Computational and Structural Biotechnology Journal 23 (2024) 77–86

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Database article

SCInter: A comprehensive single-cell transcriptome integration database for human and mouse

Jun Zhao [a,1], Yuezhu Wang [j,1], Chenchen Feng [c,1], Mingxue Yin [b,d,e,f,1], Yu Gao [a], Ling Wei [h,i], Chao Song [b,c,d], Bo Ai [a], Qiuyu Wang [b,c,d,e,f], Jian Zhang [a,*], Jiang Zhu [a,*], Chunquan Li [b,c,d,e,f,g,**]

[a] School of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing, 163319, China
[b] The First Affiliated Hospital, Cardiovascular Lab of Big Data and Imaging Artificial Intelligence, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China
[c] School of Computer, University of South China, Hengyang, Hunan 421001, China
[d] Hunan Provincial Key Laboratory of Multi-omics And Artificial Intelligence of Cardiovascular Diseases, University of South China, Hengyang, Hunan, 421001, China
[e] Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China
[f] Department of Cell Biology and Genetics, School of Basic Medical Sciences, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China
[g] National Health Commission Key Laboratory of Birth Defect Research and Prevention, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China
[h] Institute of Medical Innovation and Research, Peking University Third Hospital, Beijing 100191, China
[i] Cancer Center, Peking University Third Hospital, Beijing 100191, China
[j] School of Artificial Intelligence, Jilin University, Changchun 130012, China

ARTICLE INFO

Keywords:
Single cell integration database
Cell heterogeneity
Multi-method automatic cell-type annotation
Cell to cell communication

ABSTRACT

Single-cell RNA sequencing (scRNA-seq), which profiles gene expression at the cellular level, has effectively explored cell heterogeneity and reconstructed developmental trajectories. With the increasing research on diseases and biological processes, scRNA-seq datasets are accumulating rapidly, highlighting the urgent need for collecting and processing these data to support comprehensive and effective annotation and analysis. Here, we have developed a comprehensive *S*ingle-*C*ell transcriptome *inte*gration database fo*r* human and mouse (SCInter, https://bio.liclab.net/SCInter/index.php), which aims to provide a manually curated database that supports the provision of gene expression profiles across various cell types at the sample level. The current version of SCInter includes 115 integrated datasets and 1016 samples, covering nearly 150 tissues/cell lines. It contains 8016,646 cell markers in 457 identified cell types. SCInter enabled comprehensive analysis of cataloged single-cell data encompassing quality control (QC), clustering, cell markers, multi-method cell type automatic annotation, predicting cell differentiation trajectories and so on. At the same time, SCInter provided a user-friendly interface to query, browse, analyze and visualize each integrated dataset and single cell sample, along with comprehensive QC reports and processing results. It will facilitate the identification of cell type in different cell subpopulations and explore developmental trajectories, enhancing the study of cell heterogeneity in the fields of immunology and oncology.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) is a cutting-edge technology that enables high-throughput sequencing of genes at the individual cell level [1]. The scRNA-seq can perform unbiased, reproducible, high-resolution and high-throughput transcription analysis of single

* Corresponding authors.
** Corresponding author at: The First Affiliated Hospital, Cardiovascular Lab of Big Data and Imaging Artificial Intelligence, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China.
E-mail addresses: hmudqzj@163.com (J. Zhang), hydzhujiang@126.com (J. Zhu), lcqbio@163.com (C. Li).
[1] The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors

https://doi.org/10.1016/j.csbj.2023.11.024
Received 27 September 2023; Received in revised form 12 November 2023; Accepted 13 November 2023
Available online 15 November 2023

cells [2]. Compared to traditional transcriptome analysis, scRNA-seq effectively identifies the heterogeneity among morphologically indistinguishable cells [3–5]. It also contributes to the discovery of novel and rare cell types and gain insight into the gene regulatory mechanisms of cell development and cell fate decisions [6–8]. In the past few years, there have been significant improvements in the quality and quantity of scRNA-seq data [9–12]. These massively generated single-cell data have been deposited on some public resource platforms, such as Human Cell Atlas (HCA) and NCBI/Gene Expression Omnibus (GEO) [13,14]. However, platforms lack user-friendly interfaces and fail to adequately assist single-cell researchers [15]. Therefore, it's necessary to integrate a large number of single-cell datasets and conduct comprehensive analysis for better understanding cellular identity and function [16,17].

Currently, several scRNA-seq related databases have been developed. For example, ScRNASeqDB (Cao et al., Genes (Basel), 2017) provided the gene expression of different cell types for only 36 human datasets from GEO [18]. ColorCells (Zheng et al., Brief Bioinform, 2020) supported comparative analysis of lncRNA expression, classification and function in single-cell RNA-Seq data without annotating cell subtypes [19]. SCovid (Qi et al., Nucleic Acids Research, 2022) [20], DISCO (Li et al., Nucleic Acids Research, 2022) [21], AgeAnno (Huang,et al., Nucleic Acids Research, 2023) [22] and HTCA (Pan et al., Nucleic Acids Research, 2023) [23] only store some human single cell data, but they do not fully consider the analysis of integrated datasets and single-sample data. Multi-sample integration can effectively reveal tissue heterogeneity, and the collection and processing of gene expression profiles across cell types at the level of single sample is also helpful to explore the potential specificity within cells [24,25]. Therefore, we have collected and screened thousands of single cell samples from multiple sequencing platforms, integrating the datasets. Additionally, annotating data information at the single-sample level is quite complicated for most biological researchers. Therefore, there is an urgent need to comprehensively collect and integrate these rapidly accumulating scRNA-seq datasets, and provide a platform for storing gene expression profiles and analytical results of single cell datasets.

To this end, we developed a comprehensive ***S***ingle-***C***ell transcriptome ***Inte***gration database fo***r*** human and mouse (SCInter, https://bio.liclab.net/SCInter/index.php). SCInter aims to provide a comprehensive manually curated single cell transcriptome integration database. Additionally, it supports the provision of gene expression profiles across cell types at the sample level. The current version of SCInter includes a total of 115 integrated datasets and 1016 single-cell samples covering nearly 150 tissues/cell lines. It is worth noting that each sample represents a single-cell expression map confirmed by high-throughput experiments. We have annotated 12,925 clusters and 8016,646 cell markers in 457 cell types using the gene expression profile matrix. At the same time, we have implemented a series of uniform download and multiscale identification on each integrated dataset and single cell sample. This includes accurate multi-method cell type annotation, functional gene expressions, data dimension reduction, cell clustering, screening of specific cell markers, cell differentiation trajectories and etc. In addition, our database supports scRNA-seq data generated from various protocols and platforms, such as 10x Chromium, Microwell-seq, Drop-seq, SMART-seq2 and CEL-Seq2 [26–29]. In summary, SCInter is a user-friendly database to query, browse, analysis and visualize information associated with single cell. We believe that SCInter could become a useful and effective platform for exploring developmental trajectories and heterogeneity of cells in diseases and biological processes.

2. Materials and methods

2.1. Data collection and quality control

We manually collected thousands of publicly available human and mouse scRNA-seq data from GEO based on the following keywords "scRNAseq" OR "single-cell RNA-seq" OR "scRNA-seq" OR "single-cell RNA-sequencing" OR "single cell RNA sequencing" (Fig. 1). All data were examined in the sample description text of GEO, and only data containing barcode, genes/features and matrix were retained. Given the technical noise inherent in scRNA-seq data, we performed data filtering and quality control. First, we filtered out cells with less than two hundred expressed genes, and genes with detectable expression in fewer than three cells. To further ensure the quality of cells, we used the popular open-source R package Seurat to plot violin graphs of three features: number of feature genes, gene counts and mitochondrial ratio of each cell. We removed cells with mitochondrial contamination and an abnormal number of feature genes by following the quality control principle [30,31]. Moreover, we considered potential experimental errors and variations in data accuracy resulting from different times, experimenters, and instruments during the single-cell sample integration process. Therefore, we used Harmony software to remove batch effect by applying the iterative clustering method [32].

2.2. Data integration

For each integration dataset, we integrate multiple GSM samples from the same sequencing platform and the same GSE dataset through the "merge" function. In order to ensure the quality of cells for each integration dataset, Seurat was used to plot violin graphs of three features, including the number of mitochondrial ratios, gene counts and feature genes of each dataset. To remove lowly quality cells and lowly expressed genes, we excluded cells with mitochondrial contamination and abnormal gene numbers by observation. Considering the influence of technical noise, we adopt "Harmony" function to remove the batch effect between samples. Then we use the same R package for data integration, clustering visualization and other analysis. For these analyses, the function "SCTransform" was used to integrate and scale data.

2.3. Cell clustering

To cluster the cells, for each sample, we first use linear dimension reduction principal component analysis (PCA) to reduce the dimension of data. Then, the function "ElbowPlot" is used to identify PCs, and ElbowPlot ranks the main components based on the percentage of variance explained by each component. In the ElbowPlot results, the sharp decline of the PC variance curve represents the "real" signal related to the biological differences between cell groups. When the PC variance curve gradually flattens, it represents technical variation or random noise of a single cell. Therefore, we take the PC cutoff point whose cumulative contribution is greater than 90% variance as the inflection point of the curve, and select the most suitable number of PCs for each data sample [33]. Immediately after that, we applied the FindClusters function with clustering resolution tocluster cells based on the Louvain algorithm. Then, based on the above analysis, we use two nonlinear dimensionality reduction methods umap and tsne to realize cell clustering, and provide a comparative display. Its purpose is to help better retain the structural characteristics of data in single cell clustering, accurately capture the similarities and differences between cells, and thus achieve more accurate clustering results.

2.4. Differential gene expression

To identify the positive markers for each cluster, we employed the FindAllMarkers function of Seurat to identify the positive markers with specific expression levels by comparing current cluster with all remaining clusters [34]. We used three methods to find the markers of each cluster for single sample data processing, including Wilcoxon rank sum test [35], Student's t-test [36] and logistic regression framework [37]. In addition, SCInter also provides significant differential gene analysis from any two clusters through the FindMarkers function.

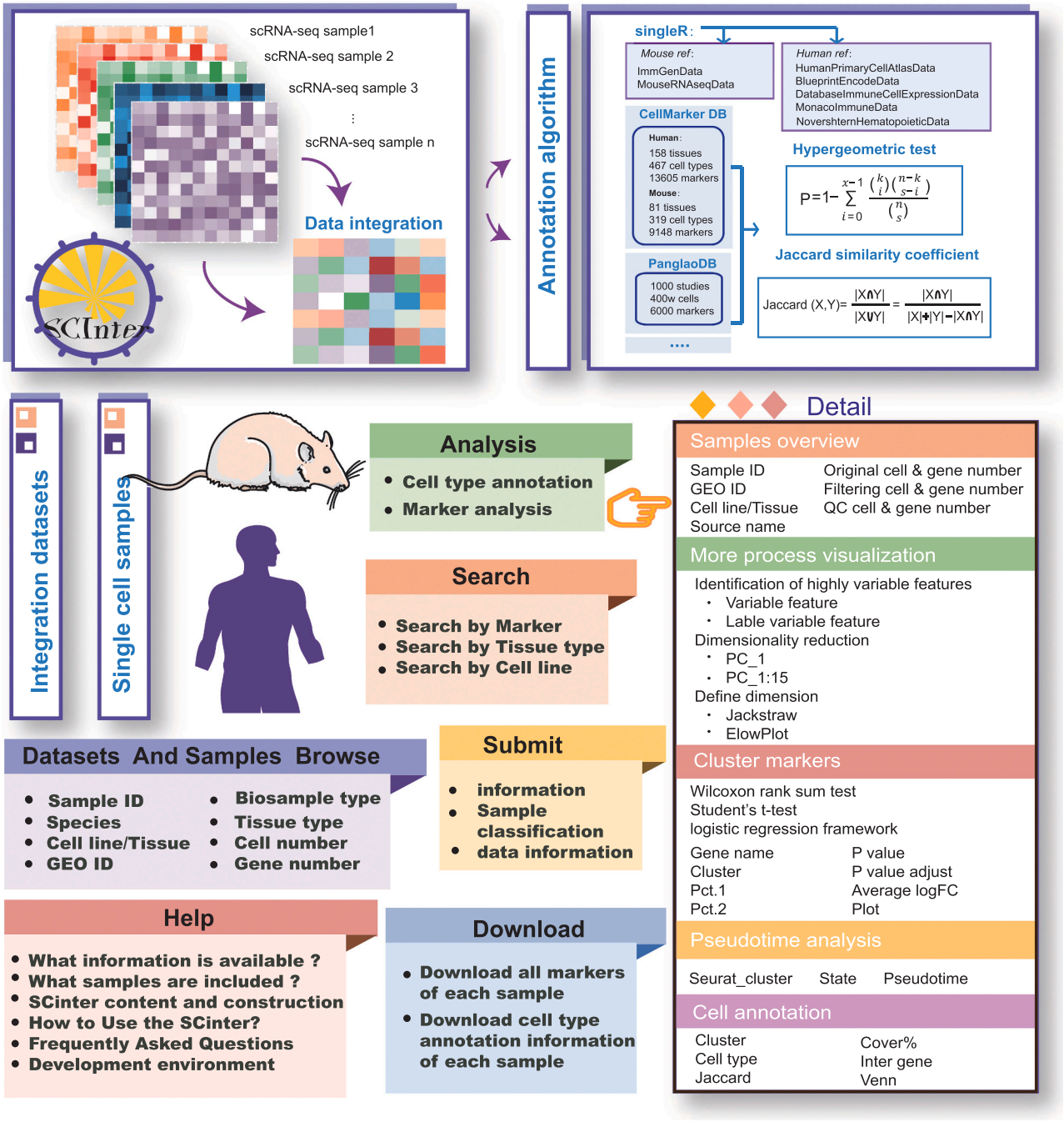

Fig. 1. Database content and construction. SCInter is a comprehensive single-cell transcriptome database for human and mouse with a large number of datasets. Importantly, SCInter provides useful single-cell clustering analysis to identify specific cell markers, available pseudotime analysis and cell type automatic annotation. Users can query to determine the integration datasets and single-cell samples through three paths. SCInter provides a user-friendly interface to query, browse, analyze and visualize each integration dataset and single-cell sample with comprehensive QC reports and processing results.

2.5. Selection of hypervariable genes

Considering cell-to-cell variation in gene expression, many studies have demonstrated that genes with the power to distinguish cell heterogeneity showed higher variation [38]. Therefore, we used the FindVariableFeatures function to calculate the mean expression values and variance expression values, and finally generated the top 2000 genes as hypervariable genes.

2.6. Pseudotime analysis

To study and simulate the dynamic biological processes of cells, SCInter adopted the popular software Monocle to produce a trajectory of cell's development [38,39]. Firstly, we used DDRTree to reduce the dimensionality of the dataset [40]. The orderCells function of Monocle was then performed to re-order the cells. Finally, to visualize the changes of in different clusters with the development of state and pseudotime, we used the plot_cell_trajectory function. It is worth noting that we provided markers and hypervariable genes for pseudotime analysis, respectively.

2.7. Cell to cell communication

Cells of multicellular organisms often communicate with each other

through cytokines and membrane proteins, so as to regulate life activities and ensure the efficient and orderly operation of living organisms. Among them, receptor-ligand mediated intercellular communication is very important for coordinating various biological processes such as development, differentiation and diseases. Therefore, in order to identify the receptor and ligand proteins that play a significant role in each cell cluster, we used CellPhoneDB to perform cell to cell communication network analysis in Linux environment [41,42]. This enabled us to understand the differences between cell types in physiological and pathological conditions, thereby providing new ideas and methods for disease diagnosis and treatment. The principle of CellPhoneDB to recognize cellular protein interactions by predicting the type and mechanism of intercellular communication. This is achieved by calculating the probability and intensity of intercellular communication based on the gene expression data of single cell transcriptome [43]. CellPhoneDB contains 978 types of proteins, of which 501 are secreted proteins and 585 are membrane proteins, participating in 1396 kinds of interactions. The database also considers the subunit structure of ligands and receptors, and accurately represents 466 heteromer complexes, which plays a vital role in the study of cell communication mediated by multi-subunit complexes. In addition, as the database is limited to human, we compared the marker genes identified by mouse single cell data to the human genome through homologous gene transformation. This allowed us to complete the exploration of cell communication mechanism among various cell types [44].

2.8. Automatic cell-type annotation

It is a key step to annotate cell clusters into specific cell types according to the expression of cell-specific marker gene and the biological functions performed by cells. The common strategy for defining cell attributes is to extract specific expression marker genes from different cell clusters based on verified literature and experiments. Therefore, the selection of reference background genes is particularly important for cell annotation.

2.8.1. Reference datasets collection

Cell marker was used for defining and distinguishing cell types [45]. Although the use of cell markers is very important in defining and distinguishing cell types, the key aspect of cell type annotation is to identify the characteristic genome of single cells and the reference genome of known cell types. For our annotation analysis, we selected reference databases containing marker lists associated with well-established cell types namely, CellMarker and PanglaoDB. CellMarker is a comprehensive and accurate cell marker resource that provides multiple cell types of human and mouse tissues [46]. We collected 13,605 cell markers across hundreds of cell types in human and mouse recorded in CellMarker. From these markers, we further selected more accurate cell markers as our reference for automatic cell type annotation analysis. PanglaoDB is a marker database for annotation of cell population, which contains information about tissue and organs of more than 4 million cells, involving more than 6000 marker genes. We incorporated PanglaoDB as another cell type reference gene set for users to choose.

Method1 SCInter supports automatical cell types annotation for clusters based on cell markers. We jointly used hypergeometric tests and Jaccard similarity to determine cell types of different clusters [47,48]. The significance of cluster A being annotated as cell type B was measured by the hypergeometric test *P*-value. The *P*-value was measured as:

$$P = 1 - \sum_{i=0}^{x-1} \frac{\binom{k}{i}\binom{n-k}{s-i}}{\binom{n}{s}}$$

where, n represents the number of all human or mouse genes, k represents the number of all markers interacting with cluster A, s represents the number of genes interacting with the cell markers of cell type B, and x represents the number of genes shared between cluster A and cell type B.

In addition, we also calculated the Jaccard similarity between cell type and cluster. The Jaccard coefficient is a probability used to compare the similarity and dispersion in the set. The Jaccard score was calculated as:

$$\text{Jaccard}\ (\mathrm{X},\ \mathrm{Y}) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

where, X represents the marker set of cluster A and Y represents the genes set of cell type B.

Method2 SCInter added SingleR automatic cell type annotation to the single cell integration datasets to help us better understand the function and characteristics of each cell [49]. SingleR is a software package based on the R environment. Its principle is based on the heterogeneity and similarity of gene expression in scRNA-seq data. By calculating the similarity between single cell expression patterns and cell type patterns in the reference dataset, the cell type and subtype to which each single cell belongs is determined. It is worth noting that SingleR can annotate not only cells on the new scRNA-seq datasets, but also the existing public datasets [50]. SingleR annotation can effectively reveal previously undiscovered subtypes and new cell types, and help identify and study rare cell types in the human body [51].

2.8.2. Cell marker analysis

Because the marker genes in cells are important sign of cell definition and classification, the markers in the same cluster often have similar expression patterns. In order to measure the consistency among major marker genes. Pearson correlation coefficients were used to calculate the correlations between the genes entered by users and other markers in the cluster where the genes belong. The Pearson calculation is as follows:

$$\rho(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma\mathrm{X}\sigma\mathrm{Y}} = \frac{\mathrm{E}[(\mathrm{X}-\mu\mathrm{X})(\mathrm{Y}-\mu\mathrm{Y})]}{\sigma\mathrm{X}\sigma\mathrm{Y}}$$

where, Pearson correlation number $\rho(X, Y)$ is equal to the covariance $cov(X, Y)$ between two continuous variables (X, Y) divided by the product (X, Y) of their respective standard deviations. μX is the average value of X, μY is the average value of Y, and E is the expectation. The calculated correlation coefficient is a number between -1 and 1. If the coefficient is close to 0, there is no correlation between the calculated two variables, and if it is close to 1 or -1, there is a strong correlation between the two variables. At last, we use R language function cor.test to calculate Pearson correlation coefficient R expressed between marker genes and statistical P value of correlation test.

3. Database use and access

3.1. A search interface for retrieving single-cell data

SCInter is a powerful platform that provides users with an intuitive search interface to access single-cell data. Users can determine the scope of single cell integration datasets and single samples query through four paths including “Search by tissue type (input tissue name of interest)”, “Search by cell type (input cell type name of interest)”, “Search by cell marker (input marker gene name of interest)” and “Search by disease (input disease of interest)” (Fig. 2A). Based on the cell type query, users can obtain all datasets and sample information by the cell type through submiting a cell type of interest. For marker-based query, the search results provide information about integration datasets or single samples where the input gene is a cluster marker, including cluster information and fold-change values. In the tissue based query, users can select species for customizing filters. Based on the disease query, users can obtain all datasets and sample information through submiting a disease type of

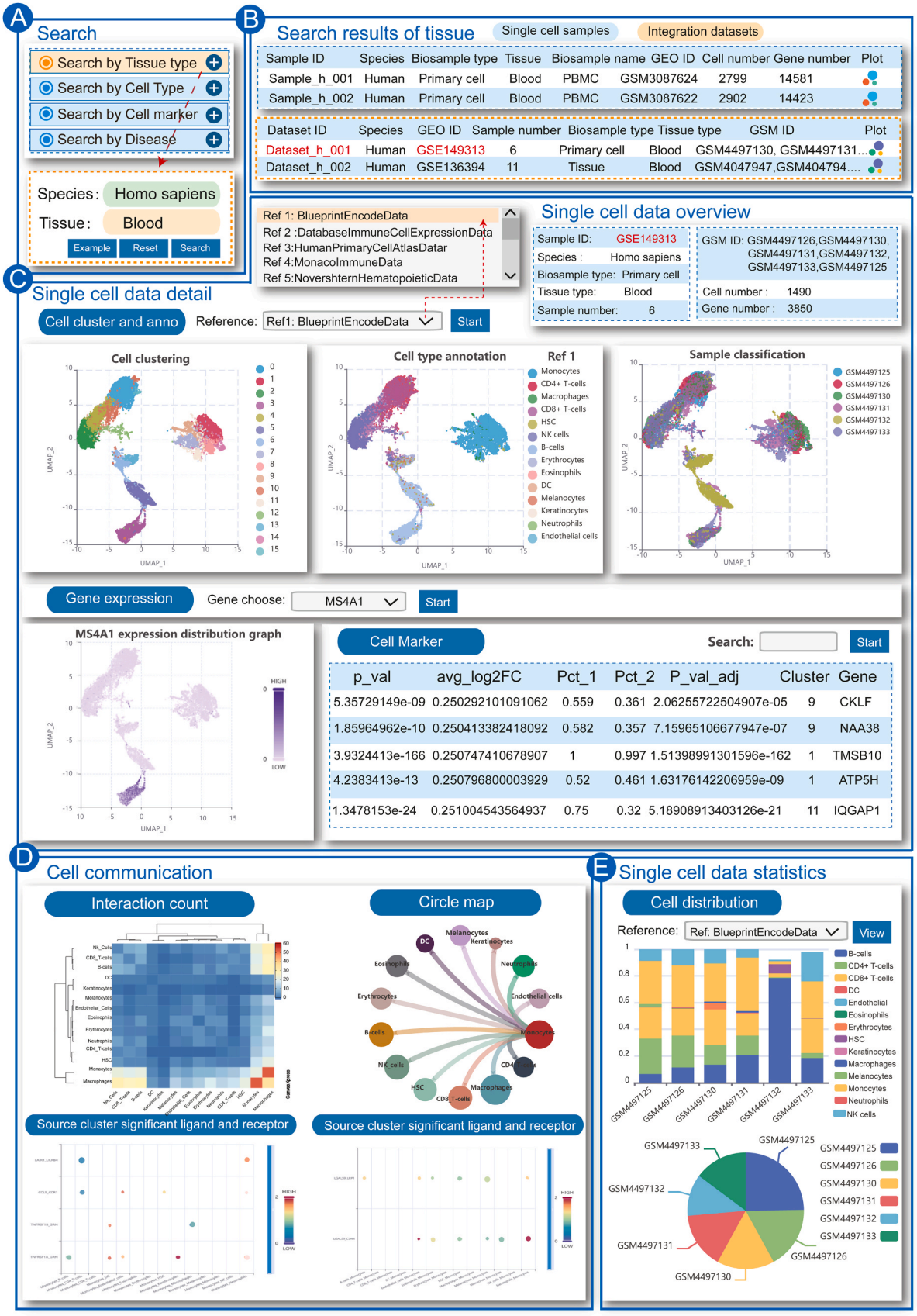

Fig. 2. The main functions and usages of SCInter. (A) Three inquiry modes are available. (B) Table of search results including sample id/dataset id, biosample type and name, tissue type, GEO ID, cell/gene number and plot diagram. (C) Cell clustering, accurately multi-angle cell annotation and functional gene expressions. (D) Cell communication of receptor and ligand protein between different cell types. (E) Cell type correlation statistics.

interest. The brief information from the search results is displayed in a table on the results page. The table describes sample id/dataset id, biosample type and name, tissue type, GEO ID, cell/gene number and plot diagram (Fig. 2B). Upon finding a sample or dataset of interest, users can click on the "Sample ID" or "Dataset ID" to access detailed information, including: (i) sample overview; (ii) cell clustering and accurately multi-angle cell annotation; (iii) functional gene expressions; (iv) cell to cell communication; (v) cluster markers; and (vi) Functional enrichment. By sample overview, the table for the sample overview describes dataset ID, species, biosample type, tissue type, integration sample number, cell number and gene number. In the cell clustering and annotation module, SCInter uses multiple reference geneset annotation method to display cell clustering results. Users can freely choose the reference geneset to compare the cell types of specific cell cluster after annotation. In the gene expression module, we provide the gene expression of all markers in each cell cluster. The distribution graph of marker gene expression in each cell can effectively reveal the significantly high expression of marker genes in each cluster, and also help us to further explore the potential functional characteristics of each specific cell cluster (Fig. 2C). In the process of identifying the cell markers of single samples, we also provide extra-information about the differences between any two cell clusters in order to compare the heterogeneity of each cell cluster. The resulting table of marker gene recognition displays statistical information including pct.1, pct.2 and P value. In the cell communication module, receptor and ligand proteins between various cell types could be shown by clicking their respective drop-down boxes and buttons (Fig. 2D). In the statistics module, we made statistics on the number of cells in the integrated sample, the number of cells in each cell type and the marker gene of each cell cluster (Fig. 2E).

3.2. User-friendly interface for browsing single-cell data

The "Browse" page was designed as an interactive and alphanumerically sortable table that allowed users to quickly browse single-cell samples. Users could customize filters by selecting options such as "Species", "Biosample type", "Tissue type" and "Biosample name". Users could use the "Show entries" drop-down menu to get different numbers of records per page. To view the single-cell data for a specific dataset, users only needed to click on the "Sample ID" and "Dataset ID".

3.3. Effective online analysis tools

In the "**Cell type annotation**" tool, users submitted a gene list for cell type analysis. SCInter performed hypergeometric tests and calculated the Jaccard similarity between the submitted gene list and the gene set of each cell type from CellMarker or PanglaoDB to identify possible related cell types. Users have the option to select the reference genome of interest. The results table displayed cell types, intersection genes, the number of intersection genes and hypergeometric test *P* values. Additionally, SCInter provides Venn diagrams as the results of hypergeometric enrichment. When selecting the cell type of interest, all samples of relevant cell types were listed, and the user further observed detailed information (FigureS1A). In addition, users can explore the correlation between these input genes and identified markers related cell types through a co-expression network.

Using the "**Marker analysis**" tool, users can analyze any sample and any gene of interest. SCInter provides two visualization methods, UMAP and tSNE, for user selection. The analysis results provide detailed information about the input genes in the specific sample, including the gene expression distribution and co-expression with other markers. We use heatmap and network map to display the co-expression results of each marker gene. Notably, asterisks are used to highlight the genes of interest, and users can freely choose the top 20 or top 30 marker genes for viewing (FigureS1B).

3.4. Case studies

3.4.1. Case study of cell type annotation

To provide examples of how SCInter can be used to explore which cells that the interested genes associated with, we selected differential genes ($log_2FC>2$, $P < 0.05$) of hypertrophic cardiomyopathy (HCM) as input (Fig. 3A; Table S1). The results identified several cell types associated with HCM, such as cardiomyocyte cell (Rank=1, P = 1.0E-6) and ventricular cardiomyocyte cell (Rank=2, P = 2.73E-4) (Fig. 3B). Studies have shown that progressive cardiomyocyte enlargement leads to increased cell death and fibrosis, and is related to the severity of hypertrophic cardiomyopathy [52]. By clicking on cardiomyocyte cell, we obtained related single-cell samples (such as Sample_m_150, Sample_m_394) with the tissue type of heart. (Fig. 3C). SCInter also provided the detailed analysis for these annotated samples, including cell clustering analysis, pseudotime analysis and so on (Fig. 3D). The differential expressed genes (Tnni1, Csrp3, Myl2, etc.) for sample scRNA-seq_m_394 are closely associated with hypertrophic cardiomyopathy. For example, MYL2, which reveals molecular differences between dominant and recessive forms of hypertrophic cardiomyopathy [53]. Previous studies have also shown that CSRP3 mutations lead to HCM [54]. If the user wanted to conduct a more in-depth study of cardiac hypertrophy, they could download the sample. In addition, we used the differential genes of Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBCL) from TCGA as input and performed cell type analysis (Fig. S2A; Table S2) [55–57]. SCInter can also accurately identify B cells as the related cell types for these genes, while displaying samples related to B cells and their detailed information (Fig. S2B-S2D). Above all, the validation of these results demonstrates the usefulness and value of SCInter in the cell type investigation.

3.4.2. Case study of marker analysis

Atherosclerosis is a prevalent vascular disease that can result in conditions such as myocardial ischemia and cerebral ischemia. Severe plaque rupture will form thrombosis, myocardial infarction, cerebral infarction and other serious consequences. Therefore, atherosclerosis can be the root cause of cardiovascular and cerebrovascular diseases. Previous studies have shown that *TPM2* is a potential biomarker of atherosclerosis, so we input *TPM2* for marker analysis [58]. Selected cardiac-cerebral disease related tissue samples from SCInter, for example, we analyzed Sample_h_237 from cardiomyocytes tissue (Fig. 4A-4B). The expression and distribution of this gene in this sample can be known from the results of cell clustering analysis (Fig. 4C). We observed that there was obvious high expression in cluster 0, which was labeled as cardiomyocyte. However, we know that most cardiomyopathy is mainly caused by atherosclerosis, which is the leading cause of death. In addition, according to the heatmap results, we found that *ACTA2, PDLIM3* and *CALD1* were co-expressed with *TPM2*, so we think that this gene may also cause the occurrence and development of atherosclerosis (Fig. 4D). Especially, the *TPM* family genes have played a role in stabilizing myosin filaments and endothelial cells. It suggests that the low expression of *TPM2* in the diseased arteries might damage the function of endothelial cells, and lead to the occurrence of atherosclerosis. Meanwhile, the dysfunction of smooth muscle cells is also one of the factors leading to atherosclerosis. However, the *TPM* family genes can maintain the stability of smooth muscle activity and indirectly promote vasoconstriction. The users can find other equally important genes to further explore problems related to single cells.

3.5. Data download

SCInter provided data downloading in ".csv" and ".txt" format, including all marker profiles, cell annotation information and visualization graphs for each integration dataset and single cell sample. In addition, SCInter supported export query results for each search result page. Users can download all visual vector diagrams and tables in the

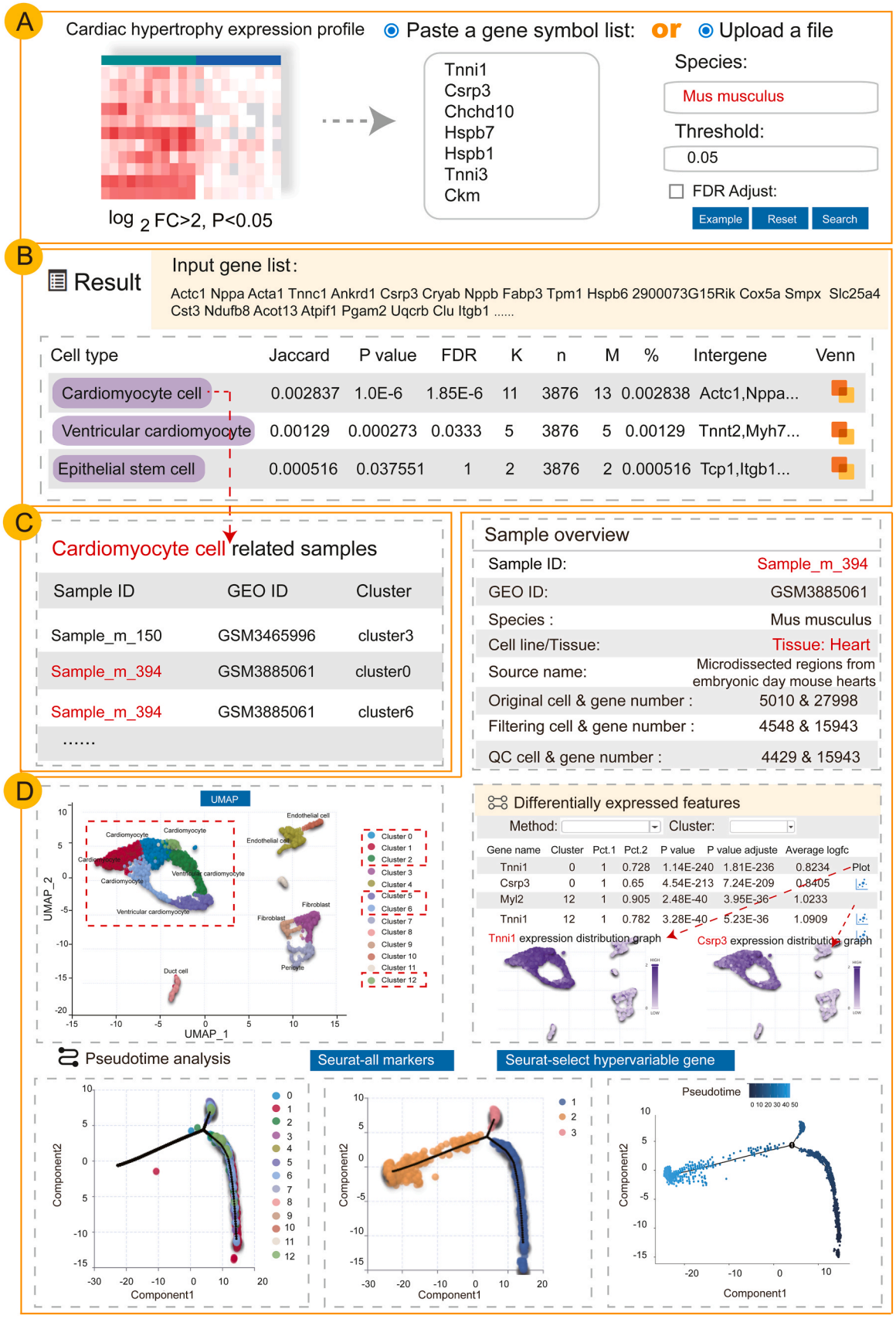

Fig. 3. The online cell type annotation tool of SCInter. (A) User submits a gene list of interest. (B) Results table for cell type analyses. (C) All single-cell data of related specific types. (D) General information and detailed analysis results of the data sample.

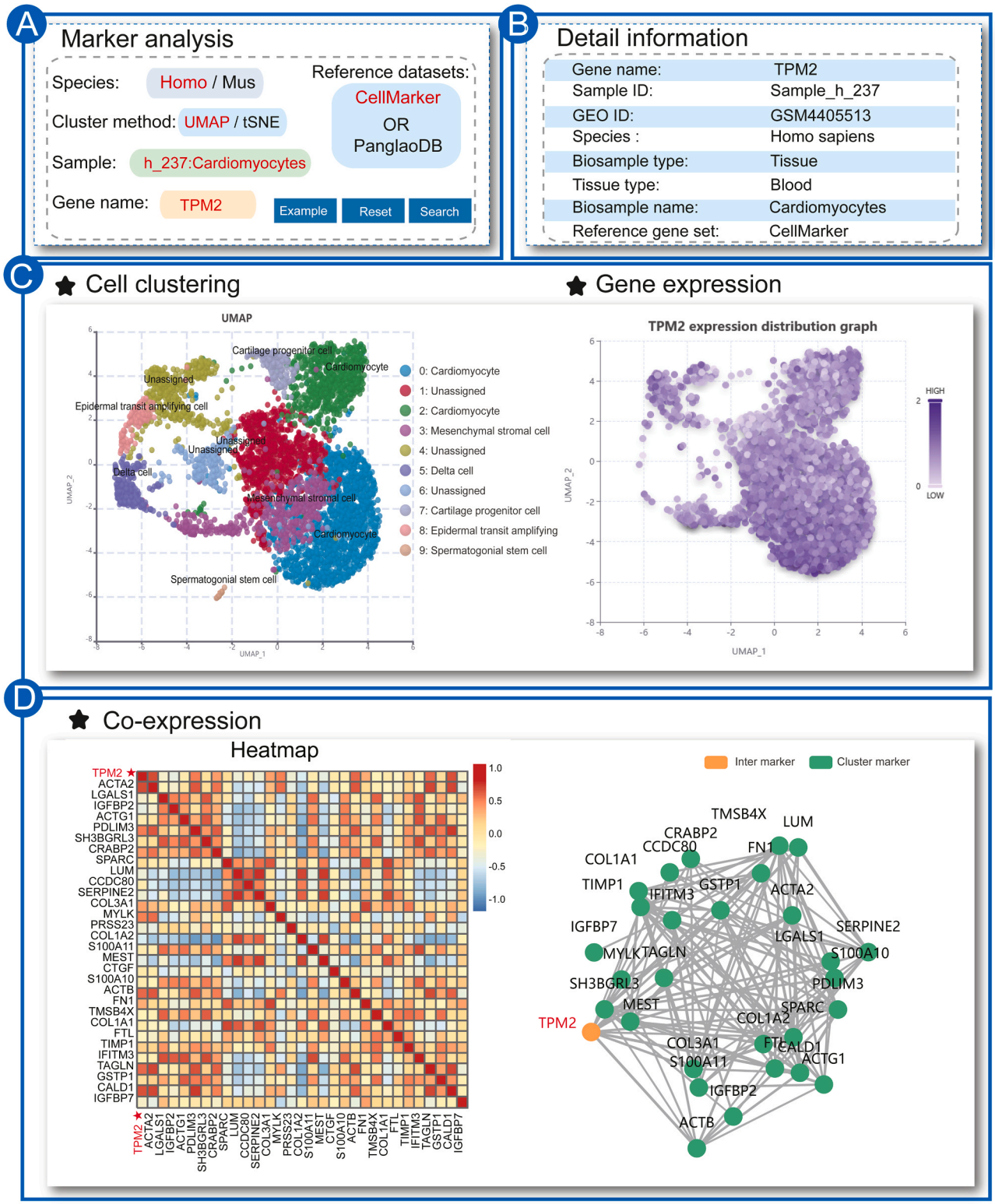

Fig. 4. The online marker analysis tool of SCInter. (A) User submits any sample and any gene of interest. (B) Detailed information about the input gene in a specific sample. (C) Sample clustering and gene expression distribution. (D) Heatmap and network visualization of co-expression results of marker genes.

database as valuable supplementary data for in-depth experimental research.

3.6. Data submission

SCInter encourages sharing single-cell data. We recommend that users submit a sample's GEO ID, tissue type, cell line and platform, as well as a link to their data source. To ensure sample quality, we will check the submitted data before updating. Finally, we will update the database dynamically based on the comprehensive analysis results of new datasets to ensure timely data release.

4. System design and implementation

The SCInter website runs on the Linux-based Apache Web server 2.4.6 (http://www.apache.org) and is developed using MySQL 5.7.27 (http://www.mysql.com). The current version of SCInter uses PHP 5.6.40 (http://www.php.net) server-side scripts. In addition, the SCInter web interface is developed and constructed based on Bootstrap v3.3.7 (https://v3.bootcss.com) and JQuery v2.1.1 (http://jquery.com). Furthermore, ECharts (http://echarts.baidu.com) used as a graph visualization framework. This database has been tested with Mozilla Firefox, Google Chrome and Internet Explorer web browsers. The research community can freely access the information in the SCInter database without registering or logging in. The URL of SCInter is https://bio.liclab.net/SCInter/index.php.

5. Conclusions and future extensions

At present, single-cell transcriptomics has become one of the most significant areas of interest for researchers, leading to the rapid accumulation of single-cell data. Notably, clustering of cells, identifying markers, annotating cell type and analyzing pseudotime can elucidate important roles in cell development and regulatory mechanisms for single cells. SCInter provided comprehensive analyses for the cataloged single-cell data, including cluster analysis for specific cell markers identification, pseudotime analysis and cell type automatic annotation. SCInter also provided cell type annotation and marker analysis for users submitting gene lists. The current version of SCInter cataloged a total of 115 integration datasets and 1016 samples involving 164,253,165 cells which are far more than the existing databases.

SCInter provides a user-friendly interface to query, browse, analyze and visualize each integration dataset and single cell sample with comprehensive QC reports and processing results. The advantages of SCInter include: (i) mammalian single-cell data (human and mouse); (ii) a total of 115 integration datasets and 1016 single cell samples from various protocols and platforms covering nearly 150 tissues/cell lines; (iii) the detailed preprocessing processes including data filtering, quality control and normalization; (iv) user-friendly clustering visualization for each integration dataset and single sample; (v) accurately the development trajectory of cells, cell type automatic annotation and identification of gene co-expression network; (vi) multiple differential gene analysis methods to identify cell markers for each cluster, such as “Wilcoxon rank sum test”, “Student’s *t*-test” and “logistic regression framework”; (vii) Recognition receptor and ligand interaction for cell-to-cell communication; (viii) two representative gene sets for pseudotime analysis for each single sample; (ix) useful and full-featured online analysis tool; and (x) user-friendly displays of integration dataset/single-cell sample information and cell type annotation information with interactive tables.

To constantly follow up new datasets to expand our database, we will consider the newly added transcriptome and single-cell Assay for Transposase-Accessible Chromatin sequencing (scATAC-seq) datasets. scATAC-seq data has recently been extended to thousands of cells for complex diseases [59]. Through analyzing scATAC-seq data, we can predict the activity of transcription factors and explore deep molecular mechanisms [60,61]. In future, we will perform integrated analysis based on scRNA-seq and scATAC-seq data [62,63]. And we believe that SCInter can help reveal intrinsic mechanisms underlying complex biological processes, so as to better understand cellular identities and functions.

Funding

This work was supported by the National Natural Science Foundation of China [62171166, 62001145, 62272212, and 62301194]; 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]; China Postdoctoral Science Foundation [2019M661311]; Key Discipline Construction Project of Harbin Medical University-Daqing [HD-ZDXK-202004]; Construction Project of Scientific Research and Innovation Team of Harbin Medical University-Daqing [HD-CXTD-202002]; Fundamental Research Funds for the Provincial Universities of Harbin Medical University-Daqing [JFQN202309].

CRediT authorship contribution statement

Jun Zhao: 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. Yuezhu Wang: Mainly responsible for building the database background. Chenchen Feng: Helped to check the English. Mingxue Yin: Mainly responsible for the analysis of some data 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 authors declare no competing interests.

Data availability

The research community can access information freely in the SCInter without registration or logging in. The URL of SCInter is https://bio.liclab.net/SCInter/index.php.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.11.024.

References

[1] Linnarsson S, Teichmann SA. Single-cell genomics: coming of age. Genome Biol 2016;17:97.
[2] Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell 2015;58:610–20.
[3] Johnson MB, Wang PP, Atabay KD, Murphy EA, Doan RN, Hecht JL, Walsh CA. Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. Nat Neurosci 2015;18:637–46.
[4] Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 2013;498:236–40.
[5] Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. Science 2017;358:58–63.
[6] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161:1187–201.
[7] Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, Mazutis L. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc 2017;12:44–73.
[8] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–14.
[9] Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol 2015;33:503–9.
[10] Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 2016;167:1853–66. e1817.
[11] Yu Y, Tsang JC, Wang C, Clare S, Wang J, Chen X, Brandt C, Kane L, Campos LS, Lu L, et al. Single-cell RNA-seq identifies a PD-1(hi) ILC progenitor and defines its development pathway. Nature 2016;539:102–6.
[12] Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539:309–13.
[13] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al. NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 2011;39:D1005–10.

[14] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. The human cell atlas. Elife 2017:6.
[15] Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.
[16] Sokolowski DJ, Faykoo-Martinez M, Erdman L, Hou H, Chan C, Zhu H, Holmes MM, Goldenberg A, Wilson MD. Single-cell mapper (scMappR): using scRNA-seq to infer the cell-type specificities of differentially expressed genes. NAR Genom Bioinform 2021;3:lqab011.
[17] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888–902. e1821.
[18] Cao Y, Zhu J, Jia P, Zhao Z. scRNASeqDB: a database for RNA-seq based gene expression profiles in human single cells. Genes (Basel) 2017:8.
[19] Zheng LL, Xiong JH, Zheng WJ, Wang JH, Huang ZL, Chen ZR, Sun XY, Zheng YM, Zhou KR, Li B, et al. ColorCells: a database of expression, classification and functions of lncRNAs in single cells. Brief Bioinform 2021:22.
[20] Qi C, Wang C, Zhao L, Zhu Z, Wang P, Zhang S, Cheng L, Zhang X. SCovid: single-cell atlases for exposing molecular characteristics of COVID-19 across 10 human tissues. Nucleic Acids Res 2022;50:D867–74.
[21] Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, Ginhoux F, Chen J. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Res 2022;50:D596-D602.
[22] Huang K, Gong H, Guan J, Zhang L, Hu C, Zhao W, Huang L, Zhang W, Kim P, Zhou X. AgeAnno: a knowledgebase of single-cell annotation of aging in human. Nucleic Acids Res 2023;51:D805–15.
[23] Pan L, Shan S, Tremmel R, Li W, Liao Z, Shi H, Chen Q, Zhang X, Li X. HTCA: a database with an in-depth characterization of the single-cell human transcriptome. Nucleic Acids Res 2023;51:D1019–28.
[24] Zheng H, Pomyen Y, Hernandez MO, Li C, Livak F, Tang W, Dang H, Greten TF, Davis JL, Zhao Y, et al. Single-cell analysis reveals cancer stem cell heterogeneity in hepatocellular carcinoma. Hepatology 2018;68:127–40.
[25] Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, Qin Q, Fan J, Qiu X, Xie Y, et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol 2020;21:198.
[26] Gao C, Zhang M, Chen L. The comparison of two single-cell sequencing platforms: BD rhapsody and 10x genomics chromium. Curr Genom 2020;21:602–9.
[27] Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct comparative analyses of 10X genomics chromium and smart-seq2. Genom Proteom Bioinforma 2021;19:253–66.
[28] Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 2016;17:77.
[29] Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, et al. Mapping the mouse cell atlas by microwell-seq. Cell 2018;172:1091–107. e1017.
[30] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502.
[31] Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, Marguerat S, Shahrezaei V. bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. Bioinformatics 2020;36:1174–81.
[32] Uddin MR, Howe G, Zeng X, Xu M. Harmony: a generic unsupervised approach for disentangling semantic content from parameterized transformations. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022;2022:20614–23.
[33] Ben Salem K, Ben Abdelaziz A. Principal component analysis (PCA). Tunis Med 2021;99:383–9.
[34] Tan Z, Chen X, Zuo J, Fu S, Wang H, Wang J. Comprehensive analysis of scRNA-Seq and bulk RNA-Seq reveals dynamic changes in the tumor immune microenvironment of bladder cancer and establishes a prognostic model. J Transl Med 2023;21:223.
[35] Li H, Johnson T. Wilcoxon's signed-rank statistic: what null hypothesis and why it matters. Pharm Stat 2014;13:281–5.
[36] Kim TK. T test as a parametric statistic. Korean J Anesth 2015;68:540–6.
[37] Choi SH, Labadorf AT, Myers RH, Lunetta KL, Dupuis J, DeStefano AL. Evaluation of logistic regression models and effect of covariates for case-control study in RNA-Seq analysis. BMC Bioinforma 2017;18:91.
[38] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
[39] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6.
[40] Shi M, Wang J, Zhang C. Integration of cancer genomics data for tree-based dimensionality reduction and cancer outcome prediction. Mol Inf 2020;39:e1900028.
[41] Katritch V, Rueda M, Abagyan R. Ligand-guided receptor optimization. Methods Mol Biol 2012;857:189–205.
[42] Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 2020;15:1484–506.
[43] Chen Y, Wang H, Yang Q, Zhao W, Chen Y, Ni Q, Li W, Shi J, Zhang W, Li L, et al. Single-cell RNA landscape of the osteoimmunology microenvironment in periodontitis. Theranostics 2022;12:1074–96.
[44] Eyquem J, Poirot L, Galetto R, Scharenberg AM, Smith J. Characterization of three loci for homologous gene targeting and transgene expression. Biotechnol Bioeng 2013;110:2225–35.
[45] He X, Liu L, Chen B, Wu C. Using cell type-specific genes to identify cell-type transitions between different in vitro culture conditions. Front Cell Dev Biol 2021;9:644261.
[46] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8.
[47] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[48] Baharav TZ, Kamath GM, Tse DN, Shomorony I. Spectral jaccard similarity: a new approach to estimating pairwise sequence alignments. Patterns (N Y) 2020;1:100081.
[49] Lee S, Chen D, Park M, Kim S, Choi YJ, Moon SJ, Shin DM, Lee JH, Kim E. Single-Cell RNA sequencing analysis of human dental pulp stem cell and human periodontal ligament stem cell. J Endod 2022;48:240–8.
[50] Zhao X, Wu S, Fang N, Sun X, Fan J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. Brief Bioinform 2020;21:1581–95.
[51] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20:163–72.
[52] Frustaci A, Chimenti C, Doheny D, Desnick RJ. Evolution of cardiac pathology in classic Fabry disease: Progressive cardiomyocyte enlargement leads to increased cell death and fibrosis, and correlates with severity of ventricular hypertrophy. Int J Cardiol 2017;248:257–62.
[53] Manivannan SN, Darouich S, Masmoudi A, Gordon D, Zender G, Han Z, Fitzgerald-Butt S, White P, McBride KL, Kharrat M, et al. Novel frameshift variant in MYL2 reveals molecular differences between dominant and recessive forms of hypertrophic cardiomyopathy. PLoS Genet 2020;16:e1008639.
[54] Geier C, Gehmlich K, Ehler E, Hassfeld S, Perrot A, Hayess K, Cardim N, Wenzel K, Erdmann B, Krackhardt F, et al. Beyond the sarcomere: CSRP3 mutations cause hypertrophic cardiomyopathy. Hum Mol Genet 2008;17:2753–65.
[55] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Conte Oncol (Pozn) 2015;19:A68–77.
[56] Martelli M, Ferreri AJ, Agostinelli C, Di Rocco A, Pfreundschuh M, Pileri SA. Diffuse large B-cell lymphoma. Crit Rev Oncol Hematol 2013;87:146–71.
[57] Li S, Young KH, Medeiros LJ. Diffuse large B-cell lymphoma. Pathology 2018;50:74–87.
[58] Wu ZY, Wu ZG, Qi HM, Chang ZT, Zhou YZ, Hong L. Correlation between MRAS gene polymorphism and atherosclerosis. Eur Rev Med Pharm Sci 2020;24:5644–9.
[59] Giansanti V, Tang M, Cittaro D. Fast analysis of scATAC-seq data using a predefined set of genomic regions. F1000Res 2020;9:199.
[60] Trevino AE, Muller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Pasca AM, Kundaje A, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell 2021;184:5053–69. e5023.
[61] Murphy LJ, Friesen HG. Differential effects of estrogen and growth hormone on uterine and hepatic insulin-like growth factor I gene expression in the ovariectomized hypophysectomized rat. Endocrinology 1988;122:325–32.
[62] Kartha VK, Duarte FM, Hu Y, Ma S, Chew JG, Lareau CA, Earl A, Burkett ZD, Kohlway AS, Lebofsky R, et al. Functional inference of gene regulation using single-cell multi-omics. Cell Genom 2022:2.
[63] Lu ZJ, Ye JG, Wang DL, Li MK, Zhang QK, Liu Z, Huang YJ, Pan CN, Lin YH, Shi ZX, et al. Integrative single-cell RNA-Seq and ATAC-seq analysis of mouse corneal epithelial cells. Invest Ophthalmol Vis Sci 2023;64:30.