Precision and Accuracy of Single-Cell/Nuclei RNA Sequencing Data

Rujia Dai[1], Ming Zhang[2], Tianyao Chu[2], Richard Kopp[1], Chunling Zhang[3], Kefu Liu[2], Yue Wang[4], Xusheng Wang[5,6], Chao Chen[2,7,8*], Chunyu Liu[1,2,3*]

[1]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA

[2]MOE Key Laboratory of Rare Pediatric Diseases & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, and Department of Psychiatry, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China

[3]Department of Neuroscience & Physiology, SUNY Upstate Medical University, Syracuse, NY, USA

[4]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, VA, USA

[5]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

[6]Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN, USA

[7]Furong Laboratory, Changsha, Hunan, China

[8]Hunan Key Laboratory of Animal Models for Human Diseases, Central South University, Changsha, China

Abstract

Single-cell/nuclei RNA sequencing (sc/snRNA-Seq) is widely used for profiling cell-type gene expressions in biomedical research. An important but underappreciated issue is the quality of sc/snRNA-Seq data that would impact the reliability of downstream analyses. Here we evaluated the precision and accuracy in 18 sc/snRNA-Seq datasets. The precision was assessed on data from human brain studies with a total of 3,483,905 cells from 297 individuals, by utilizing technical replicates. The accuracy was evaluated with sample-matched scRNA-Seq and pooled-cell RNA-Seq data of cultured mononuclear phagocytes from four species. The results revealed low precision and accuracy at the single-cell level across all evaluated data. Cell number and RNA quality were highlighted as two key factors determining the expression precision, accuracy, and reproducibility of differential expression analysis in sc/snRNA-Seq. This study underscores the necessity of sequencing enough high-quality cells per cell type per individual, preferably in the hundreds, to mitigate noise in expression quantification.

Introduction

Sc/snRNA-Seq is a powerful technology developed for measuring gene expression in individual cells. The first scRNA-Seq study was published in 2009 by Tang et al[1]. Smart-seq was developed, enabling the amplification and sequencing of full-length mRNA transcripts from individual cells, characterizing transcriptomes at single-cell resolution. Since then, more technologies have been developed for single-cell profiling[2], with 10X chromium and Smart-seq being the two most commonly used methods.

Sc/snRNA-Seq has been used in various applications, including identifying novel transcriptional regulatory mechanisms[3], characterizing cell types and tissue compositions[4], studying developmental dynamics and trajectory of different cell types[5,6], and identifying cell-type-specific changes as biomarkers for disease or treatment responses[7-9]. All these studies rely on accurate and precise measures of gene expression in each cell type. Accuracy refers to consistency between the measured expression and corresponding true expression, which is indicative of bias in measurement, while precision refers to the consistency among repeated measurements. Only when gene expression is quantified precisely and accurately in each sample, can the results of downstream analyses be reproducible and meaningful.

Random and systematic technical variability adds noise to the expression measures in sc/snRNA-Seq[10]. Many zero values are observed in sc/snRNA-Seq data, called ‘dropouts’[11]. Dropouts can be caused by target genes truly not being expressed, or due to technical error caused by factors such as low mRNA input, mRNA degradation, capture efficiency, amplification efficiency, and sequencing depth. Technical errors can add noise to precision and bias to accuracy. Previous studies attempted to quantify technical noise in scRNA-Seq data using Spike-ins[12] or sample-matched bulk-tissue RNA-Seq data[13] as references. However, these two methods have rarely been used due to practical limitations.

We evaluated the precision and accuracy of expression measures with 18 published sc/snRNA-Seq datasets produced on three different platforms (Table 1) in the framework as illustrated in Fig. 1. These datasets were from large consortium efforts in the field, such as BRAIN Initiative Cell Census Network (BICCN), which represent current technology standards. Initially, we surveyed the cell numbers and missing rates in these sc/snRNA-Seq data, followed by calculating precision in each dataset. Additionally, we explored the impact of several technical factors, including RNA quality, saturation rate, total reads, and sequencing platform on expression precision. We also evaluated the expression accuracy with four datasets of cultured mononuclear phagocytes from sample-matched pooled-cell RNA-Seq and scRNA-Seq data. Lastly, we evaluated the effect of cell number and other factors on the reproducibility of downstream differential expression analysis.

Results

Existing sc/snRNA-Seq data have high missing rate

We measured the missing rate for each gene at both the individual-cell and pseudo-bulk levels. Pseudo-bulk aggregates were created from single-cell gene expression of a specific cell type within an individual to mimic bulk RNA-Seq data. The missing rate was defined as the proportion of cells with zero expression for a given gene across all individual cells or pseudo-bulks of the same cell type. Individual cells had an average missing rate of 90% (Fig. 2A), while the pseudo-bulks reduced the average missing rate to 40% (Fig. 2B). Including more cells in the pseudo-bulks resulted in a lower observed missing rate (Fig. S1). The use of pseudo-bulk addressed the missing data problem to a good degree.

Though each project sequenced many cells, we noticed that the number of cells sequenced per cell type per individual were actually very small, particularly for minor cell types. Across the 14 brain datasets, the average total cell count was 247,190, whereas the average cell number per

individual was 34,483 (Fig. 2C). The number was even much smaller for specific cell types per individual. For instance, the BICCN_HVS study sequenced 353,194 cells and categorized 24 cell classes (Fig. 2D). The largest group of cells in this data comprised an average of 1,021 intratelencephalic (IT) neurons from layers 2 and 3, while the smallest group had only an average of 4 Somatostatin (SST) Chodl inhibitory neurons across the samples, a difference of three orders of magnitude.

Low expression precision in sc/snRNA-Seq data

Expression precision was evaluated by the expression variability across technical replicates in sc/snRNA-Seq data. First, we generated pseudo-technical replicates by randomly grouping cells of the same type from the same individual into three groups and totaling expression values of each gene from all cells within each group (Fig. S2). We then calculated the coefficient of variation (CV) for each gene to measure the variability of gene expression across the pseudo-technical replicates in each cell type. To avoid sampling bias, we calculated the CV 100 times and used the averaged CV to represent the overall precision in the data.

Our analysis revealed that the median CV of detected genes across pseudo-technical replicates was 0.68 ± 0.24 in the 14 brain datasets (Fig. 3), which is much higher than the median CV observed in bulk-tissue RNA-Seq[14] (ranging from 0.11 to 0.39) and microarray data[15] (ranging from 0.1 to 0.2). Utilizing cell classification in the BICCN_HVS study, we calculated the CV values at both cell type and subtype levels. The CVs were not significantly different at these two resolution levels, suggesting the observed variability was not driven by heterogeneity in a higher-level cell classification (Fig. S3). A similar pattern was noted in independent mouse brain data, indicating that low precision is a technical challenge in single-cell data across various sample sources (Fig. S4).

Major cell types exhibited lower CV than minor cell types. For example, excitatory neurons, as the most abundant cell type, had a CV of 0.19±0.20 across datasets. In contrast, the other cell types had a median CV of 0.55±0.40 across datasets, indicating that the precision problem is particularly severe for low-abundance cell types. Additionally, expression CV was negatively correlated with expression abundance (Correlation coefficient=-0.88, p-value<2.2e-16, Fig. S5A). Notably, marker genes have lower CV than other genes (Fig. S5B).

We calculated the percentage of samples achieving a designated precision threshold, a CV of 0.1 or lower for each cell type. There was a striking disparity: only 3% to 25% of samples from five distinct datasets satisfied this precision criterion, as illustrated in Fig. S6A. For example, every sample representing upper-layer IT neurons in the BICCN adult dataset successfully passed the precision assessment (Fig. S6B). In the case of the BICCN_HVS dataset, 67% of samples pertaining to IT neurons in layers 2 and 3 met the established quality benchmarks (Fig. S6C). Conversely, in the other nine datasets, not a single sample reached the requisite levels of precision. This indicates a prevalent problem with gene expression noise in individual samples of these datasets.

Gene expression precision is correlated with number of cells sequenced

We expected that the expression precision would be associated with the number of cells sequenced. To prove the expectation by actual data, we generated technical replicates with varying numbers of cells, ranging from single cell to the maximum cell number divided by three. The sample with the largest number of total cells in each dataset was utilized for testing. As the number of cells pooled into the technical replicates increased, the overall variability decreased until it reached a plateau for major cell types (Fig. 4A and 4C). With the small total number of cells sequenced, the minor cell types did not reach a stable CV (Fig. S7). Similar correlation

coefficients of -0.66±0.07 and -0.78±0.14 were observed between the number of cells in each replicate and median CV in excitatory neuron and oligodendrocyte respectively (p-value<0.05, Fig. 4B and 4D).

The minimum number of cells required for delivering acceptable precision is suggested by data of excitatory neurons. Based on five datasets (BICCN_adult, BICCN_dev, BICCN_trimester1, M1_10X, and BICCN_HVS), approximately 500 cells were required to achieve a median CV close to 0.1 for neurons (Fig. 4A). None of the other cell types attained CV values as low as 0.1 and they all had fewer than 500 cells sequenced.

RNA integrity is correlated with expression precision

The cell numbers required for achieving an acceptable precision level in sc/snRNA-Seq data vary across studies, suggesting expression precision may not be solely dependent on the number of cells. We examined the effects of four technical factors, including RNA integrity, sequencing depth, sequencing saturation, and sequencing platform, on expression precision of excitatory neurons.

We tested the relationship between RNA integrity, as measured by the RNA integrity number (RIN), and median CV in technical replicates. Two datasets with RIN information available for analysis were used. Samples with higher RIN value tended to have lower CV values (Fig. 5A and Fig. 5B). By zooming into replicates with 200 cells, negative correlations were observed between RIN and median CV in the ROSMAP (R^2=0.26, p-value=0.06, Fig. 5C) and autism_PFC (R^2=0.60, p-value=0.04, Fig. 5D) datasets, suggesting that RNA integrity is another factor contributing to expression precision.

In the autism_PFC data, we also explored the correlation between median CV and total sequencing depth (p-value=0.89) and saturation rates (p-value=0.76), but no significant

correlation was found (Fig. S8A, Fig. S8B). We also compared expression variability across technical replicates in data generated from two different sequencing platforms, 10X Chromium (autsim_PFC) and Smart-seq (MTG). The median CV across detected genes in replicates constructed by 200 cells was used for comparison. No significant difference in gene expression variability was observed between the two technologies (p-value=0.56, Wilcoxon signed-rank test, Fig.S7C), indicating the precision problem is not unique to a specific sequencing platform.

Low expression accuracy in scRNA-Seq data associated with the number of cells sequenced

To evaluate the accuracy of gene expression, we compared pooled-cell RNA-Seq data with single-cell RNA-Seq data of cultured mononuclear phagocytes from matched samples (Fig. 6A). The gene expressions from pooled-cell RNA-Seq were considered as the ground truth. We used Pearson correlation and linear regression to assess the expression accuracy. In the linear regression model, the ground truth was treated as the independent variable, while the pseudo-bulk from sample-matched scRNA-Seq was the dependent variable. We tested the significance of the slope deviating from one. The significance of the linear regression, combined with the Pearson correlation coefficient, was used to measure expression accuracy. We calculated the expression accuracy independently for each of the four species. To illustrate the relationship between the number of cells and expression accuracy, we performed downsampling experiments, ranging from 1,000 to 1 cell for each sample.

The number of genes with good accuracy decreased in down sampling (Fig. 6B). We observed 3,450 out of 13,907 detected genes with good accuracy as defined by criteria of regression slope of 1 (p-value of 0.05) and correlation coefficient of 0.9 when 1000 cells were analyzed for each sample in mouse data. When each sample contained a single cell, only 100 genes

showed good accuracy. When data have 500 cells in each sample, the gene accuracy tends to reach a stable value. Similar patterns were observed in data from rat, pig, and rabbit, though pig and rabbit data showed overall worse performance than mouse and rat data.

Effect of cell number, effect size of differential expression, and expression level on the reproducibility of differential expression analysis in scRNA-Seq data

To assess the impact of data quality on downstream analysis, we conducted a differential expression (DE) analysis in sample-matched scRNA-Seq and pooled-cell RNA-Seq datasets independently, comparing Lipopolysaccharide (LPS)-treated and untreated groups using the edgeR algorithm. Genes were considered significantly differentially expressed (DEGs) when their false discovery rate (FDR)-corrected p-value was less than 0.05. Since we already showed that both precision and accuracy were positively correlated with the number of cells sequenced, we employed a downsampling strategy to investigate the influence of cell number on DE results. By utilizing the DE results in pooled-cell RNA-Seq data as ground truth, we evaluated the overall reproducibility of DE results in scRNA-Seq data with true positive rate, the proportion of actual positive instances that are correctly identified as positive. Notably, as the number of cells increased, true positive rate improved and had a plateau at about 500 cells (Fig. 7A). The true positive rates were 0.72, 0.63, 0.62, 0.44 in data from mouse, rat, pig and rabbit, when 500 cells were included in each sample.

We also investigated the impact of effect size and expression levels on the reproducibility of DE results (Fig. 7B). Effect size, which reflects the differences between the compared groups, plays a crucial role in determining the statistical power of a DE analysis. Using mouse data as an example, our findings suggest that DEGs with larger effect sizes (logFC>1) exhibit greater reproducibility compared to those with smaller effect sizes (logFC<=1, Fig. 7C). For example,

when 500 cells were included in each sample, DEGs with large effect sizes (over two-fold changes) had a true positive rate of 0.73, whereas DEGs with small effect sizes had a true positive rate of only 0.38. When only 50 cells were included in each sample, the true positive rates were 0.41 and 0.09 for DEGs with large and small effect sizes, respectively. This trend was consistent with DEGs showing higher expressions tending to have better reproducibility (Fig. 7D).

Discussion

The use of sc/snRNA-Seq in biological studies has become a common practice, necessitating meticulous evaluation of data quality. The current investigation assesses the precision and accuracy of published sc/snRNA-Seq data. By analyzing 18 representative datasets, we demonstrated that the gene expressions per individual measured for most cell types are of low precision and accuracy. We found a robust correlation between the number of cells sequenced and the precision, accuracy, and reproducibility of downstream DE analysis. Only cell types have enough cells sequenced delivered relatively accurate and precise quantification of gene expression and, consequently, credible results of downstream analyses for case-control comparisons, QTL mapping, cell classifications, and others.

Our study advances existing efforts in evaluating scRNA-Seq data quality by quantitatively assessing expression precision and accuracy at the gene level, which offers a deeper understanding of sc/snRNA-Seq data and its representation of cell types. Previous studies have explored technical noise and expression accuracy in scRNA-Seq methods using spike-ins[12], qPCR[16], and RNA-Seq[13] of cell populations as benchmarks. However, two significant limitations have constrained the breadth and strength of prior evaluations. Firstly, some previous studies often relied on qualitative evaluations rather than quantitative assessments. For example,

previous research reported that 42.5% of genes in bulk tissue RNA-Seq could be detected by the scRNA-Seq C1 platform[16]. However, the reliability of the expressions of these detected genes was not evaluated, limiting the assessment of downstream analytical results. Secondly, some evaluations were conducted at the sample or cell level, not at the gene level[12,13]. The sample-level evaluation metrics can be biased by the highly-expressed genes. For instance, Wu et al.[16] reported a correlation coefficient exceeding 0.8 between scRNA-Seq and qPCR expression, but a small number of genes with high expressions drove this overall "high" accuracy in their sample-level evaluation. Our study overcomes these limitations by evaluating gene precision and accuracy on a gene-by-gene basis, providing a quality metric for each gene. These quality metrics allow for the determination of the proportion of qualified genes and inference of the overall sample quality.

The criteria used for evaluating precision and accuracy in this study are standard statistical techniques. We used CV < 0.1 as the cutoff for the precision in this evaluation. This is based on previous quality evaluations of bulk RNA-Seq data[15]. We believe that holding sc/sn RNA seq data to the same standard as bulk RNA seq is appropriate and lowering the standard will lead to noisy results and poor reproducibility. We combined correlation test and linear regression test to assess accuracy. In this composite criterion, the correlation test captures distance from ground truth, while the linear regression test assesses the bias based on the fitness to an expected linear relationship. However, when constructing the pseudo-technical replicates, we assumed homogeneity within one cell type. Such an assumption can be violated by heterogeneity caused by cell subtypes and states, which may explain the minimum CV that is observed. Our results indicated that the cell subtype was not the major cause of poor precision in the cell types we evaluated, as the subtype data does worsen the precision problem.

The low precision and accuracy in sc/snRNA-Seq data raise significant concerns regarding the findings reported based on sc/snRNA-Seq technology. All 18 examined datasets demonstrated

low precision and accuracy in technical replicates, indicating a high level of both noise and bias in expression quantification in sc/snRNA-Seq studies, especially when small numbers of cells were sequenced. From these findings, two pivotal issues come to the forefront. First, the utilization of data of individual cells may significantly exacerbate errors in downstream analyses. This concern aligns with the findings of several previous studies[17-19]. The second the issue is about false findings in downstream analyses based on low-quality data. Most of the samples in published data fall short of meeting the proposed quality standard, a precision threshold of 0.1. Consequently, the reliability of population-level analyses based on these data, such as DE analysis and eQTL mapping, is questionable. By performing DE analysis on the data from sample-matched scRNA-Seq and pooled-cell RNA-Seq experiments, we observed a marked decrease in the reproducibility of DEGs as the number of sequenced cells was reduced. When 10 cells were sequenced from each sample, the average true positive rate of DEGs was 0.38 in four datasets. This low reproducibility highlights the sensitivity of DE analysis to the number of cells analyzed, further underscoring the importance of high-quality data in obtaining robust and reproducible results, especially for minor cell types.

The quality of sc/snRNA-Seq data can be notably enhanced by sequencing a larger number of cells. In addition to the observed inverse relationship between the number of cells and data quality, we also noted that the precision, accuracy, and reproducibility of DE results tend to stabilize when the cell count reaches around 500 cells. This observation suggests that data quality could be significantly improved by increasing the number of sequenced cells.

RNA quality is another crucial factor impacting expression precision. Notably, in the BICCN_HVS study[20], the minimum number of cells required to achieve a CV of 0.1 was 500, which is lower than the other datasets. A key factor contributing to this difference may be the use of surgical samples in the BICCN_HVS study, as these samples tend to be less degraded than frozen postmortem brain samples. Ensuring high RNA quality in samples, such as using

RNA with RIN values greater than 7, will likely reduce the number of cells required for quality quantification.

The signal-to-noise ratio emerges as a pivotal determinant of the reproducibility of DE analysis. Our investigation revealed that DEGs with large effect sizes exhibit superior reproducibility compared to those with smaller effect sizes. When 500 cells were included in each sample where the noise level was low, DEGs with large effect sizes had a true positive rate of 0.73, whereas DEGs with small effect sizes had a true positive rate of only 0.38. When the cell number decreased to 50 cells where the noise level was high, the true positive rates were 0.41 and 0.09 for DEGs with large and small effect sizes, respectively. This comparison indicates that the technical noise matters more for the smaller biological effects and technical noise may be manageable for phenotypes associated with pronounced expression changes. Improving data quality becomes more critical in scenarios where the effect size of the phenotype approaches the noise level. This is particularly relevant for many complex diseases, including neuropsychiatric disorders, where the effect size is typically small[21], necessitating an increase in the number of cells to minimize technical noise.

Our findings hold significant implications across multiple domains. They underscore the vital necessity of sequencing a wider array of minor cell types in the tissues. The accurate identification and comprehensive study of these cells necessitate sequencing more cells. Beyond elucidating the nuances of DE analysis, our results imply the impacts on other analytical methodologies, including cell classification[22], eQTL mapping[23], and the construction of co-expression networks[24]. For example, in a fetal brain study, a reduction in the number of cells sequenced has resulted in diminished accuracy in cell type classification[25].

Conclusion

In this study, we conducted an extensive evaluation of expression precision and accuracy within 18 representative sc/snRNA-Seq datasets. Our analysis uncovered a notable deficiency in the precision and accuracy of gene expression measurements, especially when sequencing a limited number of cells. The reproducibility of DE analysis is also tightly correlated with the number of cells. To improve the reproducibility of sc/snRNA-Seq study, we recommend to sequence more cells, with a particular emphasis on minor cell types and the number of cells per individual donor, using high-quality RNA preparations. We advocate for implementing quality assessment on sc/snRNA-Seq data before engaging in further downstream quantitative analyses. To support these efforts, we have developed VICE, a tool specifically designed to calculate technical variability of genes in sc/snRNA-Seq data. This tool is readily available to the public via GitHub (https://github.com/RujiaDai/VICE).

Methods

Collection of sc/snRNA-Seq data from cortex

A total of 14 brain sc/snRNA-Seq datasets were obtained for analysis. The collected datasets were derived from human brain studies published between 2012 and 2023[8,9,20,26-37]. Samples from individuals with brain disorders were excluded from the analysis to prevent biasing the expression profiles. The raw count and cell annotation data were obtained from the original studies. Due to differences in cell classification across studies, we harmonized cell identities into eight major cell types present in the adult brain, namely excitatory neurons, inhibitory neurons, oligodendrocytes, oligodendrocyte precursor cells, astrocytes, microglia, endothelial cells, and pericytes. The annotation of data from BICCN 2023 collection was retained to evaluate cell subtypes.

Collection of sample-matched data from four species

To assess expression accuracy, we utilized four sample-matched scRNA-Seq and pooled-cell RNA-Seq data. These datasets were sourced from Hagai et al.[38], encompassing bone marrow-derived mononuclear phagocytes derived from mouse, rat, pig, and rabbit, all subjected to stimulation with either lipopolysaccharide or poly-I:C for a duration of four hours. Within each species, a total of three samples received lipopolysaccharide treatment, while three additional samples were designated as control groups. We employed the preprocessed data provided by Squair et al.[18], which has been made available at https://doi.org/10.5281/zenodo.5048449.

Processing of sc/snRNA-Seq data

The sc/snRNA-Seq data underwent processing using Seurat version 4[39]. The raw count matrix and cell annotation matrix were used as input to Seurat. Gene expressed in more than 1% of cells were kept. The proportion of transcripts mapped to mitochondrial genes was calculated for each cell, and cells with 10% or more mitochondrial gene expression were removed to prevent the inclusion of dead cells. Additionally, cells with less than 200 detected genes or those with more than three standard deviations from the mean number of detected genes were excluded. The count data were normalized based on library size and were scaled with a factor of 10000. The normalized data were then log-transformed.

Marker gene identification

To identify marker genes in the ROSMAP dataset, we utilized a one versus second high strategy at both the cell and pseudo-bulk level. At the cell level, marker genes were identified using Seurat. Genes with a proportion of zero expression greater than 15% in the target cell type were removed prior to marker gene identification. The Wilcoxon signed-rank test was used to assess the expression difference, and genes with a log2FC greater than 1 and FDR-corrected p-value less than 0.05 were defined as marker genes. At the pseudo-bulk level, pseudo-bulks were constructed by aggregating gene expression for the same cell type from the

same individual. Marker genes were then tested using DESeq2[40], and the likelihood ratio test was utilized to evaluate the expression difference between the two cell groups. Marker genes with a log2FC greater than 2 and FDR-corrected p-value less than 0.05 were defined as marker genes at the pseudo-bulk level. Finally, marker genes supported by both cell- and pseudo-bulk-level tests were selected as final marker genes.

Technical replicate construction and CV calculation

To generate technical replicates for each cell type, cells in the count matrix were randomly divided into three groups by the same individual. The count expression for each gene was then summed within each group. The CV value was calculated for each gene using the following formula:

$$Formula\ 1: \quad CVi = \frac{sd(x)}{mean(x)}$$

, where x represents the gene expression of gene i across three replicates of a specific cell type. To ensure the robustness of technical replicates, the cell grouping and CV calculation were repeated 100 times, and the average CV across 100 samplings was used.

Difference of CV-cell-number relationship in data with cell classes and subclasses annotations

To compare the relationship between CV values and the number of sequenced cells in the class and subclass data, we used a two-sample Z-test. The null hypothesis was that the slope in the regression model testing the relationship between the number of cells and CV values was the same in the class and subclass data. We calculated the difference in slope between the two datasets using the following formula:

$$Formula\ 2: \quad diff = \frac{b1-b2}{\sqrt{se1^2+se2^2}}$$

, where b1 and b2 were the coefficients, and se1 and se2 were the standard errors from the

regression model in the class and subclass data, respectively. We then used the area of the standard normal curve corresponding to the calculated difference to determine the probability in a two-tailed manner.

Evaluation of expression accuracy

Expression accuracy was evaluated by Pearson correlation and linear regression model. The scRNA-Seq data was summarized by pseudo-bulks firstly. Then the batch effect between pseudo-bulks and pooled-cell RNA-Seq data was corrected by combat function in sva package[41]. Pearson correlation between sample-matched scRNA-Seq and pooled-cell RNA-Seq was calculated for each gene. In linear regression model, the expression in pooled-cell RNA-Seq was treated as independent variable and expression in scRNA-Seq was treated as dependent variable. The intercept of linear regression model was set to 0. By setting offset in function lm in R, the significance of slope deviating from 1 was tested. Good accuracy was defined as correlation coefficient over 0.9 and p-value of linear regression over 0.05. Since no single statistic is sufficient to describe accuracy, we proposed this composite criterion that capture the bias and distance from ground truth simultaneously.

DE analysis

DE analysis was carried out on both scRNA-Seq and pooled-cell RNA-Seq datasets to examine the expression disparities between samples treated with lipopolysaccharide and the control samples. In the case of scRNA-Seq data, DE analysis was performed on pseudo-bulk data using the likelihood ratio test approach provided by edgeR[42]. Genes exhibiting an FDR corrected p-value of less than 0.05 were classified as DEGs. We assessed the consistency between DE results obtained from single-cell and pooled-cell RNA-Seq with true positive rate which denotes the proportion of DEGs identified in pooled-cell RNA-Seq that were also replicated in the scRNA-Seq data.

Data availability

The data supporting the findings of this study are publicly available, with details for accessing the datasets provided in Table 1.

Acknowledgment

We thank all the participants in studies for making the data available. We also thank Sidney Wang for invaluable guidance and expertise in the experimental design and data analysis. This work was supported by NIH grants U01MH122591, U01MH116489, R01MH110920, U01MH103340, and R01MH126459; the National Natural Science Foundation of China 82022024 and 31970572; the science and technology innovation Program of Hunan Province 2021RC4018 and 2021RC5027. This work was supported in part by the Bioinformatics Center, Xiangya Hospital, Central South University.

Competing interests

The authors declare no competing interests.

Author contributions

R.D, C.L, and C.C conceived and designed the study. R.D, T.C, M.Z performed the experiments and analyzed the data. R.D and C.L wrote the manuscript. C.Z, R.K, K.L, X.W, Y.W suggested relevant analysis and contributed to the manuscript revision. All authors reviewed the manuscript.

Reference

1 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).
2 Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599-604, doi:10.1038/nprot.2017.149 (2018).
3 Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat Rev Genet* **24**, 535-549, doi:10.1038/s41576-023-00599-5 (2023).
4 Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86-102, doi:10.1038/s41586-021-03950-0 (2021).
5 Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat Neurosci* **24**, 584-594, doi:10.1038/s41593-020-00794-1 (2021).
6 Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318-1323, doi:10.1126/science.aap8809 (2017).
7 Ruzicka, W. B. *et al.* Single-cell multi-cohort dissection of the schizophrenia transcriptome. *medRxiv*, 2022.2008.2031.22279406, doi:10.1101/2022.08.31.22279406 (2022).
8 Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685-689, doi:10.1126/science.aav8130 (2019).
9 Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337, doi:10.1038/s41586-019-1195-2 (2019).
10 Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17**, 63, doi:10.1186/s13059-016-0927-y (2016).
11 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31, doi:10.1186/s13059-020-1926-6 (2020).
12 Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093-1095, doi:10.1038/nmeth.2645 (2013).
13 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562-578, doi:10.1093/biostatistics/kxx053 (2018).
14 Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**, 915-925, doi:10.1038/nbt.2972 (2014).
15 Consortium, M. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151-1161, doi:10.1038/nbt1239 (2006).
16 Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11**, 41-46, doi:10.1038/nmeth.2694 (2014).
17 Murphy, A. E., Fancy, N. & Skene, N. Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset. *Elife* **12**, doi:10.7554/eLife.90214 (2023).
18 Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692, doi:10.1038/s41467-021-25960-2 (2021).
19 Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255-261, doi:10.1038/nmeth.4612 (2018).
20 Johansen, N. *et al.* Interindividual variation in human cortical cell type abundance and expression. *Science* **382**, eadf2359, doi:10.1126/science.adf2359 (2023).
21 Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, doi:10.1126/science.aat8127 (2018).
22 Saliba, A. E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**, 8845-8860, doi:10.1093/nar/gku555 (2014).

23 van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, doi:10.7554/eLife.52155 (2020).
24 Su, C. *et al.* Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nat Commun* **14**, 4846, doi:10.1038/s41467-023-40503-7 (2023).
25 Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290, doi:10.1073/pnas.1507125112 (2015).
26 Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80, doi:10.1038/nbt.4038 (2018).
27 Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat Neurosci* **23**, 771-781, doi:10.1038/s41593-020-0621-y (2020).
28 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).
29 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119, doi:10.1038/s41586-021-03465-8 (2021).
30 Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, doi:10.1126/science.aat7615 (2018).
31 Agarwal, D. *et al.* A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat Commun* **11**, 4183, doi:10.1038/s41467-020-17876-0 (2020).
32 Zhou, Y. *et al.* Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat Med* **26**, 131-142, doi:10.1038/s41591-019-0695-9 (2020).
33 Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet* **53**, 1143-1155, doi:10.1038/s41588-021-00894-z (2021).
34 Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222-3241 e3226, doi:10.1016/j.cell.2021.04.021 (2021).
35 Velmeshev, D. *et al.* Single-cell analysis of prenatal and postnatal human cortical development. *Science* **382**, eadf0834, doi:10.1126/science.adf0834 (2023).
36 Braun, E. *et al.* Comprehensive cell atlas of the first-trimester developing human brain. *Science* **382**, eadf1226, doi:10.1126/science.adf1226 (2023).
37 Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046, doi:10.1126/science.add7046 (2023).
38 Hagai, T. *et al.* Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197-202, doi:10.1038/s41586-018-0657-2 (2018).
39 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).
40 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
41 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).
42 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

Table 1 sc/snRNA-Seq datasets assessed in this study

Data label	PMID	Brain region/tissue	#Individuals/ samples	#Cells	#Genes	#Cell types	Platform

ROSMAP#	31042697	Prefrontal cortex	24	75,060	17,926	8	10X
autism_PFC#	31097668	Prefrontal cortex	10	29,900	27,563	8	10X
autism_ACC#	31097668	Anterior cingulate cortex	16	22,065	27,072	8	10X
MTG#	31435019	Medial temporal gyrus	8	15,583	43,474	7	Smart-Seq
M1_10X#	34616062	Primary motor cortex	2	76,533	27,933	8	10X
GSE97930_vc#	29227469	Visual cortex	3	19,368	21273	8	Drop-seq
GSE97930_fro#	29227469	Frontal cortex	1	19,368	18751	8	Drop-seq
GSE140231#	32826893	Cortex	5	10,581	24702	6	10X
TREM#	31932797	Prefrontal cortex	11	36,671	36,601	6	10X
GSE174367#	34239132	Prefrontal cortex	8	21,996	25,392	6	10X
BICCN_adult#	37824663	Cortex	4	130,5075	19,762	30	10X
BICCN_HVS#	37824649	Cortex	78	353,194	18,797	24	10X
BICCN_trimester1#	37824650	Cortex	21	789,139	33,538	12	10X
BICCN_dev#	37824647	Cortex	106	709,372	19,005	9	10X
Mouse*	30356220	Mononuclear phagocytes	6	17,776	15,319	1	Smart-seq2
Rat*	30356220	Mononuclear phagocytes	6	13,277	15,150	1	Smart-seq2
Rabbit*	30356220	Mononuclear phagocytes	6	17,097	9,263	1	Smart-seq2
Pig*	30356220	Mononuclear phagocytes	6	12,753	8,906	1	Smart-seq2

*With sample-matched sorted-cell RNA-Seq data; #Data that from human brain tissue

Figures

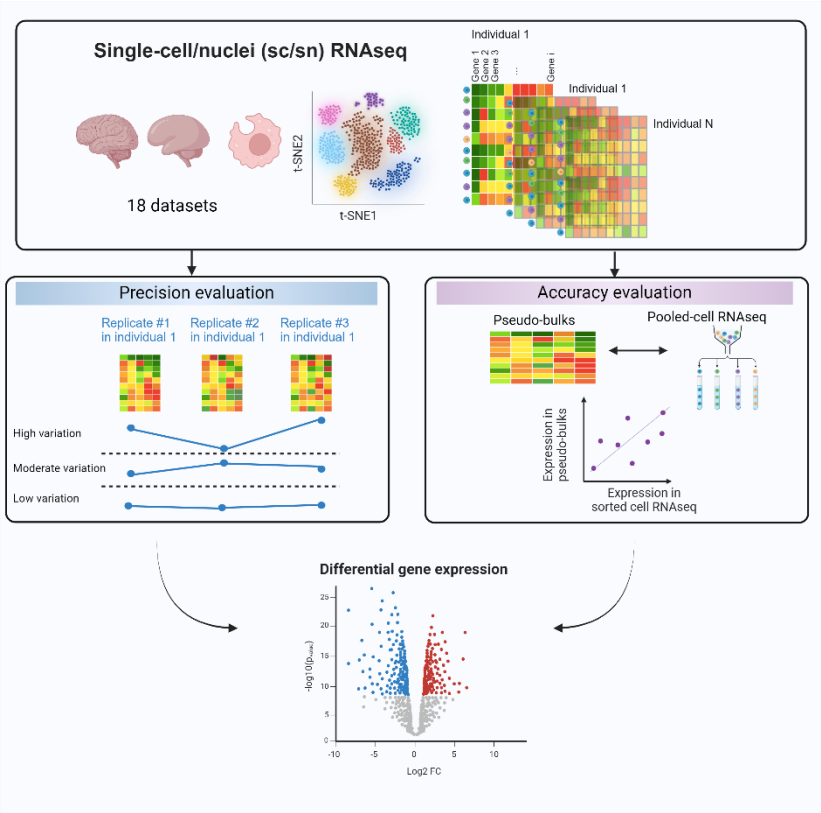
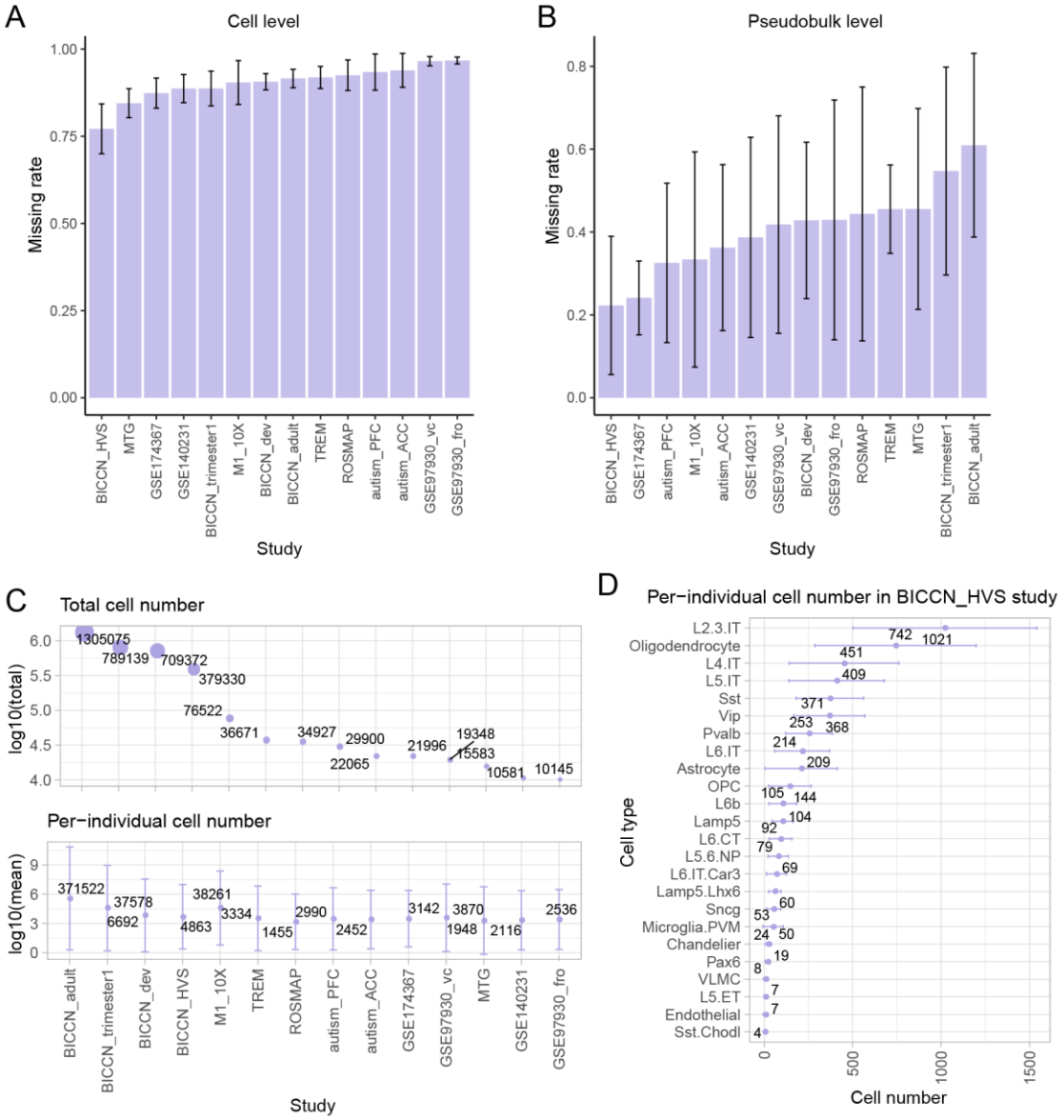


Fig. 1. Overview of the study. This figure was created with BioRender.com.



548

549 **Fig. 2. Missing rates and cell numbers in the 14 datasets.** (A) Missing rate per gene at individual cell

550 level. (B) Missing rate per gene at the pseudo-bulk level. (C) Total number of cells studied, along with

551 the average number of cells per individual. The cell count per individual was calculated by dividing the

552 total number of cells by the number of individuals in the study. Cell numbers were log10-transformed

553 for better visualization. (D) An illustrative example showcasing the average number of cells per

554 individual, specifically drawn from the BICCN_HVS study.

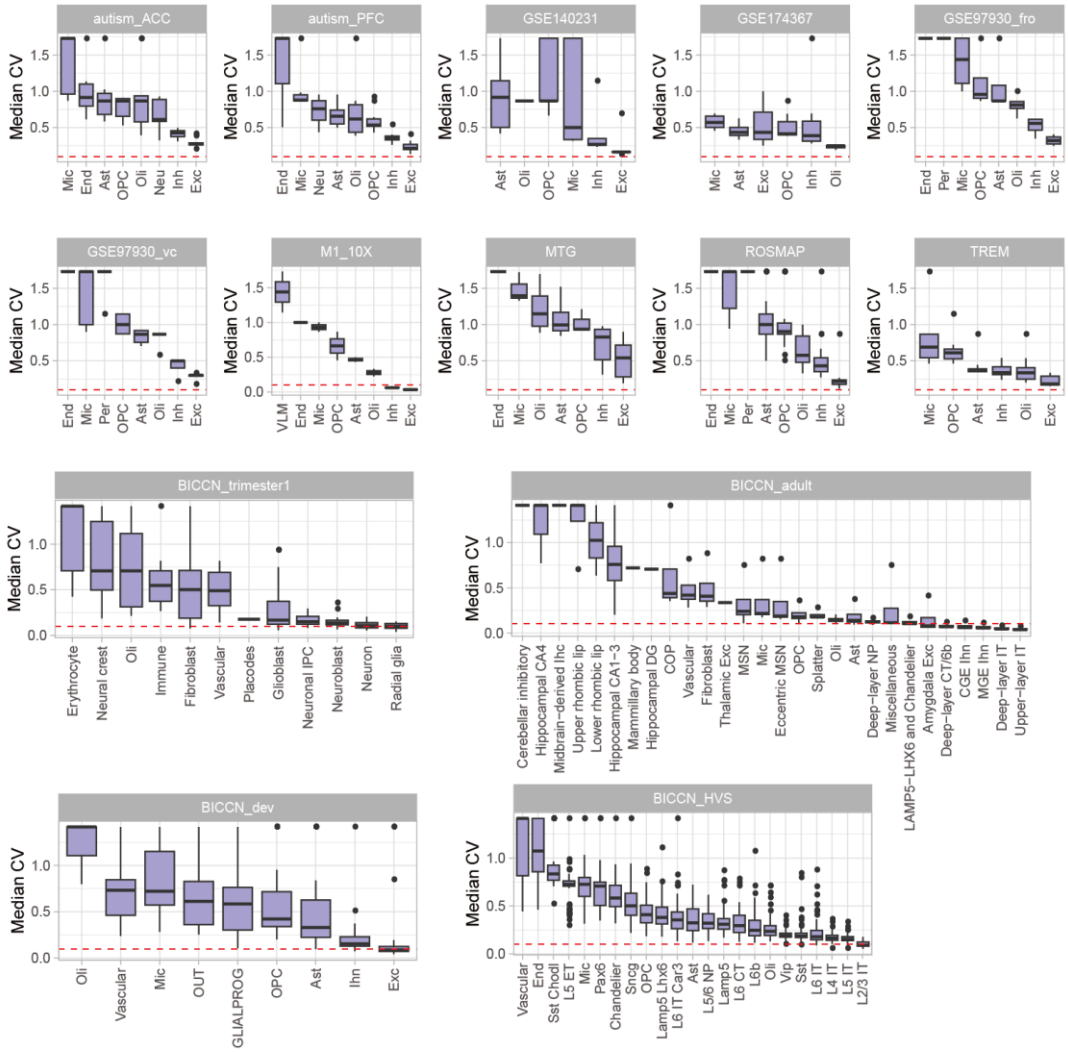
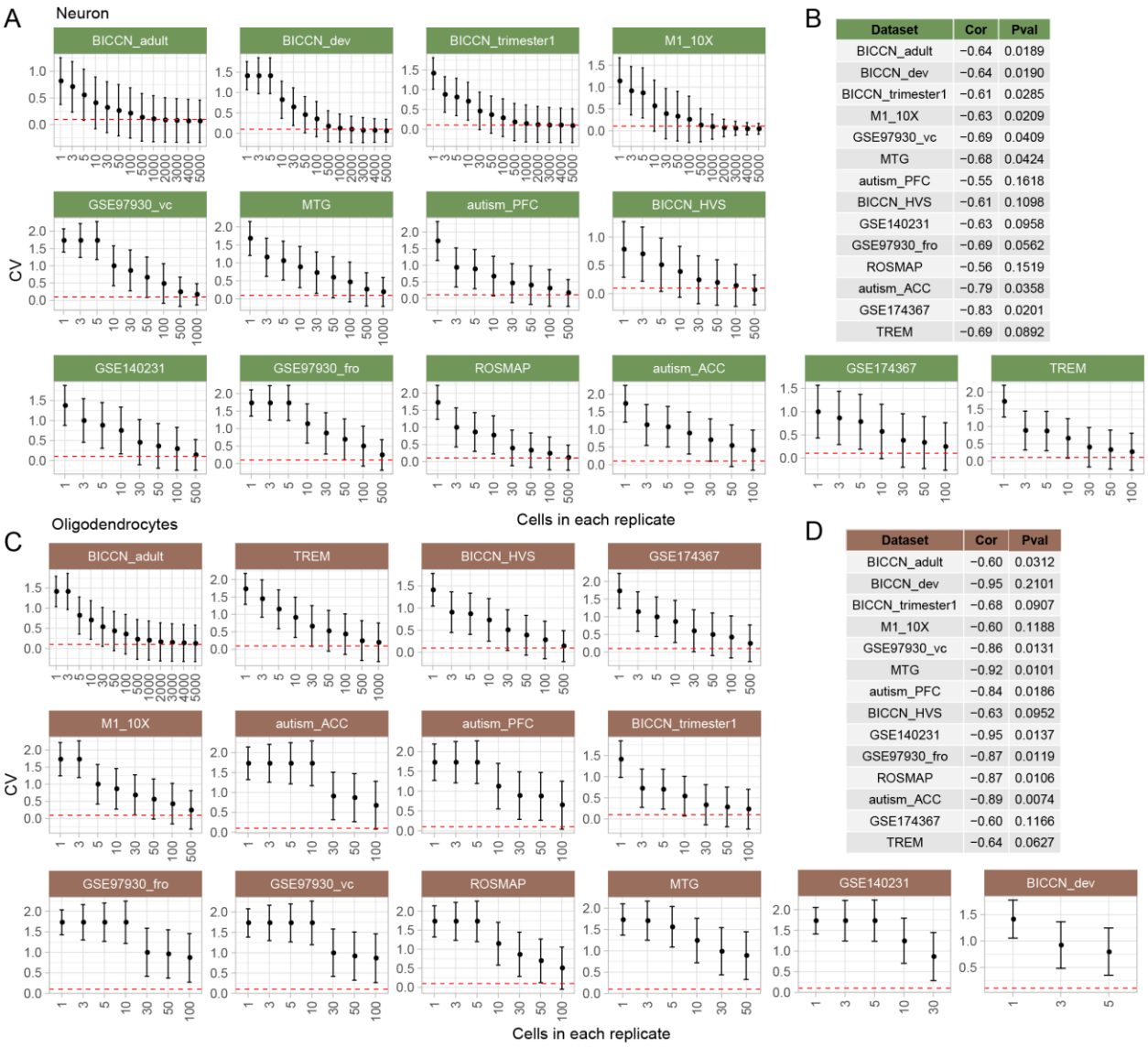


Fig. 3. Gene expression precision evaluated by technical replicates in sc/snRNA-Seq data. We calculated gene expression variability as the CV of gene expression in three technical replicates for each gene. The sample with the largest number of total cells in each dataset was used for illustration. The red dashed line denotes the CV thresholds of 0.1, which is a threshold recommended for bulk-tissue gene expression QC processing[15].



Dataset	Cor	Pval
BICCN_adult	−0.64	0.0189
BICCN_dev	−0.64	0.0190
BICCN_trimester1	−0.61	0.0285
M1_10X	−0.63	0.0209
GSE97930_vc	−0.69	0.0409
MTG	−0.68	0.0424
autism_PFC	−0.55	0.1618
BICCN_HVS	−0.61	0.1098
GSE140231	−0.63	0.0958
GSE97930_fro	−0.69	0.0562
ROSMAP	−0.56	0.1519
autism_ACC	−0.79	0.0358
GSE174367	−0.83	0.0201
TREM	−0.69	0.0892

Dataset	Cor	Pval
BICCN_adult	−0.60	0.0312
BICCN_dev	−0.95	0.2101
BICCN_trimester1	−0.68	0.0907
M1_10X	−0.60	0.1188
GSE97930_vc	−0.86	0.0131
MTG	−0.92	0.0101
autism_PFC	−0.84	0.0186
BICCN_HVS	−0.63	0.0952
GSE140231	−0.95	0.0137
GSE97930_fro	−0.87	0.0119
ROSMAP	−0.87	0.0106
autism_ACC	−0.89	0.0074
GSE174367	−0.60	0.1166
TREM	−0.64	0.0627

Fig. 4. Relationship between cell number and gene expression precision. (A) CV values in down-sampled neurons. (B) Correlation coefficient (Cor) and p value (Pval) between neuron cell numbers in replicate and median CV. (C) CV values in down-sampled oligodendrocytes. (D) Correlation coefficient (Cor) and p-value (Pval) between oligodendrocyte cell numbers in replicate and median CV. To enhance visual clarity, the number of cells in each replicate was capped at 5000. The Red dashed line denotes a CV of 0.1.

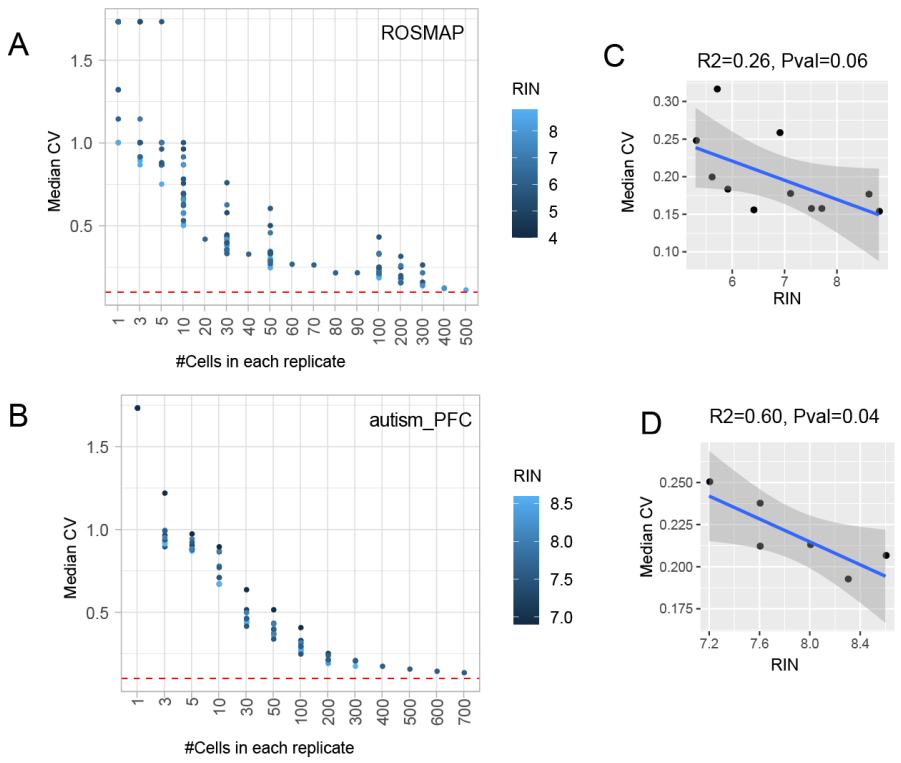


Fig. 5. Association between RNA integrity and gene expression variability across technical replicates in snRNA-Seq data. The relationship between number of cells in replicate and median CV was tested in ROSMAP (A) and autism_PFC (B) dataset. Samples were colored by RIN value. Red dashed line denotes the CV threshold of 0.1. (C) & (D) illustrate the relationship between RIN and median CV when replicate containing 200 cells in ROSMAP and autism_PFC data respectively. Linear regression model was used.

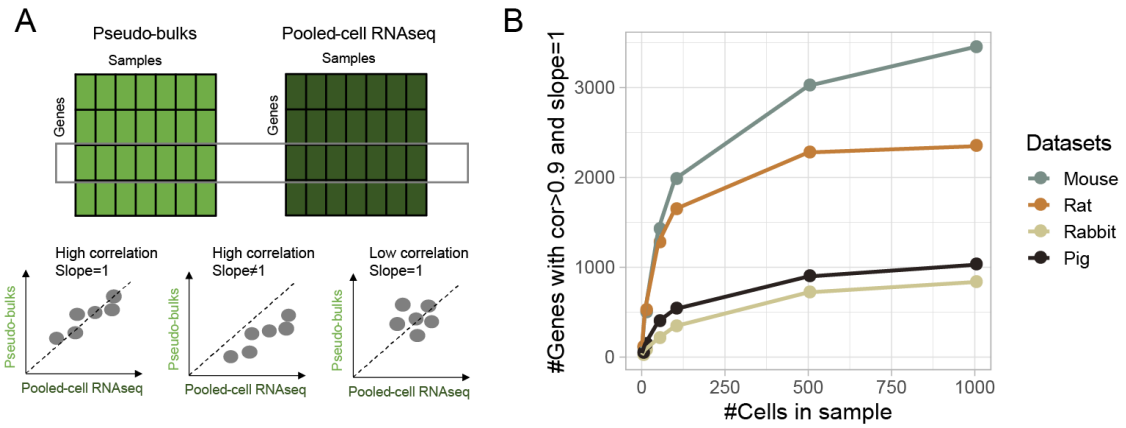


Fig. 6. Relationship between expression accuracy and cell number. (A) Illustration of expression accuracy assessment. (B) The relationship between cell number and number of genes with good accuracy which defined by Pearson correlation and linear regression model.

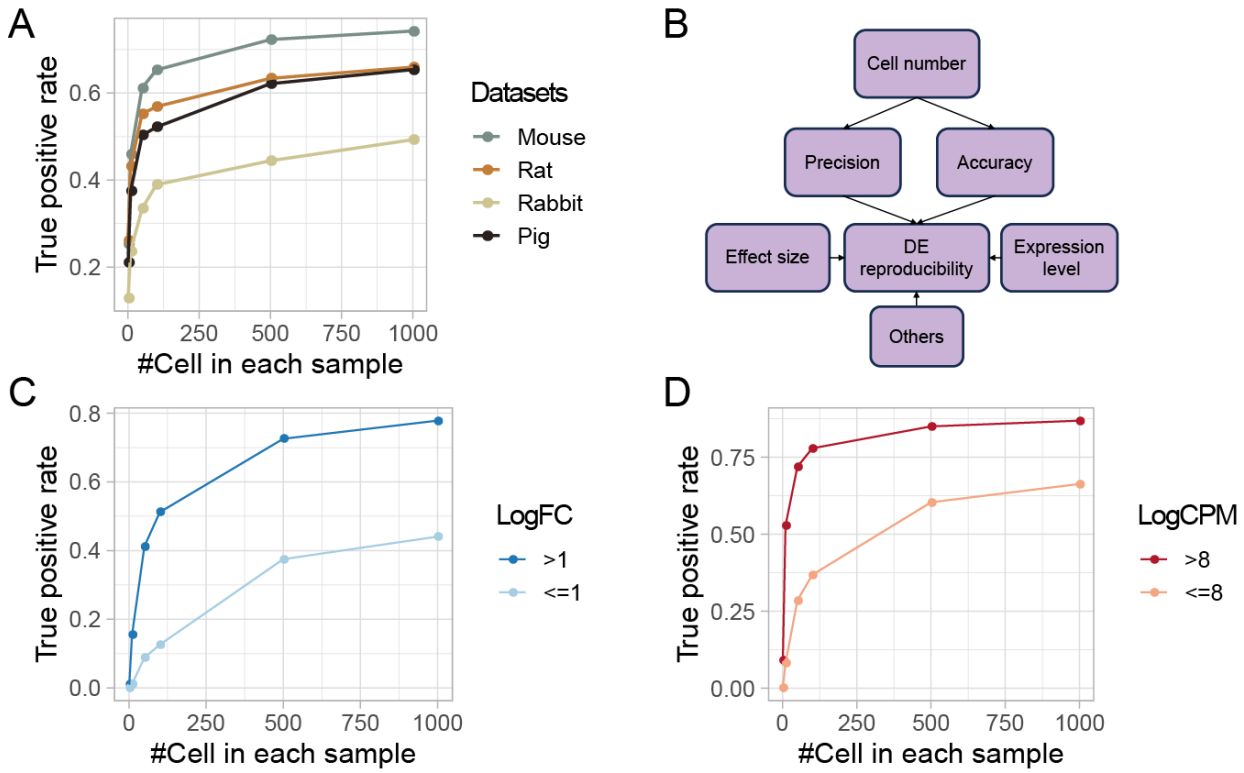


Fig. 7. Reproducibility of DE analysis in scRNA-Seq data. (A) True positive rate of DEGs in data with different numbers of cells. Datasets from different species were colored. (B) Model of DE reproducibility. (C) True positive rate of DEGs with different effect sizes defined by logFC of differential expression. (D) Replication rate of DEGs with different expression levels.

Supplemental Materials

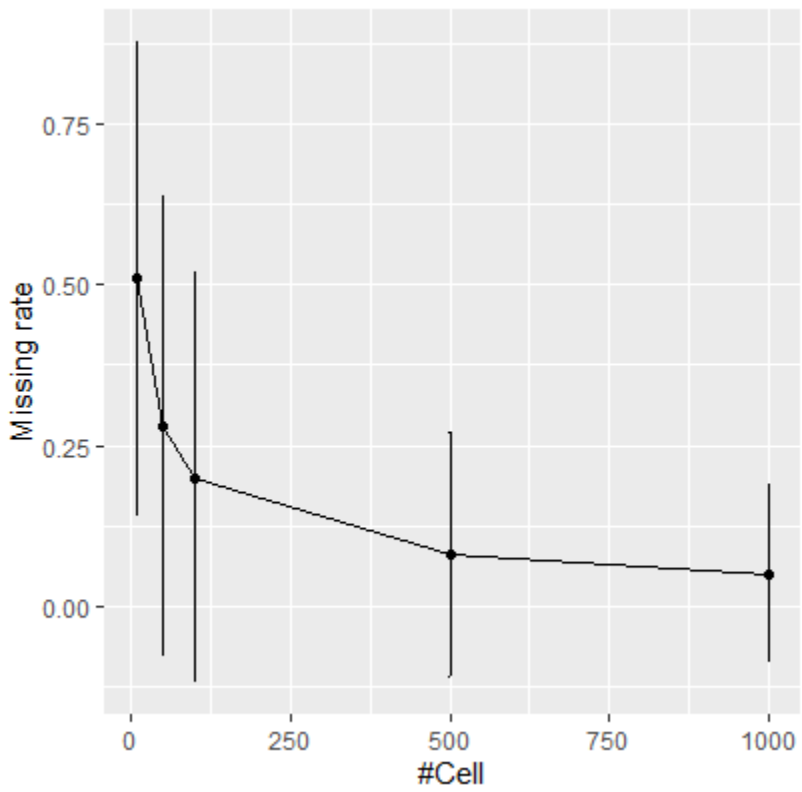


Fig. S1. Per gene missing rate in pseudo-bulk data with different number of cells. Missing rate was defined as proportion of zeros in all samples. Data from ROSMAP was used for demonstration.

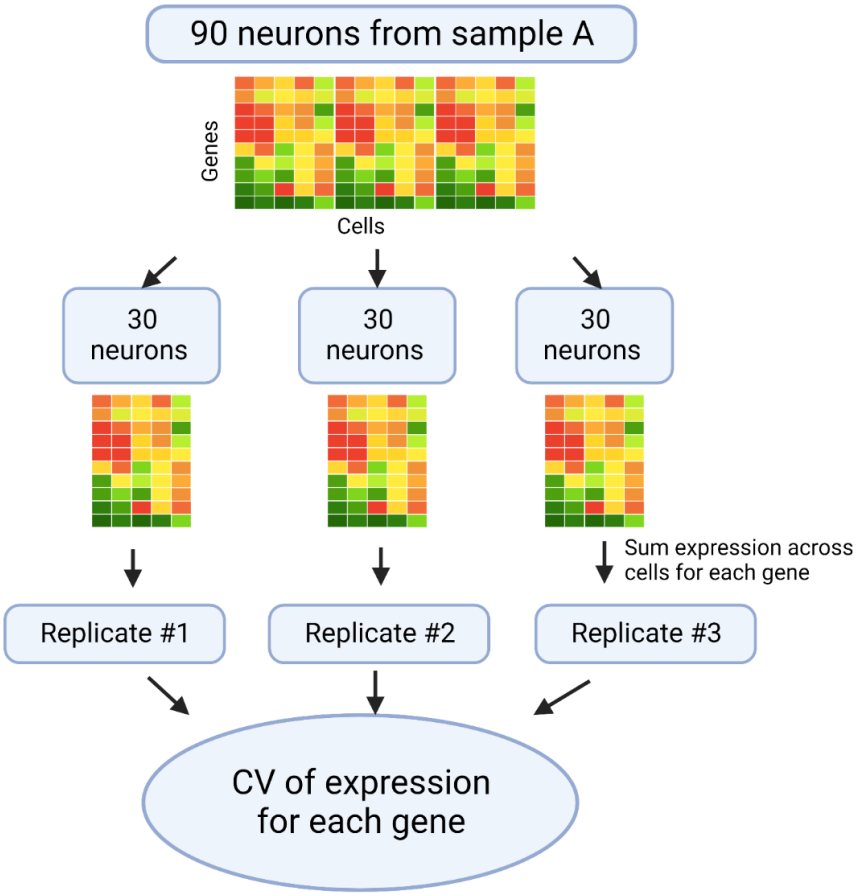


Fig. S2. Illustration of constructing technical replicates and calculating the CV of expression across technical replicates.

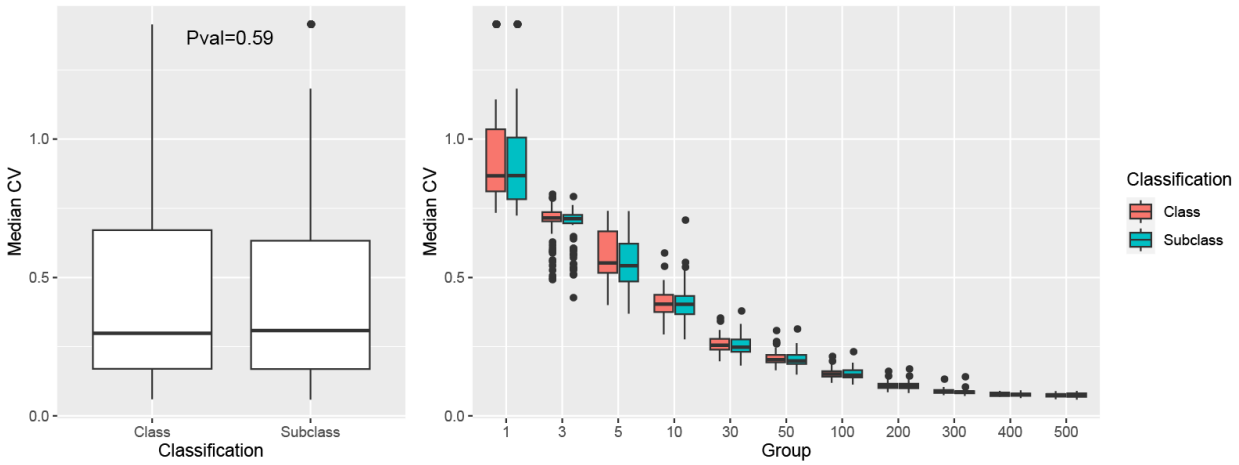
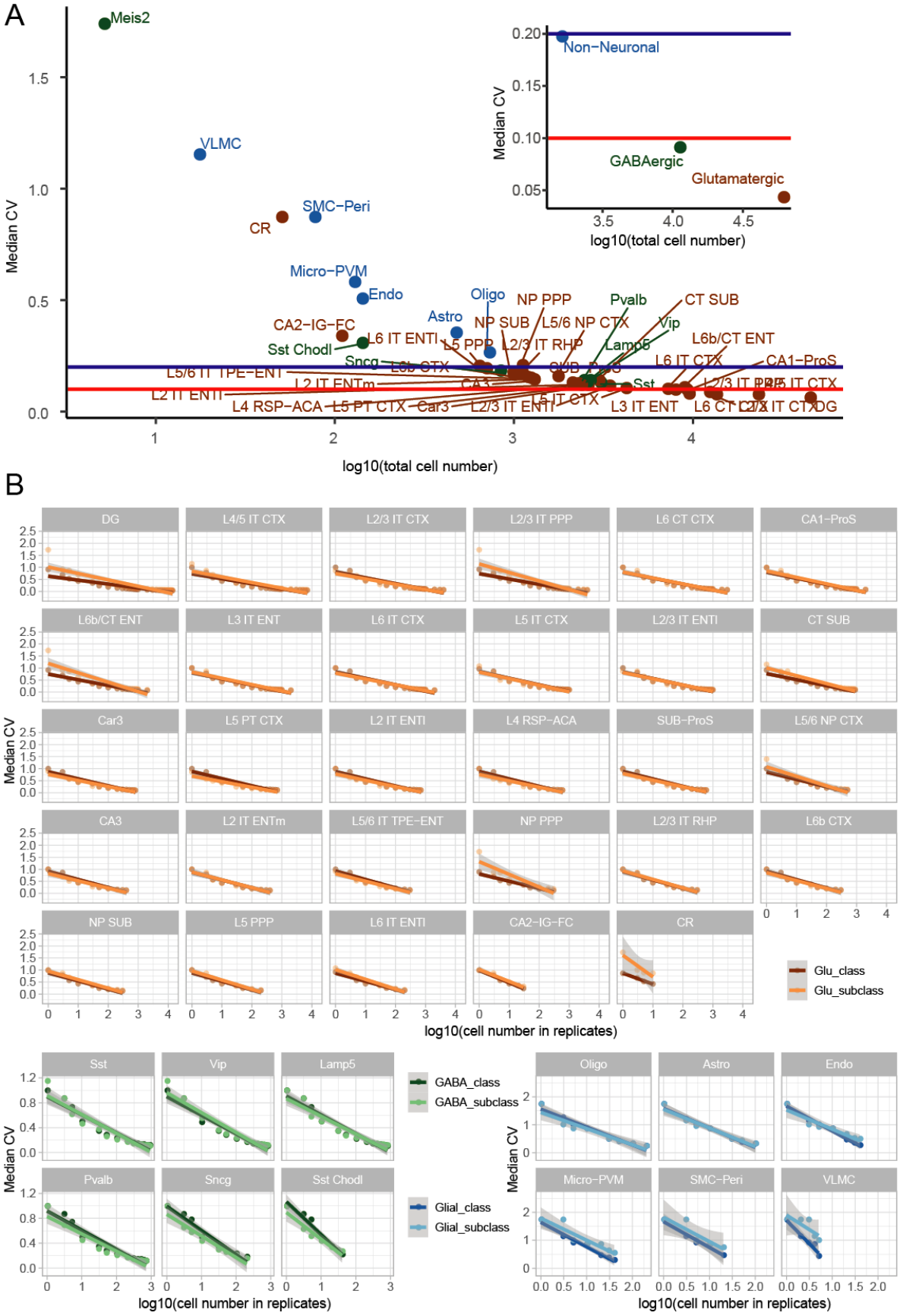


Fig. S3. Expression variability at class and subclass level for excitatory neurons. Data from BICCN_HVS was used. P value was calculated with Student's T-Test.


A
Meis2
VLMC
SMC−Peri
CR
Micro−PVM
Endo
Oligo
Astro
NP PPP
Pvalb
CT SUB
NP SUB
L5/6 NP CTX
Vip
CA2−IG−FC
L6 IT ENTI
L5 PPP
L2/3 IT RHP
L6b/CT ENT
Sst Chodl
Sncg
Lamp5
L6 IT CTX
CA1−ProS
L5/6 IT TPE−ENT
L2 IT ENTm
Sst
L2 IT ENTI
L4 RSP−ACA
L5 PT CTX
Car3
L2/3 IT ENTI
L3 IT ENT
Median CV
log10(total cell number)
Non−Neuronal
GABAergic
Glutamatergic
B
DG
L4/5 IT CTX
L2/3 IT CTX
L2/3 IT PPP
L6 CT CTX
CA1−ProS
L6b/CT ENT
L3 IT ENT
L6 IT CTX
L5 IT CTX
L2/3 IT ENTI
CT SUB
Car3
L5 PT CTX
L2 IT ENTI
L4 RSP−ACA
SUB−ProS
L5/6 NP CTX
CA3
L2 IT ENTm
L5/6 IT TPE−ENT
NP PPP
L2/3 IT RHP
L6b CTX
NP SUB
L5 PPP
L6 IT ENTI
CA2−IG−FC
CR
Glu_class
Glu_subclass
log10(cell number in replicates)
Sst
Vip
Lamp5
GABA_class
GABA_subclass
Pvalb
Sncg
Sst Chodl
Glial_class
Glial_subclass
Oligo
Astro
Endo
Micro−PVM
SMC−Peri
VLMC

607

Fig. S4. Gene expression variability across technical replicates in mouse brain snRNA-Seq data. (A) Median CV across genes in replicates constructed at class and subclass level. (B) Linear regression model of cell numbers in replicates and median CV identified at class and subclass level.

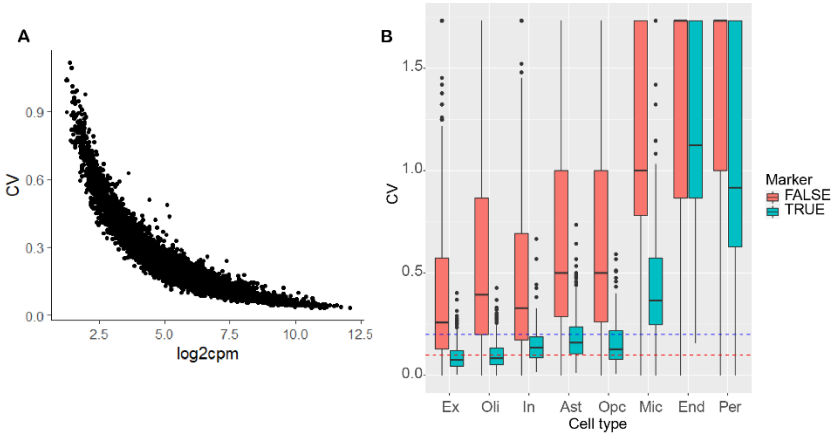


Fig. S5. Expression abundance and CV in ROSMAP data. (A) Relationship of expression and CV in excitatory neuron from ROSMAP study. (B) CV of marker genes. Blue and red line denote CV of 0.2 and 0.1.

A

Dataset	Mean	Max
M1_10X	25%	100%
BICCN_adult	21%	100%
BICCN_trimester1	13%	52%
BICCN_dev	8%	67%
BICCN_HVS	3%	58%
autism_ACC	0	0
autism_PFC	0	0
GSE140231	0	0
GSE174367	0	0
GSE97930_fro	0	0
GSE97930_vc	0	0
MTG	0	0
ROSMAP	0	0
TREM	0	0

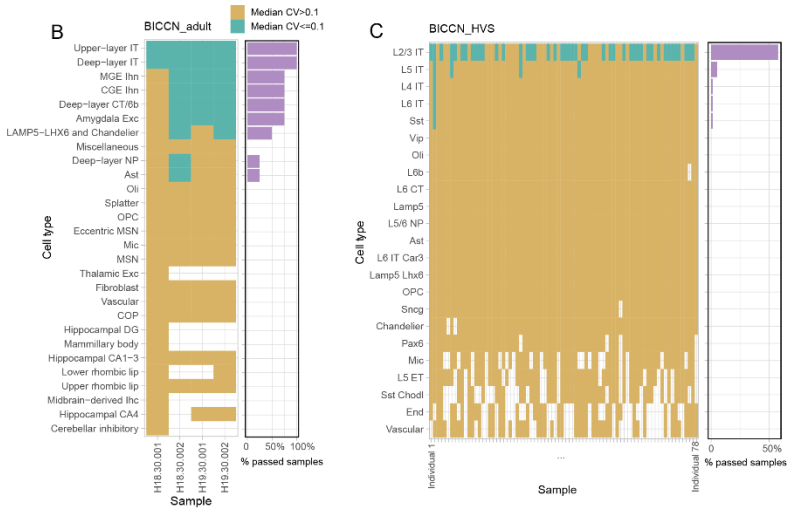


Fig. S6. Proportions of samples with acceptable expression variability. (A) The average and maximum percentages of samples that satisfy the precision criterion. Examples from BICCN_adult (B) and BICCN_HVS(B) dataset were used from illustration. Samples achieving the precision threshold, defined by a coefficient of variation (CV) of 0.1 or lower are indicated in green, signifying acceptable expression precision, while those failed to meet the threshold are marked in yellow, indicating low precision. Instances where a cell type is not represented in a sample are left blank. The accompanying bar plot provides a detailed breakdown of the exact proportion of samples that satisfy the precision criterion.

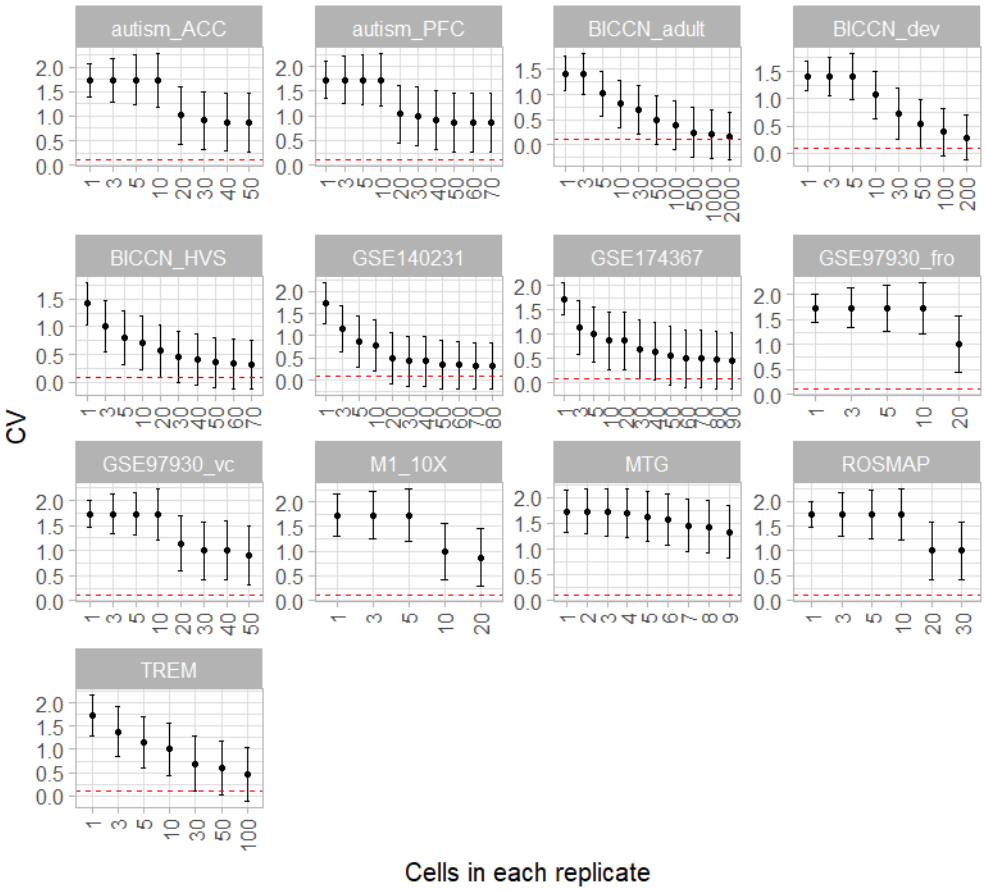


Fig. S7. Expression variability across replicates in microglia.

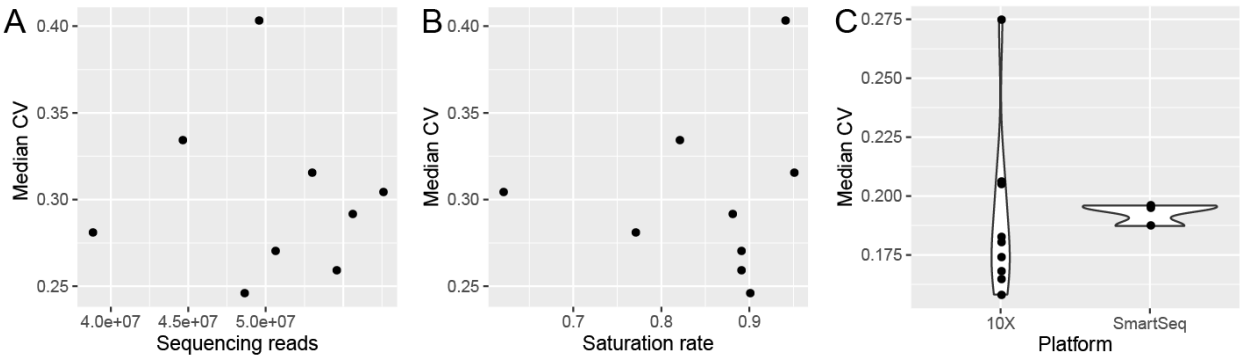


Fig. S8. Association between technical factors and gene expression variability across technical replicates in snRNA-Seq data. (A) Median CV across detected genes in replicates versus total sequencing depth. (B) Median CV across detected genes in replicates versus total sequencing

saturation rate. (C) Comparison of median CV across detected genes in replicates between data sequenced by 10X Chromium and Smart-Seq platforms.