Computational and Structural Biotechnology Journal 23 (2024) 1978–1989

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Software/web server article

NTCdb: Single-cell transcriptome database of human inflammatory-associated diseases

Chaochao Wang [1], Ting Huyan [1], Wuli Guo, Qi Shu, Qi Li [2,*], Jianyu Shi [*]

School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Single cell transcriptome
Inflammatory-associated disease
Immunocytes
Cell communication
Database

ABSTRACT

With both the advancement of technology and the decline in costs, single-cell transcriptomics sequencing has become widespread in the biomedical area in recent years. It can facilitate the pathogenic characteristics at the single-cell level, which will assist clinical researchers in exploring the mechanism of diseases. As a result, single-cell transcriptome data based on clinical samples grew exponentially. However, there is still a lack of a comprehensive database about immunocytes in inflammatory-associated diseases. To address this deficiency, we propose a human inflammatory-associated disease-based single-cell transcriptome database, NTCdb (www.ntcdb.org.cn). NTCdb integrates the open-source data of 1,023,166 cells derived from 11 tissues of 17 inflammatory-associated diseases in a uniform pipeline. It provides a set of analyzing results, including cell communication analysis, enrichment analysis, and Pseudo-Time analysis, to obtain various characteristics of immune cells in inflammatory-associated disease. Taking COVID-19 as a case study, NTCdb displays important information including potentially significant functions of certain cells, genes, and signaling pathways, as well as the commonalities of specific immunocytes between different inflammatory-associated disease.

1. Introduction

Inflammation is part of the complex biological response of body tissues relevant to a series of immune-related reaction, which is to eliminate the initial cause of cell injury, clear out necrotic cells and initiate tissue repair, this process involving immune cells, blood vessels, and molecular mediators and play very important role in keeping human healthy [1]. Inflammatory- associated diseases is the disease associated with the inflammation which has the characteristics of pain and swelling, redness, heat, and loss of function [2]. Inflammation can be induced by physical factors such as burns, frostbite, trauma, foreign bodies (splinters, dirt and debris) invasion, and ionizing radiation [3]; as well as the biological causes including infection by pathogens, hypersensitivity, stress and chemical irritants [4]. Inflammatory- associated disease can be roughly divided into acute inflammation and chronic inflammation. Acute inflammation occurs immediately upon injury, lasting only a few days, presented by migration of neutrophils and macrophages to the site of inflammation [5]. Chronic inflammation is inflammation that lasts for months or years which relevant to dysregulated immune response and interaction between immune cells and local tissues [6]. With further research, people realized that diverse diseases, such as atherosclerosis, autoimmune diseases, diabetes, Alzheimer's disease, psoriasis, asthma, chronic lung diseases, inflammatory bowel disease, multiple viral infection-associated diseases, even cancer are all related to inflammation [6–10].

Immunocyte is important responder and performer of inflammation. The dysfunction of immunocyte would cause unexpected acute or chronic inflammations. In some worse cases, uncontrolled inflammation responses severely damage human health and even cause death [7,11]. Thus, the uncovering of underlying mechanisms of dysregulated immune cell and inflammatory response contributes to exploring the pathogeny of inflammatory-associated diseases and develop therapeutic targets.

Sequencing, one of the most important modern biotechnologies, provides full-scale information on genome, transcriptome, and epigenome in tissues and cells. Compared to traditional biological assays, sequencing technologies enable researchers to uncover the underlying mechanism of diseases in a fast, high-throughput, and automatic

[*] Corresponding authors.
E-mail addresses: liqi_1111@nwpu.edu.cn (Q. Li), jianyushi@nwpu.edu.cn (J. Shi).
[1] These authors contributed equally to this work
[2] School of Life Sciences, Northwestern Polytechnical University, Xi'an, Shaanxi Province, 710129, China

https://doi.org/10.1016/j.csbj.2024.04.057
Received 15 January 2024; Received in revised form 22 April 2024; Accepted 22 April 2024
Available online 25 April 2024

manner. The first generation of sequencing technologies, called bulk sequencing, analyzes a population of cells with the assumption that these cells are homogeneous. For example, bulk transcriptome sequencing can easily obtain differentially expressed genes (DEGs) with respect to (w.r.t.) diseases between patients and healthy controls [12]. However, bulk sequencing cannot capture the significant heterogeneity between cell individuals, especially the heterogeneity related to the spatial and temporal developments of diseases. The urgent need to explore the characteristics of individual cell impels the development of single-cell sequencing (sc-Seq), which can acquire information of individual cell with high resolution and enables researchers to capture individual changes of cells and obtain more details in life activities. scRNA-seq, which focuses on sequencing the transcriptome of single cell, has obtained abundant data since its inception in 2009 [13]. The massive amount of data has increased the difficulty for researchers to extract target information to some extent. Consequently, establishing self-built databases to manage data has been proposed. In the following years, varied scRNA-seq databases have been built [14]. According to the purpose of analysis, these databases can be roughly classified into cell identification and pathological mechanism identification. The former aims to realize the identification of cell types and states by analyzing gene expression profiles during a certain physiological process. For example, Cellmaker records tens of thousands of marker genes, which can differentiate human and mouse cells [15]. SCDevDB can identify the developmental stage of cells on a single-cell level by tracking the gene expression during human body developmental processes [16]. The latter mainly identify pathological characteristics by comparing the DEGs of pathological conditions with those of healthy states. For instance, based on the integration of 76 tumor datasets across 27 cancer types and three peripheral blood mononuclear cell (PBMC) datasets, Tumor Immune Single Cell Hub (TISCH) identifies heterogeneity of tumor microenvironment by comparing the expression characteristics of malignant cells and non-malignant cells [17]. CancerSEA analyzed and displayed 49 cancer-related scRNA-seq datasets and distinguished the different status of tumor cells (e.g., stemness, invasion and metastasis) according to the gene expression patterns [18]. DISCO integrated the scRNA-seq data from several human diseases (COVID-19, breast cancer, and colorectal cancer) to provide the most critical pathological information [14]. Nowadays, these scRNA-seq databases have improved researchers' understanding of the biological characteristics of cells in pathological status, which will help in both developing targeted therapy and improving prognosis [19]. Nevertheless, a scRNA-seq database containing comprehensive information on inflammatory-associated diseases is still missing up to now.

To fill this gap, this work constructs a novel scRNA-Seq database "Notable single-cell Transcriptome inflammation Computational database (NTCdb)" to partially reveal the mechanism of inflammation-associated diseases (see the details in Supplementary Table 1) by profiling single cell's characteristics and facilitate the discover of potential therapeutic targets. The overview of sample information is showed in Fig. 1.

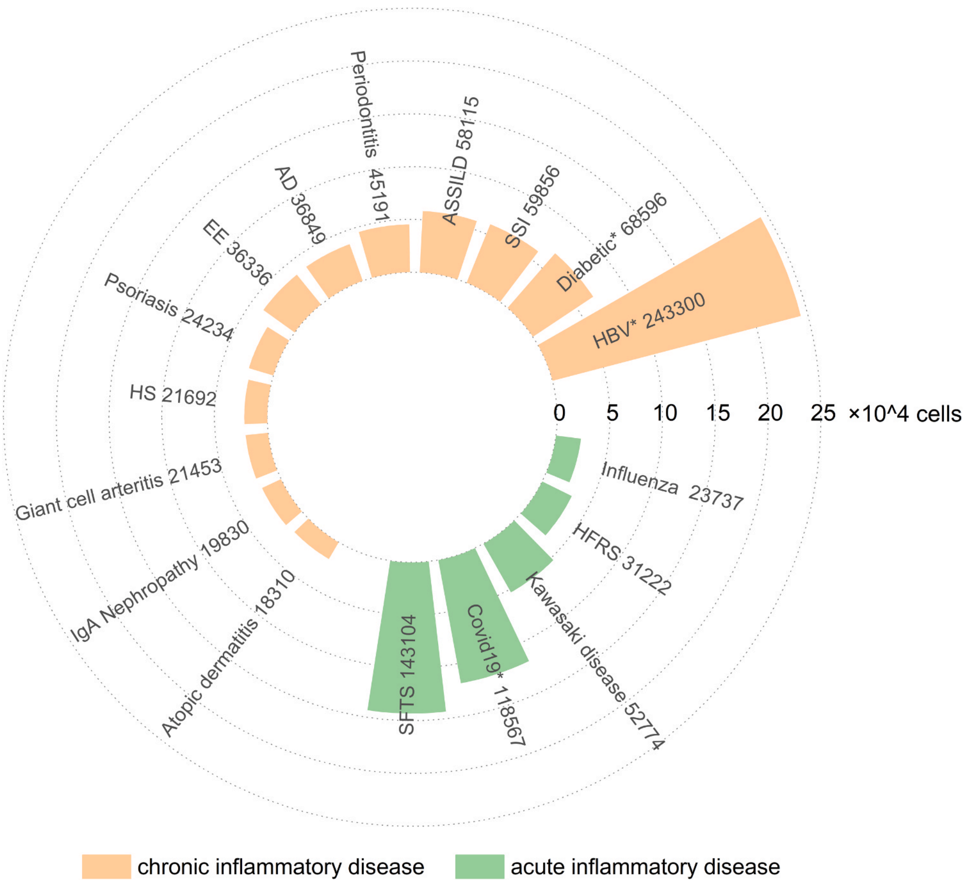


Fig. 1. Overview of Sample information. The involving inflammation-related diseases falls into chronic (orange) and acute (green) diseases in terms of the characteristics of inflammatory inducing responses. There are 12 chronic diseases, including Atopic dermattis, IgA Nephropathy, Giant cell arteritis, Hidradenitis suppurativa (HS), Psoriasis, Eosinophilic esophagitis (EE), Alzheimer's disease (AD), Periodontitis, Anti-Synthetase Syndrome-associated Interstitial Lung Disease (ASSILD), Systemic Sclerosis-associated Interstitial lung disease (SSI), Diabetic, HBV. Acute diseases involve Severe Fever with Thrombocytopenia Syndrome (SFTS), Covid 19, Kawasaki Disease, Hemorrhagic Fever with Renal Syndrome (HFRS), and, Influenza. The numbers of collected cells w.r.t. diseases follow disease names, while the radiuses of dotted circles denote the scale of cell numbers. In addition, the asterisk (*) next to the disease name indicates that samples were collected from two tissues.

2. Results

The interface of NTCdb provides four main components: 'Home', 'Query', 'Shared DEGs', and 'Miscellaneous' (shown in the left menu of Fig. 2). The 'Home' page only displays the statistics of diseases and samples number in NTCdb. The 'Query' component contains a search page which display a full list of data entries, covering a set of basic items, external links, and internal links. The 'Shared DEGs' page provides a gene query function to find common genes which are differentially expressed among various diseases. The 'Miscellaneous' component provides three pages, including the way of data preprocessing, the used R packages, and the annotation mapping between cell abbreviations and their full names.

To illustrate the functions of NTCdb, we select Corona Virus Disease 2019 (COVID-19), a severe pandemic in recent years, as a case in the following sections.

2.1. Searching

Once we enter the disease's name (i.e., "Covid", this website support search keyword with capital/lowercase mode) in the search box, NTCdb retrieves a list of single-cell data entries w.r.t. "Covid" (Fig. 3. a). Each Covid record contains basic items, external links, and internal function links. Taking the first entry as a demonstration.

Its basic items include Disease Name (Covid), Sampling Source (peripheral blood), #Cells (the number of cells, 42862 evaluated cells under quality assurance), and Metadata (14 records). Specifically, the value in Metadata indicates the number of clinical records and its hyperlink leads to an extra page that lists the clinical details collected by the original literature (Fig. 3-b).

The external links contain its PMID (i.e., 32514174) and its Data Access identity (i.e., GSE150728). The PMID shows as a URL to the source publication in NCBI literature [20], where both the abstract and the full text can be usually downloaded. The accession identity also provides a URL to the page of the data w.r.t. the publication in GEO, where the data details can be found (e.g., the contributor, the submission, the generating platform, and the raw data).

The internal function links provide the specific function operations of data entry, involving Cell Population, #Cell Types, DEG, and Pseudo-Time as follows.

- The Cell Population entry provides a two-dimensional spatial distribution profile of all the cells included in this dataset. The user can click the button 'View' to obtain the cell distribution (Section 2.2).
- The #Cell Types entry indicates the number of cells types and also provides a URL to a functional page, which lists cell proportions, group, source w.r.t. samples. More importantly, the page contains three functional buttons, 'Visualization', 'PCA', and 'Cell Talk' for finding potential target cells which are significantly altered or have cell communication changes (Section 2.3).
- The DEG entry provides a hyperlink to a functional page, which indicates the DEGs obtained by comparing diseased and healthy statuses. Similarly, the page contains three functional buttons, 'Visualization', 'GO', and 'KEGG' for finding potential targeted genes based on DEGs, enriched pathways, and their associations (Section 2.4).
- The Pseudo-Time entry shows a functional button 'View' to infer the dynamic changes in gene expression from the healthy state to the diseased state based on cell type-independent DEGs (Section 2.5).

In addition, three buttons in the column "Download" enable users to download supplementary files, including source codes in R language, the RDS file (an R object containing all cells and annotation information), and the MON file (an R object consisting of randomly sampled cells used in the Pseudo-Time analysis).

2.2. Cell population

After clicking the 'View' button in the Cell Population column, NTCdb provides the two-dimensional distribution of single-cell data points (Fig. 4), which is produced by the method of Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction. The distribution is displayed by four plots in terms of sample origins/tissues (Fig. 4-a), disease status (Fig. 4-b), cell clusters (Fig. 4-c), and cell types (Fig. 4-d) respectively.

Fig. 4-a comprehensively displayed the distribution of each sample according to sampling sources, allowing for a thorough examination of sample uniformity, thereby reflecting the level of technical noise inherent in the sequencing data. Fig. 4-b offers a holistic representation of cell distribution in both health and disease conditions. It elucidates three distinctive categories: a prominent class characterized by a significantly higher abundance of healthy cells (highlighted in yellow box 1), another class exhibiting a noteworthy enrichment of diseased cells (highlighted in yellow box 2), and the third class shows the distributions of patient and healthy cells are neighboring (highlighted in yellow box 3). Fig. 4-c shows cell communities generated by the Louvain algorithm. Lastly, Fig. 4-d annotates the types of cell individuals.

According to Fig. 4-b and Fig. 4-d, the class predominantly composed of healthy cells corresponds to B cells, whereas the class characterized by a larger number of diseased cells is primarily annotated as Pre-B cells. The class exhibiting a close resemblance between patient and healthy cells is mainly associated with monocytes. Fig. 4-c reveals that this particular region encompasses five distinct cell clusters, indicating that gene expression in monocytes may have significantly deviated in

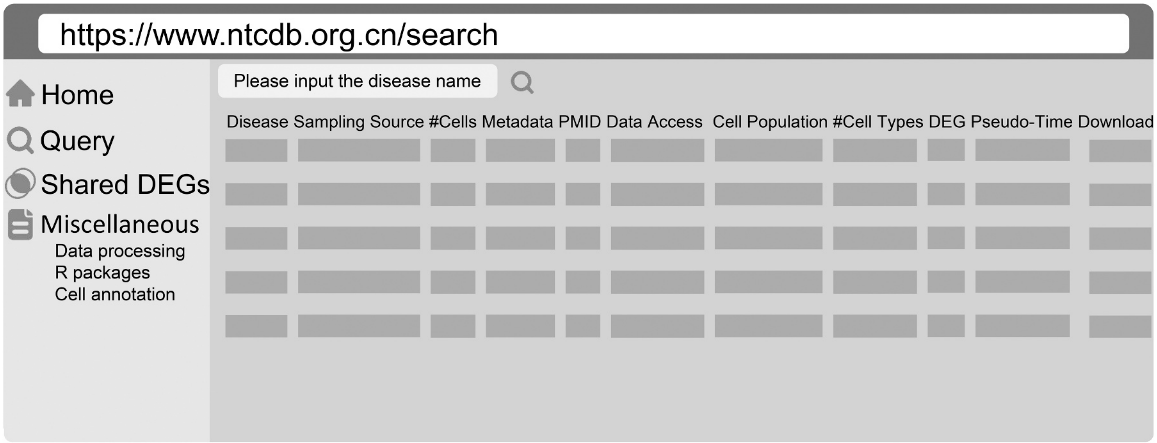
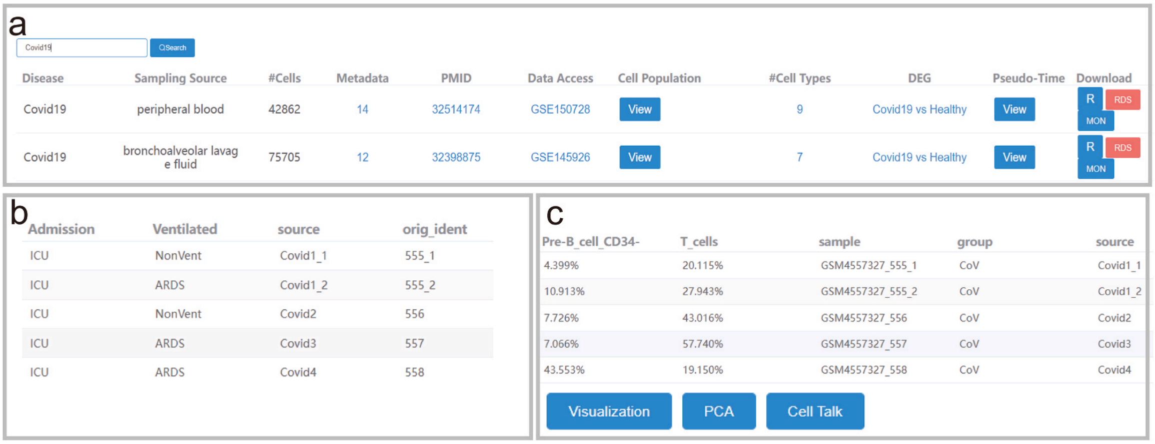

https://www.ntcdb.org.cn/search
Home
Query
Shared DEGs
Miscellaneous
Data processing
R packages
Cell annotation
Please input the disease name
Disease Sampling Source #Cells Metadata PMID Data Access Cell Population #Cell Types DEG Pseudo-Time Download

Fig. 2. User interface of NTCdb.



Disease	Sampling Source	#Cells	Metadata	PMID	Data Access	Cell Population	#Cell Types	DEG	Pseudo-Time	Download
Covid19	peripheral blood	42862	14	32514174	GSE150728	View	9	Covid19 vs Healthy	View	R RDS MON
Covid19	bronchoalveolar lavage fluid	75705	12	32398875	GSE145926	View	7	Covid19 vs Healthy	View	R RDS MON

Admission	Ventilated	source	orig_ident
ICU	NonVent	Covid1_1	555_1
ICU	ARDS	Covid1_2	555_2
ICU	NonVent	Covid2	556
ICU	ARDS	Covid3	557
ICU	ARDS	Covid4	558

Pre-B_cell_CD34-	T_cells	sample	group	source
4.399%	20.115%	GSM4557327_555_1	CoV	Covid1_1
10.913%	27.943%	GSM4557327_555_2	CoV	Covid1_2
7.726%	43.016%	GSM4557327_556	CoV	Covid2
7.066%	57.740%	GSM4557327_557	CoV	Covid3
43.553%	19.150%	GSM4557327_558	CoV	Covid4

Fig. 3. Searching results. (a) Retrieval. Two records of Covid are retrieved. Each record contains eleven items, including four basic items (i.e., the name of the disease of interest, its sampling source, the number of its cells, the number of involving clinical records (i.e., Metadata)), two external links (i.e., the PMID and the identity of Data Access in GEO), four internal function links (i.e., its cell population, cell types, DEGs, and Pseudo-Time trajectory), and the additional download links from left to right. (b) Metadata. It contains clinical records w.r.t. sample. Each row represents a sample, while columns represent the clinical features of the sample. (c) Cell Types. Each row represents a sample, displaying the cell type and sample source, and the percentage represents the cell type's proportion in the sample (see also Section 2.3 for details).

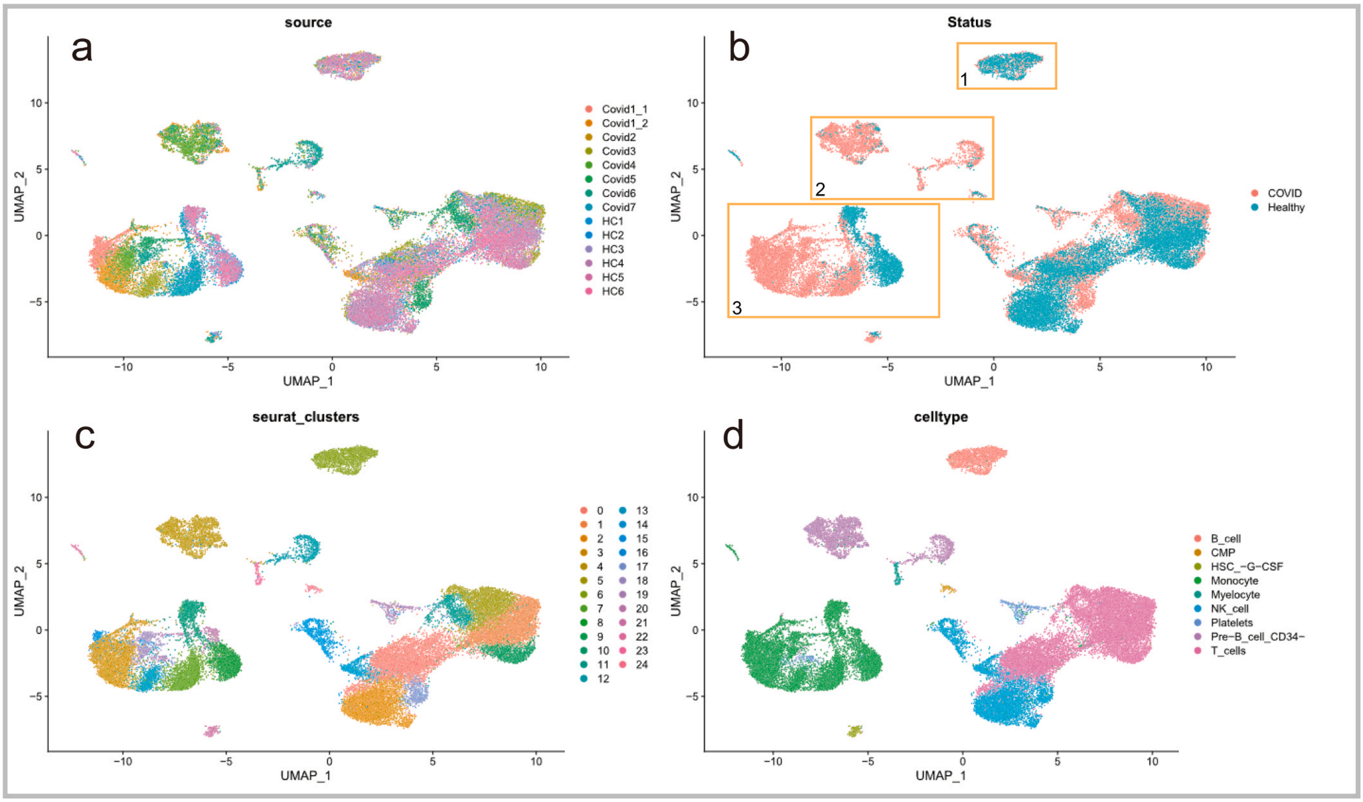


Fig. 4. Single-cell distribution plots. Four maps of single-cell data on a two-dimensional space are produced by UMAP. Points in the figure represent individual cells. (a) Sample sources. Different colors represent different sample sources, including 7 healthy controls and 6 patients. (b) Physiological status. Two colors indicate the presence and absence of Covid respectively. (c) Cell clusters. Totally 25 distinct cell clusters are found. (d) Cell types. Nine cell types are rendered in appropriate colors, including B cell (B_cell), Common myeloid progenitor cell (CMP), Hematopoietic stem cells mobilized with G-CSF (HSC_-G-CSF), Monocyte, Myelocyte, Natural killer cell (NK_cell), Platelet, Pre-B cell (Pre-B_cell_CD34-), and T cell (T_cells).

patients compared to the healthy controls or monocytes have undergone differentiation induced by the disease.

2.3. *Finding potential target cells of disease*

NTCdb provides an approach to investigate the variance of cell types since they are the key players in the disease process [21]. In the case of "Covid", nine types of cells are involved. After clicking the URL in the column of #Cell Types, a functional page (Fig. 3-b), which lists the detailed records of cell types, including percentages over cell types, group categories (including Cov and HC), and sample IDs in GEO was shown. More importantly, it provides three functional buttons:

'Visualization', 'PCA', and 'Cell Talk', which help find potential target cells. Their functions are illustrated as follows.

- The 'Visualization' button provides stacked bars to show the percentages of cells in samples, and the statistical change of proportion of each cell type by comparing diseased samples with healthy control (Fig. 5. b). In the example of "Covid", as shown in cell proportion bars (Fig. 5. a), Pre-B cells (rendered in purple) show the most significantly different proportions between COVID-19 patients (marked with the prefix 'Covid') and healthy control (marked with the prefix 'HC'). Meanwhile, the statistical significance (p-value = 0.0013) also demonstrates such a difference between Covid and HC (Fig. 5. b). Thus, Pre-B cells can be regarded as potential targeting cells in COVID-19.
- The 'PCA' button provides the result of Principal Component Analysis (PCA) dimensionality reduction on the cell proportion data. Fig. 5.c-1 reveals the relationship between cell types (indicated by blue arrows), samples (represented by small dots), and the first and second principal components (x and y axes). The values of the cell types projected onto the axes indicate their correlation with the principal components. In PCA, the cos2 value is used to measure the contribution of variables (cell types) to the principal components. Fig. 5.c-2 provides a visual depiction of this relationship. Therefore, considering the most important variables in the main principal components, pre-B cells, platelets, and T cells exhibited considerable alteration, indicating that they merit a comprehensive investigation (Fig. 5. c).
- The 'Cell Talk' feature showed by a Circos plot depicts the enhanced or weakened interactions between cells, primarily based on ligand-receptor interactions under pathological conditions. Hence, with regards to natural killer (NK) cells, there were two sets of ligand-receptor combinations, COL11A1-ITGB1 and EREG-ERBB2 deserve special consideration, since their interactions appear to be greatly amplified and decreased, respectively (Red box part in Fig. 5.d). The independent clear images of Fig. 5 were showed in Supplementary Figure 2-5.

2.4. Potential target genes in the disease

NTCdb provides a list of DEGs over all the cell types as they are the primary focus of single-cell transcriptome study [22]. It calls the function FindMarkers in the Seurat package to fine 925 COVID-19 related DEGs, of which gene names (shown as URLs), p adjust (p_val_adj), log2FoldChange (avg_log2FC) are listed and organized in 93 subpages. Users can inspect the specific cells by the filter of cell type in the drop box (Fig. 7-a). Then, detailed gene information can be found by clicking the URL to the GeneCards website.

More importantly, it provides three functional buttons: 'Visualization', 'GO', and 'KEGG', which help find potential target genes. Their functions are illustrated as follows:

- The 'Visualization' button generates a volcano plot, which helps the selection of genes with the most notable expression discrepancies (Fig. 6-b). According to Fig. 6.b, Hemoglobin Subunit Beta (HBB) which is involved in oxygen transport [23] has the highest differential expression fold change. Study has shown that mutations in HBB can lead to sickle cell disease and relevant to heightened risk for development and severe outcomes of COVID-19 caused pneumonia [24]. Therefore, COVID-19 possibly affect respiratory system

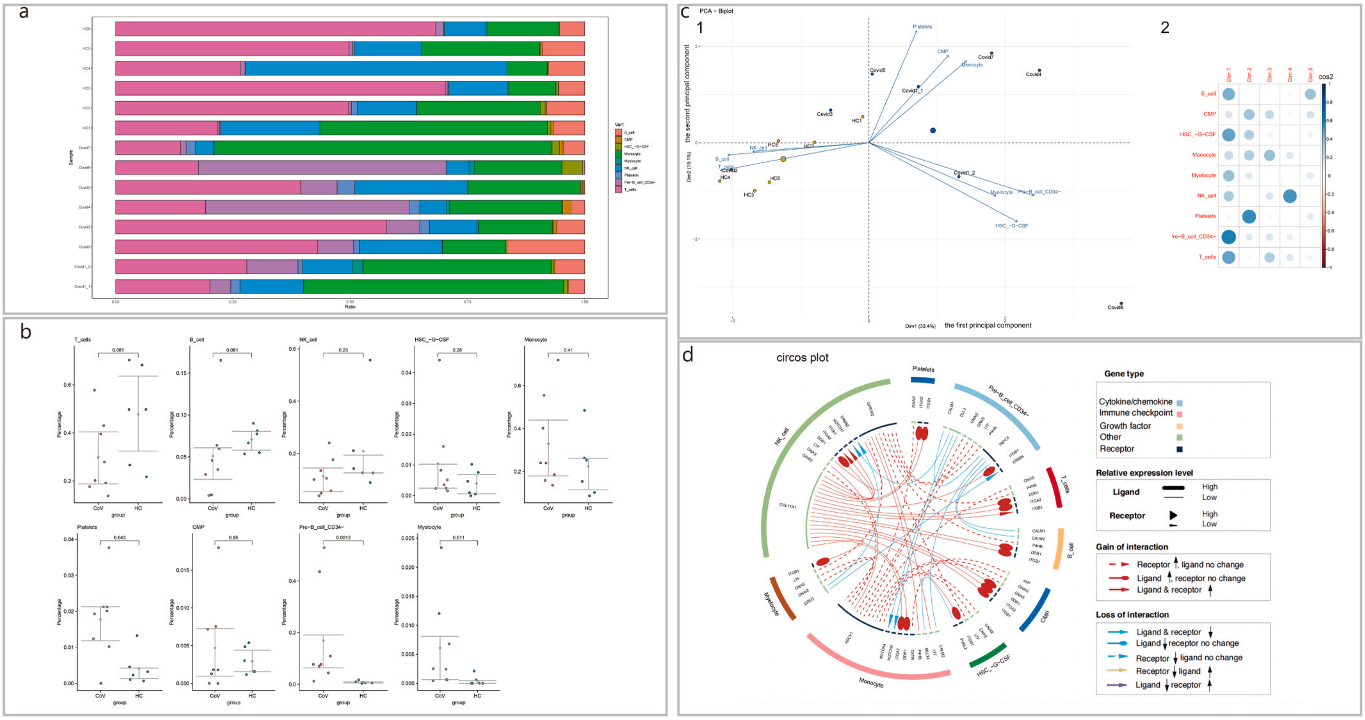
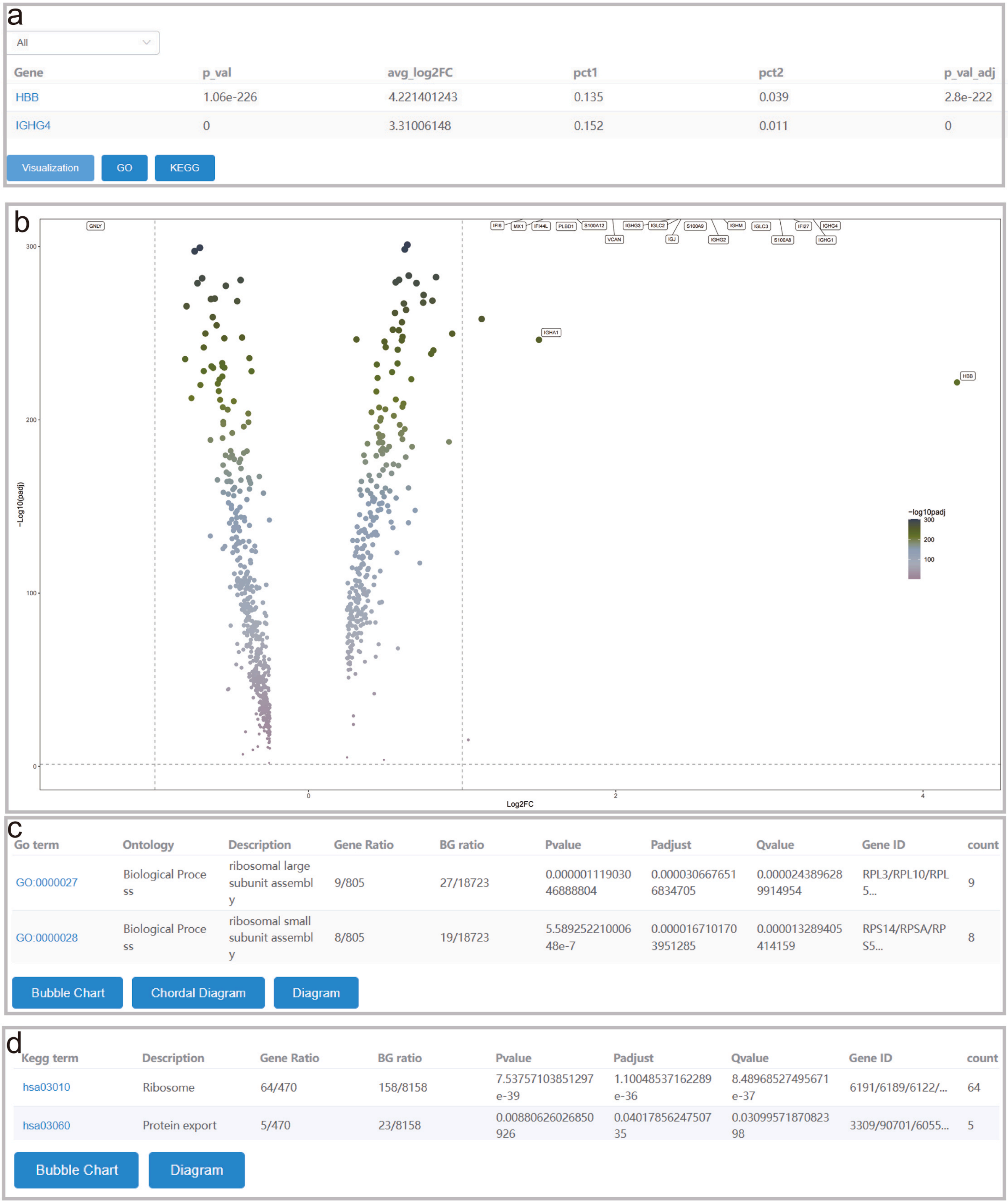


Fig. 5. (a) Proportions of different cell types in sample sources. Sample sources are represented by the vertical axis, while cell proportions are indicated by the horizontal axis. Cell types are highlighted by colors. (b) The proportion of each cell type in all samples. Sample are represented by points, while the vertical axis indicates cell proportions. The patient and the heathy control are compared. (c) PCA analysis. Figure c-1 shows the correlation of samples, variables, and principal components. Samples are points, while cell types are represented by arrows. The horizontal and vertical axes account for the two most important principal components, which are represented by ' Dim '. Figure c-2 consists of the variables and the corresponding 'cos2' of the variables in the principal component. The cos2 value represents the quality of variables. The cell types are represented by the vertical axis. The main components are also represented by ' Dim.', and the size of the circle indicates the cos2 value. (d) Circos plot of cell communications. The Circos plot obtained by ITALK package [34], from the outside to the inside, represents the cell type and the gene type, respectively. The genes are categorized into four types: cytokine/chemokine, immune checkpoint, growth factor, and others. Ligands are represented by line segments, while receptors are represented by arrows. Possible ligand-receptor changes are distinguished by different lines and arrows.



Gene	p_val	avg_log2FC	pct1	pct2	p_val_adj
HBB	1.06e-226	4.221401243	0.135	0.039	2.8e-222
IGHG4	0	3.31006148	0.152	0.011	0

Go term	Ontology	Description	Gene Ratio	BG ratio	Pvalue	Padjust	Qvalue	Gene ID	count
GO:0000027	Biological Process	ribosomal large subunit assembly	9/805	27/18723	0.00000111903046888804	0.0000306676516834705	0.0000243896289914954	RPL3/RPL10/RPL5...	9
GO:0000028	Biological Process	ribosomal small subunit assembly	8/805	19/18723	5.58925221000648e-7	0.0000167101703951285	0.000013289405414159	RPS14/RPSA/RPS5...	8

Kegg term	Description	Gene Ratio	BG ratio	Pvalue	Padjust	Qvalue	Gene ID	count
hsa03010	Ribosome	64/470	158/8158	7.53757103851297e-39	1.10048537162289e-36	8.48968527495671e-37	6191/6189/6122/...	64
hsa03060	Protein export	5/470	23/8158	0.00880626026850926	0.0401785624750735	0.0309957187082398	3309/90701/6055...	5

(caption on next page)

Fig. 6. Summary of genetic information. (a)Difference analysis results. The DEGs obtained by Seurat's FindMarkers function. The value of avg log2FC represents the differential multiple. P_val is the corrected p-value, which indicated the significance more accurately. Click the "Gene column" to access the GeneCards database which contains gene annotation information. (b) DEGs volcano plot. This volcano map was constructed using avg log2FC as the horizontal axis and -log10 (p_val_adj) as the vertical axis. The arbitrarily established threshold is shown by a dashed line, meanwhile, genes are represented by a point. The ordinate denoted the point's color and size. At the top, the genes have a p_val_adj of zero, displaying the name but not a specific value, possibly due to the computer's default precision retention setting. (c) Results of GO enrichment analysis. (d) The outcomes of KEGG enrichment analysis. The ID of the pathway in the KEGG database was denoted in the "KEGG term column" which also links to GenomeNet, and provides an exhaustive explanation of the KEGG pathway. The GO term denotes the path's identifier in the GO database. Regardless of enrichment outcome, the "Description column" refers to the pathway description; "Gene Ratio" reflects the ratio of genes enriched in this pathway to the total number of genes in the pathway, and the "BG ratio" represented the ratio of the genes in this term to all genes. The "P-value" refers to the statistical significance of enrichment analysis. The corrected P-value, "Padjust," is greater than the P-value. In general, a term with a Padjust < 0.05 is considered enrichment. The "Qvalue" is derived from the P-value as well which is defined as the probability that the p-value produces false positives. The "count" reflected the number of genes that have been differentially enriched for this phrase.

function by impacting the expression of HBB and consequently affecting oxygen transport.

- The GO button provides detailed information about enriched pathways, including GO terms, descriptions, p-values, and gene ratios (Fig. 6-c). Additionally, it includes three subpages: Bubble Chart, Chordal Diagram, and Diagram. The Bubble Chart visually presents the most significant pathways using bubble and bar charts. The Chordal Diagram illustrates the overall upregulation and downregulation of pathways based on the expression of DEGs. The Diagram shows the relationship between pathways and genes. In this case, most DEGs were enriched into a pathway related to the "structural constituent of the ribosome" which plays a role in maintaining the integrity of the ribosome (Figs. 7-a1). These results indicated that COVID-19 may impact normal cellular function by affecting the synthesis of ribosomes. Furthermore, Bubble Chart (Figs. 7-a2) was employed as well to better display the results. From the perspective of the Chordal Diagram (Fig. 7-b), the "cytoplasmic translation pathway" may deserve further investigation as it exhibits the most significant down-regulated trend in COVID-19 patients. For a specific gene, the more pathways it influences, the more importance it might have in regulating cellular function. From this standpoint, RPS15A may play an important role in COVID-19 because it participates in the two pathways with the most significance (Fig. 7-c). In addition, similarly to the GO enrichment analysis, the results of the KEGG enrichment analysis (Bubble Chart, Diagram) were also recorded in the database (Fig. 6.d).

2.5. Trajectory analysis via Cell type-independent DEGs

Trajectory analysis is an additional crucial technique for interpreting single-cell data [25]. First, the cells of healthy control and patient with equal number (5000 cells each) were randomly sampled. Monocle software [26] was then used to generate a Pseudo-Time trajectory, and all cells were classified into different branches. NTCdb can draw the Pseudo-Time trajectory and the changes of DEGs along the trajectory. Also, it performs a Branched Expression Analysis Modeling (BEMA) on the DEGs. Thus, the dynamic changes in gene expression from health to disease states can be uncovered and illustrated.

By clicking the 'View' button under the 'Pesudo-Time' column (Fig. 3-a), users can choose from three sub-pages: Trajectory, Pseudo-Time Gene, and BEMA (Fig. 8). Among them, the Trajectory page visually displays the starting point of the Pseudo-Time trajectory (selecting the state with the healthiest cells) (Figs. 8–a1–3), the Pseudo-Time Gene page provides the dynamic changes of DEGs from the trajectory's starting point to the endpoint (Fig. 8-b), and the BEMA page reveals the dynamic changes of genes before and after branching (Fig. 8-c).

The DEGs most relevant to the Pseudo-Time trajectory were depicted by Fig. 8. The genes exhibiting the same expression pattern in the trajectory may be tightly related, and they may regulate each other and influence particular life processes. The first is to pick all genes inside a cluster for later verification, and the second is to investigate a subset of genes from distinct clusters.

In this example, users can pick all genes in cluster 4 (displayed in purple): RPS27, RPL34, CD52, RPL31, RPL37A and RPL15, as well as another group of genes: TRFC, and DUSP1 (displayed in red box), to further study, because their expressions in the trajectory are gradually falling or increasing, respectively (Fig. 8-b).

According to BEMA (Fig. 8-c), the genes in cluster 2 and cluster 3 were up-regulated and down-regulated, respectively in diseased state compared to the healthy cell populations, suggesting that these genes have the potential to change cell fate and deserve further investigation.

2.6. Common DEGs across diseases

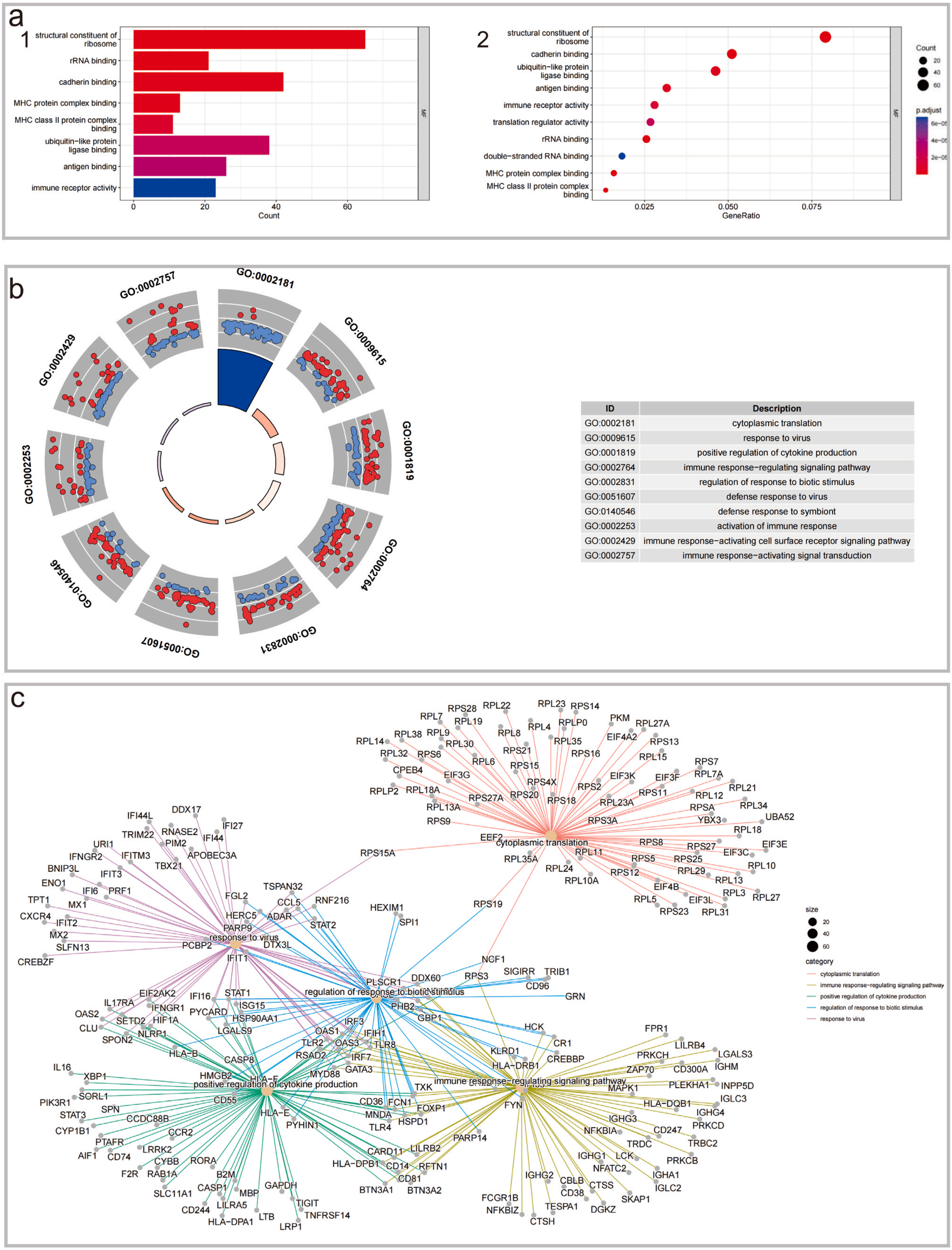
Shared DEGs page is capable to search for DEGs shared by several diseases (Fig. 9). Users can freely pick up to ten distinct diseases in a single search session. After selecting the specific cell type of the selected disease, the common DEGs can be listed in web server. Through browse the differential expression multiple of interested gene, users can and select the interested DEGs and export the list by clicking the 'Export' button. Fig. 9 depicts the DEGs shared by COVID-19 and Influnza included in NTCdb.

Overall, NTCdb intuitively shows the potential targets of genes and cells in COVID-19 patients (Table 1), such as the proportion of Pre-B cells and T cells, as well as the ligand-receptor pairs interaction of COL11A1-ITGB1 and EREG-ERBB2 between NK cells and other cells were both significantly changed in COVID-19 patients. Furthermore, the expression of HBB was affected by COVID-19 significantly, which may be one of the potential targets waiting for subsequent verification. In addition, genes enriched in the "structural constituent of the ribosome" and "cytoplasmic translation pathway" may play important roles in COVID-19. From the results of Pseudo-Time trajectory analysis, genes such as RPS27, RPL34, and RPL31 showing significant expression differences between healthy cell and cells in a diseased state.

3. Discussions

The NTCdb is a human single-cell transcriptomic database that encompasses the single-cell data of inflammatory-associated diseases with great visualization capabilities. In addition to exhibiting the differential and enrichment analysis results, it contains prospective targeted cells, changes of gene expression and intercellular communication intensity under pathological conditions. Ultimately, it provides a means to acquire the significant DEGs shared by many diseases, which may represent the common characteristics of inflammatory-associated diseases. In conclusion, NTCdb can intuitively and conveniently provide users the information about both cells and genes with significant changes in diseases from multiple analytical perspectives.

The NTCdb still requires additional enhancements to be finalized, which will be continuously updated and periodically maintained along with the update of sequencing data and annotation tools. Cell annotation is the key to providing sequencing data with biological meaning. In this stage, the SingleR package was used to annotate cell types automatically. Although, for instance, SingleR annotation can identify natural killer (NK) cells, it is difficult to distinguish CD56 ^{bright} NK cells and



ID	Description
GO:0002181	cytoplasmic translation
GO:0009615	response to virus
GO:0001819	positive regulation of cytokine production
GO:0002764	immune response−regulating signaling pathway
GO:0002831	regulation of response to biotic stimulus
GO:0051607	defense response to virus
GO:0140546	defense response to symbiont
GO:0002253	activation of immune response
GO:0002429	immune response−activating cell surface receptor signaling pathway
GO:0002757	immune response−activating signal transduction

Fig. 7. Visual representation of enrichment results. GO enrichment analysis results can be categorized as molecular function (MF), cellular component (CC), and biological process (BP). (a) A bar graph (a1) and bubble chart (a2) of enriched GO keywords. (b)GO circle. GO term is represented by the perimeter, while the difference between genes expression is shown by the circle in the middle. (c) Gene pathway diagram. The GO term is depicted with colored lines and bright yellow dots. The size of the dots represents the number of enriched genes, whereas the gray dots represent the genes themselves.

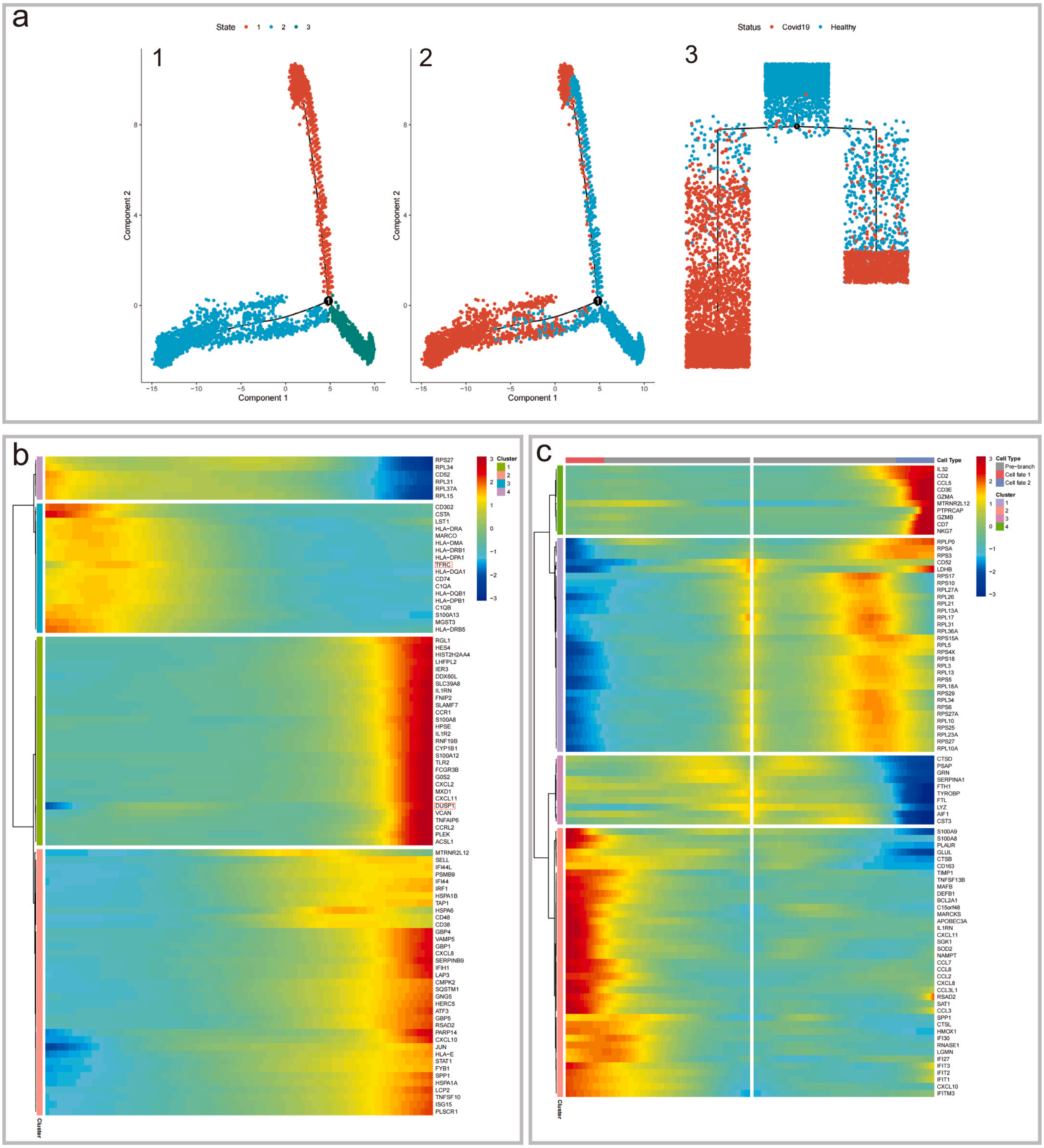
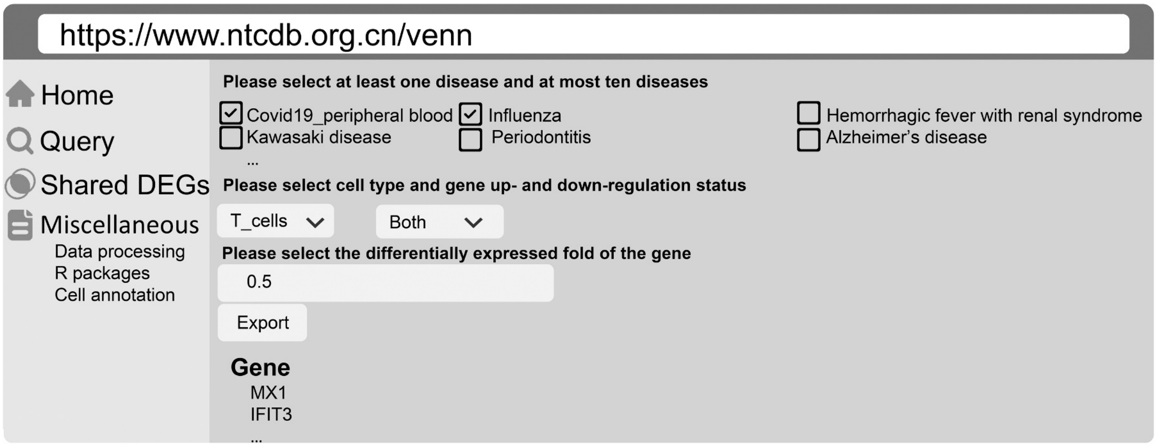


Fig. 8. The Pseudo-Time analysis results. Monocle2 infers the Pseudo-Time of healthy cells and patients' cells. Dots symbolize cells. (b) The DEGs (rows) along the Pseudo-Time (columns) are aggregated hierarchically into four profiles. (c)The DEGs (rows) along the Pseudo-Time branch (columns).

CD56 ^{dim} NK cells because of its low precision. We hope that SingleR will soon be replaced by a more precisely annotated package. By then, NTCdb will be updated to provide more accurate results.

4. Methods

An initial search with the keyword ‘scRNA diseases’ was run in both GENE EXPRESSION OMNIBUS (GEO) and Pubmed databases and resulted in ~5000 scRNA-disease-related entries. Subsequently, a precise filtration was processed on the initial resulting entries to obtain refined entries. We considered three filtering criteria as follows: 1) entries should be derived from human tissues or cell lines; 2) they should originate from inflammatory-associated diseases; 3) they should contain healthy control records. Eventually, we obtained 1023,166 cells derived from 5 acute inflammations and 12 chronic inflammations in total (Fig. 10). The collection was finished on August 10, 2022.



Please select at least one disease and at most ten diseases

☑ Covid19_peripheral blood ☑ Influenza □ Hemorrhagic fever with renal syndrome
□ Kawasaki disease □ Periodontitis □ Alzheimer's disease
...

Please select cell type and gene up- and down-regulation status

T_cells ▾ Both ▾

Please select the differentially expressed fold of the gene

0.5

Gene
MX1
IFIT3
...

Fig. 9. Shared DEGs page function. Users can select numerous diseases. The first dropdown menu is used to select the cell type, and the second dropdown menu allows you to choose up-regulated, down-regulated, or both DEGs.

Table 1
Key information about COVID-19 in the database.

Basis	Conclusion
Fig. 4	The proportions of Pre-B and B cells among COVID-19 patients differ from healthy controls, indicating a potential trend of monocyte differentiation during the course of disease.
Fig. 5	Pre-B cells exhibit significant differences; PCA analysis indicated the importance of pre-B cells, platelets, and T cells; the interaction between NK cells and other cells is most affected by COVID-19, with particular attention warranted for COL11A1-ITGB1 and EREG-ERBB2 interaction.
Fig. 6	HBB exhibited the greatest differential fold-change in COVID-19 patients.
Fig. 7	"Structural constituent of the ribosome" and "cytoplasmic translation pathway," along with RPS15A, may be key the factors involved in COVID-19 prevention.
Fig. 8	The expression of RPS27, RPL34, CD52, RPL31, RPL37A and RPL15 exhibit a significant downward trend during the transition of cells from a healthy state to a diseased state.
Fig. 9	COVID-19 and influenza may have similar immune response mechanisms.

Due to the diversity of data sources, the collected single-cell entries should be standardized for further data analysis. As suggested in SC2disease [27], a uniform processing pipeline was adopted, which contains two modules, Data Preprocess and Target Analysis.

The Data Preprocess module includes three steps: Quality Control, Cell Clustering and Cell Annotation. Regarding that the cells having a high proportion of mitochondrial genes are apoptotic or dead cells [28], the first step is to remove them from single-cell entries. The second step categorizes the remaining single-cell entries (characterized by gene expressions) into cell groups by both PCA and UMAP [29,30]. Seurat, a popular R package [31], was applied to perform the cell clustering. After that, by using a well-known annotation R-package SingleR [32], the third step identifies cell clusters by consulting the datasets provided by a data R-package HumanPrimaryCellAtlasData [33]. HumanPrimaryCellAtlasData contains Chip-sequencing and RNA-sequencing data of specific cell types, while SingleR is an automatic annotated tool by comparing single-cell data with reference data.

Once the data preprocess was done, multiple approaches were employed to examine potential targets (e.g., cells and genes) in disease. The Target Analysis module focuses on two types of disease targets. The first type is specific types of cells, whose proportions in body fluids or tissues are dramatically altered. To identify potential targeted cells, this module discriminates patients from healthy controls in terms of cell proportion and intercellular communication, respectively. First, both Wilcoxon test and PCA were used to determine whether there are significant differences between cell proportion derived from different sources across patients and healthy controls. Then, iTALK package, an analytical package for studying the gains or losses of cellular interactions, was used to reveal ligand-receptor mediated cell communication [34]. The second type refers to the genes with significantly variable expressions. The Target Analysis module indicate the cell type-free DEGs and cell type-specific DEGs between healthy controls and patients as potential targeted genes. During enrichment analysis, the association between DEGs and inflammation-related processes can be dug out. Furthermore, cell type-free DEGs were utilized to construct Pseudo-Time trajectories during disease progression by using Monocle, a R package with reverse graph embedding algorithms for trajectory inference [26]. In the aforementioned process, the main algorithms and tools we use are listed as follows : .

Finally, our database NTCdb includes rich information of inflammation-related scRNA entries, including DEGs, enriched pathways, cell communication outcomes, Pseudo-Time trajectories, and cell proportions. Moreover, it contains literature sources, data entry sources, sample metadata, as well as source codes and RDS (a file containing R objects) files for replicating the Data Preprocess and the Target Analysis.

We have developed a front-end and back-end separated web application to facilitate user access to the information stored in NTCdb. This program is structured using the MVC (Model-View-Controller) architecture, with Vue.js employed for the front-end development. Vue.js provides a multitude of components to effectively display information. We utilize MySQL to store the back-end data, where we have designed seven tables to manage and organize data efficiently. The interaction between the front-end and back-end is implemented using Java. Finally, the website is deployed on Alibaba Cloud server and is proxied using nginx. In Supplementary Figure 1, the design logic of the entire web application is illustrated. Taking the top portion of the figure as an example, when a user accesses the path "www.ntcdb.org.cn/annotation", the back-end triggers the getAnnotation function in the AnnotationController, written in Java. Consequently, all the information stored in the "annotation" table within our MySQL database, which contains cell annotations, is returned.

CRediT authorship contribution statement

Jianyu Shi: Funding acquisition, Methodology, Supervision, Validation. **Qi Li:** Conceptualization, Funding acquisition, Project administration, Writing – review & editing. **Chaochao Wang:** Data curation, Formal analysis, Writing – original draft. **Ting Huyan:** Data curation, Formal analysis, Writing – original draft. **Qi Shu:** Resources, Validation. **Wuli Guo:** Resources, Validation.

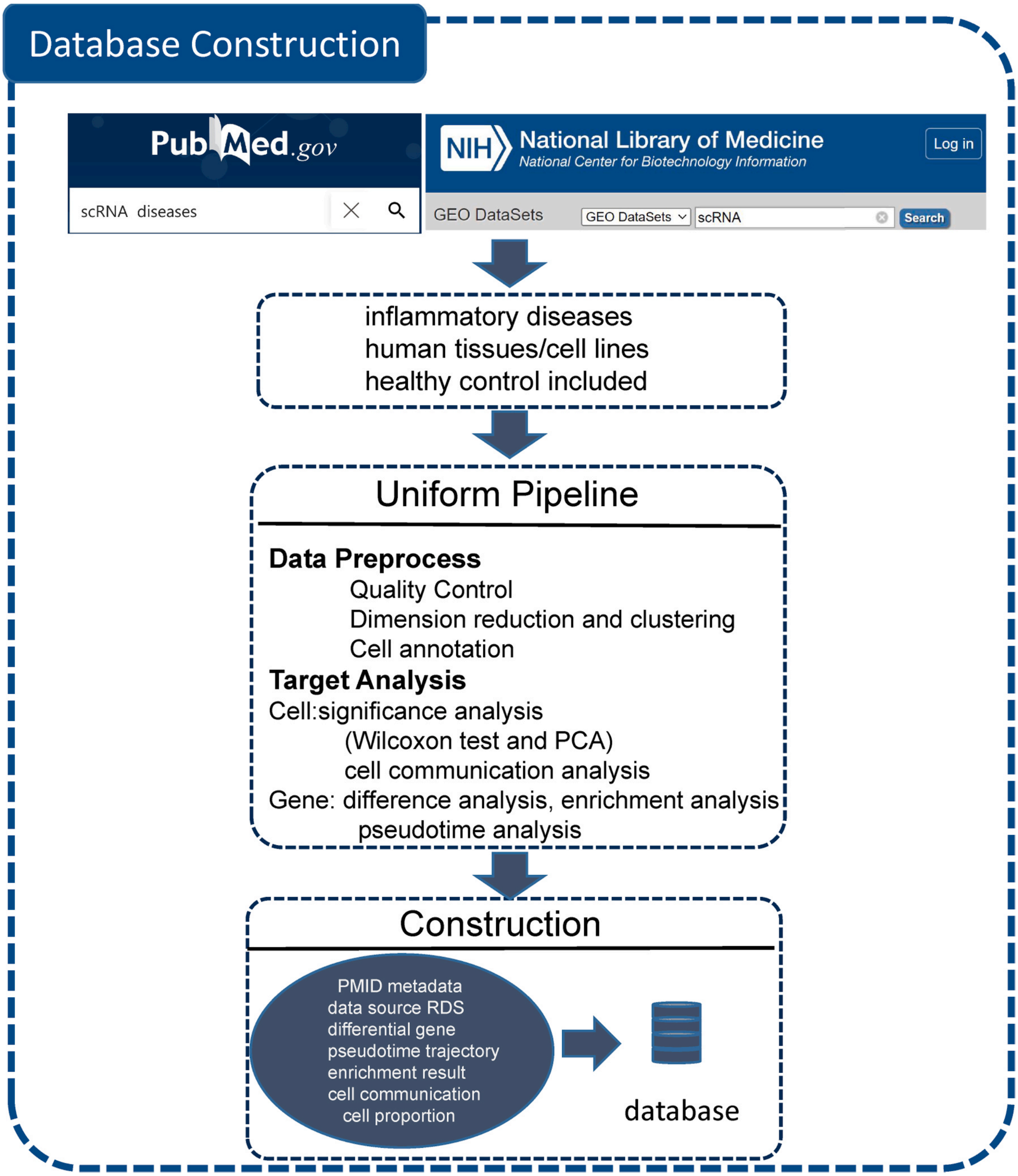


Fig. 10. Database construction. The data obtained from PubMed and GEO were filtered and put into a uniform pipeline. The results of target analysis were stored in the database.

Table 2
Main tools and algorithms.

Tools	Algorithms	Abstract	Reference
Seurat/ FactoMineR	PCA	Dimensionality reduction to preserve core data information	[31,35]
Seurat	UMAP		
	Louvain	Clustering based on core data information	
SingleR		Cell annotation	[32]
iTalk		Cellular communication analysis	[30]
Monocle	Reversed Graph Embedding	Constructing Pseudo-Time trajectories	[26]
Goplot		Visualization of GO enrichment information	[36]
clusterProfiler		Enrichment analysis	[37]

Declaration of Competing Interest

The authors (Chaochao Wang, Ting Huyan, Wuli Guo, Qi Shu, Qi Li, Jianyu Shi) have declared no conflict of interest.

Acknowledgements

This work was supported by the Natural Science Basic Research Program of Shaanxi [grant number 2022JM-117 and 2023JC-YB-190], the National Nature Science Foundation of China [62372375], the Shaanxi Province Key R&D Program [2023-YBSF-114], and the CAAI-Huawei MindSpore Open Fund [CAAIXSJLJJ-2022–035A].

Code Availability

The code used in the analysis process can be accessed at https://www.scidb.cn/en/anonymous/blVWN0Ji and www.ntcdb.org.cn.

Author Contributions

Chaochao Wang and Ting Huyan analyzed the data and drafted the manuscript, Wuli Guo and Qi Shu collected and visualized the data, Qi Li formulated the overarching research goals and aims, and edited the manuscript, Yujian Shi provided overall project guidance and financial support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.057.

References

[1] Tian Y, Cheng C, Wei Y, Yang F, Li G. The role of exosomes in inflammatory diseases and tumor-related inflammation. Cells 2022;11.

[2] Ferrero-Miliani L, Nielsen OH, Andersen PS, Girardin SE. Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1beta generation. Clin Exp Immunol 2007;147:227–35.

[3] Hall J.E., Guyton A.C.. Guyton and Hall textbook of medical physiology. Philadelphia, Pa.: Saunders/Elsevier, 2011.

[4] Granger D.N., Senchenkova E. Inflammation and the Microcirculation. San Rafael (CA), 2010.

[5] Hannoodee S., Nasuruddin D.N. Acute Inflammatory Response. StatPearls. Treasure Island (FL) ineligible companies. Disclosure: Dian Nasuruddin declares no relevant financial relationships with ineligible companies., 2024.

[6] Pahwa R., Goyal A., Jialal I. Chronic Inflammation. StatPearls. Treasure Island (FL) ineligible companies. Disclosure: Amandeep Goyal declares no relevant financial relationships with ineligible companies. Disclosure: Ishwarlal Jialal declares no relevant financial relationships with ineligible companies., 2024.

[7] Lin Y, Qiu T, Wei G, Que Y, Wang W, Kong Y, et al. Role of histone post-translational modifications in inflammatory diseases. Front Immunol 2022;13:852272.

[8] Caughey GH. Mast cell tryptases and chymases in inflammation and host defense. Immunol Rev 2007;217:141–54.

[9] Libby P. Inflammation in atherosclerosis. Nature 2002;420:868–74.

[10] Coussens LM, Werb Z. Inflammation and cancer. Nature 2002;420:860–7.

[11] Greten FR, Grivennikov SI. Inflammation and cancer: triggers, mechanisms, and consequences. Immunity 2019;51:27–41.

[12] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13.

[13] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82.

[14] Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, et al. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Res 2022;50:D596–602.

[15] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8.

[16] Wang Z, Feng X, Li SC. SCDevDB: a database for insights into single-cell gene expression profiles during human developmental processes. Front Genet 2019;10:903.

[17] Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res 2021;49:D1420–30.

[18] Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res 2019;47:D900–8.

[19] Lei Y, Tang R, Xu J, Wang W, Zhang B, Liu J, et al. Applications of single-cell sequencing in cancer research: progress and perspectives. J Hematol Oncol 2021;14:91.

[20] Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, McKechnie JL, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat Med 2020;26:1070–6.

[21] Kist M, Vucic D. Cell death pathways: intricate connections and disease implications. EMBO J 2021;40:e106700.

[22] Wang M, Song WM, Ming C, Wang Q, Zhou X, Xu P, et al. Guidelines for bioinformatics of single-cell sequencing data analysis in Alzheimer's disease: review, recommendation, implementation and application. Mol Neurodegener 2022;17:17.

[23] Manu Pereira MD, Ropero P, Loureiro C. Vives Corrons JL. Low affinity hemoglobinopathy (Hb Vigo) due to a new mutation of beta globin gene (c200 A>T; Lys>Ile). A cause of rare anemia misdiagnosis. Am J Hematol 2017;92:E38–40.

[24] Chen HH, Shaw DM, Petty LE, Graff M, Bohlender RJ, Polikowsky HG, et al. Host genetic effects in pneumonia. Am J Hum Genet 2021;108:194–201.

[25] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:1–14.

[26] Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017;14:309–15.

[27] Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res 2021;49:D1413–9.

[28] Balzer MS, Ma Z, Zhou J, Abedini A, Susztak K. How to get started with single cell RNA sequencing data analysis. J Am Soc Nephrol 2021;32:1279–92.

[29] Ben Salem K, Ben Abdelaziz A. Principal component analysis (PCA). Tunis Med 2021;99:383–9.

[30] Wang C, Huyan T, Zhou X, Zhang X, Duan S, Gao S, et al. Development of single-cell transcriptomics and its application in COVID-19. Viruses 2022;14.

[31] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, et al. Comprehensive integration of single-cell data. Cell 2019;177. 1888-902 e21.

[32] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20:163–72.

[33] Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genom 2013;14:632.

[34] Wang Y, RW, SZ, SS, CJ, GH, et al. iTALK: an R Package to Characterize and Illustrate Intercellular Communication. bioRxiv 2019.

[35] Jedroszka D, Orzechowska M, Hamouz R, Gorniak K, Bednarek AK. Markers of epithelial-to-mesenchymal transition reflect tumor biology according to patient age and Gleason score in prostate cancer. PLoS One 2017;12:e0188842.

[36] Walter W, Sanchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics 2015;31:2912–4.

[37] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.