Computational and Structural Biotechnology Journal 23 (2024) 2746–2753

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Database article

scBlood: A comprehensive single-cell accessible chromatin database of blood cells

Yu Zhao [a,b,1], Zheng-Min Yu [b,1], Ting Cui [a,c,1], Li-Dong Li [b,1], Yan-Yu Li [h,1], Feng-Cui Qian [f], Li-Wei Zhou [i], Ye Li [c], Qiao-Li Fang [b], Xue-Mei Huang [b], Qin-Yi Zhang [c], Fu-Hong Cai [b], Fu-Juan Dong [b], De-Si Shang [f,*], Chun-Quan Li [a,b,c,d,e,f,g,**], Qiu-Yu Wang [b,c,d,e,***]

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

ARTICLE INFO

Keywords:
ScATAC-seq
Single cell
Blood cells
Transcriptional regulation
Chromatin
Accessibility

ABSTRACT

The advent of single cell transposase-accessible chromatin sequencing (scATAC-seq) technology enables us to explore the genomic characteristics and chromatin accessibility of blood cells at the single-cell level. To fully make sense of the roles and regulatory complexities of blood cells, it is critical to collect and analyze these rapidly accumulating scATAC-seq datasets at a system level. Here, we present scBlood (https://bio.liclab.net/scBlood/), a comprehensive single-cell accessible chromatin database of blood cells. The current version of scBlood catalogs 770,907 blood cells and 452,247 non-blood cells from ~400 high-quality scATAC-seq samples covering 30 tissues and 21 disease types. All data hosted on scBlood have undergone preprocessing from raw fastq files and multiple standards of quality control. Furthermore, we conducted comprehensive downstream analyses, including multi-sample integration analysis, cell clustering and annotation, differential chromatin accessibility analysis, functional enrichment analysis, co-accessibility analysis, gene activity score calculation, and transcription factor (TF) enrichment analysis. In summary, scBlood provides a user-friendly interface for searching, browsing, analyzing, visualizing, and downloading scATAC-seq data of interest. This platform facilitates insights into the functions and regulatory mechanisms of blood cells, as well as their involvement in blood-related diseases.

1. Introduction

Blood cells derived from hematopoietic stem cells (HSCs) play a critical role in defending against foreign pathogens, maintaining vital functions, and stabilizing the immune system [1–4]. The emergence of scATAC-seq provides a unique opportunity to explore the heterogeneity

* Corresponding author.
** Corresponding author at: The First Affiliated Hospital & MOE Key Lab of Rare Pediatric Diseases, Hengyang Medical School, University of South China, Hengyang, Hunan 421001, China.
*** Corresponding author at: School of Computer, University of South China, Hengyang, Hunan 421001, China.
E-mail addresses: sds_46@163.com (D.-S. Shang), lcqbio@163.com (C.-Q. Li), wangqiuyu900490@163.com (Q.-Y. Wang).
[1] The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

https://doi.org/10.1016/j.csbj.2024.06.015
Received 16 April 2024; Received in revised form 17 June 2024; Accepted 18 June 2024
Available online 20 June 2024

of blood cells, offering detailed insights into the genomic characteristics and chromatin accessibility of individual blood cells [5–9]. At present, scATAC-seq has been widely used to reveal extensive and complex regulation in blood cells, aiding to characterize the functional characteristics of various cell subpopulations at the single cell level [10–15]. Satpathy et al. analyzed the chromatin profiles of individual cells in human blood, where they identified cell type-specific cis-and trans-regulatory elements, delineated disease-associated enhancer activity, and reconstructed cell differentiation trajectories [16]. In another study on acute myeloid leukemia (AML), Corces et al. investigated chromatin accessibility and revealed the unique regulatory evolution in cancer cells and highlighting the pivotal role of HOX factors in the characteristics of preleukemic hematopoietic stem cells [17]. Evidently, scATAC-seq holds enormous potential to reveal cellular regulatory mechanisms, identify disease biomarkers, and discover new therapeutic approaches [18–23]. The advancement of single-cell technologies has facilitated the rapid accumulation of scATAC-seq data related to blood from various sources [24]. To fully exploit the wealth of information in these datasets, efficient collection, integration, and analysis are urgently needed, which will shed new insights for biologists and researchers in the field of blood-related diseases.

Currently, several single-cell databases focused on blood cells have been developed, including BloodSpot [25–27] and ABC portal [28]. The ABC portal is a repository comprising 198 single-cell transcriptomic datasets of blood cells. The original BloodSpot database integrated gene expression data of haematopoietic cells from bulk datasets [26]. The updated BloodSpot provides several single-cell RNA-seq datasets and single-cell proteome datasets [25,27]. Additionally, some single-cell databases also provide limited blood cell scATAC-seq data, such as CATLAS [29,30], HTCA [31], AgeAnno [32] and scEnhancer [33]. However, these databases mainly focus on broadly single-cell datasets and fail to cover comprehensive blood-related scATAC-seq data (Table S1). The explosive growth of blood cell scATAC-seq data has made it possible to thoroughly characterize blood cells based on large-scale analysis. The scATAC-seq technique has been widely used to calculate gene activity and TF enrichment, which contributes to the understanding of blood cell heterogeneity and molecular regulatory mechanisms [19,34–38]. Obviously, the development of a scATAC-seq database specifically targeting blood cells is becoming increasingly urgent.

Here, we developed scBlood (https://bio.liclab.net/scBlood/), a comprehensive single-cell accessible chromatin database of blood cells. This database covers changes in blood cells in various disease states and tissues of human and mouse. The current version of scBlood documents a total of 1223,154 cells from 377 high-quality scATAC-seq samples (Table S2). Among these, 292 blood cell samples encompass 15 disease types, such as chronic lymphocytic leukemia (CLL), ankylosing spondylitis (AS), clear cell renal cell carcinoma (ccRCC), and MSSA bacteremia, etc. The other 85 non-blood samples from 24 tissues such as brain, lung, and kidney were used for comparative analysis. To ensure data quality and consistency, after preprocessing to obtain the raw peak matrixs, we performed multiple standard quality control on all datasets using unified software parameters. Furthermore, we also performed comprehensive downstream analyses, including multi-sample integration analysis, cell clustering and annotation, differential chromatin accessibility analysis, functional enrichment analysis, co-accessibility analysis, gene activity score calculation, and TF enrichment analysis. scBlood provides a user-friendly interface supporting interactive exploration and visualization of blood cell scATAC-seq data. We anticipate that scBlood will become an important resource platform for hematological research, advancing the study of blood cell heterogeneity and function, and promoting new discoveries and therapeutic strategies.

2. Materials and methods

2.1. Data collection and pre-processing

We searched single-cell scATAC-seq datasets from NCBI GEO/SRA [39,40] and 10X Genomics website with keywords '(single-cell ATAC-sequencing) OR (scATAC-seq) OR (single cell ATAC sequencing) OR (single-cell ATAC-seq) or (single cell ATAC-seq)' (Retrieved on June 2023). Through proofreading the sample descriptions of $>$ 8000 entries, we then manually curated blood cell-related samples, such as blood tissue, hematopoietic stem cells, etc. Meanwhile, we also collected various non-blood samples for comparative analysis, such as brain, lung, and kidney. Overall, 292 blood cells samples and 85 non-blood cells samples originating from various tissues were retained for subsequent quality control and downstream analysis. We further manually compiled the detailed information of each sample from databases or original research, including publications, species, cell types, tissues, disease states, cell numbers, and sequencing platforms.

The raw FASTQ files of the retained samples were downloaded from the SRA database using sratoolkit (v.3.0.0). Considering that Cell Ranger ATAC is a comprehensive tool, specifically designed to process data generated by the 10X Genomics platform, we employed Cell Ranger ATAC (v.2.1.0) to process FASTQ files from the 10X Genomics platform to construct peak-by-cell matrices [41]. For data from other sequencing platforms, such as HyDrop-ATAC-seq [42], sciATAC-seq [43], and snATAC-seq, we used the scATAC-pro (v.1.5.0) [41] to construct peak-by-cell matrices. Subsequently, we performed unified quality control and downstream analyses based on the raw peak-by-cell matrices obtained from these processes. The quality assessment metrics include the fraction of fragments in the mitochondrial genome, the fraction of fragments in peaks, as well as the fraction of fragments located within transcription start sites (TSSs), enhancers, and promoters. Based on these metrics, we filtered out low-quality cells with unique fragments counts $<$ 1000 or $>$ 50,000, with fractions of fragments in peaks less than 15 % and fractions of fragments in promoters less than 20 %. After filtering, a total of 770,907 blood cells were retained, covering 6 tissues, including PBMC, blood, bone marrow, spleen, lymph node, coronary artery. For non-blood cells samples, a total of 452,247 cells were retained, covering 24 tissues, including cortex, brain, lung, kidney, embryonic, breast, pancreatic, adrenal gland, cerebellum, cerebrum, heart, intestine, gut endoderm, forelimb, retina, ventricle, liver, eye, muscle, placenta, stomach, thymus, frontal cortex, islet (Table S2).

2.2. Normalization, dimension reduction, and cell clustering

We employed the Term Frequency-Inverse Document Frequency (TF-IDF) to normalize the peak-by-cell matrix [34,43,44]. Subsequently, Principal Component Analysis (PCA) was then performed on the TF-IDF matrix, generating a low-dimensional representation of the scATAC-seq data. Following this, we selected the first 30 principal components as input for the Seurat object. The clusters were then identified using the Seurat's (v.3.2.3) SNN graph cluster 'FindClusters' function, with a default resolution set to 0.5 [45]. Finally, the Uniform Manifold Approximation and Projection (UMAP) [46] were utilized to visualize the clustering results.

2.3. Differential chromatin accessibility and functional enrichment analysis

Peaks with differential accessibility across different cell subpopulations may often be potentially cell-specific gene regulatory elements [11,34]. Thus, we employed the Wilcoxon test to calculate the significance between each cluster and the rest of the clusters. The differential accessibility regions (DARs) were further determined on the basis of the false discovery rate (FDR) $<$ =0.05 [41]. Then, each DAR was annotated with its nearest gene. To characterize the functions of

different cell clusters, we performed Gene Ontology (GO) and KEGG pathway enrichment analyses with the nearest genes of DARs using clusterProfile package (v.4.6.2) [47–49]. Finally, we visualized the enrichment analysis results with histogram and bubble diagram, and pathways with significance P value < 0.05 were provided on the webpage of scBlood.

2.4. Co-accessibility and gene activity score

To deepen the understanding of transcriptional regulatory mechanisms and reveal potential interactions between peaks, we used Cicero (v.1.4.4) to predict co-accessibility interactions [36,50] by setting the co-accessibility cutoff value > 0.25. We then assigned the nearest genes to co-accessible chromatin regions. Based on these co-accessibility interactions, gene activity scores in each cell were then calculated using the 'build_gene_activity_matrix' function. Finally, we applied the 'normalize_gene_activities' function to normalize the gene activity scores to account for different sequencing depths among cells, thereby eliminating biases caused by these differences and enabling a direct comparison of gene activity scores across different cells.

2.5. TF/motif enrichment analysis

TFs play a crucial role in the regulation of gene expression by recognizing and binding to specific DNA sequences, thereby initiating or inhibiting the transcription process of genes [51]. We utilized the chromVAR tool to quantify the activity of transcription factors (v.1.20.2) [37]. Specifically, chromVAR uses a peak-by-cell matrix derived from the scATAC-seq dataset and motif position weight matrices (PWMs) as input. We performed motif matching within these peaks using the 'matchMotifs' function from the motifmatch package, combined with the PWMs ('human_pwms_v2′ and 'mouse_pwms_v2′). chromVAR then computed the raw accessibility deviation for each motif in each cell. Additionally, we calculated GC content of the peaks based on 'BSgenome.Hsapiens.UCSC.hg38′ and 'BSgenome.Mmusculus.UCSC.mm10′. To correct biases caused by external factors such as sequencing techniques, chromVAR constructed a background peak set for each motif, which consists of an equal number of peaks matched for GC content and average accessibility. The raw accessibility deviations of the background sets were used to compute a bias corrected deviation and deviation z-score for each TF/motif, reflecting their variations within and between different cell types. Finally, differential TFs/motifs between one cell cluster and the rest were identified using a two-sample Wilcoxon test with a P value < 0.05. The TF/motif z -score and differential TFs / motifs are visualized using scatter plot and heatmap, respectively.

2.6. Cell type annotation

The peak-by-cell matrix obtained from scATAC-seq data is highly sparse, making it difficult for cell types annotation. Therefore, we first converted the peak-by-cell matrix into a gene activity matrix using Cicero (v.1.4.4). We then used SingleR (v.2.2.0) for cell type annotation [52], leveraging its comprehensive built-in reference datasets. Additionally, we provided an alternative cell type annotation method, SCINA, which is a semi-supervised cell type labeling algorithm [53]. Specifically, we first compiled a list of reference marker genes related to the intended cell type from CellMarkers [54]. Cell type labels were then assigned to individual cells based on the expression levels of these marker genes. Cells that do not fit well with known markers were annotated as 'Unknown', thus providing an opportunity to discover new cell types or subtypes.

2.7. Integration of multiple scATAC-seq samples

To integrate multiple samples in the scATAC-seq dataset, we initially merged the peaks from each sample using the mergePeaks module in scATAC-pro, specifying a minimum distance interval of 500 bp. Subsequently, we reconstructed the peak-by-cell matrix for each sample based on the merged peaks and selected the top 5000 variable features along with the top 30 PCA dimensions for downstream integration using Harmony (v.1.0.0) [55]. After adding harmony embeddings, RunUMAP was used to generate a UMAP dimensionality reduction from the harmony reduction. Finally, we clustered the cells using FindNeighbors (with reduction set to 'harmony' and dims set to 1:30), followed by FindClusters with a resolution parameter of 0.6.

2.8. Database implementation

The current version of the scBlood database operates on a server based on Centos 7.7.1908. We use the IntelliJ IDEA 2021.3 integrated development environment to deploy and release the project to the remote server via a Dockerfile. Prior to this, Docker 19.03.5 (https://www.docker.com/) was installed on the remote server, and the backend API and frontend pages were reverse proxied through Nginx 1.22.0 (https://nginx.org/). This project is implemented with a technology architecture that separates the frontend and the backend. On the backend, business logic processing is constructed with the Spring Boot 3.0.5 framework (https://spring.io/projects/spring-boot) based on Java 17.0.1 (https://dev.java/). For database management, MyBatis 3.0.2 (https://blog.mybatis.org/) serves as the ORM framework connecting to the MySQL structured database (https://www.mysql.com/), which is set up through a Docker container version mysql:8.0.32. To enhance system performance, we introduced Redis 6.2.11-alpine Docker (https://redis.io/) container version as a caching mechanism. The frontend is constructed using the Vue 3.2.4 framework (https://vuejs.org/) in a Node.js v16.13.0 environment (https://nodejs.org/en). For the development of the frontend pages, Axios 0.21.4 (https://www.axiosdev.com.au/) is used for data interaction with the backend API, while Element-UI (element-plus 2.2.0) (https://element-plus.gitee.io/en-US/) and Bootstrap v5.1.3 (https://getbootstrap.com/) for page layout and style design. Font Awesome 6.1.1 (https://fontawesome.com/) provides icon style support, while Echarts 5.3.1 (https://echarts.apache.org/en/index.html), Plotly 2.23.0 (https://plotly.com/), and CanvasXpress 38.4.1 (https://canvasxpress.org/) were used for graph visualization. To ensure the best browsing experience, it is recommended that users access the website with modern web browsers that support the HTML5 standard, such as Firefox, Google Chrome, and Edge.

3. Result

3.1. Overview of scBlood

The current version of scBlood encompasses 292 blood cells scATAC-seq samples and 85 non-blood samples sourced from 30 tissues and 21 disease types of human and mouse (Fig. 1). These blood cells are mainly derived from bone marrow, peripheral blood mononuclear cells (PBMCs), umbilical cord blood, lymph nodes, hematopoietic stem cells (HSCs), and the spleen. We obtained these samples from four sequencing platforms, including 10X Genomics, HyDrop-ATAC-seq, snATAC-seq and sciATAC-seq. After preprocessing to obtain the raw peak-by-cell matrices of scATAC-seq data, we used a uniform pipeline and software parameters to process and analyze the data, including quality control and filtering, multi-sample integration analysis, cell clustering and annotation, differential chromatin accessibility analysis, functional enrichment analysis, co-accessibility analysis, gene activity score calculation, and TF enrichment analysis. scBlood will be a user-friendly resource platform for exploring blood-related scATAC-seq data. It displays all analysis results and allows public access to seven functional areas: 'Home', 'Data-Browse', 'Search', 'Analysis', 'Download', 'Contact Us', and 'Help', via the navigation bar at the top of the page.

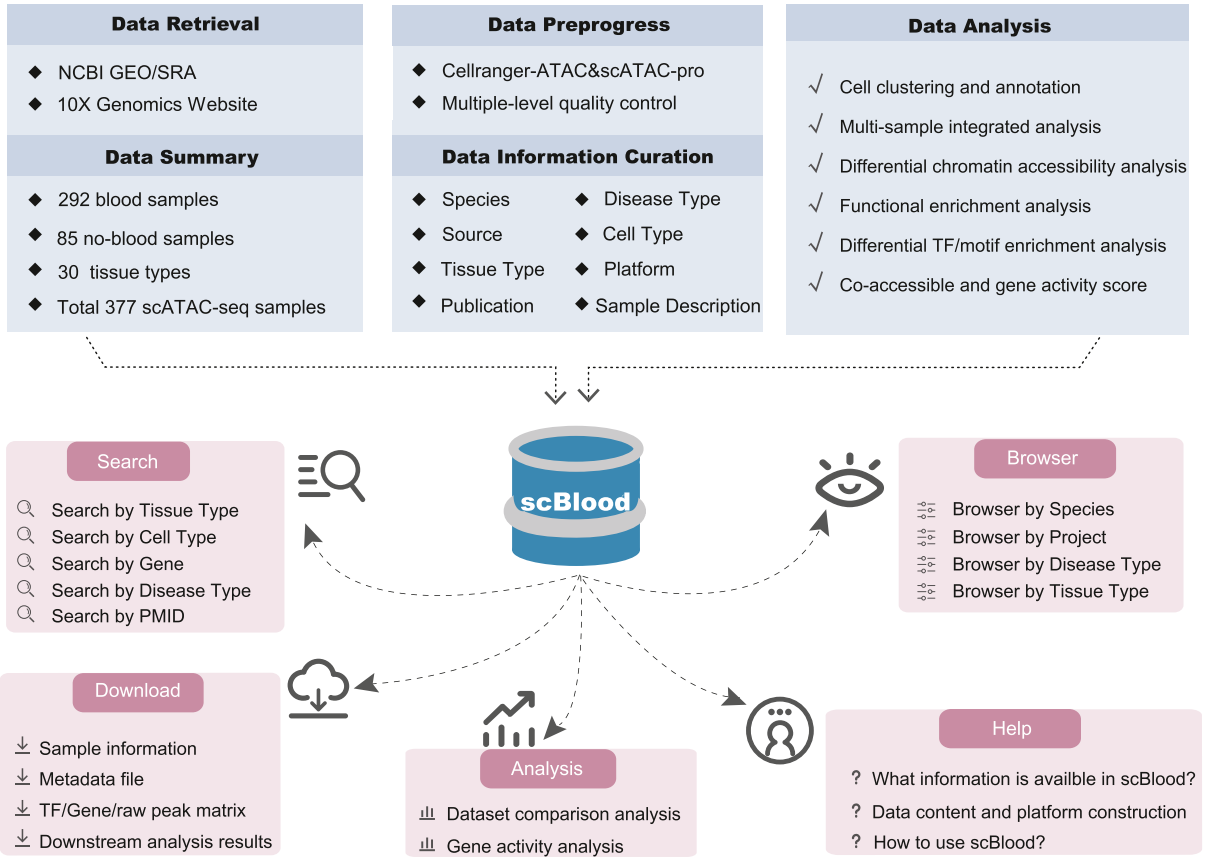


Fig. 1. Construction of scBlood. scBlood is a comprehensive single-cell accessible chromatin database of blood cells. After preprocessing to obtain the raw peak matrixs, all datasets in scBlood were used unified processes and software parameters for multiple standard quality control and comprehensive downstream analysis. scBlood also provided user-friendly browsing, search, analysis, downloading and visualization functions.

3.2. Search interface for retrieving scATAC-seq data

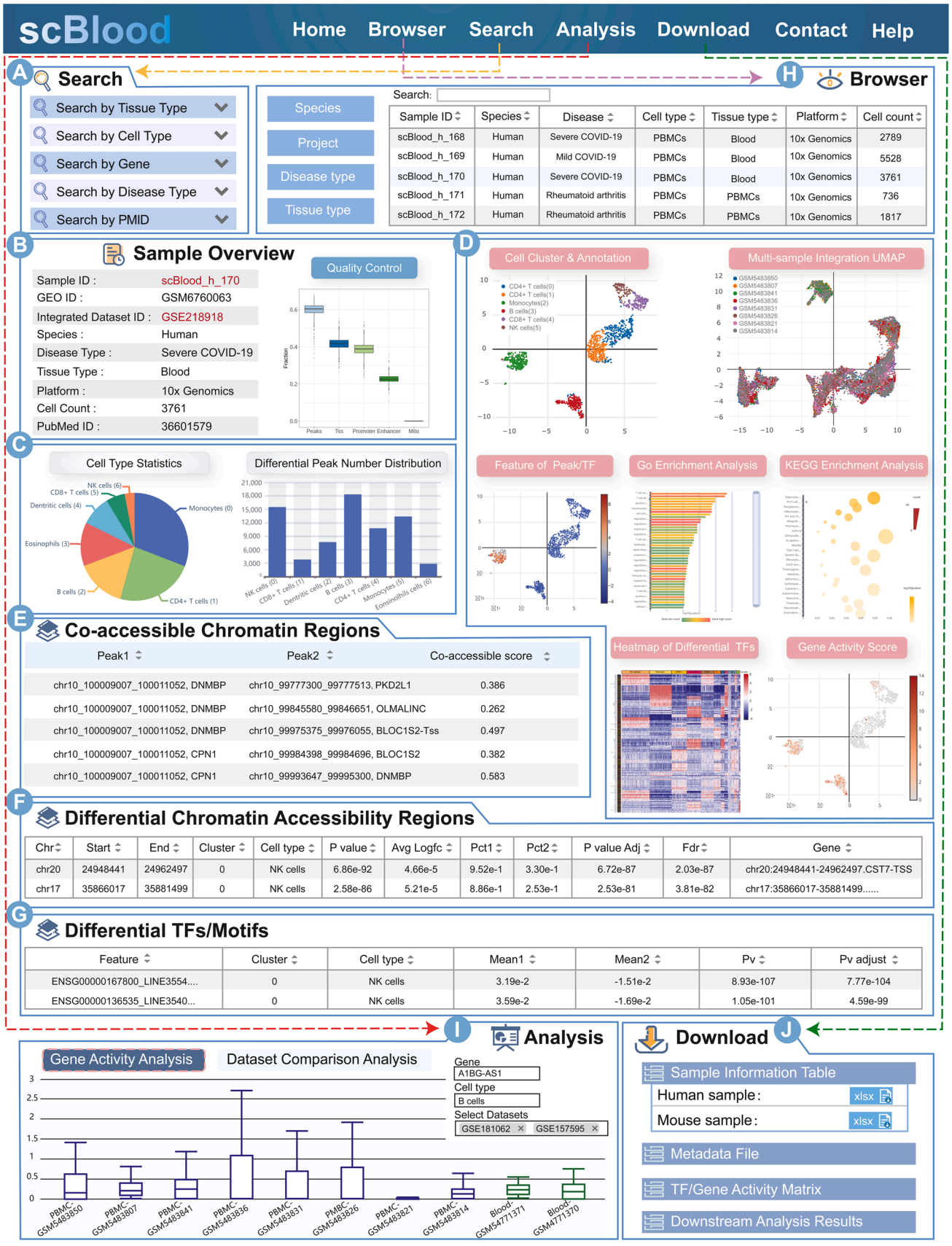
scBlood provides five user-friendly search interfaces for exploring scATAC-seq data: 'Search by Tissue Type', 'Search by Cell Type', 'Search by Disease Type', 'Search by Gene' and 'Search by PMID' (Fig. 2A). In the tissue type-based query, users can search for scATAC-seq data by specifying the species and tissues of interest to retrieve scATAC-seq samples of interest. The search results are presented in a summary table featuring sample ID, sample description, integrated dataset ID, species, disease type, cell type, tissue type, cell number, type and PubMed ID. By clicking on the 'Sample ID', users can access detailed information and analysis results for a specific sample (Fig. 2B). Similarly, users can click on the 'Integrated Dataset ID' they are interested in to access the details page. This page provides more comprehensive information about the selected integrated dataset, including integrated dataset ID, species, source, disease type, tissue type, and PubMed ID, among others. Within this integrated dataset, the proportions of various cell types are presented in the form of a pie chart, while the number of differential chromatin accessibility regions in each cell cluster is displayed via bar charts (Fig. 2C). A sample information table displays relevant details for all samples within the integrated dataset. The analysis results section includes scatter plots for cell clustering and annotation UMAP on the left side, and five modules on the right: differential peak, gene activity score, TFs enrichment, GO and KEGG pathway enrichment. Differential peak, gene activity, and TFs enrichment are visualized through scatter plots, enabling users to explore chromatin regions' accessible across different cell cluster. Additionally, a heatmap visualization is available for TF enrichment. The GO and KEGG pathway enrichment modules display enrichment results using bar charts and bubble charts, respectively, with a significance threshold of P-value$<$ 0.01 (Fig. 2D). Towards the bottom of the details page, tables for differential chromatin accessibility regions (Fig. 2F), co-accessible chromatin regions (Fig. 2E), and differential TFs/motifs are provided (Fig. 2G). In the cell type-based query, users can select or input preferred species and cell type to locate relevant scATAC-seq data, returning the sample table containing the input cell type. In the disease type-based query, users can select or input a species and disease type of interest, scBlood will return a table containing the input disease type. By clicking on the 'Sample ID', users can explore scATAC-seq data for information on specific disease types. In the gene-based query, users can select or input a species and one or more genes of interest. scBlood will then return a table featuring sample ID, sample description, integrated dataset ID, species, disease type, cell type, gene, average activity, tissue type, cell number and PubMed ID, allowing users to select multiple samples of interest to compare and view scatter plots depicting the selected gene activity. In PMID-based queries, users can select a species and a PubMed ID of interest to locate specific datasets more quickly.

3.2.1. A user-friendly interface for browsing scATAC-seq data

The 'Data-Browse' page is structured as an interactive and alpha-numerically sortable table that allows users to quickly browse scATAC-seq data and customize filters including 'Species', 'Project', 'Disease type', 'Tissue Type' (Fig. 2H). Users can use the 'Show entries' drop-down menu to get different number of records per page. To further view the details of a given sample or integrated dataset, users can click on 'Sample ID' and 'Integrated Dataset ID' respectively.

3.2.2. Online analysis tool

We implemented two analysis functions, including 'Dataset comparative analysis' and 'Gene activity analysis' (Fig. 2I). Specifically,



(caption on next page)

Fig. 2. Main functions and usage of scBlood. (A) Five search modes are provided, including 'Search by Tissue Type', 'Search by Cell Type', 'Search by Gene', 'Search by Disease Type' and 'Search by PMID'. (B) Detailed information and quality control of each sample. (C) Cell type statistics and differential peak number distribution for each sample. (D) Visualization of downstream analysis results, including cell clustering and annotation UMAP, multi-sample integration UMAP, scatter plot of accessibility of differential chromatin regions, KEGG/GO term enrichment bubble charts and bar charts, TFs/motifs enrichment scatter plot and heatmap. (E) Table of co-accessible chromatin regions. (F) Table of differential chromatin accessibility regions. (G) Table of differential TFs/motifs. (H) Browse page of scBlood. (I) Two analysis functions are provided, including 'Dataset comparative analysis' and 'Gene activity analysis', the box plot represents the results of gene activity analysis. (J) scBlood allows users to download sample information table, metadata file, raw peak-by-cell matrix, TF activity matrix, and gene activity matrix, tables and visualization files of all downstream analysis results.

'Dataset comparative analysis' aims to provide comparisons between blood cell datasets as well as between blood cells and non-blood cells datasets from tissues such as the brain, lung, and kidney. Firstly, users need to specify the species and select two datasets for analysis. Clicking 'Start Analysis' produces a dataset comparative analysis result that allows the user to make a comprehensive comparison of the selected datasets. Through 'Gene activity analysis', users need to select or input a gene and choose two or three datasets of interest for analysis. Subsequently, clicking 'Start analysis' will return a box plot of specific gene activity across all samples in these integrated datasets, enabling users to intuitively compare the activity scores of this gene among different tissues and disease types.

3.2.3. Data download and help interface

The 'Download' page allows users to download sample information sheets for human and mouse. scBlood also provided download support for metadata file, raw peak-by-cell matrix, TF activity matrix, and gene activity matrix for each dataset. Additionally, table files of all downstream analysis results, such as differential chromatin accessibility analysis, TF enrichment analysis, functional enrichment analysis are also available for download (Fig. 2J). For further guidance, the 'Help' page provides a detailed tutorial for users.

3.2.4. Case study

Rheumatoid arthritis (RA) is a complex autoimmune and inflammatory disease characterized by systemic inflammation, and its disease progression is closely associated with various immune cell types [56, 57]. To delve deeper into the roles of these cells in RA, we used the "Search by Disease Type" function to query scATAC-seq samples related to RA. We selected species as "Human" and entered "Rheumatoid arthritis" in the "Disease Type" input box. Upon clicking the "Start search" button, the system returned relevant scATAC-seq samples "scBlood_h_171″ (Figs. S1A and S1B). We further explored the sample detail page by clicking on the sample ID. Our analysis result showed that the cells within the scBlood_h_171 sample were clustered into four distinct cell types, with cluster2 annotated as monocytes (Fig. S1C). We performed TF enrichment analysis on these monocytes and discovered a significant enrichment of key factors, including CEBPA, CEBPB, CEBPD, CEBPE, CEBPG, SPIB and SPI1 (Fig. S1E and S1G). These factors play crucial regulatory roles in the development of monocytes [58–60]. We then examined the differential chromatin accessibility regions in monocytes, where chr2:112836341–112837268 (P-value= 1.76e-17) and chr1:154404846–154405747 (P-value= 2.39e-09) showed significant differences (Fig. S1F). Detailed analysis revealed that the nearest genes of these regions were IL1B and IL6R. Studies have indicated that IL1B, linked to the function of pro-inflammatory monocytes, is a key mediator in the pathogenesis of RA [57]. Additionally, the abnormal expression of IL6 and its receptor IL6R is closely related to the pathogenic mechanism of RA, and the humanized monoclonal antibody Tocilizumab, targeting IL6R-α, has potential therapeutic efficacy in the treatment of RA and other autoimmune diseases [61–63]. Consistent with existing research, results of gene activity analysis showed that these genes have higher activity in monocytes (Fig. S1D). Furthermore, among these nearest genes of differential chromatin accessibility regions of monocytes, we also discovered SRGN and TNFAIP3, etc. These genes have also been shown to be associated with RA. These results not only validate the reliability of scBlood but also emphasizes the significance of these genes in RA pathogenesis.

4. Discussion

scBlood offers a vital resource for exploring transcriptional regulation mechanisms in blood cells, addressing the limitations present in existing scATAC-seq databases related to blood cells. Currently, most single-cell databases focus primarily on transcriptomic data, lacking specific chromatin accessibility data for blood cells. While some databases contain partial scATAC-seq data about blood cells, they are not enough to broadly cover more blood cells with different conditions. In contrast, scBlood provides a wealth of chromatin accessibility data for blood cells, including 292 scATAC-seq samples of blood cells from six tissues. Additionally, we processed 85 non-blood scATAC-seq samples from 24 tissues such as brain, kidney, and lung for comparative analysis with blood cells samples. All data in scBlood underwent uniform multiple standards of quality control. We performed a comprehensive analysis of scATAC-seq data in scBlood from multiple perspectives. Differential chromatin accessibility analysis helped us identify peak regions in certain cell clusters that significantly differ from others, potentially representing cell-specific gene regulatory elements. Functional enrichment analysis revealed the functional characteristics of different cell populations, uncovering pathways and biological processes associated with various diseases. Additionally, by co-accessibility analysis, we uncovered interactions between chromatin regions and calculated gene activity scores, aiding in a deeper understanding of the transcriptional regulation mechanisms and revealing potential interactions. Through TF enrichment analysis, we identified enriched TFs and motifs in specific cell populations, providing insights for further research into cell-specific transcriptional regulation mechanisms. The integration of these functionalities provides users with a user-friendly interface to search, browse, analyze, visualize, and download scATAC-seq data of interest, facilitating a better understanding of the functionality and regulatory mechanisms of blood cells. In conclusion, as a comprehensive single-cell accessible chromatin database, scBlood provides researchers rich scATAC-seq datasets along with various functional analysis tools. We believe scBlood will significantly support and advance progress and discoveries in the field of hematology and single-cell epigenetic research.

With the rapid development of single cell techniques, we will continue to maintain and update the scBlood database. Future updates of scBlood will mainly focus on the following three directions. First, we will collect more high-quality scATAC-seq datasets from newly published studies to enrich the content of scBlood. Secondly, we plan to add more practical analysis tools to improve the user experience in data exploration and result interpretation. Finally, we will explore integrating more single-cell omics data into scBlood to provide a more comprehensive biological perspective. We believe that through continuous improvement, scBlood will be able to better adapt to the development of the scientific community and provide researchers with richer and more valuable resources.

Funding

This work was supported by National Natural Science Foundation of China [62171166, 62301246, 62272212]; Natural Science Foundation of Hunan Province [2023JJ30536]; Natural Science Foundation of

Heilongjiang Province [LH2021F044]; Research Foundation of the First Affiliated Hospital of University of South China for Advanced Talents [20210002–1005 USCAT-2021–01]; Clinical Research 4310 Program of the University of South China [20224310NHYCG05]; Research Foundation of Education Bureau of Hunan Province [22C0210].

CRediT authorship contribution statement

Ye Li: Conceptualization. **Chun-Quan Li:** Supervision. **Li-Dong Li:** Data curation. **Qiuyu Wang:** Writing – review & editing. **Yan-Yu Li:** Supervision. **Feng-Cui Qian:** Methodology, Conceptualization. **Li-Wei Zhou:** Investigation. **Qin-Yi Zhang:** Visualization. **Fu-Hong Cai:** Validation. **Yu Zhao:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Fu-Juan Dong:** Data curation. **Zheng-Min Yu:** Software. **De-Si Shang:** Writing – review & editing. **Ting Cui:** Visualization, Data curation. **Qiao-Li Fang:** Visualization. **Xue-Mei Huang:** Data curation.

Declaration of Competing Interest

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

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.06.015.

References

[1] You M, Chen L, Zhang D, Zhao P, Chen Z, Qin EQ, Gao Y, Davis MM, Yang P. Single-cell epigenomic landscape of peripheral immune cells reveals establishment of trained immunity in individuals convalescing from COVID-19. Nat Cell Biol 2021;23:620–30.
[2] Chen T, Conroy J, Wang X, Situ M, Namas RA, Vodovotz Y, Chen W, Singh H, Billiar TR. The independent prognostic value of global epigenetic alterations: An analysis of single-cell ATAC-seq of circulating leukocytes from trauma patients followed by validation in whole blood leukocyte transcriptomes across three etiologies of critical illness. EBioMedicine 2022;76:103860.
[3] Wei Q, Frenette PS. Niches for Hematopoietic Stem Cells and Their Progeny. Immunity 2018;48:632–48.
[4] Ranzoni AM, Tangherloni A, Berest I, Riva SG, Myers B, Strzelecka PM, Xu J, Panada E, Mohorianu I, Zaugg JB, et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell 2021;28:472–87. e477.
[5] Poos, A.M., Prokoph, N., Przybilla, M.J., Mallm, J.P., Steiger, S., Seufert, I., John, L., Tirier, S.M., Bauer, K., Baumann, A. *et al.* (2023) Resolving therapy resistance mechanisms in multiple myeloma by multi-omics subclone analysis. Blood.
[6] Poos AM, Mallm J-P, Tirier SM, Casiraghi N, Susak H, Giesen N, Bauer K, Baumann A, John L, Lander I. A Comprehensive Analysis of Single-Cell Chromatin Accessibility and Gene Expression Identifies Intra-Tumor Heterogeneity and Molecular Treatment Responses in Relapsed/Refractory Multiple Myeloma. Blood 2019;134:575.
[7] Avagyan S, Weber MC, Ma S, Prasad M, Mannherz WP, Yang S, Buenrostro JD, Zon LI. Single-cell ATAC-seq reveals GATA2-dependent priming defect in myeloid and a maturation bottleneck in lymphoid lineages. Blood Adv 2021;5:2673–86.
[8] Shi W, Ye J, Shi Z, Pan C, Zhang Q, Lin Y, Luo Y, Su W, Zheng Y, Liu Y. Chromatin accessibility analysis reveals regulatory dynamics and therapeutic relevance of Vogt-Koyanagi-Harada disease. Commun Biol 2022;5:506.
[9] Zhang Y, Amaral ML, Zhu C, Grieco SF, Hou X, Lin L, Buchanan J, Tong L, Preissl S, Xu X, et al. Single-cell epigenome analysis reveals age-associated decay of heterochromatin domains in excitatory neurons in the mouse brain. Cell Res 2022;32:1008–21.
[10] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486–90.
[11] Rhodes CT, Thompson JJ, Mitra A, Asokumar D, Lee DR, Lee DJ, Zhang Y, Jason E, Dale RK, Rocha PP, et al. An epigenome atlas of neural progenitors within the embryonic mouse forebrain. Nat Commun 2022;13:4196.
[12] Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, Luo W, Huang TS, Yeung BZ, Papalexi E, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol 2021;39:1246–58.
[13] Giles JR, Manne S, Freilich E, Oldridge DA, Baxter AE, George S, Chen Z, Huang H, Chilukuri L, Carberry M, et al. Human epigenetic and transcriptional T cell differentiation atlas for identifying functional T cell-specific enhancers. Immunity 2022;55:557–74. e557.
[14] Smith RJ, Zhang H, Hu SS, Yung T, Francis R, Lee L, Onaitis MW, Dirks PB, Zang C, Kim T-H. Single-cell chromatin profiling of the primitive gut tube reveals regulatory dynamics underlying lineage fate decisions. Nat Commun 2022;13:2965.
[15] Zhang L, Zhang J, Nie Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. Sci Adv 2022;8:eabl7393.
[16] Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol 2019;37:925–36.
[17] Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 2016;48:1193–203.
[18] Kourtis N, Wang Q, Wang B, Oswald E, Adler C, Cherravuru S, Malahias E, Zhang L, Golubov J, Wei Q, et al. A single-cell map of dynamic chromatin landscapes of immune cells in renal cell carcinoma. Nat Cancer 2022;3:885–98.
[19] Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, Aranyi T, Kwan A, Kondo A, Pontoglio M, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat Commun 2021;12:2277.
[20] Wang Q, Zhang Y, Zhang B, Fu Y, Zhao X, Zhang J, Zuo K, Xing Y, Jiang S, Qin Z, et al. Single-cell chromatin accessibility landscape in kidney identifies additional cell-of-origin in heterogenous papillary renal cell carcinoma. Nat Commun 2022;13:31.
[21] Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods 2022;19:1088–96.
[22] Wimmers F, Donato M, Kuo A, Ashuach T, Gupta S, Li C, Dvorak M, Foecke MH, Chang SE, Hagan T, et al. The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination. Cell 2021;184:3915–35. e3921.
[23] Ziffra RS, Kim CN, Ross JM, Wilfert A, Turner TN, Haeussler M, Casella AM, Przytycki PF, Keough KC, Shin D, et al. Single-cell epigenomics reveals mechanisms of human cortical development. Nature 2021;598:205–13.
[24] Jiang P, Zhang Z, Hu Y, Liang Z, Han Y, Li X, Zeng X, Zhang H, Zhu M, Dong J, et al. Single-cell ATAC-seq maps the comprehensive and dynamic chromatin accessibility landscape of CAR-T cell dysfunction. Leukemia 2022;36:2656–68.
[25] Gíslason MH, Demircan GS, Prachar M, Furtwängler B, Schwaller J, Schoof EM, Porse BT, Rapin N, Bagger FO. BloodSpot 3.0: a database of gene and protein expression data in normal and malignant haematopoiesis. Nucleic Acids Res 2023.
[26] Bagger FO, Sasivarevic D, Sohi SH, Laursen LG, Pundhir S, Sønderby CK, Winther O, Rapin N, Porse BT. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. Nucleic Acids Res 2016;44:D917–24.
[27] Bagger FO, Kinalis S, Rapin N. BloodSpot: a database of healthy and malignant haematopoiesis updated with purified and single cell mRNA sequencing profiles. Nucleic Acids Res 2019;47:D881–5.
[28] Gao X, Hong F, Hu Z, Zhang Z, Lei Y, Li X, Cheng T. ABC portal: a single-cell database and web server for blood cells. Nucleic Acids Res 2023;51:D792–804.
[29] Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell 2021;184:5985–6001. e5919.
[30] Li YE, Preissl S, Miller M, Johnson ND, Wang Z, Jiao H, Zhu C, Wang Z, Xie Y, Poirion O, et al. A comparative atlas of single-cell chromatin accessibility in the human brain. Science 2023;382:eadf7044.
[31] 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-d1028.
[32] 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-d815.
[33] Gao T, Zheng Z, Pan Y, Zhu C, Wei F, Yuan J, Sun R, Fang S, Wang N, Zhou Y, et al. scEnhancer: a single-cell enhancer resource with annotation across hundreds of tissue/cell types in three species. Nucleic Acids Res 2022;50. D371-d379.
[34] Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 2018;174:1309–24. e1318.
[35] Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al. A human cell atlas of fetal chromatin accessibility. Science 2020;370.
[36] Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell 2018;71:858–71. e858.
[37] Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods 2017;14:975–8.
[38] Xu H, Yu H, Liu L, Wu H, Zhang C, Cai W, Hong X, Liu D, Tang D, Dai Y. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Peripheral Mononuclear Cells in Patients With Ankylosing Spondylitis. Front Immunol 2021;12:760381.

[39] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5.
[40] Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res 2022;50. D387-d390.
[41] Yu W, Uzun Y, Zhu Q, Chen C, Tan K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. Genome Biol 2020;21:94.
[42] De Rop FV, Ismail JN, Bravo González-Blas C, Hulselmans GJ, Flerin CC, Janssens J, Theunis K, Christiaens VM, Wouters J, Marcassa G, et al. Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. Elife 2022;11.
[43] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 2015;348:910–4.
[44] Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature 2018;555:538–42.
[45] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502.
[46] Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F. and Newell, E.W. (2018) Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol.
[47] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov (Camb) 2021;2:100141.
[48] The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47. D330-d338.
[49] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61.
[50] Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, Machol I, Gordon MG, Subramaniam M, Shamim M, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet 2018;50:1140–50.
[51] Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat Neurosci 2018;21:432–9.
[52] 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.
[53] Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, Mahrt E, Guo W, Stawiski EW, Modrusan Z, et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes (Basel) 2019;10.
[54] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, Chen J, Jiang W, Yang K, Ou Q, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res 2023;51. D870-d876.
[55] Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019;16:1289–96.
[56] Yap HY, Tee SZ, Wong MM, Chow SK, Peh SC, Teow SY. Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. Cells 2018;7.
[57] Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, Goodman SM, Tabechian D, Hughes LB, Salomon-Escoto K, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat Immunol 2019;20:928–42.
[58] Kurotaki D, Sasaki H, Tamura T. Transcriptional control of monocyte and macrophage development. Int Immunol 2017;29:97–107.
[59] Zhu YP, Thomas GD, Hedrick CC. 2014 Jeffrey M. Hoeg Award Lecture: Transcriptional Control of Monocyte Development. Arterioscler Thromb Vasc Biol 2016;36:1722–33.
[60] Halene S, Gaines P, Sun H, Zibello T, Lin S, Khanna-Gupta A, Williams SC, Perkins A, Krause D, Berliner N. C/EBPepsilon directs granulocytic-vs-monocytic lineage determination and confers chemotactic function via Hlx. Exp Hematol 2010;38:90–103.
[61] Genovese MC, McKay JD, Nasonov EL, Mysler EF, da Silva NA, Alecock E, Woodworth T, Gomez-Reino JJ. Interleukin-6 receptor inhibition with tocilizumab reduces disease activity in rheumatoid arthritis with inadequate response to disease-modifying antirheumatic drugs: the tocilizumab in combination with traditional disease-modifying antirheumatic drug therapy study. Arthritis Rheum 2008;58:2968–80.
[62] Varghese JN, Moritz RL, Lou M-Z, van Donkelaar A, Ji H, Ivancic N, Branson KM, Hall NE, Simpson RJ. Structure of the extracellular domains of the human interleukin-6 receptor α-chain. Proc Natl Acad Sci 2002;99:15959–64.
[63] Pandolfi F, Franza L, Carusi V, Altamura S, Andriollo G, Nucera E. Interleukin-6 in Rheumatoid Arthritis. Int J Mol Sci 2020;21.