KLF4, Sox2, Nanog). *Tumour Biol.*, **39**, 1010428317692253.