Published online 22 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1325–D1332***
https://doi.org/10.1093/nar/gkac922

ImmCluster: an ensemble resource for immunology cell type clustering and annotations in normal and cancerous tissues

Tiantongfei Jiang[1,†], Weiwei Zhou[1,†], Qi Sheng[1,†], Jiaxin Yu[1,†], Yunjin Xie[1], Na Ding[1], Yunpeng Zhang[1,*], Juan Xu[1,*] and Yongsheng Li[2,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang150081, China and [2]College of Biomedical Information and Engineering, Hainan Women and Children's Medical Center, Hainan Medical University, Haikou, 571199, China

Received August 14, 2022; Revised September 22, 2022; Editorial Decision October 04, 2022; Accepted October 06, 2022

ABSTRACT

Single-cell transcriptome has enabled the transcriptional profiling of thousands of immune cells in complex tissues and cancers. However, subtle transcriptomic differences in immune cell subpopulations and the high dimensionality of transcriptomic data make the clustering and annotation of immune cells challenging. Herein, we introduce ImmCluster (http://bio-bigdata.hrbmu.edu.cn/ImmCluster) for immunology cell type clustering and annotation. We manually curated 346 well-known marker genes from 1163 studies. ImmCluster integrates over 420 000 immune cells from nine healthy tissues and over 648 000 cells from different tumour samples of 17 cancer types to generate stable marker-gene sets and develop context-specific immunology references. In addition, ImmCluster provides cell clustering using seven reference-based and four marker gene-based computational methods, and the ensemble method was developed to provide consistent cell clustering than individual methods. Five major analytic modules were provided for interactively exploring the annotations of immune cells, including clustering and annotating immune cell clusters, gene expression of markers, functional assignment in cancer hallmarks, cell states and immune pathways, cell–cell communications and the corresponding ligand–receptor interactions, as well as online tools. ImmCluster generates diverse plots and tables, enabling users to identify significant associations in immune cell clusters simultaneously. ImmCluster is a valuable resource for analysing cellular heterogeneity in cancer microenvironments.

INTRODUCTION

With the development of high throughput sequencing technologies, single-cell RNA sequencing (scRNA-Seq) has emerged as a powerful tool for analysing cellular heterogeneity across multiple species, tissues and cellular contexts (1). In addition, massively multiplexed single-cell transcriptomics has enabled the transcriptional profiling of thousands of immune cells in complex tissues and cancers (2). As a result, great efforts have been made to collect and curate scRNA-Seq data, such as DISCO (2), scRNASeqDB (3), JingleBells (4) and TISCH (5). The growing availability of scRNA-Seq data provides opportunities for analysing the complex tissues and diseases' immune microenvironments, such as cell–cell communications. However, subtle transcriptomic differences in immune cell subpopulations and the high dimensionality of scRNA-Seq data make the clustering and annotation of immune cells challenging (6). There is a still lack web resource for comprehensive, intuitive, and user-friendly interactive annotating single-cell transcriptomes.

A commonly used approach for cell cluster annotation consists of identifying highly expressed genes in each cluster and overlapping them with established marker-gene lists for cell types (7). Several databases have been proposed to manually inspect the marker genes for diverse cell types from available information in the literature, such as CellMarker (8), PanglaoDB (9), CancerSEA (10) and PCMDB (11). These databases provide valuable resources for annotating cell clusters derived from scRNA-Seq experiments. However, the manual annotation of cell cluster-specific marker

*To whom correspondence should be addressed. Tel: +86 13604805482; Email: liyongsheng@hainmc.edu.cn
Correspondence may also be addressed to Juan Xu. Email: xujuanbiocc@ems.hrbmu.edu.cn
Correspondence may also be addressed to Yunpeng Zhang. Email: zhangyp@hrbmu.edu.cn
†The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

genes is time-consuming and has limited reproducibility. In addition, emerging evidence has demonstrated that several marker genes are usually expressed in multiple cell types and thus make the annotations difficult. For example, CD44 is expressed in diverse immune cell populations (8), including CD4 + T cells, activated CD8 + T cells and stem cells.

To overcome the limitations of the manual annotations of cell clusters based on marker genes, there have been a growing number of computational methods for the automated annotation of cell clusters. These methods ultimately annotate single-cell gene expressions with curated marker-gene databases, reference expression data, or supervised classification (12). The first category of methods based on marker genes includes scSorter (13), scCATCH (14), SCINA (15) and Sctype (16), and relies on a set of established cell type-specific marker genes to assign the cell identity in the queried cells. The second category of methods relies on annotated bulk or single-cell RNA datasets as references (methods based on references). These methods use correlation as the most straightforward statistical method for revealing information about unlabelled cell clusters, and the most representative methods are SingleR (17), SingleCellNet (18), scPred (19) and ScClassify (20). Automated cell type annotation methods have been applied in various tissues and cancers. However, a benchmarking study has demonstrated that each method possesses specific advantages over the others (21). Furthermore, due to subtle transcriptomic differences in immune cell subpopulations, the annotations of different methods are not consistent with each other for the same single-cell transcriptome. Moreover, these methods are scattered in multiple studies or tools; it is difficult for users to obtain the reference profiles and select an appropriate method.

To address these challenges, we introduce the ImmCluster database, which is freely available at http://bio-bigdata.hrbmu.edu.cn/ImmCluster. Currently, ImmCluster integrates 346 manually curated marker genes from 61 cell types obtained from the literature. In addition, over 420 000 immune cells from nine tissues of healthy donors and over 648 000 cells from different tumour samples of 17 cancer types were used to generate stable immune marker-gene sets and develop context-specific immune cell type references. Eleven computational methods were classified into two types (reference-based and marker-gene-based). In addition, the ensemble method was integrated into ImmCluster. Furthermore, ImmCluster is equipped with five major analytic modules that allow users to explore the annotations of immune cells interactively. Finally, functional heat maps, bar plots, river plots, circos plots and tables were provided to enable the user to easily and simultaneously identify significant associations in multiple immune cell clusters. In summary, we believe that ImmCluster is a valuable resource for analysing cellular heterogeneity in complex systems and cancer microenvironments.

MATERIALS AND METHODS

Collection of marker genes for immune cell types

We queried PubMed with 'single-cell RNA-seq' or 'scRNA-seq' as keywords to collect the studies published in recent years. We manually curated 1163 articles and recorded 346 canonical marker genes for 61 immune cell types (Figure S1). Moreover, genes were scored by the number of studies in the literature in which they appeared.

Single-cell transcriptomes across normal and cancerous tissues

For scRNA-Seq datasets of various cancer types, we queried the Gene Expression Omnibus (GEO) (22), 10× Genomics, and the National Genomics Data Center (NGDC) (23) with 'scRNA seq' and 'cancer', 'carcinoma' or 'adenocarcinoma' as keywords. The species was restricted to '*Homo sapiens*'. Over 682 tumour samples within 41 datasets obtained from diverse platforms were collected (Figure 1A and Supplementary Figure S2A), including 10× Genomics, Smart-seq2, Drop-seq, inDrop, MARS-seq and Seq-Well. The 'raw count' or 'TPM' expression profiles and the meta information of each dataset were downloaded. In addition, we obtained three datasets of peripheral blood mononuclear cells (PBMCs) from healthy donors from the 10× Genomics website (Figure S2B). To obtain other normal tissues, we downloaded scRNA-Seq data that include ∼330 000 immune cells across eight tissues from a recent study (24). Moreover, we collected another 11 datasets involving >60 000 normal cells across four tissues from the Human Cell Landscape (HCL) project (25).

Computational methods for immunology cell type annotations

With the wide application of scRNA-Seq technology, several computational methods were developed to annotate cell types automatically. First, the numbers of cell clusters in each dataset were automatically determined by the 'findcluster' function in the Seurat package with resolution = 0.5. Here, we integrated two types of computational methods into ImmCluster (Figure 1B), including the reference-based and marker-gene-based methods. The reference-based methods use labelled scRNA-Seq dataset as the input for cell type annotations, which finds the best correlation between the reference and user queried dataset. SingleR (17), SingleCellNet (18), scPred (19), Garnett (26), ScClassify (20), CHETAH (27) and scmap (28) were integrated into ImmCluster. The marker-gene-based methods rely on cell type-specific marker genes that are publicly available in databases or that have been published, including scSorter (13), scCATCH (14), SCINA (15) and Sctype (16). All these methods were performed by R scripts.

Construction of reference profiles and marker atlas of immune cell types

We first constructed comprehensive reference profiles for distinct contexts for the reference-based computational methods. First, all scRNA-Seq datasets from diverse platforms were quality controlled and immune cells were identified based on known annotations and expression of marker genes (see details in Supplementary methods). The necessary processes were mainly performed using the R package Seurat (29), including the normalization, dimension reduction, unsupervised clustering, and visualisation of cell

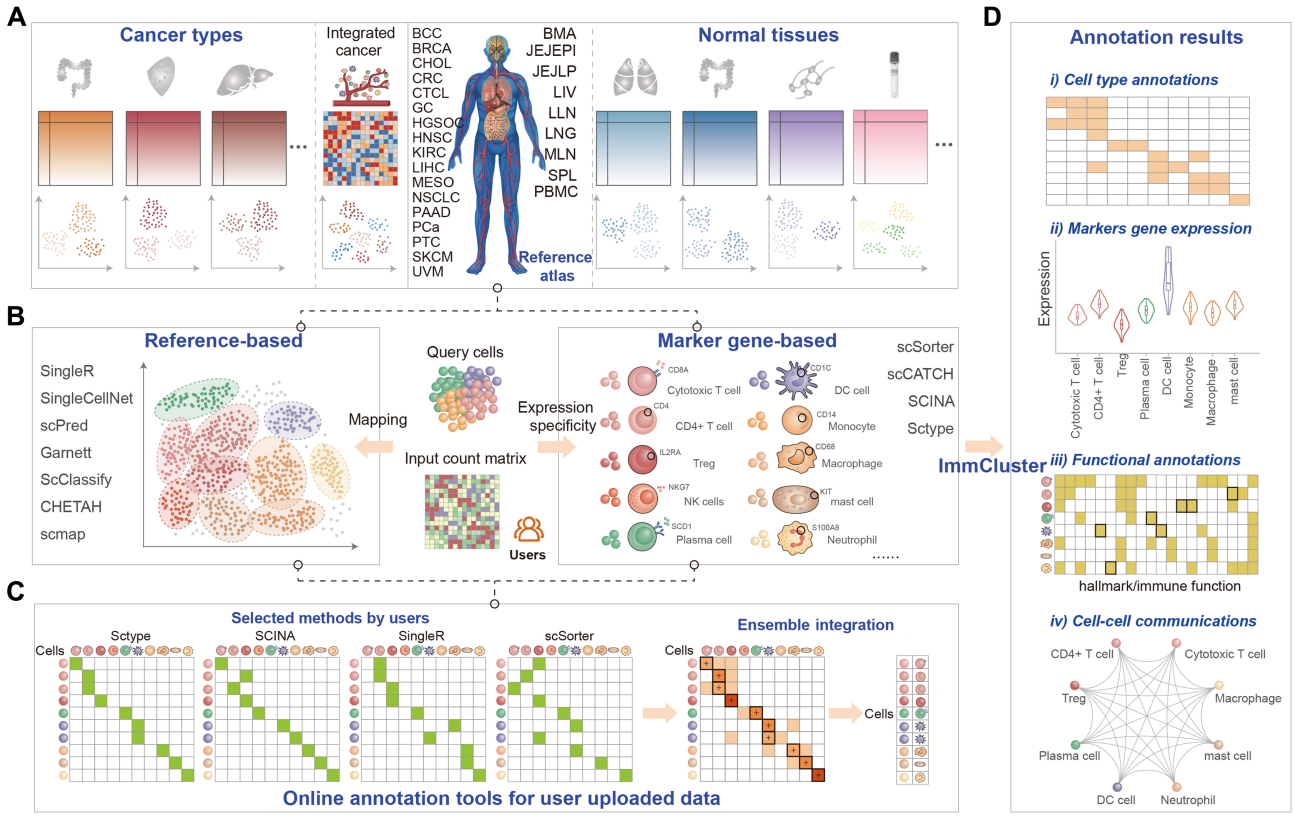

Figure 1. Schematic diagram of the overall design of ImmCluster. (**A**) Collection of reference transcriptomes from cancerous or normal tissues from published studies. (**B**) Computational methods for automatically annotating cell types. (**C**) Ensemble method is provided in ImmCluster. (**D**) Four main types of annotations are provided in ImmCluster.

clusters. Next, the batch effects from different samples or datasets were removed by Harmony (30). Then, the datasets from diverse cancer types, or normal tissues were clustered and the clusters were annotated based on canonical marker genes. Finally, based on the annotated cell types, we constructed cancer- or tissue-specific reference profiles for each computational method with corresponding formats. The reference files are available on the download page.

We next constructed the marker-gene atlas of immune cell types based on the manually curated marker genes from the literature and the reference profiles. Similar to a recent study (31), four filters were applied to this process: (i) the gene was detected in at least three cells on at least three counts across all cells; (ii) the gene was statistically significantly higher in expression in this cluster than in the complement set; to establish significance, we used the two-tailed Mann–Whitney *U* test with multiple hypothesis correction, false discovery rate (FDR) <5%; (iii) gene exhibited maximal average expression in this cluster and (iv) the max-to-second-max ratio for the gene was at least 1.1×. The marker genes for immune cell types were saved as R data and can be downloaded from ImmCluster.

Ensemble annotations of immune cell types

To obtain robust cell type annotations from diverse computational methods, we developed an ensemble annotation method to integrate the results from different methods (Figure 1C). The majority voting strategy was used in the ensemble methods. For each cell, the annotations from different methods were collected, and we labelled the cell with the cell type that had the highest number of votes. In addition, cells with more than two labels annotated as different immune lineages were defined as 'unknown'. If a cell was labelled with two cell types in the same immune lineages and each type had the same number of votes, we kept both cell types for users. Users can ultimately determine the most accurate cell type based on their knowledge of biology. The marker-gene and reference-based methods were analysed separately.

Cell type-specific expression of genes and functional assignments

Based on the annotations of immune cell types, we used COSG to identify the specifically highly expressed genes in a cell type compared to other cell types (32). COSG is a cosine similarity-based method that is faster than the 'findAllMarkers' of Seurat. The top highly expressed 100 genes were selected for each immune cell type.

To investigate the functions of cell types, we first calculated the single sample gene set enrichment analysis (ssGSEA) score for each cell (33). The cell states (10), immune signatures (34) and cancer hallmarks (35) were considered. In addition, we performed a hypergeometric test to identify the significantly enriched functions based on highly expressed genes. The genes with FDR <0.01 and

fold-change >2 were identified as highly expressed. Functions with a P <0.05 were identified as significant.

Cell–cell communications

To further explore the interactions between immune cell types and identify cell–cell communications, ImmCluster integrated iTALK (https://github.com/Coolgenome/iTALK), celltalker (https://github.com/arc85/celltalker), and ICELLNET (36). The integrated ligand-receptor interactions were collected from several databases (see the supplementary methods for further details).

Database implementation

The frontend of ImmCluster was built with HTML5, JavaScript, and CSS, and it includes the jQuery (v3.3.1), Datatable (1.10.25), ECharts (v5.5.1) and D3 (v7.6.1) plugins. The backend of ImmCluster is powered by MySQL (v5.5.21) and is queried via the Java Server Pages with Apache Tomcat container (v6.0) as the middleware. All data in ImmCluster are stored and managed using MySQL (v5.5.21) and it employs Java and R programs to perform online analyses. ImmCluster has been tested on several popular web browsers, including Google Chrome, Firefox, and Apple Safari.

DATABASE CONTENT AND ONLINE TOOLS

ImmCluster provides a comprehensive collection of canonical marker genes of immune cell types and single-cell transcriptomes. Approximately 346 marker genes from 61 cell types were manually curated from the literature. ImmCluster integrated over 420 000 immune cells from nine tissues of healthy donors and over 648 000 cells from patients with 17 different cancer types to generate stable immune marker-gene sets and develop context-specific immunology references. In total, 11 computational and the ensemble methods were integrated into ImmCluster to annotate immune cell types automatically. ImmCluster provided five major analytic modules that allow users to interactively explore the annotations of immune cells (Figure 1D): (i) clustering and annotating the immune cell clusters based on computational methods; (ii) gene expression of markers across cell clusters in the user uploaded dataset and the reference immunology profiles; (iii) functional assignment of the cell clusters in cancer hallmarks, cell states and immune pathways; (iv) cell–cell communications and corresponding ligand–receptor interactions and (v) online annotation tools for user uploaded data.

ImmCluster is organised into six main pages (Figure 2): Home, Browse, Search, Tool, Download and Help. We provided a global browse function on the 'Browse' page to allow users to query the annotations of published cancer or normal single-cell transcriptomes. Users can browse the cancerous or normal tissues of interest via the human body map or the dataset tables (Figure 2A). Annotations of the reference dataset include the tSNE, or UMAP visualisation of the clustering, details of the marker genes for each cell type, and functional annotations of the cell types and cell–cell communications mediated by ligand–receptor interactions (Figure 2B). The search page provides four types of query options for users to search for datasets, genes, cell types, or reference profiles of interest (Figure 2C). The heatmap, tSNE, or UMAP figures can be used to visualise the expression of genes across cell types (Figure 2D). The functional annotations and the cell–cell communications between cell types that the users search for are provided in the result pages (Figure 2E). In particular, ImmCluster provides useful tools for annotating the single-cell transcriptome uploaded by users (Figure 2F). After quality control of the data uploaded by users, ImmCluster automatically annotates the cell types and performs downstream analyses based on the user's selected methods. The immune cell type annotations for different computational and the ensemble method are then provided (Figure 2G). Details of marker genes, functional annotations in immune pathways, cellular states, cancer hallmarks, and cell–cell communications are provided in the forms of diverse types of figures and tables (Figure 2G). Finally, the download page allows users to download all of the integrated data, including the reference transcriptomes, marker genes and metadata of cells in RDS format. Instructions for each module are provided on the help page.

CASE STUDY

To demonstrate the application of ImmCluster, we comprehensively analysed the colon cancer data from a recent study (37). In the original study, immune cells were classified into five major classes: B cells, CD4 T cells, CD8 T cells, innate lymphoid cells (ILCs, major class of NK cells), and myeloid cells (Figure 3A). When we analysed the dataset based on ImmCluster, we found that the ensemble method can accurately recapture the cell annotations of the original study (Figure 3B). The average accuracy of cell annotations for all cell types reached 0.963. The maximum accuracy was 0.998 for myeloid cells. We found that ImmCluster can distinguish the subpopulations of the same cell type. For example, the B cells were further classified into plasma and B-Fol cells. In addition, T cells were also classified into functional subpopulations, such as naive CD4+ T cells, T helper cells and cytotoxic T cells (Figure 3B). These results suggested that ImmCluster can accurately annotate the cell types by integrating various computational methods.

Next, we identified genes highly expressed in various cell types based on the analysis module in ImmCluster. We found that most genes identified in ImmCluster were validated by previous studies (Figure 3C). For example, JCHAIN was highly expressed in plasma, which was also demonstrated in the original study (37). Moreover, IL7R and CCR7 were highly expressed in naive CD4+ T cells, and GZMK and CD160 were highly expressed in cytotoxic T cells. In particular, we found that ImmCluster identified XCL2 and PRF1 as being highly expressed in cytotoxic T cells and NK cells, respectively, which has also been shown in a recent study (38) but not in the original study (Figure 3C). These results suggested that ImmCluster can identify novel genes of interest and generate new hypothesis for investigation. In addition, we performed functional enrichment analysis based on ImmCluster. We found that genes highly expressed in several cell types, such as NK cells, were significantly enriched in the natural killer cell cytotoxicity

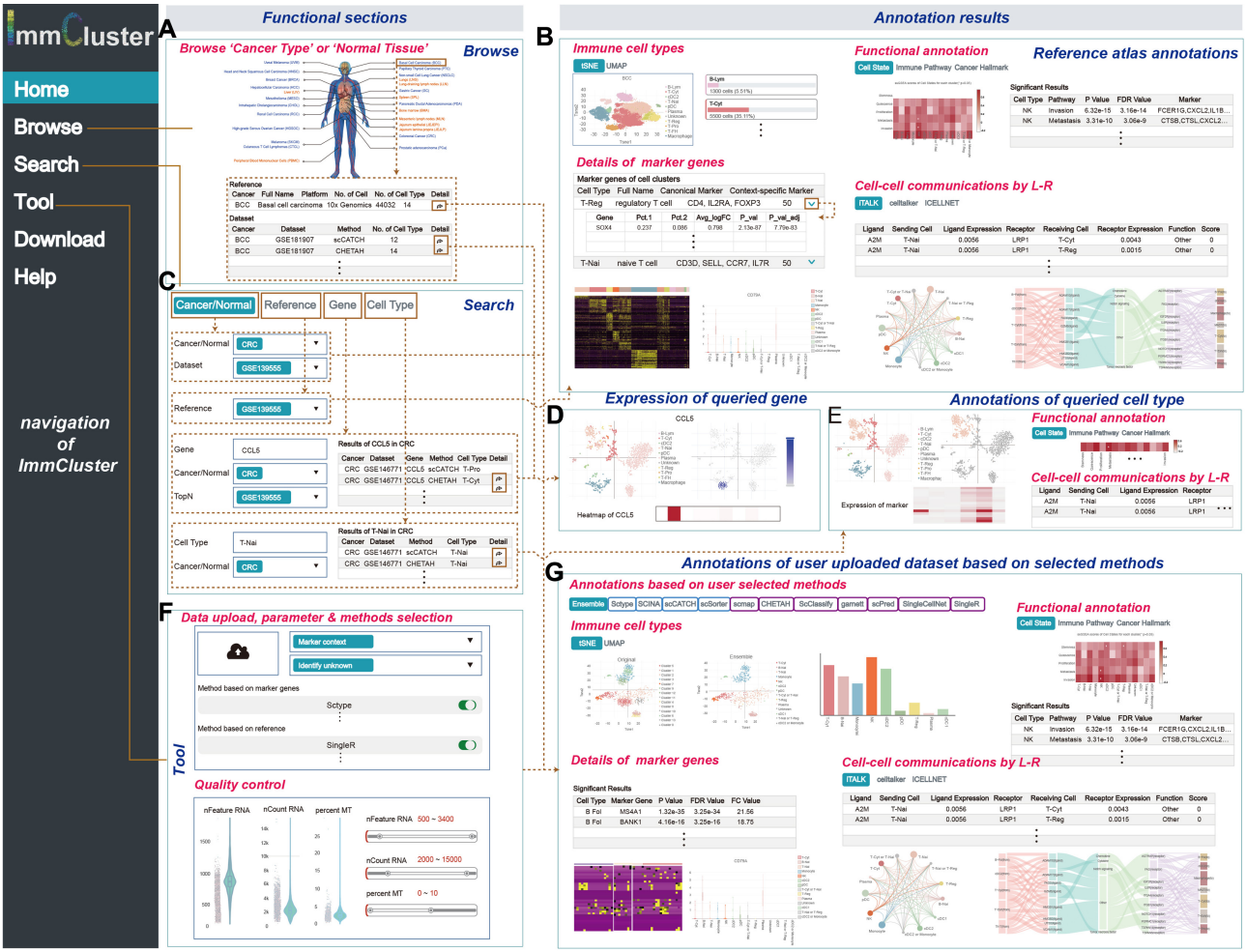

Figure 2. User interface and workflow of using ImmCluster. (**A**) Browse page of ImmCluster. (**B**) Annotation results of the reference datasets selected by users. (**C**) Search page of ImmCluster. (**D**) Expression of genes queried by users. (**E**) Annotations of cell types queried by users. (**F**) Tools provided in ImmCluster for annotating user uploaded data. Users must select parameters and methods; quality control was performed before cell type annotations. (**G**) Four main types of annotations were provided in ImmCluster, including immune cell type annotation, details of marker genes, functional annotations and cell–cell communications.

pathway (Figure 3D). Genes highly expressed in DC cells were significantly enriched in antigen processing and presentation pathway, which was consistent with their functions reported in a previous study (39). Finally, we analysed the cell–cell communications and found that the interactions were mediated by various ligand–receptor interactions (Figure 3E). In particular, ITGA4–CD44 interaction helps the communications between DCs and CD4+ T cells (39). Together, the applications demonstrated that the ImmCluster database could not only recapitulate findings from the literature, but could also generate several novel hypotheses for further functional studies.

CONCLUSIONS AND FUTURE DEVELOPMENT

Herein, we presented ImmCluster, a database for automatically clustering and annotating immune cell types based on the constructed reference profiles and numbers of computational methods. To better understand the functions of immune cell clusters, we provided diverse downstream analyses following cell clustering and annotations in single-cell analysis. The expression of highly expressed genes in each cell cluster was visualised and functional pathways of cell clusters were predicted based on gene set enrichment analysis and hypergeometric test. Moreover, the cell–cell communications were identified and visualised in diverse types. In particular, we provided several tools for automatic analysis of scRNA-Seq data uploaded by users. All the results can be returned to the webpage or sent to the users via emails. In addition, users can browse and download the reference profiles and marker atlas for more advanced analysis.

The ImmCluster database integrated manually curated cell-type marker genes and over 420 000 immune cells from nine healthy tissues and over 648 000 cells from different tumor samples of 17 cancer types. The ImmCluster database also provided comprehensive analysis and visualization functions of immune cell clusters. We compared ImmCluster with public databases and web-based analysis platforms for scRNA-Seq data, and found that ImmCluster provided more immunology datasets and functional analysis modules in immunology (Tables S1 and S2). These results suggested that ImmCluster is a valuable resource

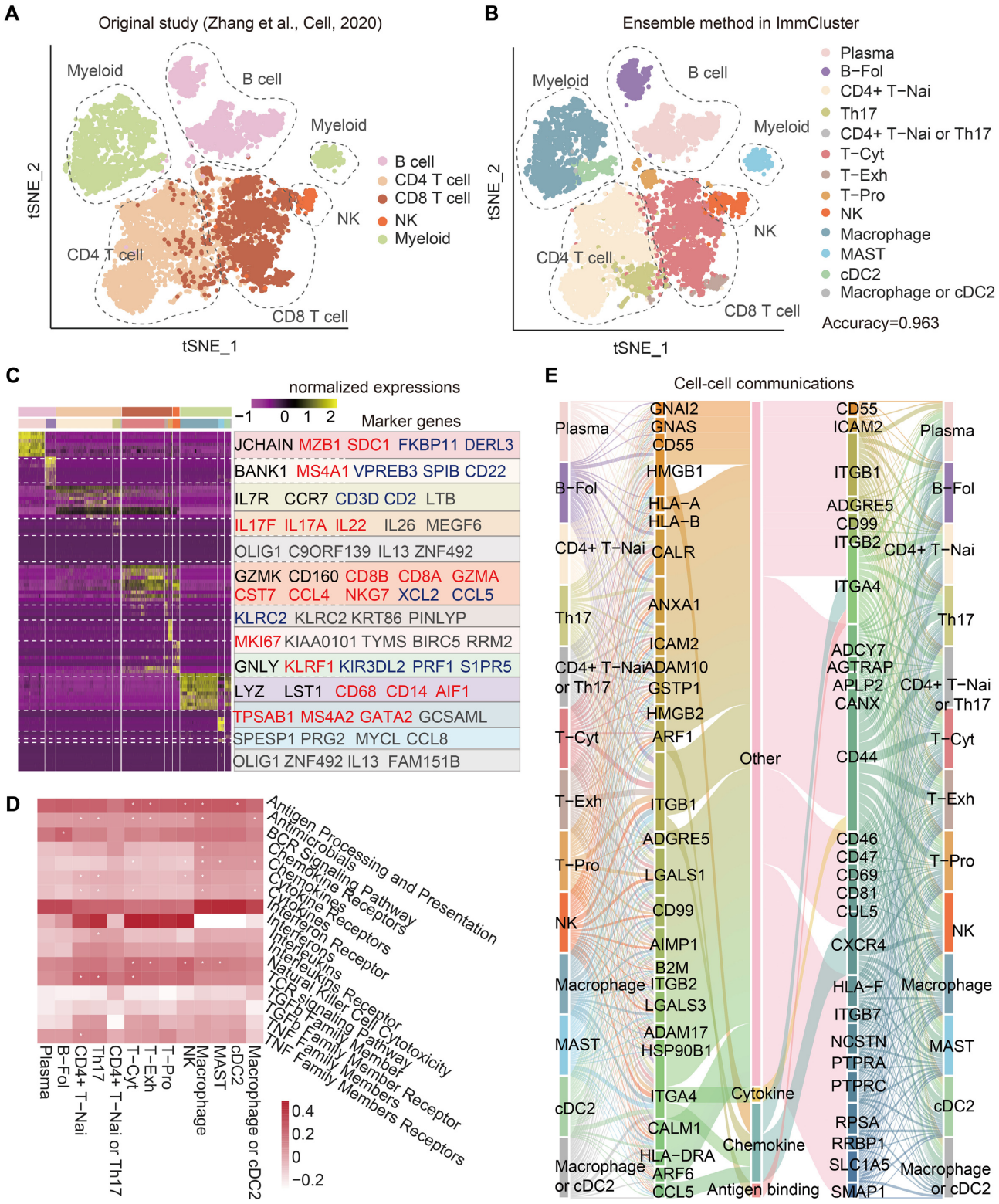

Figure 3. Case study of scRNA-Seq data analysis based on ImmCluster. (**A**) TSNE plot showing the cell annotations from the original study. (**B**) TSNE plot showing the cell annotations based on the ensemble method in ImmCluster. (**C**) Heat map showing the expressions of marker genes in various cell types. Genes marked with black are from the original study, marker genes manually curated from literature are shown in red, blue indicates genes reported by recent literature, and genes highly expressed in specific cell types are shown in grey. (**D**) Functional pathways enriched by genes highly expressed in cell types. Heat map showing the ssGSEA scores and $*P < 0.05$ for hypergeometric tests. (**E**) Cell–cell communications mediated by ligand–receptor pairs.

for analyzing cellular heterogeneity in complex systems and cancer microenvironments. ImmCluster also integrates 11 computational methods, and we developed the ensemble method to annotate the immune cell clusters robustly. Next, we evaluated the accuracy of the ensemble method compared with individual methods based on 11 PBMC scRNA-Seq datasets with cell type annotations (40). We found that the ensemble method greatly improved the accuracy of cell type annotations compared with other methods (Supplementary Figure S3).

As single-cell technologies develop and increase the scRNA-Seq data, we intend to update and upgrade ImmCluster continuously. We will integrate more scRNA-Seq data into ImmCluster as new studies are published and more datasets are available. In addition, assigning cell identities to cell clusters generated by clustering is a crucial step in scRNA-Seq data analysis (12). Manual cell annotation is time-consuming and partially subjective. Thus, emerging computational tools have been developed for automatic cell type annotations. ImmCluster currently integrates 11 widely used computational methods and the ensemble method. More computational methods will be provided in the future. We also plan to expand the scope of ImmCluster to encompass other single-cell omics data, such as scATAC-seq (41), scTCR-seq and scBCR-seq. Moreover, spatial transcriptomics can also be deployed for exploring tissue architecture and tumor microenvironment (42,43). In summary, the ImmCluster database provides comprehensive analysis and visualization functions of immune cell clusters, which is a valuable resource for analysing cellular heterogeneity in complex systems and cancer microenvironments.

DATA AVAILABILITY

ImmCluster is an open source for immunology cell type clustering and annotations in normal and cancerous tissues (http://bio-bigdata.hrbmu.edu.cn/ImmCluster).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [32060152, 32170676, 31970646, 32070673]; Hainan Province Science and Technology Special Fund [ZDYF2021SHFZ051]; Hainan Provincial Natural Science Foundation of China [820MS053]; Major Science and Technology Program of Hainan Province [ZDKJ202003]; Natural Science Foundation for Distinguished Young Scholars of Heilongjiang Province [JQ2019C004]; HMU Marshal Initiative Funding [HMUMIF-21024]; Marshal Initiative Funding of Hainan Medical University [JBGS202103]; National Key R&D Program of China [2018YFC2000100]; Bioinformatics for Major Diseases Science Innovation Group of Hainan Medical University, and Heilongjiang Touyan Innovation Team Program. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, **20**, 257–272.
2. Li,M., Zhang,X., Ang,K.S., Ling,J., Sethi,R., Lee,N.Y.S., Ginhoux,F. and Chen,J. (2022) DISCO: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res.*, **50**, D596–D602.
3. Cao,Y., Zhu,J., Jia,P. and Zhao,Z. (2017) scRNASeqDB: a database for RNA-Seq based gene expression profiles in human single cells. *Genes (Basel)*, **8**, 368.
4. Ner-Gaon,H., Melchior,A., Golan,N., Ben-Haim,Y. and Shay,T. (2017) JingleBells: a repository of immune-related single-cell RNA-sequencing datasets. *J. Immunol.*, **198**, 3375–3379.
5. Sun,D., Wang,J., Han,Y., Dong,X., Ge,J., Zheng,R., Shi,X., Wang,B., Li,Z., Ren,P. *et al.* (2021) TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.*, **49**, D1420–D1430.
6. Kiselev,V.Y., Andrews,T.S. and Hemberg,M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
7. Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
8. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
9. Franzen,O., Gan,L.M. and Bjorkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*, **2019**, baz046.
10. Yuan,H., Yan,M., Zhang,G., Liu,W., Deng,C., Liao,G., Xu,L., Luo,T., Yan,H., Long,Z. *et al.* (2019) CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.*, **47**, D900–D908.
11. Jin,J., Lu,P., Xu,Y., Tao,J., Li,Z., Wang,S., Yu,S., Wang,C., Xie,X., Gao,J. *et al.* (2022) PCMDB: a curated and comprehensive resource of plant cell markers. *Nucleic Acids Res.*, **50**, D1448–D1455.
12. Pasquini,G., Rojo Arias,J.E., Schafer,P. and Busskamp,V. (2021) Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.*, **19**, 961–969.
13. Guo,H. and Li,J. (2021) scSorter: assigning cells to known cell types according to marker genes. *Genome Biol.*, **22**, 69.
14. Shao,X., Liao,J., Lu,X., Xue,R., Ai,N. and Fan,X. (2020) scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *Iscience*, **23**, 100882.
15. Zhang,Z., Luo,D., Zhong,X., Choi,J.H., Ma,Y., Wang,S., Mahrt,E., Guo,W., Stawiski,E.W., Modrusan,Z. *et al.* (2019) SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes (Basel)*, **10**, 531.
16. Ianevski,A., Giri,A.K. and Aittokallio,T. (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.*, **13**, 1246.
17. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R. *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.
18. Tan,Y. and Cahan,P. (2019) SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Syst.*, **9**, 207–213.
19. Alquicira-Hernandez,J., Sathe,A., Ji,H.P., Nguyen,Q. and Powell,J.E. (2019) scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.*, **20**, 264.
20. Lin,Y., Cao,Y., Kim,H.J., Salim,A., Speed,T.P., Lin,D.M., Yang,P. and Yang,J.Y.H. (2020) scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol. Syst. Biol.*, **16**, e9389.
21. Abdelaal,T., Michielsen,L., Cats,D., Hoogduin,D., Mei,H., Reinders,M.J.T. and Mahfouz,A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.
22. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
23. Members,C.-N. and,Partners. (2021) Database resources of the national genomics data center, china national center for bioinformation in 2021. *Nucleic Acids Res.*, **49**, D18–D28.

24. Dominguez Conde,C., Xu,C., Jarvis,L.B., Rainbow,D.B., Wells,S.B., Gomes,T., Howlett,S.K., Suchanek,O., Polanski,K., King,H.W. *et al.* (2022) Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, **376**, eabl5197.
25. Han,X., Zhou,Z., Fei,L., Sun,H., Wang,R., Chen,Y., Chen,H., Wang,J., Tang,H., Ge,W. *et al.* (2020) Construction of a human cell landscape at single-cell level. *Nature*, **581**, 303–309.
26. Pliner,H.A., Shendure,J. and Trapnell,C. (2019) Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, **16**, 983–986.
27. de Kanter,J.K., Lijnzaad,P., Candelli,T., Margaritis,T. and Holstege,F.C.P. (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.*, **47**, e95.
28. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.
29. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
30. Tran,H.T.N., Ang,K.S., Chevrier,M., Zhang,X., Lee,N.Y.S., Goh,M. and Chen,J. (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
31. Zilionis,R., Engblom,C., Pfirschke,C., Savova,V., Zemmour,D., Saatcioglu,H.D., Krishnan,I., Maroni,G., Meyerovitz,C.V., Kerwin,C.M. *et al.* (2019) Single-Cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity*, **50**, 1317–1334.
32. Dai,M., Pei,X. and Wang,X.J. (2022) Accurate and fast cell marker gene identification with COSG. *Brief. Bioinf.*, **23**, bbab579.
33. Hanzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf.*, **14**, 7.
34. Li,Y., Jiang,T., Zhou,W., Li,J., Li,X., Wang,Q., Jin,X., Yin,J., Chen,L., Zhang,Y. *et al.* (2020) Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat. Commun.*, **11**, 1000.
35. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
36. Noel,F., Massenet-Regad,L., Carmi-Levy,I., Cappuccio,A., Grandclaudon,M., Trichot,C., Kieffer,Y., Mechta-Grigoriou,F. and Soumelis,V. (2021) Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.*, **12**, 1089.
37. Zhang,L., Li,Z., Skrzypczynska,K.M., Fang,Q., Zhang,W., O'Brien,S.A., He,Y., Wang,L., Zhang,Q., Kim,A. *et al.* (2020) Single-Cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell*, **181**, 442–459.
38. Zaitsev,A., Chelushkin,M., Dyikanov,D., Cheremushkin,I., Shpak,B., Nomie,K., Zyrin,V., Nuzhdina,E., Lozinsky,Y., Zotova,A. *et al.* (2022) Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell*, **40**, 879–894.
39. Bassler,K., Schulte-Schrepping,J., Warnat-Herresthal,S., Aschenbrenner,A.C. and Schultze,J.L. (2019) The myeloid cell compartment-cell by cell. *Annu. Rev. Immunol.*, **37**, 269–293.
40. Xie,B., Jiang,Q., Mora,A. and Li,X. (2021) Automatic cell type identification methods for single-cell RNA sequencing. *Comput. Struct. Biotechnol. J.*, **19**, 5874–5887.
41. Mimitou,E.P., Lareau,C.A., Chen,K.Y., Zorzetto-Fernandes,A.L., Hao,Y., Takeshima,Y., Luo,W., Huang,T.S., Yeung,B.Z., Papalexi,E. *et al.* (2021) Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.*, **39**, 1246–1258.
42. Rao,A., Barkley,D., Franca,G.S. and Yanai,I. (2021) Exploring tissue architecture using spatial transcriptomics. *Nature*, **596**, 211–220.
43. Lv,D., Xu,K., Jin,X., Li,J., Shi,Y., Zhang,M., Jin,X., Li,Y., Xu,J. and Li,X. (2020) LncSpA: LncRNA spatial atlas of expression across normal and cancer tissues. *Cancer Res.*, **80**, 2067–2071.