Computational and Structural Biotechnology Journal 23 (2024) 4306–4314

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Software/web server article

BrainCellR: A precise cell type nomenclature pipeline for comparative analysis across brain single-cell datasets

Yuhao Chi [a], Simone Marini [b,*], Guang-Zhong Wang [a,*]

[a] CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[b] Department of Epidemiology, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
Brain
ScRNA-seq
Cell type annotation
R package
Comparison between datasets

ABSTRACT

Single-cell studies in neuroscience require precise cell type classification and consistent nomenclature that allows for meaningful comparisons across diverse datasets. Current approaches often lack the ability to identify fine-grained cell types and establish standardized annotations at the cluster level, hindering comprehensive understanding of the brain's cellular composition. To facilitate data integration across multiple models and datasets, we designed BrainCellR. This pipeline provides researchers with a powerful and user-friendly tool for efficient cell type classification and nomination from single-cell transcriptomic data. While initially focused on brain studies, BrainCellR is applicable to other tissues with complex cellular compositions. BrainCellR goes beyond conventional classification approaches by incorporating a standardized nomenclature system for cell types at the cluster level. This feature enables consistent and comparable annotations across different studies, promoting data integration and providing deeper insights into the complex cellular landscape of the brain. All documents for BrainCellR, including source code, user manual and tutorials, are freely available at https://github.com/WangLab-SINH/BrainCellR.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the field of neuroscience by enabling comprehensive characterization of cellular heterogeneity in the brain [1–6]. This technology provides researchers with unprecedented resolution to study individual cells and uncover cell type-specific gene expression profiles. A standard step in scRNA-seq downstream analysis is cell clustering, i.e., cells with a similar gene expression profile are grouped into clusters; groups of marker genes are then extracted by clusters via statistical analysis, and given a label—typically a cell type or subtype[7,8]. However, the accurate naming of cell types at the cluster level remains challenging [9–13].

Conventionally, cell type classification in the brain has been based on major classes, such as excitatory neurons or inhibitory neurons, and subclasses, such as Lamp5 or L5 IT [14,15]. These classifications are determined through morphology, location in the brain, electrophysiological characteristics, and gene expression [16]. However, it is increasingly recognized that these broad categories do not adequately capture the full diversity of cell types within the brain [17]. To achieve a more nuanced understanding of cellular composition, researchers are now focusing on identifying and characterizing fine-grained cell types [18–22].

A major hurdle in precise cell type classification is the lack of a standardized nomenclature system at the cluster level. Due to the absence of uniform guidelines, different studies often use disparate naming conventions [23,24]. For example, cell (sub)types might be characterized by their cluster ID numbers [25,26], grouped by cluster at the subclass level (e.g., Lamp5_1, Lamp5_2), or named using a combination of cell type and marker gene (e.g., L6b P2ry12, or Sst Nts) [14]. This lack of a common nomenclature system leads to inconsistencies and difficulties in comparing cell types across datasets [27–29], hindering the integration of data from multiple studies and the generation of comprehensive cell atlases.

To address these challenges, we have developed BrainCellR (Fig. 1), a pipeline designed specifically for cell type nomenclature in brain scRNA-seq data. BrainCellR offers a comprehensive set of tools and functionalities to enable researchers to classify and nominate cell types and do comparative analysis across brain single-cell datasets.

* Corresponding authors.
E-mail addresses: simone.marini@ufl.edu (S. Marini), guangzhong.wang@picb.ac.cn (G.-Z. Wang).

https://doi.org/10.1016/j.csbj.2024.11.038
Received 2 July 2024; Received in revised form 24 November 2024; Accepted 25 November 2024
Available online 26 November 2024

2. Materials and methods

2.1. The selection of clustering methods

In evaluating clustering methods, we consider both external and internal cluster validity indicators: consistency and the ROGUE score [30]. The external cluster indicator is the consistency of clustering between two datasets, defined as the proportion of corresponding clusters between datasets (Fig. 2A). The ROGUE score, based on entropy, serves as our internal indicator for clustering (Fig. 2B) [30]. Intuitively, a pure cell cluster is defined as a population with identical function and state across all cells and no variable genes. To select the best clustering approach, we evaluated seven methods and pipelines: the Seurat package using the Louvain algorithm [22], Monocle3 package using Louvain algorithm [31], SC3 using SVM algorithm [32], scCCESS-SIMLR [33], scCCESS-Kmeans [33], scrattch-hicat [17], and our one-iteration method, Consensus1, based on improvements to scrattch-hicat.

In addition to performance metrics, we also considered the execution time (Fig. 2C)and computational memory usage (Fig. 2D)of each method. We thoroughly evaluated these methods across six brain cell datasets (Table S1), comprising three human and three mouse datasets. Among the evaluated clustering methods, the ROGUE scores [30] suggest that there may not be a significant difference in performance between most of the methods. However, when considering the external indicators, both Consensus1 and scrattch-hicat exhibit significantly higher scores compared to the other methods. Consequently, these two methods are selected as candidate methods for the clustering process in our pipeline. Scrattch-hicat [17], which shows the best performance in terms of external indicators, is a strong candidate for accurate cell type identification across datasets. On the other hand, Consensus1, with the

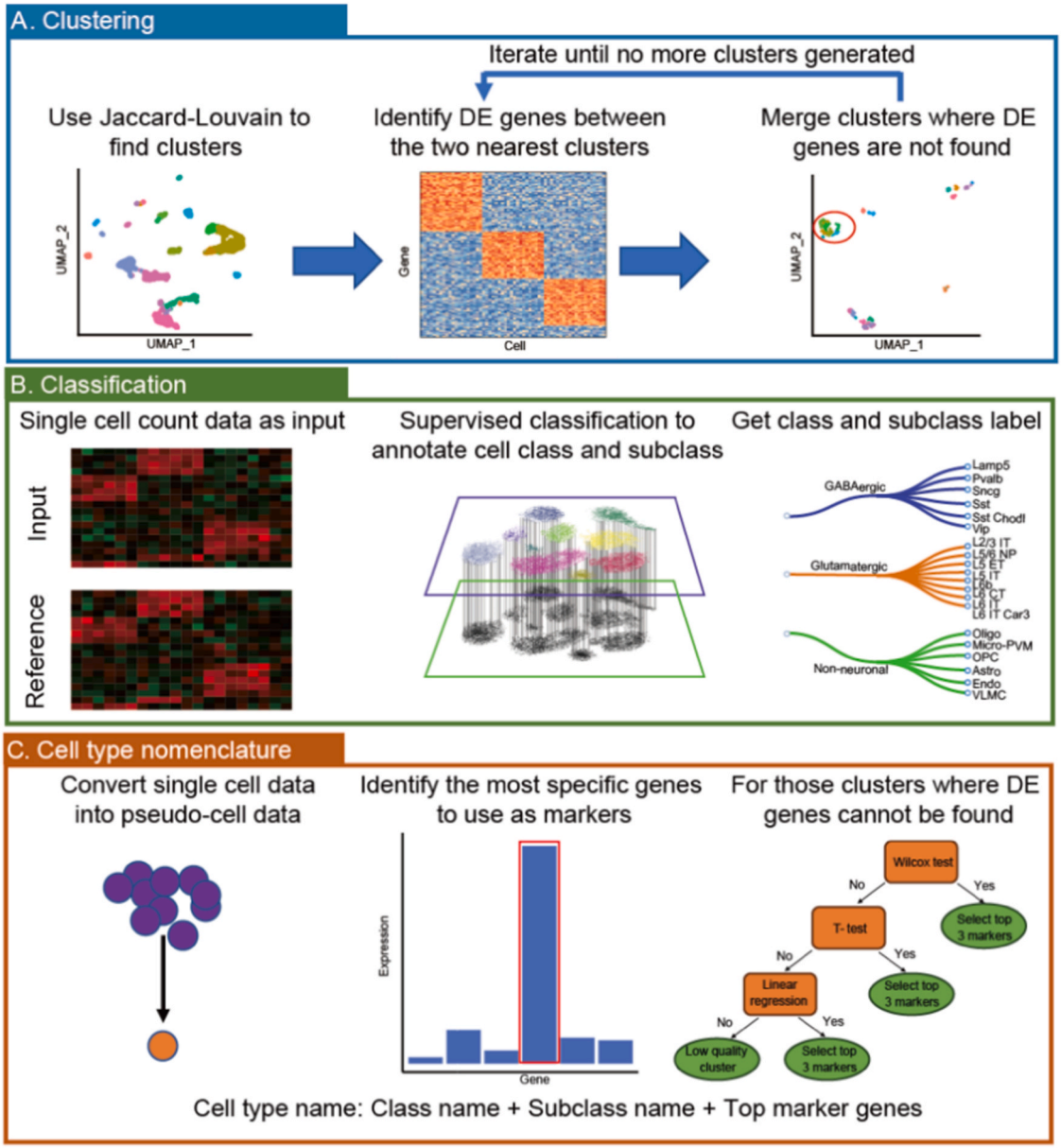

Fig. 1. Overall workflow of BrainCellR. The cell type classification and nomination pipeline can be divided into three steps: (A) We use an iterative clustering method called Consensus1 to obtain the cell clustering results; (B) We employ Seurat's FindTransferAnchors method for supervised classification of the major cell class and cell subclass; (C) After processing single-cell data into pseudo-cells, we identify differentially expressed genes using the ROC methods in Seurat and sequence them according to expression specificity scores. If we cannot find any marker gene for a cell type, we select other differential expression gene identification methods for processing.

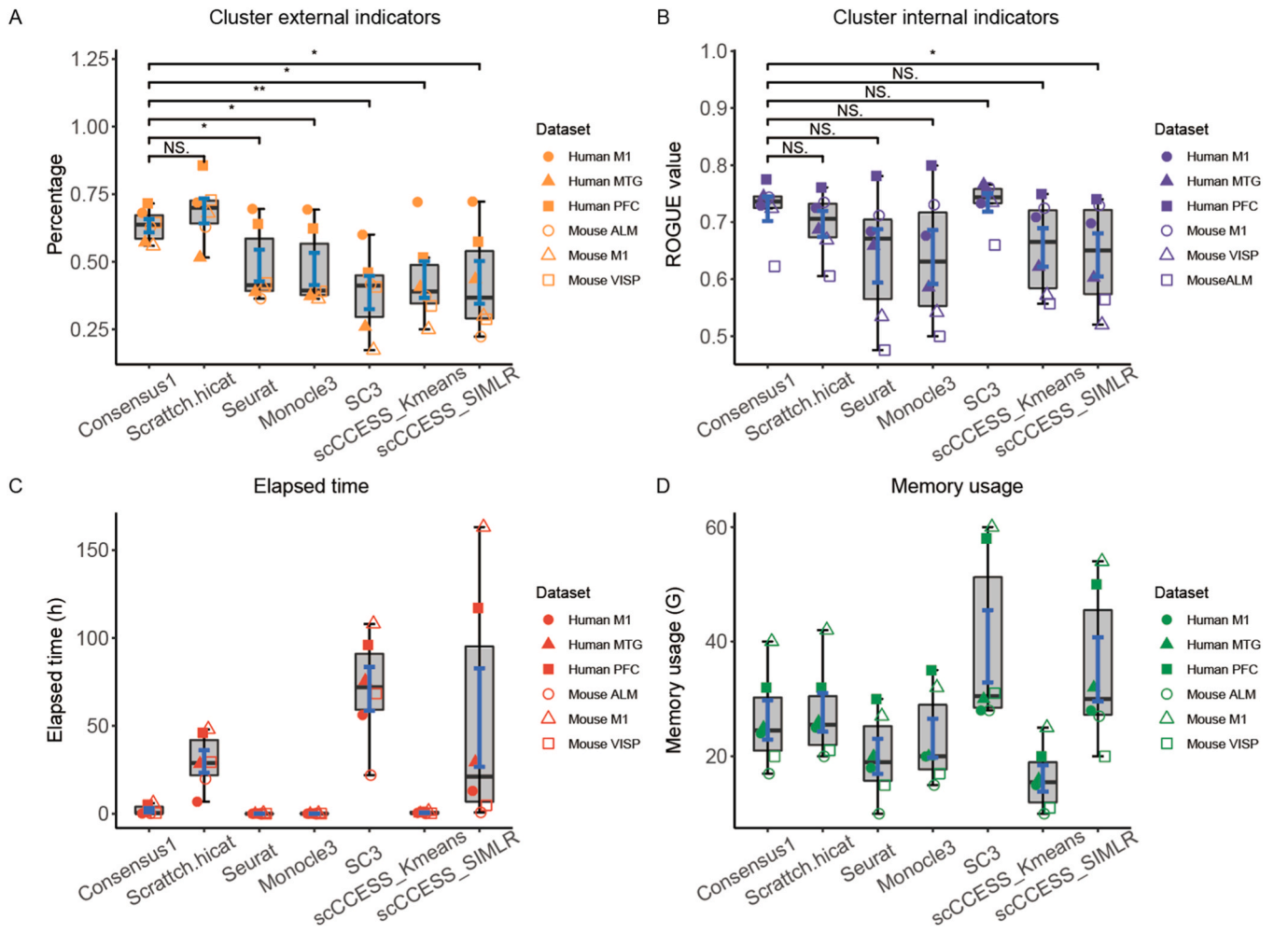

Fig. 2. Evaluation of clustering methods. (A) Percentage of common clusters between Dataset1 and Dataset2. Scrattch.hicat achieves the highest percentage, and the Consensus1 method achieves the second highest score. (B) Purity score of each clustering method. (C) Time usage (in hours) for each method. (D) Memory usage (in GB) for each method.

second-best performance, offers the advantage of significantly reduced computing time compared to scrattch-hicat (Fig. 2C). Therefore, we have chosen Consensus1 for the clustering process in our pipeline.

2.2. Evaluating the consistency of clusters between datasets

The consistency of clustering between datasets is used to evaluate whether the clustering of two data sets is similar after independent clustering. High consistency indicates that cells with similar expression characteristics are grouped into the same class in both sets of data. In this process, we first integrate the two sets of data together for unified clustering (clustering is performed using the methods described in the articles from which each dataset is sourced). Each cell can then obtain a clustering label from this integrated clustering.

Subsequently, we use the clustering algorithm to cluster the two sets of data separately. We then count the distribution of the number of integrated cluster labels contained by cells in each independent cluster and normalize this distribution by dividing it by the maximum value in the distribution. We then calculate the correlation coefficient between the quantity distribution calculated by each cluster from each dataset. If the correlation coefficients between two clusters from different datasets are the highest respectively, then we consider these two clusters to be a pair of clusters with cluster consistency. The proportion of cluster consistency is the ratio between the number of identified clusters with consistency and the total number of clusters contained in the dataset.

2.3. Evaluating the purity of clusters between datasets

The purity score is an entropy-based statistic called ROGUE [30] to quantify the purity of identified cell clusters. The entropy can be defined as

$$H(x) = -\int_{-\infty}^{+\infty} p(x)\cdot \ln p(x)dx$$

where X is the expression value and p(x) is the probability density function. Then the degree of disorder or randomness of gene expression can be represented as

$$ds = \ln E(X_i) - \frac{\sum_{j=1}^{n} \ln X_{ij}}{n}$$

$E(X_i)$ is the expectation expression of X under negative binomial distribution for gene i. The ROGUE value which represents the purity score can be defined as

$$ROGUE = 1 - \frac{\sum ds}{\sum ds + K}$$

Where K is a parameter to constrain the value between 0 and 1. Additionally, K can also serve as a reference factor to aid in interpreting the purity evaluation.

2.4. Description of the Consensus1 method

The Consensus1 method is an improvement on the scrattch.hicat [17] method. With the scrattch.hicat method, a subset of cells is selected 100 times for clustering, and then the differentially expressed genes between neighboring clusters are calculated. If no differentially expressed genes can be found between clusters, the two clusters are merged and the process is iterated continuously until no new clusters are generated. If a group of cells is grouped into the same cluster in all 100 iterations, we consider them to be a robust cluster.

The Consensus1 method modifies this process by selecting all the cells for clustering only once, instead of selecting 80 % of the cells and clustering them 100 times as in the scrattch.hicat method. The Consensus1 method also preserves the process of merging clusters from the scrattch.hicat method. In this case, the first round of clustering is performed and then the differentially expressed genes between the current cluster and its two adjacent clusters are calculated. If no differentially expressed genes are detected, the cluster is merged with the nearest cluster. This process is iterated until no new cluster is generated.

2.5. The selection of supervised classification methods

The purpose of supervised classification within our pipeline is to assign cell types based on existing annotations for the major classes and subclasses of brain cells. In the case of the cortex, there is a general consensus on the major classes and subclasses of cell types found in specific regions [15,34]. For instance, in the primary motor cortex of mice, there are two major classes of neurons: Glutamatergic and GABAergic, along with several non-neuronal classes based on the neurotransmitters they release [35]. The classification of these major classes is widely agreed upon in the mouse cortex [17]. Glutamatergic neurons can be further classified into different subclasses based on their cortical localization and projection patterns [14,36]. Similarly, GABAergic neurons can also be divided into various subclasses. Importantly, these subclasses have been found to be consistent and prevalent across different brain regions of the cortex [17].

We tested our approach using one biological replicate of each dataset (Table S1) as a training set, training the model on the training set using five methods: CHEATH [37], scmap [38], Seurat [22], SingleCellNet [39] and SingleR [40]. We then applied the trained model to another biological replicate of each dataset to evaluate the accuracy of the classification. Based on the evaluation results(Fig. 3), Seurat [22] performs best in three out of four indicators: Accuracy, Precision, Recall, and F1-score. Several studies have evaluated different methods for cell type classification, with SingleR achieving the highest classification accuracy and Seurat ranking second [41]. While SingleR [40] has shown superior performance, it requires more computational time, particularly for large datasets [41]. By incorporating Seurat and SingleR into our pipeline, we achieve precise and consistent cell type classification across brain single-cell datasets.

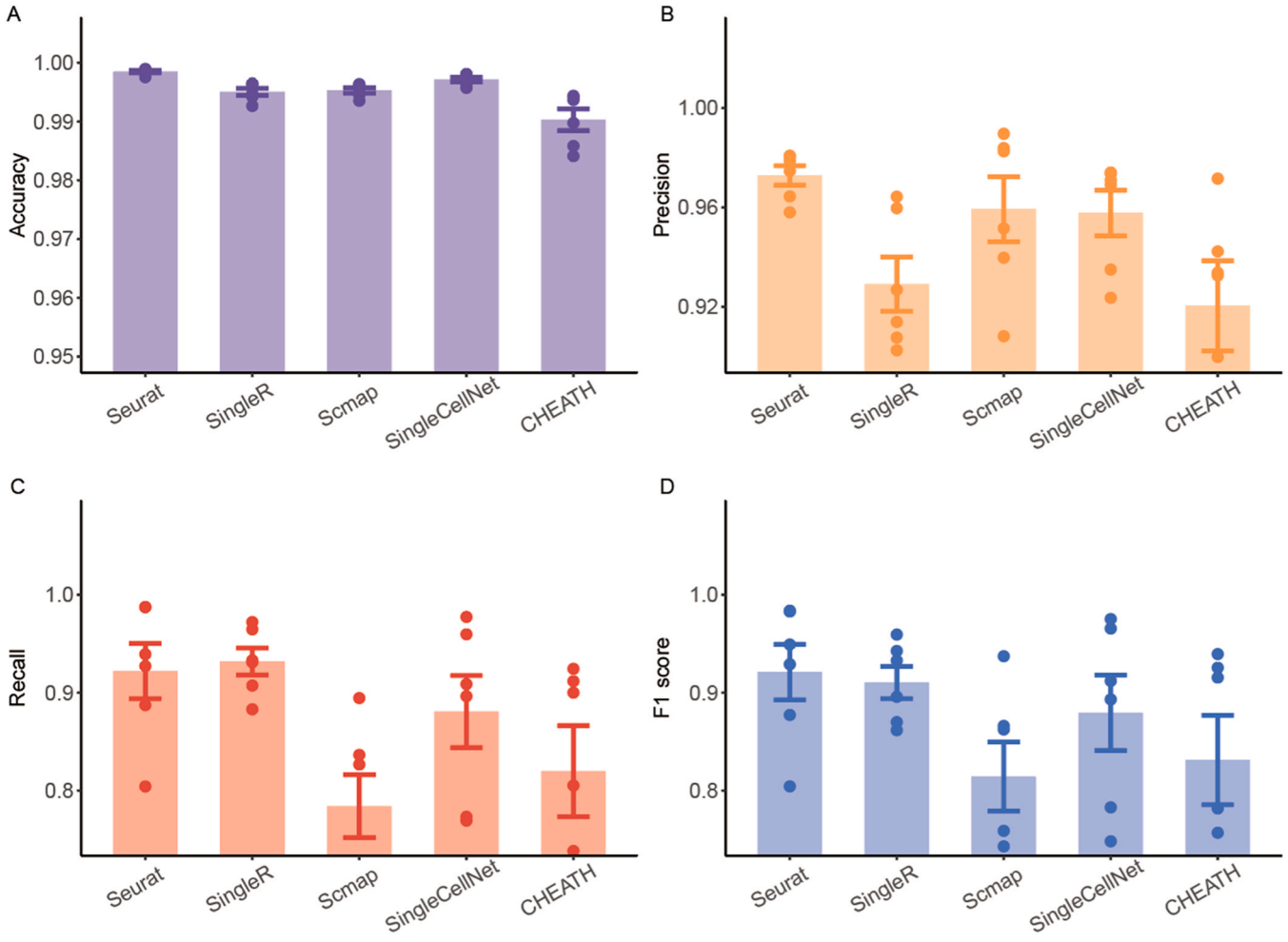

Fig. 3. Comparison of supervised methods for cell subclass classification. Evaluation indicators include (A) accuracy, (B) precision, (C) recall, and (D) F1 score.

2.6. Selection of marker gene identification methods

In BrainCellR, the identification of marker genes involves two substeps: identifying differentially expressed (DE) genes within the same cell subclass; and selecting marker genes from these differentially expressed genes. We evaluated four differential expression gene identification methods: Wilcoxon, ROC curve, T-test, and linear regression, as implemented and described in Seurat [22]. These methods were combined with three marker gene screening methods: 1) Sorting by Specific Score: The top three genes are sorted based on a specific score, which represents their expression specificity across clusters; 2) Top 10 % Ranking: Genes sorted by specific score are required to be ranked in the top 10 % of all expressed genes; 3) P-value Selection: The top three genes are selected based on the p-value identified by the differential gene detection algorithm. We also explored additional approaches for marker gene detection based on Machine Learning: Random forest [42], PCA [43], and Node2Vec+CNN [44]. Finally, we consider whether to use cells or pseudo-cells as an input. We therefore evaluated two methods based on pseudo-cells: Processing the input data by randomly selecting 10 cells of the same cell type and averaging their expression levels to create pseudo-cells; and the hdWGCNA method [45]. We tested the performance of dozens of methods by combining various input data processing approaches (Raw data, Pseudo-cell, Hdwgcna-generated), four distinct methods for identifying differential gene expression, and multiple marker gene selection methods. We found that the best performance was achieved when using pseudo-cell data as input, identifying differentially expressed genes with ROC, and then selecting the top 10 % of highly expressed genes ordered by specific score(Fig. 4). We incorporate this combination of methods into our pipeline. For cell types where differentially expressed genes could not be identified by the ROC method, we utilized Wilcox test, *t*-test, and linear regression which are ranked by our evaluation result to identify DE genes and extract marker genes. By incorporating these methods into our pipeline, we ensured to empirically select the most accurate approach for marker genes identification, which is essential for distinguishing and characterizing specific cell types.

2.7. Detailed method for identifying marker gene

The Wilcoxon, ROC, t-, and linear regression tests used to identify differentially expressed genes were the methods provided by Seurat package [22]. Specific score is used to measure the degree of specificity of genes, and the formula is as follows:

$$Score = \frac{\sum_{i=1}^{n} 1 - \frac{MED(y_i)}{MED(y_c)}}{n-1} \times PER_{y_{i=0}} \times PER_{y_c!=0}$$

Where, c represents the cell type which currently concerned, i represents other cell types except for c, MED represents the median expression level, and $PER_{y_c!=0}$ refers to the proportion of cells where the gene expression is non-zero, while $PER_{y_i=0}$ represents the proportion of cells where the gene expression is zero.

The input data for both the random forest and PCA consist of expression matrices. Marker genes are identified based on the feature extraction ability of the model. For the random forest, genes are sorted according to Gini importance [46], while for PCA, genes are sorted based on their ranking on PC1. The input for the Node2Vec+CNN method is the co-expression matrix of expression data between each gene, and the objective of the training is to classify the input genes. The training set comprises the marker genes identified after integrating the two sets of data. Since Node2Vec [44] does not limit the number of co-expressed genes that can be imported, we used the correlation coefficients among all differentially expressed genes for classification.

Fig. 4. Comparison of the percentage of common cell types between Dataset1 and Dataset2 using different marker gene identification methods. Rank index is provided for each method.

2.8. Identifying the same cell type

In our pipeline, the name of the cell type is composed of the major class of the cell type, the subclass of the cell type, and the top marker genes of the cell type. If the three marker genes of one cell type overlap with the three marker genes of another cell type within the same subclass, these two cell types are considered to be the same cell type.

3. Results

The BrainCellR pipeline can be divided into three steps (Fig. 1). The first step involves clustering single-cell data at a fine scale. The second step involves supplying the clusters to a supervised cell type classifier, which outputs major and subclass cell types. The third step is the selection of marker genes from the identified differentially expressed genes. Each step of the pipeline is systematically evaluated to select the most appropriate approach for cell type nomenclature as shown in methods section. The final cell label is then constructed by combining the major class and subclass annotations derived from the classifier with the top three marker genes associated with each cell type.

The single-cell data types obtained from different biological samples, different individuals, or different sequencing platforms often vary, presenting a challenge in single-cell classification. To evaluate the effectiveness of our pipeline, we conducted studies using (a) six datasets to assess the consistency among different biological data [1,2,17,47,48]; (b) a mouse dataset to assess the consistency among biological replications [2]; (c) a human dataset to assess the consistency across different individuals [48]; and (d) external mouse datasets [2], derived from various sequencing methods, showcasing its adaptability to technical variations. This diverse range of datasets allows for a thorough evaluation of BrainCellR's performance and applicability across species, developmental stages, and sequencing technologies, emphasizing its versatility and reliability in single-cell data analysis.

3.1. Identification of optimal parameters

In our study, we compared several widely used single-cell analysis methods and optimized their parameters to evaluate each method's performance. By comparing seven different analysis methods, including Seurat, Monocle3, SC3, etc. (Fig. 2), we found that Consensus1 and scrattch-hicat performed the best in terms of consistency and accuracy, particularly in cross-dataset cell type classification. Further optimization revealed that Consensus1 outperformed scrattch-hicat in computational efficiency, especially in terms of computing time (Fig. 2C) and memory usage (Fig. 2D).

Additionally, we systematically optimized the methods for differential gene expression detection and marker gene selection. After comparing ROC, Wilcoxon, and T-tests, we ultimately selected the ROC method to identify differentially expressed genes in pseudo-cell data, combining specificity scores to select marker genes from the top 10 % of highly expressed genes. This ensured the accuracy of marker gene selection.

3.2. Assessing consistency and comparability of cell types across diverse datasets

We aimed to evaluate whether cell types, labeled identically across different datasets, display similar expression patterns. To determine this similarity, we utilized four metrics: Euclidean distance, Spearman correlation, Pearson correlation, and Cosine similarity, comparing gene expression counts between pairs of cell types. Our analysis incorporated six datasets, which included both human and mouse data (Table S1) [1,2,17,47,48]. In five out of these six datasets, we found that the Euclidean distance between cell types with the same label was significantly smaller than the distance between cell types with different labels. A similar trend was observed for the other three metrics(Fig. 5). These results indicate that cell types identified using BrainCellR are consistent and comparable across different datasets, exhibiting significant similarities when labeled identically.

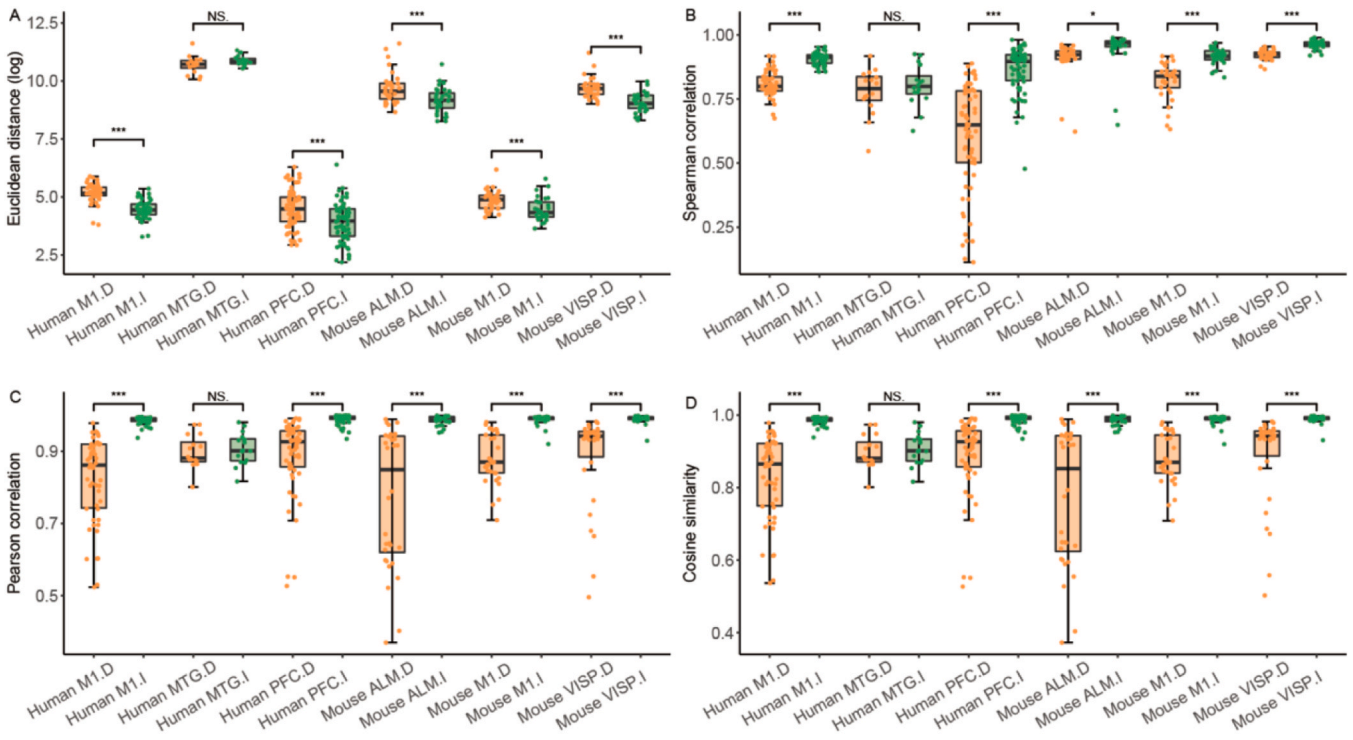

Fig. 5. Comparison of similarities across distinct biological datasets. (A) Euclidean distance comparison between identical and different cell types. (B) Spearman correlation comparison between identical and different cell types. (C) Pearson correlation comparison between identical and different cell types. (D) Cosine similarity comparison between identical and different cell types. '.I': identical cell types; '.D': different cell types.

3.3. Evaluating cell type consistency across biological replicate data

We evaluated cell type consistency in biological replication data. Using data from the primary motor cortex of mice (Table S2), we selected two sets of biological replicate data, containing 42,108 and 35,303 cells, respectively. We clustered each set independently, then proceeded with the classification of the major cell class and subclass, as well as marker gene selection using the BrainCellR pipeline with the Consensus1 and Seurat methods selected. As a result, we identified 62 cell types in one dataset and 60 cell types in the other, along with their marker genes(Fig. 6A). Remarkably, 57 cell types were found to be common between the two datasets, showcasing a high consistency level of 95 %. To exemplify this, we selected the Pvalb subclass for display (Table S3). Within the data annotated by our pipeline, the Pvalb subclass was further subdivided into seven clusters. Importantly, we observed that the distance between the two datasets for cells belonging to the same cell type was close compared to different cell types in the UMAP plot(Fig. 7). This finding indicates that BrainCellR is capable of effectively comparing cell types across different datasets, thus highlighting its utility for cross-dataset analysis.

3.4. Evaluating cell type consistency across different individual sources

Next, we evaluated the cell type consistency in different individual sources. We utilized data from the human middle temporal gyrus (MTG) as our test dataset (Table S2), and segregated the cells into two subsets based on the individuals they originated from [48]. These two subsets consisted of 7206 cells and 7421 cells, respectively. After conducting marker gene screening using BrainCellR pipeline, we successfully identified a total of 50 and 63 cell types. Remarkably, 43 of these cell types (86 %) were found to be consistent across both subsets (Fig. 6B).

3.5. Evaluating cell type consistency across different sequencing techniques

Then, we evaluated cell type consistency across data from different sequencing techniques. We conducted a test to examine the consistency of cell type identification within our pipeline across data from high-noise sequencing technology and various batches of experiments [2]. Specifically, we utilized data from 122,641 cells obtained from the primary motor cortex of mice using 10X v2 single-cell sequencing, as well as data from 76,525 cells acquired through 10X v3 single nuclei sequencing (Table S2). The analysis of the 10X v2 single-cell data revealed 77 distinct cell types.

while the 10X v3 single-nuclei data comprised 56 cell types, with an overlap of 44 cell types (79 %). It's important to note that deviations in cell extraction due to variations in sequencing techniques and cell states, as well as inconsistencies in the number of cells between the datasets, could potentially impact the extraction of certain cell types from the 10X v3 single-nuclei data. In comparison, according to the original data annotations, the two datasets were classified into 90 and 67 categories, with only 32 categories intersecting, resulting in an intersection ratio of merely 47.76 %(Fig. 6C).

3.6. Comparative analysis of cell types across developmental stages in the mouse somatosensory cortex

To test the applicability of our method to developmental data, we analyzed single-cell RNA sequencing data from the mouse somatosensory cortex [49] across several key developmental stages (E13.5, E14.5, E15.5, and E16.5). These stages represent critical periods during which progenitor cells give rise to various neuronal and glial subtypes (Table S4).

Our analysis revealed that Apical Progenitors, serving as early stem cell types, were consistently identified across all developmental stages from E13.5 to E16.5. These progenitor cells play a continuous role in generating new cells during brain development. Genes associated with proliferation and differentiation, such as *Pcna-ps2* and *Kif18b*, were highly expressed at E15.5, indicating an active proliferative state. This finding aligns with previous studies showing that apical progenitors produce large numbers of neurons and glial cells during early brain development. The generation of excitatory neurons became particularly notable after E14.5, especially at E15.5 and E16.5, where their migration was clearly observed. Marker genes such as *Sema6d* and *Sp9* were highly expressed in migrating neurons at E15.5, suggesting active differentiation from progenitor cells and migration to designated cortical

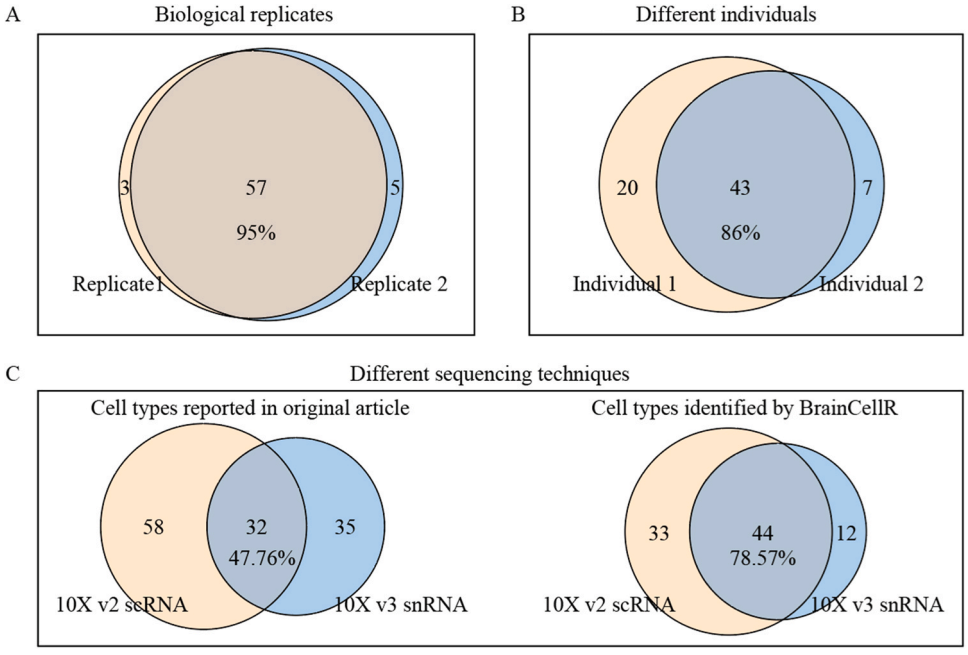

Fig. 6. Comparison of cell types from different replicates, individuals, and sequencing techniques. (A) Comparison between different biological replicates, (B) different individuals, and (C) different sequencing techniques.

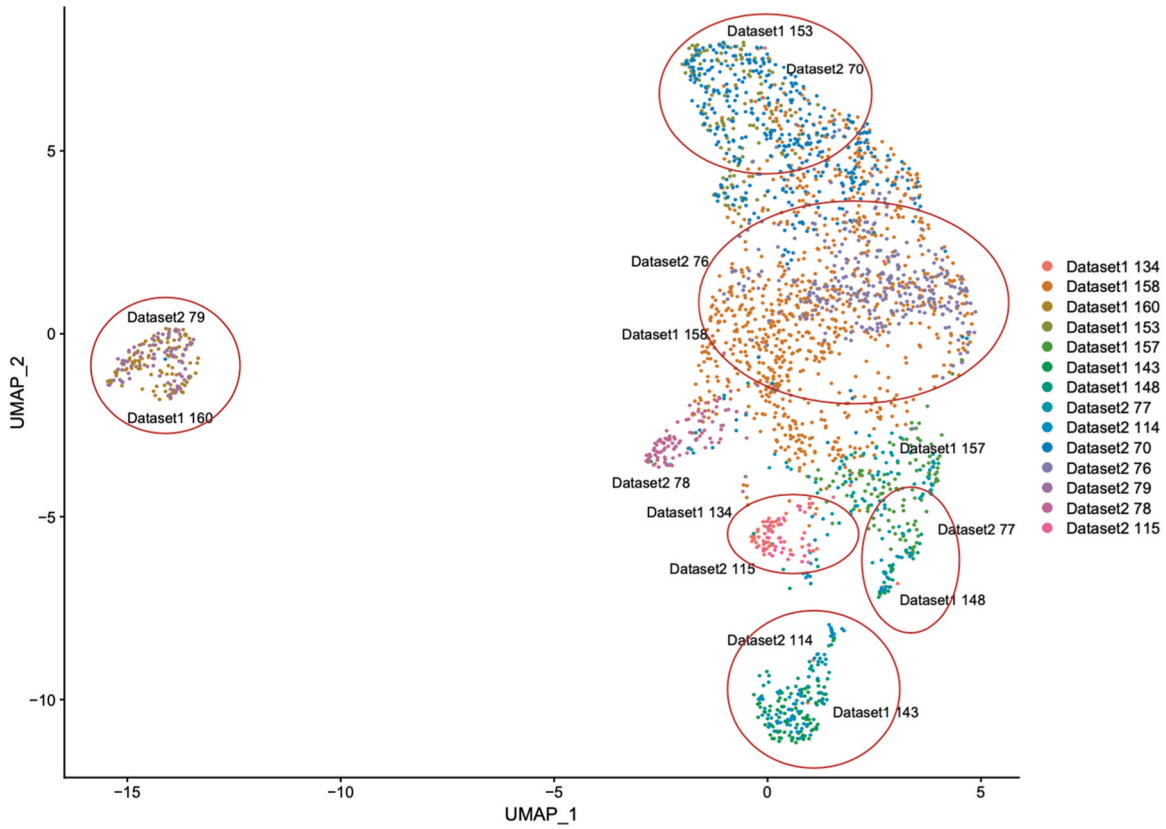

Fig. 7. UMAP plot for Pvalb subclass of two combined datasets. The red circles indicate the cell types that correspond between the two datasets, and the cells that are not indicated by the red circles are the cell types that are independent of each of the two datasets.

layers.

Lastly, the generation of interneurons was observed at E15.5 and E16.5, demonstrating their gradual emergence. Inhibitory neurons, which typically develop later than excitatory neurons, showed expression of marker genes such as *Gsx2* and *Dlk1*, indicating differentiation and migration to functional regions of the cortex. This gradual developmental pattern aligns with the known role of inhibitory neurons in regulating local circuits during the later stages of cortical formation.

4. Discussion

Understanding the diversity of cell types in the brain is a fundamental pursuit in neuroscience research. To facilitate this exploration, we have developed BrainCellR, a powerful pipeline that enables automated classification and nomination of cell types from single-cell transcriptome data. BrainCellR leverages marker genes for cell type annotation, standardizing nomenclature at the cluster level to ensure consistency and high computational efficiency. This approach facilitates comparable cell type annotations across diverse datasets, enabling comprehensive investigations into the complex cellular landscape of the brain. In addition, BrainCellR has the potential to discover new cell types, making it particularly useful for exploring datasets where novel or rare cell types may be present, and providing unique opportunities to expand our understanding of cellular diversity. Please note that the comparison of cell types across datasets is influenced by the number of clusters within the cell subclass. This suggests that our pipeline is better suited for datasets with deep sequencing, a large number of cells, and extensive sampling. BrainCellR holds great promise in advancing our understanding of brain cell diversity and its functional implications. More importantly, BrainCellR can be applied to other tissues with complex compositions of cell types, not only just to the brain.

Author contributions

G.W., and S.M. designed the study; G.W., S.M., and Y.C. wrote the manuscript. Y.C. wrote the package and documentation.

Funding

This work is supported by grants from the National Natural Science Foundation of China (Grant Nos. 81827901 and 32170567).

CRediT authorship contribution statement

Yuhao Chi: Visualization, Methodology, Investigation, Formal analysis, Data curation. **Guang-Zhong Wang:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Simone Marini:** Writing – review & editing, Supervision, Methodology.

Declaration of Competing Interest

The authors declare no competing interests.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.11.038.

Data availability

BrainCellR is freely available at https://github.com/WangLab-SINH/BrainCellR.

References

[1] Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature 2021;598:111–9.
[2] Yao Z, Liu H, Xie F, Fischer S, Adkins RS, Aldridge AI, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature 2021;598:103–10.
[3] Wang W, Wang GZ. Understanding molecular mechanisms of the brain through transcriptomics. Front Physiol 2019;10:214.
[4] Chi Y, Qi R, Zhou Y, Tong H, Jin H, Turck CW, et al. scBrainMap: a landscape for cell types and associated genetic markers in the brain. Database (Oxf) 2023;2023.
[5] Hu G, Li J, Wang GZ. Significant evolutionary constraints on neuron cells revealed by single-cell transcriptomics. Genome Biol Evol 2020;12:300–8.
[6] Li J, Wang GZ. Application of computational biology to decode brain transcriptomes. Genom Proteom Bioinforma 2019;17:367–80.
[7] Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq data. Genom Proteom Bioinforma 2021;19:267–81.
[8] Mu Q, Chen Y, Wang J. Deciphering brain complexity using single-cell sequencing. Genom Proteom Bioinforma 2019;17:344–66.
[9] Zeng H. What is a cell type and how to define it? Cell 2022;185:2739–55.
[10] Michielsen L, Lotfollahi M, Strobl D, Sikkema L, Reinders MJT, Theis FJ, et al. Single-cell reference mapping to construct and extend cell-type hierarchies. NAR Genom Bioinform 2023;5:lqad070.
[11] Zhang Y, Aevermann BD, Bakken TE, Miller JA, Hodge RD, Lein ES, et al. FR-Match: robust matching of cell type clusters from single cell RNA sequencing data using the Friedman-Rafsky non-parametric test. Brief Bioinform 2021;22.
[12] Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun 2022;13:1246.
[13] Pei G, Yan F, Simon LM, Dai Y, Jia P, Zhao Z. deCS: a tool for systematic cell type annotations of single-cell RNA sequencing data among human tissues. Genom Proteom Bioinforma 2023;21:370–84.
[14] Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 2016;19:335–46.
[15] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015;347:1138–42.
[16] Greig LC, Woodworth MB, Galazo MJ, Padmanabhan H, Macklis JD. Molecular logic of neocortical projection neuron specification, development and diversity. Nat Rev Neurosci 2013;14:755–69.
[17] Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 2018;563:72–8.
[18] Xu Y, Baumgart SJ, Stegmann CM, Hayat S. MACA: marker-based automatic cell-type annotation for single-cell expression data. Bioinformatics 2022;38:1756–60.
[19] Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of deep learning on single-cell RNA sequencing data analysis: a review. Genom Proteom Bioinforma 2022;20:814–35.
[20] Ren X, Zheng L, Zhang Z. SSCC: a novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. Genom Proteom Bioinforma 2019;17:201–10.
[21] Gonzalez-Ferrer J, Lehrer J, O'Farrell A, Paten B, Teodorescu M, Haussler D, et al. SIMS: a deep-learning label transfer tool for single-cell RNA sequencing analysis. Cell Genom 2024;4:100581.
[22] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–87. e29.
[23] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, 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-d6.
[24] Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxf) 2019;2019.
[25] Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 2016;352:1586–90.
[26] Hammond TR, Dufort C, Dissing-Olesen L, Giera S, Young A, Wysoker A, et al. Single-cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. Immunity 2019;50:253–71. e6.
[27] Miller JA, Gouwens NW, Tasic B, Collman F, van Velthoven CT, Bakken TE, et al. Common cell type nomenclature for the mammalian brain. Elife 2020;9.
[28] Andreatta M, Berenstein AJ, Carmona SJ. scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets. Bioinformatics 2022;38:2642–4.
[29] Stanojevic S, Li Y, Ristivojevic A, Garmire LX. Computational methods for single-cell multi-omics integration and alignment. Genom Proteom Bioinforma 2022;20:836–49.
[30] Liu B, Li C, Li Z, Wang D, Ren X, Zhang Z. An entropy-based metric for assessing the purity of single cell populations. Nat Commun 2020;11:3155.
[31] Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019;566:496–502.
[32] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483–6.
[33] Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol 2022;23:49.
[34] Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular diversity and specializations among the cells of the adult mouse brain. Cell 2018;174:1015–30. e16.
[35] Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular architecture of the mouse nervous system. Cell 2018;174:999–1014.e22.
[36] Hodge RD, Miller JA, Novotny M, Kalmbach BE, Ting JT, Bakken TE, et al. Transcriptomic evidence that von Economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons. Nat Commun 2020;11:1172.
[37] de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 2019;47:e95.
[38] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 2018;15:359–62.
[39] Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. Cell Syst 2019;9:207–13. e2.
[40] 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.
[41] Sun X, Lin X, Li Z, Wu H. A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. Brief Bioinform 2022;23.
[42] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[43] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst 1987;2:37–52.
[44] Grover A., Leskovec J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016:855–864.
[45] Morabito S, Miyoshi E, Michael N, Shahin S, Martini AC, Head E, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. Nat Genet 2021;53:1143–55.
[46] Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinforma 2009;10:1–16.
[47] Ma S, Skarica M, Li Q, Xu C, Risgaard RD, Tebbenkamp ATN, et al. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. Science 2022;377:eabo7257.
[48] Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature 2019;573:61–8.
[49] Di Bella DJ, Habibi E, Stickels RR, Scalia G, Brown J, Yadollahpour P, et al. Molecular logic of cellular diversification in the mouse cerebral cortex. Nature 2021;595:554–9.