

Correcting Cross-Cell UMI Bias for Improved Single-Cell Transcriptomics

Rulin Hua¹, Wenjing Yan², Xiaoyan Liu³, Suyang Xuan⁴, Xuejuan Li⁵, Shafiq Ur Rahman¹, Dekang Lv^{4*} and Feng Zheng^{1,5*}

¹Advanced Institute for Medical Sciences, Dalian Medical University, Dalian, 116044, China

²Dalian Center for Disease Control and Prevention, Dalian, 116021, China

³Department of Nephrology, the Second Hospital of Dalian Medical University, Dalian, 116023, China

⁴Institute of Cancer Stem Cell, Dalian Medical University, Dalian, 116044, China

⁵Wuhu Hospital and Health Science Center, East China Normal University, Shanghai, 200241, China

Rulin Hua and Wenjing Yan contributed equally.

ABSTRACT

Accurate quantification of mRNA at the single-cell level is crucial for understanding gene expression dynamics, cellular diversity, and disease mechanisms. Traditional normalization methods for single-cell RNA sequencing (scRNA-seq) typically measure relative RNA abundance, which limits their ability to provide absolute quantification and does not account for variations in total RNA content across cells. To address this limitation, we present the LOESS regression-corrected UMI count (LC-UMI), a method that adjusts UMI counts to correct cross-cell amplification variability, enabling more accurate gene expression quantification at the single-cell level. We demonstrate that LC-UMI accurately captures differences in RNA content across cell types. At the single-cell level, total LC-UMIs strongly correlate with transcription-related pathways, such as RNA polymerase complex and mitochondrial oxidative phosphorylation. In downstream analyses, including cell clustering, differential gene expression, and pseudotime analysis, log(LC-UMI+1) delivers results comparable to conventional normalization methods. LC-UMI is implemented in the R package LcUMI, which is freely available for any UMI-based scRNA-seq dataset. These findings underscore the potential of LC-UMI to enhance the precision and reliability of single-cell transcriptomics, providing a valuable metric for future research in single-cell gene expression.

Article Information

Received 06 June 2025

Revised 25 July 2025

Accepted 08 August 2025

Available online 29 January 2026 (early access)

Authors' Contribution

FZ and DL conceived and supervised the study. XLi performed the animal experiments, including rat maintenance, tissue dissection, and sample preparation. XLiu and SU Rahman performed sample processing following tissue collection. RH and DL developed the computational models. RH, WY, SX and DL conducted data analysis and interpretation. FZ, DL and RH wrote and revised the manuscript. All authors have read and approved the final version of the manuscript.

Key words

Single-cell RNA-seq, UMI, PCR amplification bias, Absolute quantification, Normalization, LOESS regression

INTRODUCTION

Unique molecular identifiers (UMIs) have emerged as a powerful tool for improving the accuracy of RNA quantification by mitigating PCR-induced amplification biases (Kivioja *et al.*, 2012). These short, random oligonucleotide tags uniquely label individual RNA molecules prior to amplification, enabling direct transcript

counting and mitigating distortions from variable amplification efficiency. Unlike conventional RNA-seq approaches that rely on relative expression metrics such as RPKM or TPM, UMI-based quantification establishes an absolute scale of measurement with a defined zero (Islam *et al.*, 2014). This distinction is crucial, as relative normalization methods can obscure fluctuations in total mRNA content. For example, a gene may be 'upregulated' in terms of RPKM and have a decrease in absolute expression level if the total mRNA content also changes. Given its potential to provide an absolute molecular count, UMI-based quantification warrants further refinement and broader application in transcriptomic analyses.

In single-cell transcriptomic analysis, current standardization methods for UMI counts typically overlook variations in total RNA content between cells. Most methods treat these variations as technical biases caused by sequencing depth and attempt to correct for them (Ahlmann-Eltz and Huber, 2023; Hafemeister and Satija,

* Corresponding author: dekanglv@126.com, Fzheng@hsc.ecnu.edu.cn
0030-9923/2026/0001-0001 \$ 9.00/0



Copyright 2026 by the authors. Licensee Zoological Society of Pakistan.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

2019). One common approach is log-transformed library size normalization, which scales gene counts using total UMI counts as size factors (Ahlmann-Eltze and Huber, 2023; Butler *et al.*, 2018). The scran method further refines this by grouping cells with similar total UMI counts and estimating pool-based size factors, which are deconvolved into cell-specific values (Lun *et al.*, 2016). Other methods, such as SCTransform, incorporate total molecular content as a covariate in a generalized linear model to adjust for sequencing depth, with Pearson residuals from negative binomial regression used for further adjustment (Hafemeister and Satija, 2019). However, these UMI normalization methods fail to account for true differences in RNA content between cells and do not provide an absolute measure of RNA molecules at the single-cell level.

In droplet-based single-cell transcriptomic library construction, cDNA from each cell is first generated independently within individual droplets, then pooled and subjected to PCR amplification in a shared reaction system. Similarly, subsequent steps in library preparation and sequencing involve uniform amplification of all molecules in a common mixture. Under such a shared amplification environment, each UMI-labeled molecule is theoretically exposed to comparable amplification efficiency. This provides a rationale for using raw UMI counts as a proxy for cellular transcriptional activity. In this study, we evaluate the robustness of UMI counts for absolute transcript measurement across individual cells. We find that cells with higher total UMI counts exhibit an increased PCR amplification ratio. This leads to not only a disproportionate increase of reads per UMI but also overestimation of UMI counts in RNA-rich cells under a fixed sequencing depth. To correct this bias, we applied locally estimated scatterplot smoothing (LOESS) regression (Cleveland, 1979) to adjust UMI counts based on the average UMI amplification ratio per cell, ensuring that UMI-based quantification more accurately reflects true differences in cellular RNA content. Our results demonstrate that LOESS-corrected UMI count (LC-UMI) effectively capture RNA abundance heterogeneity across cell types. Notably, genes associated with total LC-UMIs are enriched in pathways linked to transcriptional activity, including the RNA polymerase complex, Golgi vesicle transport, oxidative phosphorylation, proteasome, protein folding, and the spliceosomal complex. Finally, leveraging the LC-UMI matrix, we validate the robustness of our approach in downstream analyses, including cell clustering, differential gene expression, and pseudotime trajectory inference, demonstrating performance superior to or comparable to traditional normalization methods. Our approach is broadly applicable for any UMI-based

scRNA-seq dataset and is freely available to users through the open-source R package LcUMI (<https://github.com/RulinHua/LC-UMI>).

MATERIALS AND METHODS

Experimental animals

All animal experiments were approved by the Animal Welfare and Ethics Committee of East China Normal University (Approval No. m+R20250402). Wild-type Sprague-Dawley rats (200–300 g) were housed in a specific pathogen-free (SPF) environment under controlled conditions (temperature: 22–25°C, humidity: Appropriate, light/dark cycle: 12 h). Animals had free access to food and water, with bedding and feed sterilized by Co60 irradiation.

Parathyroid cell isolation and 10x Genomics scRNA-seq

Twelve weeks old male wild-type Sprague-Dawley rats were maintained on a standard diet for five weeks. In compliance with the guidelines of the Animal Welfare and Ethics Committee of East China Normal University, the rats were humanely euthanized. Parathyroid glands were promptly excised and preserved in ice-cold tissue preservation solution (Miltenyi Biotec) for immediate transport and processing.

The excised glands were minced and enzymatically digested at 37°C for 40 min with gentle shaking in a solution containing TL (1 mg/mL, Roche), Dispase (2 mg/mL, Miltenyi Biotec), and DNase I (0.05 mg/mL, Sigma). Every 20 min, dissociated cells were collected using a 10 mL plastic pipette containing fetal bovine serum (FBS). The resulting cell suspension was filtered through a 20 µm mesh filter (BD Biosciences), centrifuged at 300 g for 5 min, and incubated in red blood cell lysis buffer (155 mM NH₄Cl, 0.1 mM Na₂EDTA, 10 mM KHCO₃) for 3 min. The dissociated parathyroid cells were then resuspended and stored in Dulbecco's phosphate-buffered saline (DPBS).

Single-cell cDNA library preparation and 3' single-cell RNA sequencing were conducted by CapitalBio Technology. Libraries were prepared using the 10x Genomics Chromium platform, following the manufacturer's instructions with the single cell 3' Library and Gel Bead Kit V3 and the chromium single cell B chip kit. Cells were suspended in PBS containing 0.04% BSA, and approximately 8,000 cells were loaded onto the Chromium Single Cell Controller (10x Genomics) to generate single-cell gel bead emulsions (GEMs). Captured cells were lysed, and released RNA was barcoded through reverse transcription within individual GEMs. RT-PCR was performed using an S1000™ Touch Thermal Cycler (Bio-Rad) under the following conditions: 53°C for 45

min, 85°C for 5 min, followed by a hold at 4°C. cDNA was recovered, amplified, and quality-assessed using the Agilent 4200 system. Single-cell RNA-seq libraries were constructed using the single cell 3' Library and gel bead kit V3, following the manufacturer's protocol. Finally, sequencing was performed on an Illumina NovaSeq 6000 platform, generating paired-end 150 bp (PE150) reads with a minimum sequencing depth of 100,000 reads per cell (conducted by Beijing Cade Biotechnology Co., Ltd.).

scRNA-seq data processing

Raw data from 10x Genomics scRNA-seq were processed using Cell Ranger (v8.0.1) for read alignment, barcode demultiplexing, and gene expression quantification, producing a UMI count matrix. For the snDrop-seq dataset, barcode and UMI sequences were extracted from the first 12 and 8 bases of read1, respectively. Reads were aligned with STAR (v2.7.10) to generate the gene count matrix. The UMI count matrix was subsequently converted into Seurat object using the Seurat package (v4.4.0) (Hao *et al.*, 2021) in R (v4.3.1).

For the human IBD colon, rat parathyroid, and human frontal cortex datasets, UMI counts were normalized using the log(CP10K+1) method, where counts were first normalized by total UMI counts, scaled by a factor of 10,000, and then log-transformed. Batch effect correction was performed with the bbknnR package (v1.1.0). For the Tabula Sapiens and Tabula Muris datasets, UMI counts were processed using the LC-UMI method. For the mouse BM myeloid, human 68k PBMC, and mouse brain metastatic microenvironment datasets, normalization was performed using log(CP10K+1), sctransform, and log(LC-UMI+1), respectively, with comparisons made across these methods. The sctransform method was implemented using the R package sctransform (v0.4.1). Following normalization, HVGs were selected, followed by dimensionality reduction using PCA, cell clustering, and embedding with t-SNE or UMAP, all performed with the Seurat package.

Analysis of the relationship between UMI amplification ratio and total UMI counts before and after correction

To assess the relationship between UMI amplification ratio (reads per UMI molecule) and total UMI counts in scRNA-seq data, we stratified cells within each sample of the human IBD colon dataset, human frontal cortex dataset, and rat parathyroid dataset into 10 equal-sized groups based on their total UMI counts. Boxplots were used to compare the distribution of UMI amplification ratio across these groups. To model the overall trend, we applied kernel regression (ksmooth function) with a normal kernel and a bandwidth set to 20 times the value recommended by the

bw.SJ function. UMI amplification was fitted as a function of either log10-transformed raw total UMI counts or log10-transformed total LC-UMIs. For each sample, the smoothed median UMI amplification ratio was visualized using a line plot across 200 evenly spaced total UMI values, while a colored ribbon was used to represent variability within the interquartile range (25th–75th percentile).

Correction of single-cell UMI counts using LOESS regression

To correct for cross-cell amplification biases in scRNA-seq data, we employed a LOESS regression approach. For each sample, we modeled the relationship between the average UMI amplification ratio (the number of reads per UMI molecule; predictor variable) and the log10-transformed total UMI counts (response variable) across cells. The regression was performed using the loess function in R with default settings: span= 0.75 and degree= 2 (local quadratic fitting), applied consistently across all datasets in this study. The span parameter can be adjusted based on the data-specific relationship between amplification ratio and total UMI counts to ensure appropriate smoothness and avoid over or underfitting. The residuals from this regression, representing the relative deviation of total UMI counts from the fitted amplification trend, were exponentiated to obtain the corrected total UMI counts for each cell.

To maintain consistency with the raw UMI distribution, we applied a scaling factor defined as the ratio of the median raw total UMI count to the median corrected total UMI count across all cells within a sample. For datasets involving multiple samples, we further accounted for differences in sequencing depth and cell number by first computing the mean total UMI count per cell for each sample. This value was then normalized by dividing it by the median of mean total UMI counts across all samples, yielding a sample-specific normalization factor. Each cell's scaling factor was then adjusted by dividing it by the normalization factor of its corresponding sample, ensuring comparability of UMI counts across samples at the single-cell level. Finally, the corrected total UMI counts were multiplied by this adjusted scaling factor.

To generate the corrected gene expression matrix, UMI counts for each gene were proportionally scaled according to the final corrected total UMI counts of their respective cells. The resulting matrix, corrected for amplification biases, was natural log-transformed with a pseudocount of 1 for downstream analyses.

Correlation analysis between total LC-UMIs and gene expression

We analyzed 10x Genomics scRNA-seq datasets

from Tabula Sapiens and Tabula Muris. The Tabula Sapiens dataset consists of 456,101 cells spanning 312 cell types across 24 human tissues and organs, while Tabula Muris includes 55,656 cells covering 177 cell types from 12 mouse tissues and organs. To ensure data quality, we excluded cells with mitochondrial gene expression ratios exceeding 10%. The remaining UMI count matrices were processed using the LC-UMI method. For each cell type within each tissue, we computed the average gene expression profile, excluding cell types with fewer than 30 cells. This process yielded 341 tissue-specific cell types in Tabula Sapiens and 72 in Tabula Muris. For each tissue-specific cell type, we calculated the mean total LC-UMIs across all cells of that type. We then assessed the relationship between gene expression levels and total LC-UMIs across all tissue-specific cell types by computing Spearman correlation coefficients using the `cor.test` function in R. The distribution of correlation coefficients was visualized using cumulative distribution plots. In the BM myeloid dataset, we applied $\log(\text{LC-UMI}+1)$ normalization. Pathway activity scores were computed for each cell using ssGSEA, as implemented in the GSVA package (v1.50.0) in R. Spearman correlation coefficients and P-values between total LC-UMIs and pathway activity scores were calculated across cells using `cor.test`. The results were visualized using hexbin plots, with generalized additive model (GAM) regression lines overlaid to capture the overall trend.

Functional enrichment analysis

The gene ontology (GO) dataset, including molecular function (MF), biological process (BP), and cellular component (CC) gene sets, was retrieved using the `org.Hs.eg.db` (v3.16.0) and `org.Mm.eg.db` (v3.16.0) R packages. The Kyoto encyclopedia of genes and genomes (KEGG) dataset was obtained via clusterProfiler (v4.6.2). These GO and KEGG gene sets were used for functional enrichment analysis. The top 500 genes most strongly correlated with total LC-UMIs in the Tabula Sapiens and Tabula Muris datasets (determined by Spearman correlation coefficients) were subjected to over-representation analysis (ORA). ORA was performed using a hypergeometric distribution test, with multiple testing correction applied via the Benjamini-Hochberg false discovery rate (FDR) method. Enrichment analysis was conducted using the `enrichGO` and `enrichKEGG` functions in clusterProfiler. To minimize redundancy in the enrichment results, significantly enriched gene sets ($\text{FDR} < 0.05$) were further analyzed using a custom R script implementing a Cohen's kappa statistic-based similarity analysis, inspired by the ClueGO methodology. This approach grouped functionally related gene sets and selected the most representative

terms to define key biological themes. The top 10 most significantly enriched gene sets were visualized using dot plots. For GSEA, genes were ranked based on their Spearman correlation coefficients with total LC-UMIs. GSEA was performed using the `gseGO` and `gseKEGG` functions in clusterProfiler, which computed enrichment scores, P-values, and FDR. The results were visualized using `gseaplot2` from the `enrichplot` R package (v1.18.4).

Cell type classification evaluation

We used two datasets for cell type classification: the Zheng 68k-cell PBMC dataset and the mouse brain metastatic microenvironment dataset. The PBMC dataset contains 7 cell types, with the subsets of CD4+ and CD8+ T cells from the original study merged into single CD4+ and CD8+ T cell groups, respectively. The brain metastatic microenvironment dataset includes 21 cell types. Cell type annotations in both datasets were carefully established in the original publications and considered accurate, serving as ground truth for classification evaluation.

The UMI count matrices of both datasets were normalized using three methods: $\log(\text{CP10K}+1)$, `scTransform`, and $\log(\text{LC-UMI}+1)$. After normalization, we selected the top 2000 HVGs using Seurat's `FindVariableFeatures` function with default parameters. We performed PCA for dimensionality reduction and used the first 50 principal components (PCs) for t-SNE embedding. To assess cell type separation in the reduced space, we employed a kNN classifier with leave-one-out cross-validation. For each cell, we trained the classifier using the remaining cells, with the class prediction based on the majority vote of its $k=15$ nearest neighbors. kNN classification was carried out using the R package `caret` (v6.0-94). Prediction performance was quantified using the macro F1 score, which averages precision and recall across all classes to account for class imbalance. To further evaluate the classification performance of the three normalization methods, we randomly sampled cells from both datasets. For the PBMC dataset, 500 cells were sampled from each cell type, while for the brain metastatic microenvironment dataset, 200 cells were sampled per cell type (excluding 3 cell types with fewer than 200 cells). Sampling was repeated 10 times. After applying the normalization methods, the top 2000 HVGs were selected, followed by PCA to reduce dimensionality. SNN clustering was performed using Seurat's `FindNeighbors` and `FindClusters` functions with default parameters, adjusting the resolution to ensure the number of clusters matched the actual number of cell type labels. The concordance between the clustering results and predefined cell-type annotations was quantified using NMI and ARI with the R package `aricode` (v1.0.3). The results were visualized using boxplots.

Comparison of cell type marker gene identification

In the 68k PBMC dataset and mouse brain metastatic microenvironment dataset, we identified cell type marker genes using the `find all markers` function from the R package Seurat, with parameters set to: `min.pct = 0.1` and `logfc.threshold = 0`. The Wilcoxon rank-sum test was used to identify marker genes for all expressed genes across different cell types, and p-values were corrected for multiple testing using the Benjamini and Hochberg method (FDR). We used both the $\log(\text{CP10K}+1)$ matrix and the $\log(\text{LC-UMI}+1)$ matrix as input data. Cell type marker genes were defined as those with an adjusted p-value less than 0.05, a $\log_2\text{FC}$ greater than a threshold (which varied between 0.25, 0.5, 1, 2, 3, 4, 5, and 6), and the highest expression level in their respective cell type. We then calculated the overlap between the marker genes identified by the two normalized expression matrices under different $\log_2\text{FC}$ thresholds, along with the number and proportion of uniquely identified marker genes, and presented the results using bar plots.

Differentiation trajectory analysis comparison

In the mouse BM myeloid dataset, to infer the clustering and lineage relationships between monocytes, dendritic cells (DCs), and their precursors in the bone marrow, we utilized Monocle3 (v1.3.7) for trajectory analysis. The UMI count expression matrices were processed using three normalization methods: $\log(\text{CP10K}+1)$, `sctransform`, and $\log(\text{LC-UMI}+1)$. After filtering for the top 2000 HVGs using the Seurat function `find variable features`, the resulting gene expression matrix was used as input for further analysis. PCA dimensionality reduction, UMAP embedding, trajectory graph learning, and pseudo-time measurement through reversed graph embedding were performed with Monocle3. The differentiation trajectories of BM myeloid cells based on the three normalization methods, were visualized using UMAP. Lineage-specific marker genes were plotted along the differentiation trajectory.

Data availability

The raw sequencing data and processed expression matrix for the rat parathyroid dataset have been deposited in the gene expression omnibus (GEO) under accession number GSE291906. This study utilized seven publicly available single-cell datasets from both humans and mice, as summarized in [Supplementary Table SI](#). Raw sequencing data and UMI count matrices are available for download from the GEO database using their respective accession numbers. The cell annotation file for the Tabula Muris dataset can be accessed from ([https://figshare.com/articles/dataset/Single-cell_RNA-](https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2)

[seq_data_from_microfluidic_emulsion_v2_/5968960/2](https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2)). The cell annotation file for the Zheng 68k-cell PBMC dataset is available at (https://github.com/10XGenomics/single-cell-3prime-paper/tree/master/pbmc68k_analysis). The cell annotation file for the mouse brain metastatic microenvironment dataset can be obtained from (<https://doi.org/10.5281/zenodo.13863738>). Cell metadata for the remaining public datasets are directly accessible from the GEO database.

RESULTS

Correction of single-cell UMI count matrix using LOESS regression

The UMI-based molecular counting method mitigates PCR amplification biases, providing an absolute measurement scale with a defined zero ([Islam et al., 2014](#)). By incorporating single-cell indexed DNA barcodes, this approach is expected to enable parallel mRNA quantification at the single-cell level. To evaluate the robustness of UMI counts for absolute transcript measurement across individual cells, we analyzed UMI amplification number distributions in three droplet-based scRNA-seq datasets, including two publicly available datasets human inflammatory bowel disease (IBD) colon (46,698 cells) ([Garrido-Trigo et al., 2023](#)) and human frontal cortex (10,319 cells) ([Lake et al., 2018](#)) as well as our own rat parathyroid dataset (6,844 cells) ([Fig. 1A](#)). Cell identities were assigned based on canonical marker gene expression ([Supplementary Fig. S1](#)) or published annotations. Analysis of single-molecule amplification number distributions across distinct cell populations revealed a broadly consistent pattern across all datasets, though minor variations were observed ([Fig. 1B](#), [Supplementary Fig. S2A, B](#)), supporting the reliability of UMIs for absolute quantification across single cells. To further investigate variability in UMI amplification across cells, we categorized them into ten equal-sized groups based on total UMI counts and examined their single-molecule amplification number distributions. In all datasets, cells with higher total UMI counts exhibited an increased number of amplification copies ([Fig. 1C](#), [Supplementary Fig. S2C, D](#)). Curve fitting demonstrated a positive correlation between total UMI counts and single-molecule amplification numbers, with RNA-rich cells displaying higher UMI amplification numbers across all samples ([Fig. 1D](#), [Supplementary Fig. S3A, C](#)). These findings indicate that UMI counts tend to be overestimated in RNA-rich cells, reinforcing the need for correction methods to ensure accurate cross-cell quantification.

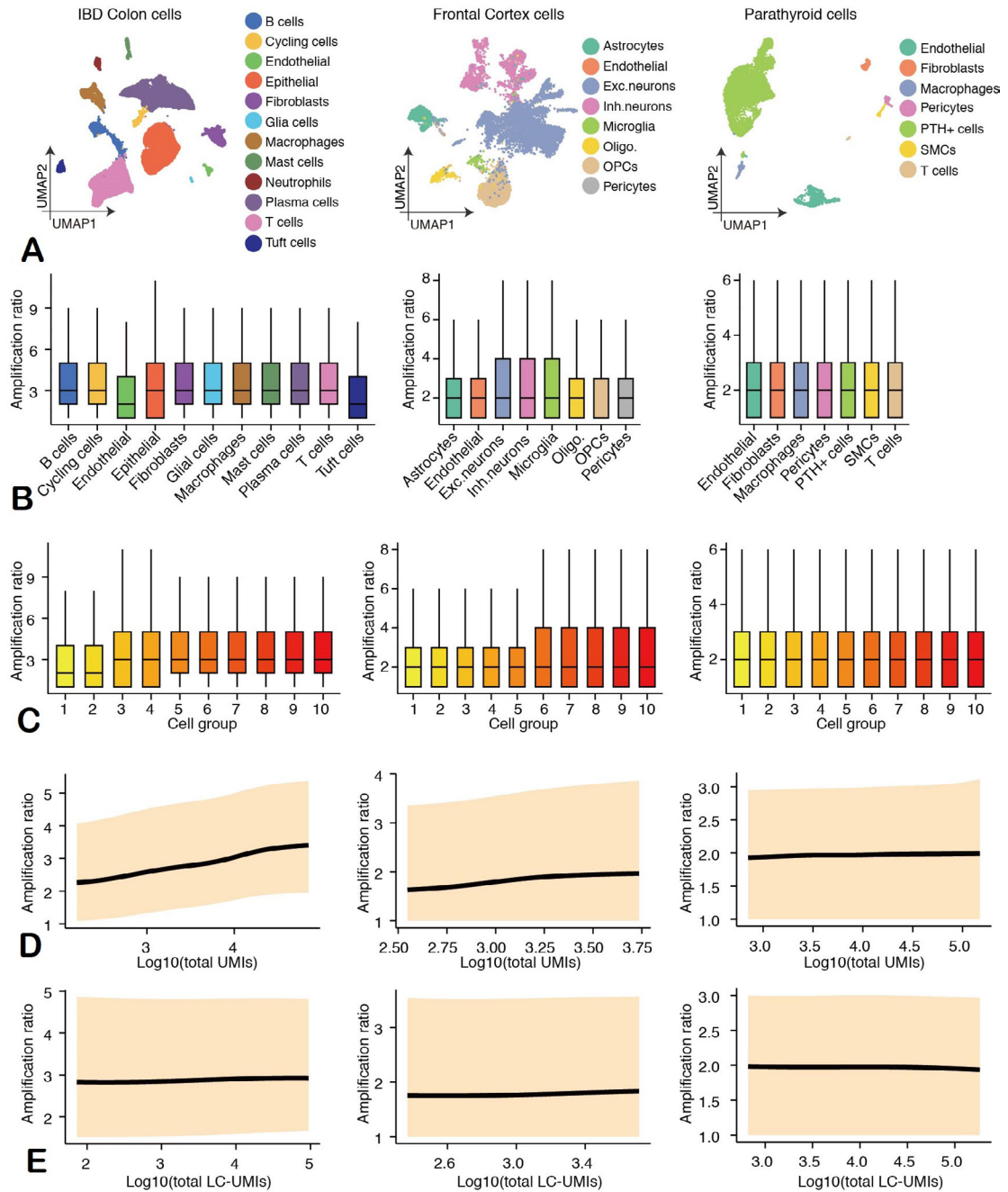


Fig. 1. Distribution of UMI amplification ratio across the human IBD colon, human frontal cortex, and rat parathyroid datasets. Each column represents data from one of the three datasets. (A) UMAP projections of cells from three datasets, shown on the left (human IBD colon, 46,698 cells), middle (human frontal cortex, 10,319 cells), and right (rat parathyroid, 6,844 cells) panels, with cells color-coded by cell type. The projections were generated using log(CP10K+1)-normalized expression data. (B-E) Data from a representative sample per dataset: GSM6614349 for human IBD colon, GSM2581279 for human frontal cortex, and the single sample for rat parathyroid. (B) Distribution of UMI amplification ratio across different cell types. (C) Cells were divided into ten equal-sized groups based on their total UMI counts within the sample, showing the distribution of UMI amplification ratio across groups, ordered from lowest to highest total UMI counts. (D) Relationship between cellular total UMI counts (log10-transformed) and UMI amplification ratio. A smoothed line represents the median UMI amplification ratio, with the colored region indicating the interquartile range. (E) Same as (D), but using total LC-UMI counts instead of total raw UMI counts.

In order to mitigate amplification variability and ensure accurate cross-cell comparisons, we assessed multiple regression models and selected LOESS to adjust UMI values (Supplementary Fig. S4), generating LOESS-corrected UMI count (LC-UMI). To evaluate our regression model, we explored the relationship between total LC-UMIs and UMI amplification numbers. Encouragingly, the correlation was minimized (Fig. 1E, Supplementary Fig. S3B, D, contrast to Fig. 1D, Supplementary Fig. S3A, C). These results demonstrate that LC-UMI enables more accurate quantification, independent of amplification variability across cells.

LC-UMI as an absolute measure of RNA abundance

RNA content varies across cell types due to their distinct physiological roles (Monaco *et al.*, 2019; Huuki-Myers *et al.*, 2023). Proliferative, secretory, and metabolically active cells typically exhibit higher transcriptional activity to support protein synthesis and secretion, leading to increased mRNA abundance (Lin and Amir, 2018; Khetchoumian *et al.*, 2019; Wiggins and Scharer, 2021; Biffo *et al.*, 2024). To evaluate

whether LC-UMI provides an absolute measure of RNA abundance, we compared total LC-UMIs among plasma cells, cycling cells, PTH-secreting cells, and neurons relative to other cell types within the same tissue. As expected, plasma cells in the IBD colon dataset exhibited the highest total LC-UMIs, followed by cycling cells, both of which significantly exceeded other cell types (Wilcoxon test, $p < 0.001$) (Fig. 2A, D; Supplementary Fig. S5A). In the frontal cortex dataset, neurons displayed significantly higher total LC-UMI counts compared to other cell types (Wilcoxon test, $p < 0.001$), with excitatory neurons exhibiting greater LC-UMIs than inhibitory neurons (Fig. 2B, E; Supplementary Fig. S5B). Similarly, in our parathyroid dataset, PTH-secreting cells exhibited markedly higher LC-UMIs than other cell types (Wilcoxon test, $p < 0.001$) (Fig. 2C, F). The consistently high LC-UMIs observed in these transcriptionally active cell types objectively reflect their increased RNA abundance. Collectively, these findings support LC-UMI as a reliable metric for capturing RNA abundance heterogeneity across diverse cell populations.

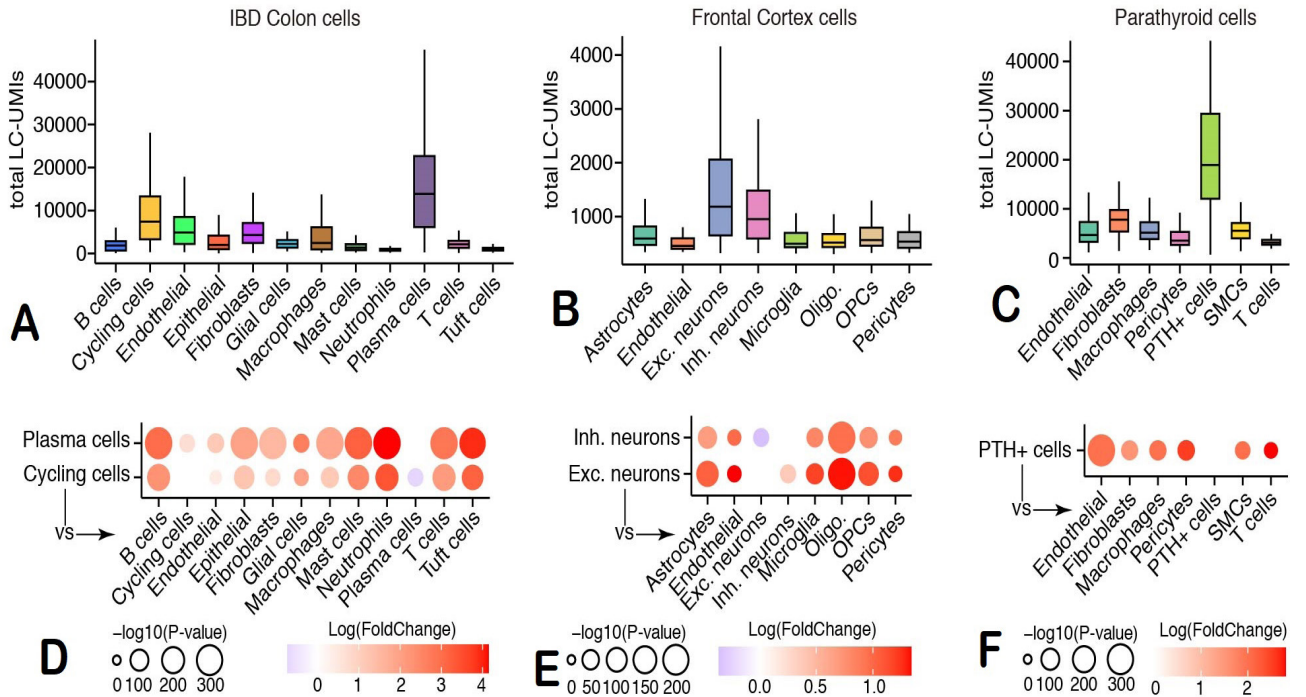


Fig. 2. LC-UMI provides an accurate measure of RNA content across cell types.

Each column corresponds to one of the three datasets, consistent with Figure 1. (A–C) Distribution of total LC-UMIs across different cell types in the human IBD colon dataset (A), human frontal cortex dataset (B), and rat parathyroid dataset (C). (D–F) Comparison of total LC-UMIs between transcriptionally active cell types and all other cell types within each dataset, assessed using the Wilcoxon test. The circle size represents $-\log_{10}(p\text{-value})$ from the Wilcoxon test, while color gradient indicates the $\log_2(\text{fold change})$ between transcriptionally active and other cell types.

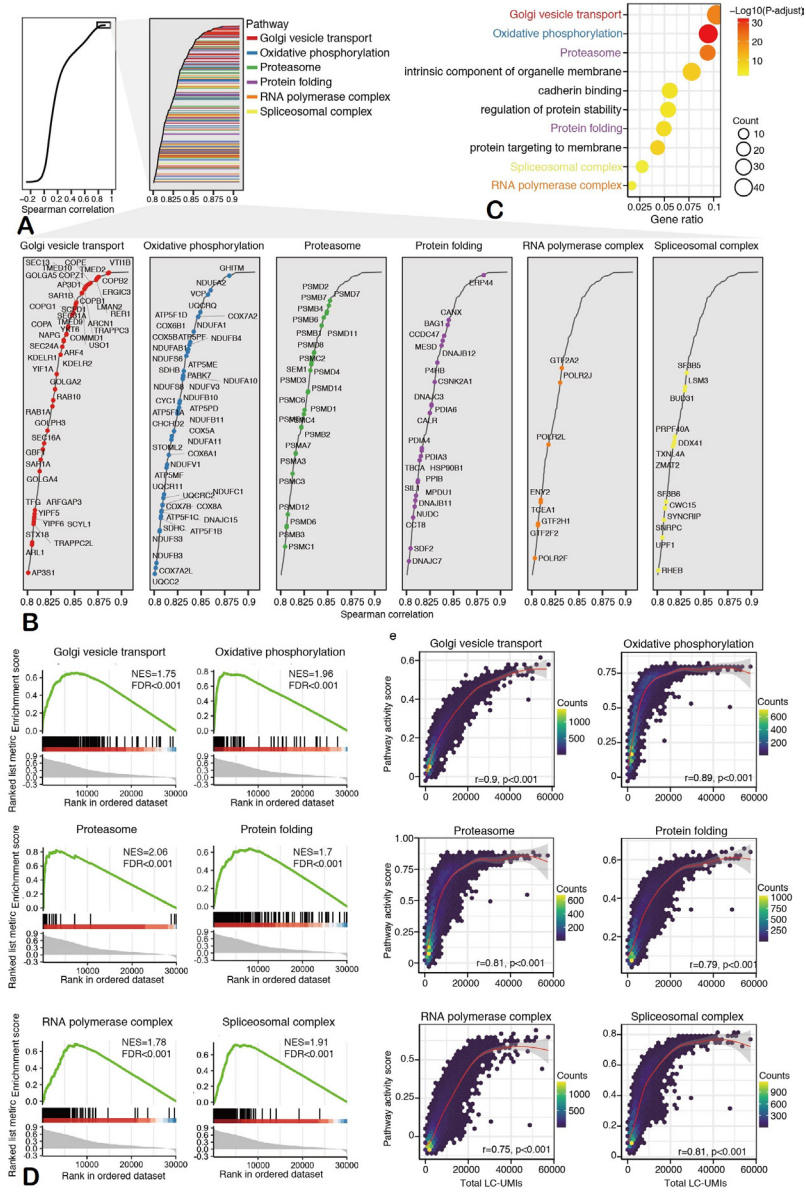


Fig. 3. Correlation analysis between total LC-UMIs and gene expression levels.

(A) Spearman correlation analysis between total LC-UMIs and gene expression levels across 341 cell types in the Tabula Sapiens dataset. The top 500 genes with the highest correlation coefficients are highlighted on the right, with transcription-related pathway genes marked using horizontal segments and color-coded by pathway identity. (B) Detailed view of transcription-related pathways identified in (A), with individual pathways displayed separately. Genes from each pathway that rank among the top 500 most correlated genes are highlighted in distinct colors. (C) Top 10 enriched pathways among the top 500 most correlated genes from (A). The x-axis represents the gene ratio, defined as the proportion of pathway-associated genes relative to the total genes analyzed. The bubble size denotes the number of genes within each pathway, while the color gradient indicates the statistical significance, represented by $-\log_{10}(\text{adjusted } p\text{-value})$. (D) GSEA results demonstrating the enrichment of transcription-related pathways. Each panel displays the enrichment score (green line) for a given pathway across the ranked gene list, ordered by Spearman correlation coefficients from (A). The normalized enrichment score (NES) and false discovery rate (FDR) are indicated. (E) Single-cell correlation analysis between total LC-UMIs and pathway activity scores computed via ssGSEA in the mouse BM myeloid dataset. Each hexagonal bin represents the density of data points, with a color gradient indicating cell density (legend shown in color bar). A generalized additive model (GAM) smoothing curve (red line) with a 95% confidence interval (shaded region) is fitted to the data. The Spearman correlation coefficient (r) and p -value are displayed in each panel.

Genes correlated with total LC-UMIs are enriched in transcription-related pathways

To further evaluate LC-UMI as an absolute measure of RNA abundance at the single-cell level, we analyzed genes correlated with total LC-UMI counts in the Tabula Sapiens and Tabula Muris datasets (The Tabula Sapiens Consortium, 2022; The Tabula Sapiens Consortium, 20), which provide scRNA-seq data across diverse human and mouse tissues. For each tissue-specific cell type, we calculated the average LC-UMIs of gene expression, resulting in transcriptomic profiles for 341 human cell types and 72 mouse cell types. We then performed correlation analyses between average total LC-UMIs and gene expression levels across all cell types in both datasets, identifying the top 500 genes with the highest correlation coefficients (Fig. 3A, F, Supplementary Fig. S6A). We hypothesized that cells with high total LC-UMIs should elevated expression of transcription-related pathways. In agreement with this idea, we found that the RNA polymerase complex was significantly enriched among the top 500 genes most correlated with total LC-UMIs in both human and mouse datasets (Fig. 3A-C, Supplementary Fig. S6A-C). In the human dataset, these genes included core RNA polymerase II (RNA Pol II) subunits (*POLR2J*, *POLR2L*, *POLR2F*), general transcription factors (*GTF2A2*, *GTF2H1*, *GTF2F2*), a transcription elongation regulator (*TCEA1*), and a transcriptional coactivator (*ENY2*) (Fig. 3B). Similarly, in the mouse dataset, we observed enrichment of RNA Pol II subunits (*Polr2L*, *Polr2K*, *Polr2C*, *Polr2F*, *Polr2I*), general transcription factors (*Gtf2a2*, *Gtf2h5*), a transcription elongation regulator (*Tcea1*), and transcriptional coactivators (*Eny2*, *Taf9*, *Psmc5*) (Supplementary Fig. S6B). In addition, we identified Golgi vesicle transport, oxidative phosphorylation, proteasome, protein folding, and the spliceosomal complex as consistently enriched among the top 500 correlated genes in both species (Fig. 3A-C, Supplementary Fig. S6A-C). These pathways collectively ensure the efficient production, processing, and maintenance of proteins and RNA, which are essential for the transcriptional machinery (Alberts *et al.*, 2002; Chen *et al.*, 2023). For instance, oxidative phosphorylation provides the necessary ATP to power transcription (Shin *et al.*, 2020), while the proteasome and protein folding pathways maintain protein homeostasis by ensuring proper protein conformation and degradation of misfolded proteins, both crucial for the function of transcription factors and RNA polymerases (Collins and Tansey, 2006). Additionally, Golgi vesicle transport and the spliceosomal complex directly contribute to mRNA processing and transport, ensuring that newly synthesized transcripts are correctly modified and delivered for translation (Alberts *et*

al., 2002; Herzel *et al.*, 2017).

Gene set enrichment analysis (GSEA) of correlation coefficients further confirmed that these pathways were preferentially activated in cells with high total LC-UMIs (Fig. 3D, Supplementary Fig. S6D). Moreover, single-sample GSEA (ssGSEA) scores in the mouse bone marrow (BM) myeloid dataset (Trzebanski *et al.*, 2024) supported a strong association between total LC-UMIs and the activity of these transcription-related pathways at the single-cell level (Fig. 3E). These findings collectively highlight the strong association between total LC-UMIs and transcriptional activity, reinforcing LC-UMI as a robust metric for quantifying RNA abundance at the single-cell level.

Applying $\log(\text{LC-UMI}+1)$ for cell classification

To assess the suitability of LC-UMI for downstream scRNA-seq analysis, we applied log transformation with a pseudocount of 1 ($\log(\text{LC-UMI}+1)$) to approximate a normal distribution. We then compared its performance against two widely used Seurat normalization methods: $\log(\text{CP10K}+1)$ (Butler *et al.*, 2018) and *scrans* form (Hafemeister and Satija, 2019). Following normalization, we identified highly variable genes (HVGs) and performed principal component analysis (PCA), followed by shared nearest neighbor (SNN) clustering for cell classification. To assess normalization effectiveness, we applied these methods to two well-annotated single-cell datasets: The Zheng 68k peripheral blood mononuclear cell (PBMC) dataset (68,579 cells) (Zheng *et al.*, 2017) and the mouse brain metastatic microenvironment dataset (31,657 cells) (Gan *et al.*, 2024). Cell classification accuracy was assessed using the macro F1 score from k-nearest neighbors (KNN) classification. To control for cell type imbalances, we performed down sampling and evaluated clustering performance using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI).

For both datasets, the t-distributed stochastic neighbor embedding (t-SNE) projections following the three normalization methods were highly similar, effectively separating distinct cell type populations (Fig. 4A, C). In the PBMC dataset, the F1 score from KNN classification using $\log(\text{LC-UMI}+1)$ was 0.789, which was nearly identical to $\log(\text{CP10K}+1)$ (0.794) and *scrans* form (0.788) (Fig. 4B). Similarly, in the mouse brain metastatic microenvironment dataset, $\log(\text{LC-UMI}+1)$ achieved an F1 score of 0.973, comparable to $\log(\text{CP10K}+1)$ (0.975) and *scrans* form (0.973) (Fig. 4D). The relatively lower classification accuracy in the PBMC dataset is likely due to the presence of functionally similar cell types, such as NK cells and CD8+ T cells, which are inherently challenging to distinguish based on transcriptional

profiles. This observation is consistent with previous

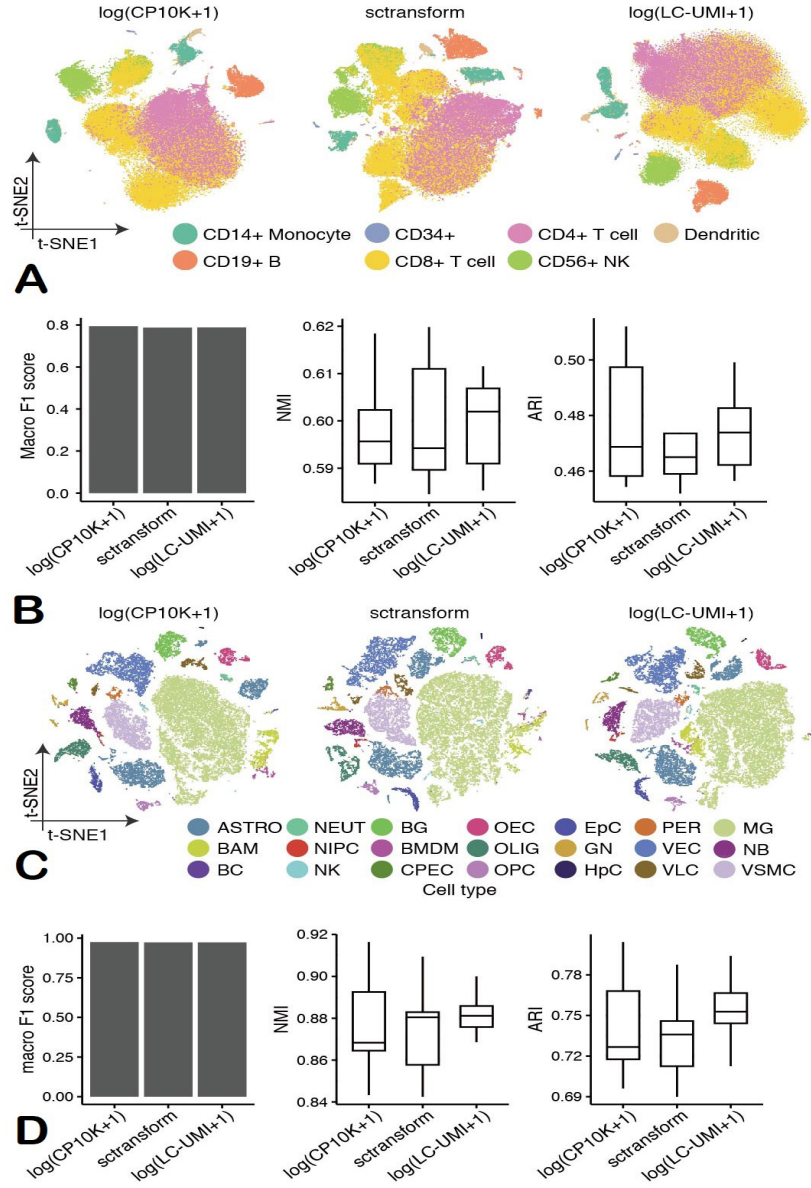


Fig. 4. Benchmarking the effect of different normalization methods on cell type classification in reduced dimensionality. (A) t-SNE embeddings of the PBMC dataset after PCA reduction, comparing three normalization methods: log(CP10K+1) (left), scransform (middle), and log(LC-UMI+1) (right). Cells are color-coded according to their original annotations. (B) Classification performance evaluation for the PBMC dataset. From left to right: macro F1 score for kNN classification (k = 15) across seven ground truth cell types for each normalization method; NMI and ARI scores measuring concordance between clustering results and ground truth labels. For robustness, 500 cells per cell type were randomly sampled and tested over 10 iterations. (C) t-SNE embeddings of the mouse brain metastatic microenvironment dataset after PCA reduction, with normalization methods shown in the same order as (A). Cells are color-coded by their original cell type labels, including astrocytes (ASTRO), border-associated macrophages (BAM), B cells (BC), Bergmann glia (BG), bone marrow-derived macrophages (BMDM), choroid plexus epithelial cells (CPEC), ependymal cells (EpC), granule neurons (GN), hypendymal cells (HpC), microglia (MG), neuroblasts (NB), neutrophils (NEUT), neuronal intermediate progenitors (NIPC), NK cells (NK), olfactory ensheathing cells (OEC), oligodendrocytes (OLIG), oligodendrocyte precursors (OPC), pericytes (PER), vascular endothelial cells (VEC), vascular leptomeningeal cells (VLC), and vascular smooth muscle cells (VSMC). (D) Classification performance evaluation for the mouse brain metastatic microenvironment dataset, with metrics presented as in (B). For robustness, 200 cells per cell type were randomly

sampled and tested over 10 iterations.

findings from the original dataset publication. In addition, the NMI and ARI evaluations indicated that $\log(\text{LC-UMI}+1)$ -normalized clustering results exhibited slightly better concordance with known cell type labels than those obtained with $\log(\text{CP10K}+1)$ and *sctransform* (Fig. 4B, D). Next, we evaluated the computational efficiency of the three normalization methods. $\log(\text{CP10K}+1)$ was the most computationally efficient, while $\log(\text{LC-UMI}+1)$ was consistently faster than *sctransform* across both datasets (Supplementary Table SII). Importantly, similar to $\log(\text{CP10K}+1)$, the $\log(\text{LC-UMI}+1)$ transformation preserves zero values in the expression matrix and maintains matrix sparsity, enabling memory-efficient storage and computation especially critical for large-scale single-cell datasets. In contrast, *sctransform* generates a dense matrix of Pearson residuals, substantially increasing memory usage during downstream analysis. These findings demonstrate that $\log(\text{LC-UMI}+1)$ normalization achieves superior or comparable performance to $\log(\text{CP10K}+1)$ and *sctransform* in preserving biological variability and distinguishing cell types in scRNA-seq data. Given its simplicity and computational efficiency, $\log(\text{LC-UMI}+1)$ serves as a robust alternative for single-cell normalization.

Comparison of marker gene identification using $\log(\text{CP10K}+1)$ and $\log(\text{LC-UMI}+1)$ normalization

To evaluate the impact of $\log(\text{LC-UMI}+1)$ normalization on marker gene identification, we performed differential gene expression analysis on the Zheng 68k PBMC dataset (Zheng *et al.*, 2017) and the mouse brain metastatic microenvironment dataset (Gan *et al.*, 2024), comparing its performance with $\log(\text{CP10K}+1)$ normalization. Marker genes for each cell type were identified using both normalized expression matrices. A gene was considered a marker gene for a specific cell type if it met the following criteria: (1) an adjusted *p*-value < 0.05 from a Wilcoxon test comparing its expression between the target cell type and all other cell types, and (2) the highest expression level in that cell type relative to all others. To further refine marker gene selection, we applied varying $\log_2$ fold change ($\log_2\text{FC}$) thresholds to filter genes based on their differential expression levels. We then compared the similarities and differences in marker genes identified using the two normalization methods across different thresholds. At lower fold change thresholds (e.g., $\log_2\text{FC} < 0.25$), distinct differences were observed between the two normalization methods, with a noticeable fraction of marker genes being method-specific (Fig. 5, Supplementary Fig. S7, Supplementary Table SIII). However, as the threshold increased, the proportion of overlapping marker genes identified by $\log(\text{CP10K}+1)$

and $\log(\text{LC-UMI}+1)$ also increased, indicating strong agreement in the detection of high-expression markers.

Consistently, most canonical marker genes were identified as shared markers by both normalization methods, supporting the biological validity of $\log(\text{LC-UMI}+1)$ normalization. In contrast, method-specific markers often resulted from borderline expression differences across multiple cell types, where moderate expression in more than one population led to shifts in marker assignment depending on normalization strategy (Supplementary Table SIV). For example, in the PBMC dataset, *FCGR3A* was identified as a marker for CD56⁺ NK cells under $\log(\text{CP10K}+1)$ normalization, but as a CD14⁺ monocyte marker under $\log(\text{LC-UMI}+1)$, likely due to its intermediate expression in both lineages and shifts in expression ranking introduced by normalization. Together, these results support $\log(\text{LC-UMI}+1)$ as a reliable normalization method for differential gene expression analysis in single-cell RNA-seq data.

$\log(\text{LC-UMI}+1)$ supports robust cell differentiation trajectory analysis

To assess the effectiveness of $\log(\text{LC-UMI}+1)$ normalization on cell differentiation trajectory inference, we analyzed a mouse BM myeloid dataset comprising 4,705 myeloid immune cells and their progenitors, as reliably annotated in the original study (Trzebanski *et al.*, 2024). Using Monocle3 (Cao *et al.*, 2019), we reconstructed differentiation trajectories based on expression matrices normalized by $\log(\text{CP10K}+1)$, *sctransform*, and $\log(\text{LC-UMI}+1)$. Uniform manifold approximation and projection (UMAP) results demonstrated a high degree of concordance across all three normalization methods (Fig. 6A-C), indicating that $\log(\text{LC-UMI}+1)$ preserves the underlying differentiation structure. In the trajectory inferred using $\log(\text{LC-UMI}+1)$, hematopoietic stem cells (HSCs) were positioned at the root. The differentiation landscape resolved into three major branches, corresponding to the monocyte, plasmacytoid dendritic cell (pDC), and granulocyte lineages (Fig. 6C). These differentiation trajectories were further validated by the expression patterns of lineage-specific marker genes (Fig. 6F). Specifically, HSCs differentiated into common myeloid progenitors (CMPs) and granulocyte-monocyte progenitors (GMPs). GMPs gave rise to two distinct lineages: One branch generated granulocyte progenitors (GPs), while the other progressed through common monocyte progenitors (CMoPs) to pre-monocytes (preMo), which subsequently differentiated into neutrophil-like monocytes (NeuMo), intermediate monocytes (IntMo), and dendritic cell-like monocytes

(DCMo), ultimately giving rise to conventional dendritic cells (cDCs) and other antigen-presenting cell populations.

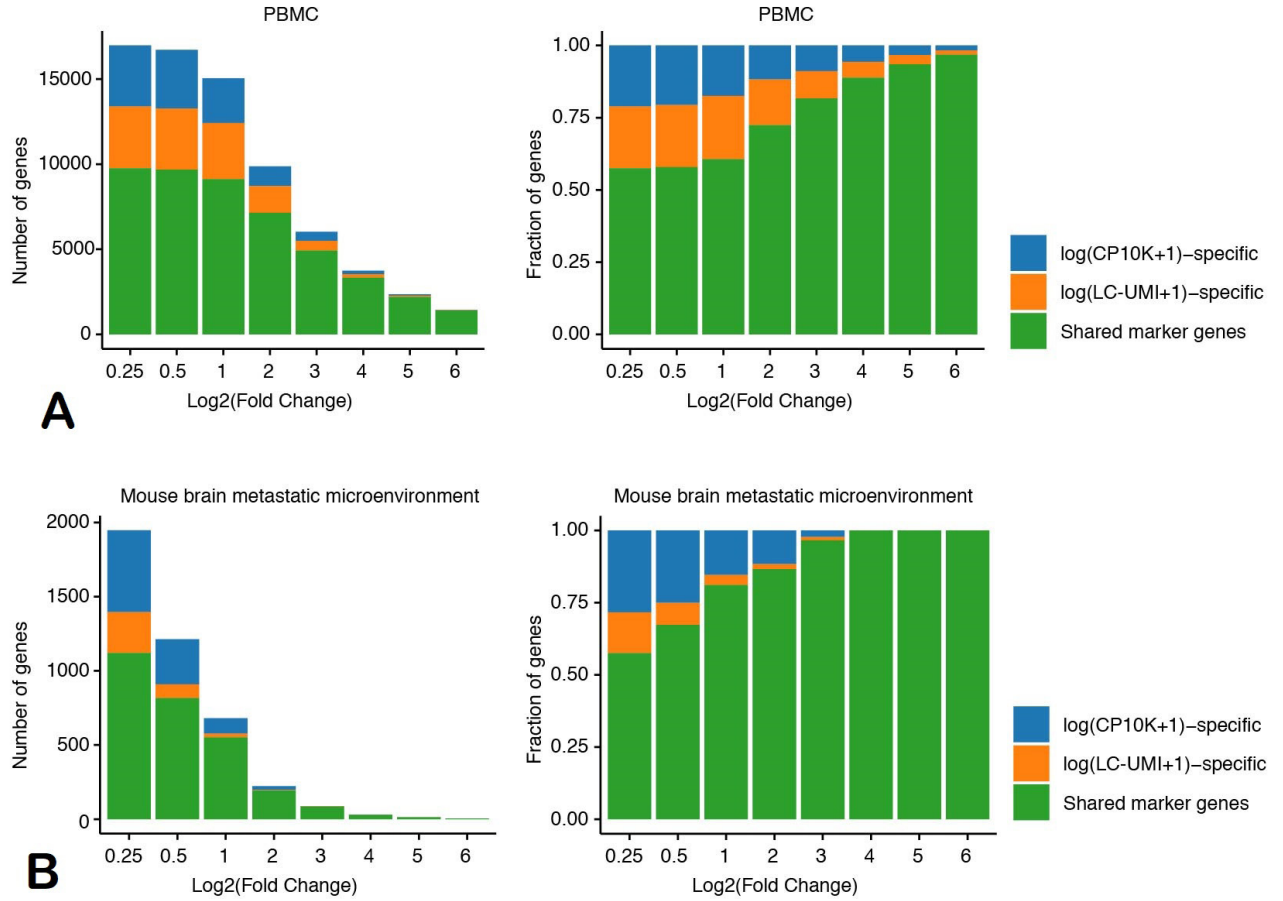


Fig. 5. Comparison of marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization across different fold change thresholds.

Stacked bar plots comparing marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization in the PBMC dataset (A) and the mouse brain metastatic microenvironment dataset (B). Differential gene expression analysis was performed using log(CP10K+1) and log(LC-UMI+1) normalized expression matrices. Marker genes for each cell type were identified using Wilcoxon tests (adjusted p-value < 0.05), required to have the highest expression in the corresponding cell type compared to all others, and further filtered by varying log2FC thresholds. Stacked bar plots depict the number (left) and proportion (right) of marker genes at different log2FC thresholds. Shared marker genes between both methods are shown in green, while those specific to log(CP10K+1) and log(LC-UMI+1) are represented in blue and orange, respectively.

In parallel, an alternative pathway originated from HSCs differentiating into common lymphoid progenitors (CLPs), which further progressed through CLP-pre-dendritic cells (CLP-preDCs) and pre-dendritic cells (preDCs), ultimately generating pDCs (Fig. 6C). This reconstructed trajectory closely recapitulates the known hierarchical differentiation of BM myeloid immune cells, aligning well with established hematopoietic differentiation models (Trzebanski *et al.*, 2024; Jacobsen and Nerlov, 2019; Ma *et al.*, 2022). These findings indicate that the log(LC-UMI+1) is well-suited for single-cell differentiation

trajectory analysis, yielding performance comparable to that of log(CP10K+1) and sctransform.

DISCUSSION

UMIs have emerged as a powerful tool for enhancing RNA quantification accuracy by correcting PCR amplification biases, providing an absolute measurement scale with a defined zero (Islam *et al.*, 2014). By integrating single-cell indexed DNA barcodes, UMIs facilitate parallel mRNA quantification at the single-cell level. However, challenges such as low capture efficiency

and technical variability continue to limit their widespread application for absolute quantification, highlighting the need for further refinement. In this study, we introduce

LC-UMI as a novel method for absolute gene expression

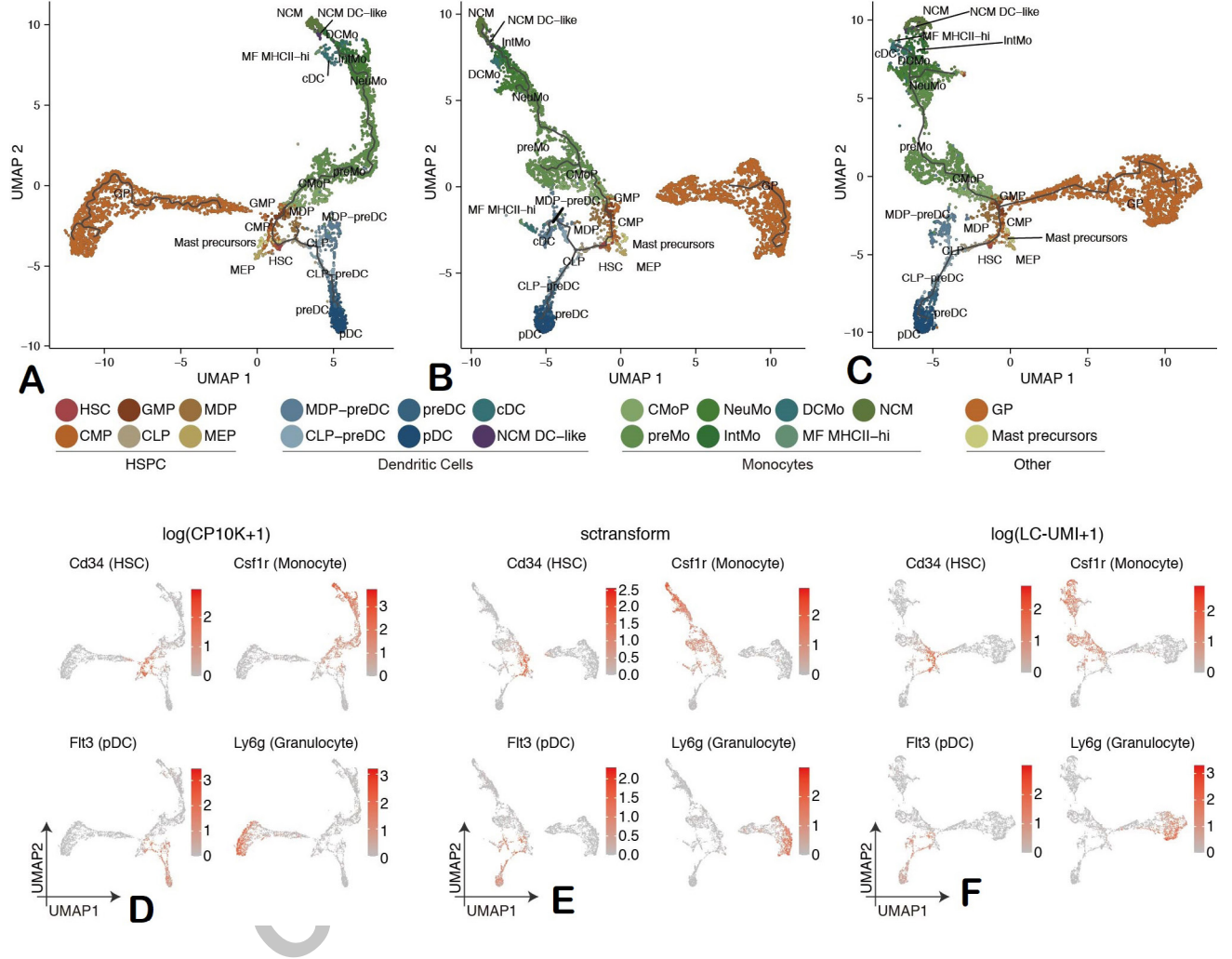


Fig. 6. Differentiation trajectories of the mouse bone marrow myeloid dataset using different normalization methods. (A–C) UMAP projections of differentiation trajectories inferred using Monocle3 with expression matrices normalized by (A) $\log(\text{CP10K}+1)$, (B) sctransform , and (C) $\log(\text{LC-UMI}+1)$. Cells are color-coded by their annotated cell types from the original study. (D–F) Expression patterns of lineage-specific marker genes mapped onto the differentiation trajectories shown in (A–C). For each normalization method, the expression levels of *Cd34* (HSCs), *Csf1r* (monocytes), *Flt3* (pDCs), and *Ly6g* (granulocytes) is visualized, with red indicating higher expression levels.

quantification at the single-cell level. Unlike conventional scRNA-seq normalization strategies, which typically treat total RNA counts as a technical artifact of sequencing depth (Ahlmann-Eltze and Huber, 2023; Hafemeister and Satija, 2019), LC-UMI recognizes total UMI counts as a biologically meaningful signal that reflects the actual RNA content of individual cells. Therefore, LC-UMI provides a more informative metric for gene expression quantification, which may facilitate a deeper understanding of cell type-specific transcriptional programs and underlying biological

mechanisms.

While UMIs effectively reduce PCR amplification bias, we observed a systematic cross-cell bias in UMI amplification, wherein cells with higher RNA content exhibited increased read-to-UMI ratios. This bias likely arises from the non-uniform nature of PCR amplification, influenced by factors such as GC content, transcript length, and the proportion of newly synthesized RNA (Kebschull and Zador, 2015; Gorin and Pachter, 2023). Consequently, UMI counts are overestimated in RNA-

rich cells under a fixed sequencing depth, leading to a disproportionate increase in total UMI counts relative to RNA-poor cells, thereby distorting their true RNA molecule ratios. To address this issue, we evaluated five regression models: linear regression, second-degree polynomial regression, natural spline ($df=4$), LOESS ($span=0.75$), and generalized additive models (GAM) across three representative datasets (Supplementary Fig. S4). Natural spline, LOESS, and GAM consistently outperformed linear and polynomial regression in terms of lower RMSE and higher R^2 . Although GAM achieved slightly better fit metrics, it showed minor oscillations (wiggles) in regions with sparse data, suggesting potential overfitting. In contrast, LOESS produced smooth, stable, and biologically plausible trends across all datasets, without introducing edge artifacts or instability. Given its empirical performance and minimal parameter tuning, LOESS offers a favorable balance between accuracy, interpretability, and robustness, making it particularly suitable for correcting UMI-related biases at the single-cell level. Unlike GAM or spline-based methods, LOESS does not require predefined basis functions or smoothing penalties, relying instead on a simple span parameter. This makes it both flexible and interpretable, allowing accurate UMI correction while preserving biological heterogeneity. By adjusting UMI counts based on the average amplification ratio per cell, LC-UMI enhances RNA quantification accuracy and better reflects true differences in cellular RNA content.

Applying LC-UMI to multiple single-cell datasets confirmed that transcriptionally active cell types including plasma cells, neurons, and PTH-secreting cells exhibited significantly higher total LC-UMIs than other cell populations. Furthermore, total LC-UMIs showed a strong positive correlation with pathways associated with high transcriptional activity, including RNA polymerase complex, Golgi vesicle transport, oxidative phosphorylation, proteasome, protein folding, and spliceosomal complex (Alberts *et al.*, 2002; Chen *et al.*, 2023; Shin *et al.*, 2020; Collins and Tansey, 2006; Herzel *et al.*, 2017). This strong association underscores the potential of LC-UMI as a proxy for absolute RNA content at the single-cell level.

Despite its advantages, LC-UMI has certain limitations. Single-cell absolute quantification is influenced by multiple factors, including cell lysis efficiency (Homburger *et al.*, 2023), RNA capture efficiency (Tang *et al.*, 2023), and sequencing saturation (Zhang *et al.*, 2020), which cannot be corrected by a single LOESS regression model. Furthermore, LC-UMI operates at the cell level and does not account for gene-specific amplification biases. While the method adjusts total UMI

counts per cell to correct for cross-cell variability, gene-level UMI counts are proportionally rescaled based on each cell's corrected total UMI count, thereby preserving the original within-cell gene expression ratios. This approach minimizes the introduction of gene-level distortions and avoids disproportionately affecting genes with specific transcript features, such as GC content or transcript length. Notably, we observed that highly expressed genes tend to exhibit elevated UMI amplification ratios (reads per UMI), suggesting the presence of expression-dependent amplification bias (Supplementary Fig. S8). Although LC-UMI does not directly correct for such gene-level effects, it could be complemented in future studies by gene-aware normalization strategies such as negative binomial regression or Bayesian hierarchical models to further improve quantification accuracy.

In conclusion, we present LC-UMI as a simple, computationally efficient, and high-throughput method for correcting UMI-based scRNA-seq data, offering a robust framework for absolute RNA quantification at the single-cell level. By accurately capturing differences in total RNA content and transcriptional activity across cell types, LC-UMI provides a valuable tool for single-cell studies. It is well suited for datasets exhibiting substantial cross-cell UMI amplification bias or containing heterogeneous cell populations with pronounced differences in RNA abundance. Furthermore, the $\log(LC-UMI+1)$ matrix serves as an improved input for downstream analyses, including cell clustering, differential gene expression analysis, and pseudotime trajectory inference. Future advancements in statistical modeling and experimental protocols will further enhance the accuracy and reproducibility of single-cell RNA absolute quantification.

DECLARATIONS

Acknowledgement

The authors thank colleagues from East China Normal University and Dalian Medical University for their valuable assistance, and appreciate the reviewers and editors for their helpful comments.

Funding

This work was supported by the East China Normal University Health Next Generation Fund (No. 2023-06) and the Youth Talent Cultivation Fund Project of Dalian Medical University (No. 508005).

IRB approval

Not applicable; only publicly available, de-identified human datasets from previously approved studies were analyzed. No new research involving human participants

was conducted.

Ethical statement

All animal experiments were conducted in accordance with institutional guidelines and were approved by the Animal Welfare and Ethics Committee of East China Normal University (Approval No. m+R20250402).

Generative AI and AI-assisted technology statement

Generative AI (ChatGPT, OpenAI) was used solely for language editing.

Supplementary material

There is supplementary material associated with this article. Access the material online at: <https://dx.doi.org/10.17582/journal.pjz/20250606044826>

Statement of conflict of interest

The authors have declared no conflict of interests regarding the publication of this article.

REFERENCES

- Ahlmann-Eltze, C. and Huber, W., 2023. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods*, **20**: 665–672. <https://doi.org/10.1038/s41592-023-01814-1>
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., 2002. Transport from the ER through the Golgi Apparatus. In: *Molecular biology of the cell*, 4th edn (ed. B. Alberts). Garland Science, New York.
- Biffo, S., Ruggero, D. and Santoro, M.M., 2024. The crosstalk between metabolism and translation. *Cell Metab.*, **36**: 1945–1962. <https://doi.org/10.1016/j.cmet.2024.07.022>
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**: 411–420. <https://doi.org/10.1038/nbt.4096>
- Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C. and Shendure, J., 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**: 496–502. <https://doi.org/10.1038/s41586-019-0969-x>
- Chen, X., Shi, C., He, M., Xiong, S. and Xia, X., 2023. Endoplasmic reticulum stress: Molecular mechanism and therapeutic targets. *Signal Transduct. Target. Ther.*, **8**: 352. <https://doi.org/10.1038/s41392-023-01570-w>
- Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**: 829–836. <https://doi.org/10.1080/01621459.1979.10481038>
- Collins, G.A. and Tansey, W.P., 2006. The proteasome: a utility tool for transcription? *Curr. Opin. Genet. Dev.*, **16**: 197–202. <https://doi.org/10.1016/j.gde.2006.02.009>
- Gan, S., Macalinao, D.G., Shahoei, S.H., Tian, L., Jin, X., Basnet, H., Bibby, C., Muller, J.T., Atri, P., Seffar, E., Chatila, W., Karacay, A., Chanda, P., Hadjantonakis, A.K., Schultz, N., Brogi, E., Bale, T.S., Moss, N.S., Murali, R., Pe'er, D. and Massagué, J., 2024. Distinct tumor architectures and microenvironments for the initiation of breast cancer metastasis in the brain. *Cancer Cell*, **42**: 1693–1712. <https://doi.org/10.1016/j.ccell.2024.08.015>
- Garrido-Trigo, A., Corraliza, A.M., Veny, M., Dotti, I., Melón-Ardanaz, E., Rill, A., Crowell, H.L., Corbí, Á., Gudiño, V., Esteller, M., Álvarez-Teubel, I., Aguilar, D., Masamunt, M.C., Killingbeck, E., Kim, Y., Leon, M., Visvanathan, S., Marchese, D., Caratù, G., Martín-Cardona, A., Esteve, M., Ordás, I., Panés, J., Ricart, E. and Salas, A., 2023. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat. Commun.*, **14**: 4506. <https://doi.org/10.1038/s41467-023-40156-6>
- Gorin, G. and Pachter, L., 2023. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophys. Rep.*, **3**: 100097. <https://doi.org/10.1016/j.bpr.2022.100097>
- Hafemeister, C. and Satija, R., 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**: 296. <https://doi.org/10.1186/s13059-019-1874-1>
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. *Cell*, **184**: 3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>
- Herzel, L., Ottoz, D.S.M., Alpert, T. and Neugebauer, K.M., 2017. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.*, **18**: 637–650.

- <https://doi.org/10.1038/nrm.2017.63>
- Homberger, C., Hayward, R.J., Barquist, L. and Vogel, J., 2023. Improved bacterial single-cell RNA-seq through automated MATQ-seq and Cas9-based removal of rRNA reads. *mBio*, **14**: e03557-22. <https://doi.org/10.1128/mbio.03557-22>
- Huuki-Myers, L.A., Montgomery, K.D., Kwon, S.H., Page, S.C., Hicks, S.C., Maynard, K.R. and Collado-Torres, L., 2023. Data-driven identification of total RNA expression genes for estimation of RNA abundance in heterogeneous cell types highlighted in brain tissue. *Genome Biol.*, **24**: 233. <https://doi.org/10.1186/s13059-023-03066-w>
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**: 163–166. <https://doi.org/10.1038/nmeth.2772>
- Jacobsen, S.E.W. and Nerlov, C., 2019. Haematopoiesis in the era of advanced single-cell technologies. *Nat. Cell Biol.*, **21**: 2–8. <https://doi.org/10.1038/s41556-018-0227-8>
- Kebschull, J.M. and Zador, A.M., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucl. Acids Res.*, **43**: e143. <https://doi.org/10.1093/nar/gkv717>
- Khetchoumian, K., Balsalobre, A., Mayran, A., Christian, H., Chénard, V., St-Pierre, J. and Drouin, J., 2019. Pituitary cell translation and secretory capacities are enhanced cell autonomously by the transcription factor Creb3l2. *Nat. Commun.*, **10**: 3960. <https://doi.org/10.1038/s41467-019-11894-3>
- Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J., 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**: 72–74. <https://doi.org/10.1038/nmeth.1778>
- Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V. and Zhang, K., 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.*, **36**: 70–80. <https://doi.org/10.1038/nbt.4038>
- Lin, J. and Amir, A., 2018. Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.*, **9**: 4496. <https://doi.org/10.1038/s41467-018-06714-z>
- Lun, A.T.L., Bach, K. and Marioni, J.C., 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**: 75. <https://doi.org/10.1186/s13059-016-0947-7>
- Ma, S., Wang, S., Ye, Y., Ren, J., Chen, R., Li, W., Sun, W., Li, J., Zhao, L., Zhao, Q., Sun, G., Jing, Y., Zuo, Y., Xiong, M., Yang, Y., Wang, Q., Lei, J., Sun, S., Long, X., Song, M., Yu, S., Chan, P., Wang, J., Qi, Z., Belmont, J.C.I., Qu, J., Zhang, W. and Liu, G.H., 2022. Heterochronic parabiosis induces stem cell revitalization and systemic rejuvenation across aged tissues. *Cell Stem Cell*, **29**: 990–1005. <https://doi.org/10.1016/j.stem.2022.04.017>
- Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Magalhães, J.P. and Larbi, A., 2019. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.*, **26**: 1627–1640. <https://doi.org/10.1016/j.celrep.2019.01.041>
- Shin, B., Benavides, G.A., Geng, J., Koralov, S.B., Hu, H., Darley-Usmar, V.M. and Harrington, L.E., 2020. Mitochondrial oxidative phosphorylation regulates the fate decision between pathogenic Th17 and regulatory T cells. *Cell Rep.*, **30**: 1898–1909. <https://doi.org/10.1016/j.celrep.2020.01.022>
- Tang, W., Jørgensen, A.C.S., Marguerat, S., Thomas, P. and Shahrezaei, V., 2023. Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. *Bioinformatics*, **39**: btad395. <https://doi.org/10.1093/bioinformatics/btad395>
- The Tabula Muris Consortium, Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T. and van Weele, L.J., 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**: 367–372. <https://doi.org/10.1038/s41586-018-0590-4>
- The Tabula Sapiens Consortium, Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R. and Giudice, L.C., 2022. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, **376**: eabl4896.
- Trzebanski, S., Kim, J.S., Larossi, N., Raanan, A., Kancheva, D., Bastos, J., Haddad, M., Solomon, A., Sivan, E., Aizik, D., Kralova, J.S., Gross-Vered, M., Boura-Halfon, S., Lapidot, T., Alon, R., Movahedi, K. and Jung, S., 2024. Classical monocyte ontogeny dictates their functions and fates as tissue macrophages. *Immunity*, **57**: 1225–1242. <https://doi.org/10.1016/j.immuni.2024.04.019>
- Wiggins, K.J. and Scharer, C.D., 2021. Roadmap to a plasma cell: Epigenetic and transcriptional cues that guide B cell differentiation. *Immunol. Rev.*, **300**: 54–64. <https://doi.org/10.1111/imr.12934>
- Zhang, M.J., Ntranos, V. and Tse, D., 2020. Determining

sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.*, **11**: 774. <https://doi.org/10.1038/s41467-020-14482-y>

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Luz Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J. and Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**: 14049. <https://doi.org/10.1038/ncomms14049>

Online First Article



Supplementary Material

Correcting Cross-Cell UMI Bias for Improved Single-Cell Transcriptomics

Rulin Hua¹, Wenjing Yan², Xiaoyan Liu³, Suyang Xuan⁴, Xuejuan Li⁵, Shafiq Ur Rahman¹, Dekang Lv^{4*} and Feng Zheng^{1,5*}

¹Advanced Institute for Medical Sciences, Dalian Medical University, Dalian, 116044, China

²Dalian Center for Disease Control and Prevention, Dalian, 116021, China

³Department of Nephrology, the Second Hospital of Dalian Medical University, Dalian, 116023, China

⁴Institute of Cancer Stem Cell, Dalian Medical University, Dalian, 116044, China

⁵Wuhu Hospital and Health Science Center, East China Normal University, Shanghai, 200241, China

Rulin Hua and Wenjing Yan contributed equally.

Supplementary Table I. Overview of public scRNA-seq data for analysis.

Reference	GEO accession no.	Protocol	Cell number	Species	Description
Garrido-Trigo <i>et al.</i> , 2023	GSE214695	10X	46,698	Human	Inflammatory bowel diseases colon
Lake <i>et al.</i> , 2018	GSE97930	snDrop-Seq	10,319	Human	Frontal cortex
The Tabula Sapiens Consortium, 2022	GSE201333	10X	456,101	Human	Tabula Sapiens
The Tabula Muris Consortium, 2018	GSE109774	10X	55,656	Mouse	Tabula Muris
Trzebanski <i>et al.</i> , 2024	GSE261003	10X	4,705	Mouse	Bone marrow myeloid cells
Zheng <i>et al.</i> , 2017	SRX1723926	10X	68,579	Human	Peripheral blood mononuclear cells
Gan <i>et al.</i> , 2024	GSE223309	10X	31,657	Mouse	Brain metastatic microenvironment

References can be found in the main article.

Supplementary Table II. Runtimes for different normalization methods.

Dataset	log (CP10K+1)	Scrans-form	log (LC-UMI+1)
68k PBMC	29 s	703 s	506 s
32k brain metastatic cells	21 s	294 s	138 s

The table presents the runtimes of different normalization methods on the Zheng 68k PBMC dataset and the mouse brain metastatic microenvironment dataset. All runtimes were measured on a server with 256 GB RAM and 64 CPU threads at 2.4 GHz.

* Corresponding author: dekanglv@126.com, Fzheng@hsc.ecnu.edu.cn
0030-9923/2026/0001-0001 \$ 9.00/0



Copyright 2026 by the authors. Licensee Zoological Society of Pakistan.

This article is an open access  article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Supplementary Table III. Comparison of the number of marker genes identified using log(CP10K+1) and log(LC-UMI+1) normalization across different fold change thresholds. This table summarizes the number of marker genes identified using log(CP10K+1) and log(LC-UMI+1) normalization in the 68k PBMC and brain metastatic microenvironment datasets across a range of log2 fold change thresholds. Marker genes were classified into CP10K-specific (identified only by log(CP10K+1)), LC-UMI-specific (identified only by log(LC-UMI+1)), and Shared (identified by both methods). In the cell type column, “combined” represents the total number of marker genes across all cell types in each dataset.

68k PBMC				
log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
0.25	552	276	1120	Combined
0.5	304	93	816	Combined
1	105	24	552	Combined
2	26	4	193	Combined
3	2	1	84	Combined
4	0	0	29	Combined
5	0	0	13	Combined
6	0	0	5	Combined
0.25	252	2	77	CD14+ Monocyte
0.5	214	2	76	CD14+ Monocyte
1	76	1	66	CD14+ Monocyte
2	19	0	20	CD14+ Monocyte
3	0	0	10	CD14+ Monocyte
4	0	0	2	CD14+ Monocyte
5	0	0	2	CD14+ Monocyte
6	0	0	0	CD14+ Monocyte
0.25	28	0	63	CD19+ B
0.5	10	0	46	CD19+ B
1	3	0	30	CD19+ B
2	2	0	6	CD19+ B
3	2	0	4	CD19+ B
4	0	0	1	CD19+ B
5	0	0	0	CD19+ B
6	0	0	0	CD19+ B
0.25	32	1	414	CD34+
0.5	32	0	411	CD34+
1	15	0	349	CD34+

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
2	4	0	137	CD34+
3	0	0	57	CD34+
4	0	0	23	CD34+
5	0	0	11	CD34+
6	0	0	5	CD34+
0.25	3	16	39	CD4+ T cell
0.5	0	1	9	CD4+ T cell
1	0	0	0	CD4+ T cell
2	0	0	0	CD4+ T cell
3	0	0	0	CD4+ T cell
4	0	0	0	CD4+ T cell
5	0	0	0	CD4+ T cell
6	0	0	0	CD4+ T cell
0.25	206	0	165	CD56+ NK
0.5	45	0	116	CD56+ NK
1	10	0	52	CD56+ NK
2	1	0	15	CD56+ NK
3	0	0	7	CD56+ NK
4	0	0	1	CD56+ NK
5	0	0	0	CD56+ NK
6	0	0	0	CD56+ NK
0.25	30	1	71	CD8+ T cell
0.5	2	0	12	CD8+ T cell
1	0	0	2	CD8+ T cell
2	0	0	0	CD8+ T cell
3	0	0	0	CD8+ T cell
4	0	0	0	CD8+ T cell
5	0	0	0	CD8+ T cell
6	0	0	0	CD8+ T cell
0.25	1	256	291	Dendritic
0.5	1	90	146	Dendritic
1	1	23	53	Dendritic
2	0	4	15	Dendritic
3	0	1	6	Dendritic
4	0	0	2	Dendritic
5	0	0	0	Dendritic
6	0	0	0	Dendritic
Mouse brain metastatic microenvironment				
0.25	3588	3631	9757	combined
0.5	3444	3597	9680	combined
1	2627	3292	9123	combined

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
2	1165	1561	7148	combined
3	534	569	4920	combined
4	212	205	3328	combined
5	79	75	2199	combined
6	25	21	1403	combined
0.25	367	19	507	ASTRO
0.5	332	19	501	ASTRO
1	245	16	462	ASTRO
2	110	11	384	ASTRO
3	49	3	273	ASTRO
4	2	1	146	ASTRO
5	0	0	52	ASTRO
6	0	0	19	ASTRO
0.25	78	1	56	BAM
0.5	77	1	56	BAM
1	54	1	55	BAM
2	25	1	42	BAM
3	13	0	30	BAM
4	4	0	25	BAM
5	2	0	19	BAM
6	0	0	9	BAM
0.25	103	0	52	BC
0.5	103	0	52	BC
1	100	0	52	BC
2	75	0	52	BC
3	51	0	50	BC
4	26	0	40	BC
5	14	0	38	BC
6	9	0	31	BC
0.25	102	168	378	BG
0.5	97	164	372	BG
1	58	130	351	BG
2	11	75	252	BG
3	6	40	177	BG
4	1	1	96	BG
5	0	0	43	BG
6	0	0	19	BG
0.25	52	242	233	BMDM
0.5	51	242	233	BMDM
1	26	226	227	BMDM
2	4	115	131	BMDM

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
3	1	55	80	BMDM
4	1	31	56	BMDM
5	1	18	46	BMDM
6	0	6	33	BMDM
0.25	26	411	645	CPEC
0.5	23	411	645	CPEC
1	12	404	645	CPEC
2	0	165	494	CPEC
3	0	60	223	CPEC
4	0	23	130	CPEC
5	0	9	91	CPEC
6	0	3	71	CPEC
0.25	163	190	1079	EpC
0.5	151	189	1079	EpC
1	109	145	1053	EpC
2	33	83	883	EpC
3	11	42	717	EpC
4	5	16	557	EpC
5	1	7	436	EpC
6	0	3	345	EpC
0.25	252	8	436	GN
0.5	252	8	436	GN
1	247	8	436	GN
2	139	7	404	GN
3	66	5	349	GN
4	31	4	291	GN
5	11	1	216	GN
6	1	0	144	GN
0.25	131	56	371	HpC
0.5	131	56	371	HpC
1	123	55	371	HpC
2	69	36	316	HpC
3	34	20	204	HpC
4	14	10	141	HpC
5	7	2	104	HpC
6	3	1	65	HpC
0.25	171	111	950	MG
0.5	137	109	922	MG
1	55	90	789	MG
2	8	39	475	MG
3	0	8	261	MG

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
4	0	1	94	MG
5	0	0	29	MG
6	0	0	1	MG
0.25	561	0	169	NB
0.5	561	0	169	NB
1	531	0	169	NB
2	229	0	161	NB
3	101	0	134	NB
4	41	0	111	NB
5	12	0	81	NB
6	2	0	42	NB
0.25	191	0	111	NEUT
0.5	191	0	111	NEUT
1	189	0	111	NEUT
2	158	0	111	NEUT
3	81	0	105	NEUT
4	36	0	85	NEUT
5	20	0	60	NEUT
6	10	0	51	NEUT
0.25	1	1476	1263	NIPC
0.5	1	1476	1263	NIPC
1	0	1466	1263	NIPC
2	0	674	1067	NIPC
3	0	170	499	NIPC
4	0	52	286	NIPC
5	0	17	158	NIPC
6	0	1	39	NIPC
0.25	94	27	197	NK
0.5	94	27	197	NK
1	83	27	197	NK
2	28	26	172	NK
3	11	18	138	NK
4	7	14	119	NK
5	2	9	98	NK
6	0	6	84	NK
0.25	107	29	283	OEC
0.5	103	28	283	OEC
1	79	21	271	OEC
2	42	11	233	OEC
3	24	1	188	OEC
4	9	1	125	OEC

Table continues on next column.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type
5	2	0	79	OEC
6	0	0	49	OEC
0.25	40	516	699	OLIG
0.5	34	516	699	OLIG
1	12	470	699	OLIG
2	1	190	544	OLIG
3	0	77	370	OLIG
4	0	27	253	OLIG
5	0	9	152	OLIG
6	0	0	94	OLIG
0.25	419	178	943	OPC
0.5	396	159	909	OPC
1	226	88	661	OPC
2	51	41	330	OPC
3	19	24	241	OPC
4	11	9	162	OPC
5	4	1	104	OPC
6	0	0	56	OPC
0.25	85	6	95	PER
0.5	85	6	95	PER
1	85	6	95	PER
2	56	5	91	PER
3	27	4	66	PER
4	9	1	40	PER
5	1	0	23	PER
6	0	0	11	PER
0.25	432	6	472	VEC
0.5	429	6	472	VEC
1	276	6	463	VEC
2	85	4	381	VEC
3	24	0	296	VEC
4	9	0	211	VEC
5	0	0	137	VEC
6	0	0	90	VEC
0.25	85	145	417	VLC
0.5	69	138	414	VLC
1	28	94	362	VLC
2	9	51	295	VLC
3	6	27	251	VLC
4	4	11	191	VLC
5	2	2	131	VLC

Table continues on next page.....

log2 (fold change)	CP10K specific	LC-UMI specific	Shared	Cell type					
6	0	1	92	VLC	2	32	27	330	VSMC
0.25	128	42	401	VSMC	3	10	15	268	VSMC
0.5	127	42	401	VSMC	4	2	3	169	VSMC
1	89	39	391	VSMC	5	0	0	102	VSMC
					6	0	0	58	VSMC

Supplementary Table IV. Marker genes identified by log(CP10K+1) and log(LC-UMI+1) normalization in two scRNA-seq datasets. This table presents representative marker genes identified by log(CP10K+1) and log(LC-UMI+1) normalization in the 68k PBMC and brain metastatic microenvironment datasets using a log2 fold change (log2FC) threshold > 0.5. For each cell type, marker genes were classified into CP10K-specific (identified only by log(CP10K+1)), LC-UMI-specific (identified only by log(LC-UMI+1)), and Shared (identified by both methods). When the number of marker genes in a given category exceeded 20, the top 20 genes ranked by log2FC were retained. “None” indicates that no marker genes met the specified criteria in that category.

68k PBMC			
Cell type	CP10K-specific	LC-UMI-specific	Shared
CD14+ Mono-cyte	FCER1G, SPI1, IFI30, CTSS, LGALS1, S100A11, PYCARD, CFP, S100A4, LGALS2, TSPO, IGSF6, BLVRB, GSTP1, COTL1, MS4A7, HCK, S100A6, NPC2, CEBPD	FCGR3A, IFITM2	S100A9, S100A8, AIF1, LST1, TYROBP, FCN1, SAT1, CFD, TYMP, PSAP, FTL, SERPINA1, TMEM176B, RP11-290F20.3, CD14, IFITM3, STXBP2, NEAT1, PILRA, LINC00936
CD19+ B	HLA-DRA, CD74, IRF8, FCGR2B, PKIG, LAT2, HHEX, STX7, LTB, LAPTM5	None	CD79A, CD79B, IGLL5, TCL1A, MS4A1, CD37, LINC00926, HVCN1, VPREB3, FCER2, NCF1, EAF2, MZB1, SPIB, CYB561A3, BANK1, PPAPDC1B, HLA-DOB, BLK, CD40
CD34+	ARPC1B, TIMP1, CD68, FTH1, TUBA1B, MYL6, ACTB, SNX3, TMSB4X, ALDOA, GAPDH, SUPT4H1, C7orf50, POLR2E, LYN, VASP, ATF4, MAP1LC3B, PARK7, RTN4	None	PPBP, PF4, GNG11, SDPR, CLU, GPX1, HIST1H2AC, ACRBP, NRG1, NCOA4, TUBB1, GP9, SPARC, RUFY1, TREML1, MYL9, PTCRA, CMTM5, ITGA2B, FERMT3
CD4+ T cell	None	LTB	CD4, IL32, AQP3, GSTK1, CORO1B, PLP2, TNFRSF4, TRADD, MAL
CD56+ NK	FCGR3A, RHOC, ABI3, CD300A, ITGB2, IFITM2, ID2, CLIC1, MIR142, AOA1, CTSC, PFN1, FAM49B, ARPC2, PPP1CA, DBI, ANXA1, OSTF1, CXXC5, CDK2AP2	None	GNLY, NKG7, GZMB, FGFBP2, GZMH, CST7, PRF1, CLIC3, GZMA, CCL5, CTSW, SPON2, HOPX, PTGDS, CCL4, CD63, PRSS23, IGFBP7, MATK, CD247
CD8+ T cell	IL7R, LDHB	None	GZMK, CD8B, CD27, CD8A, DUSP2, CD3E, NOSIP, CCR7, CD3D, TMEM66, LEF1, PIK3IP1
Dendritic	ITM2C	HLA-DRA, CD74, FTH1, FCER1G, COTL1, GSTP1, CTSS, S100A4, LGALS1, CYBA, SPI1, S100A11, MYL6, S100A6, PYCARD, ARPC1B, CFP, ACTB, NPC2, GAPDH	LYZ, CST3, HLA-DRB1, HLA-DPB1, HLA-DPA1, HLA-DQA1, HLA-DRB5, FCER1A, HLA-DQB1, HLA-DQA2, HLA-DMA, CPVL, GRN, FCGRT, ANXA2, CLEC10A, HLA-DMB, C1QA, AMICA1, LY86

Table continues on next page.....

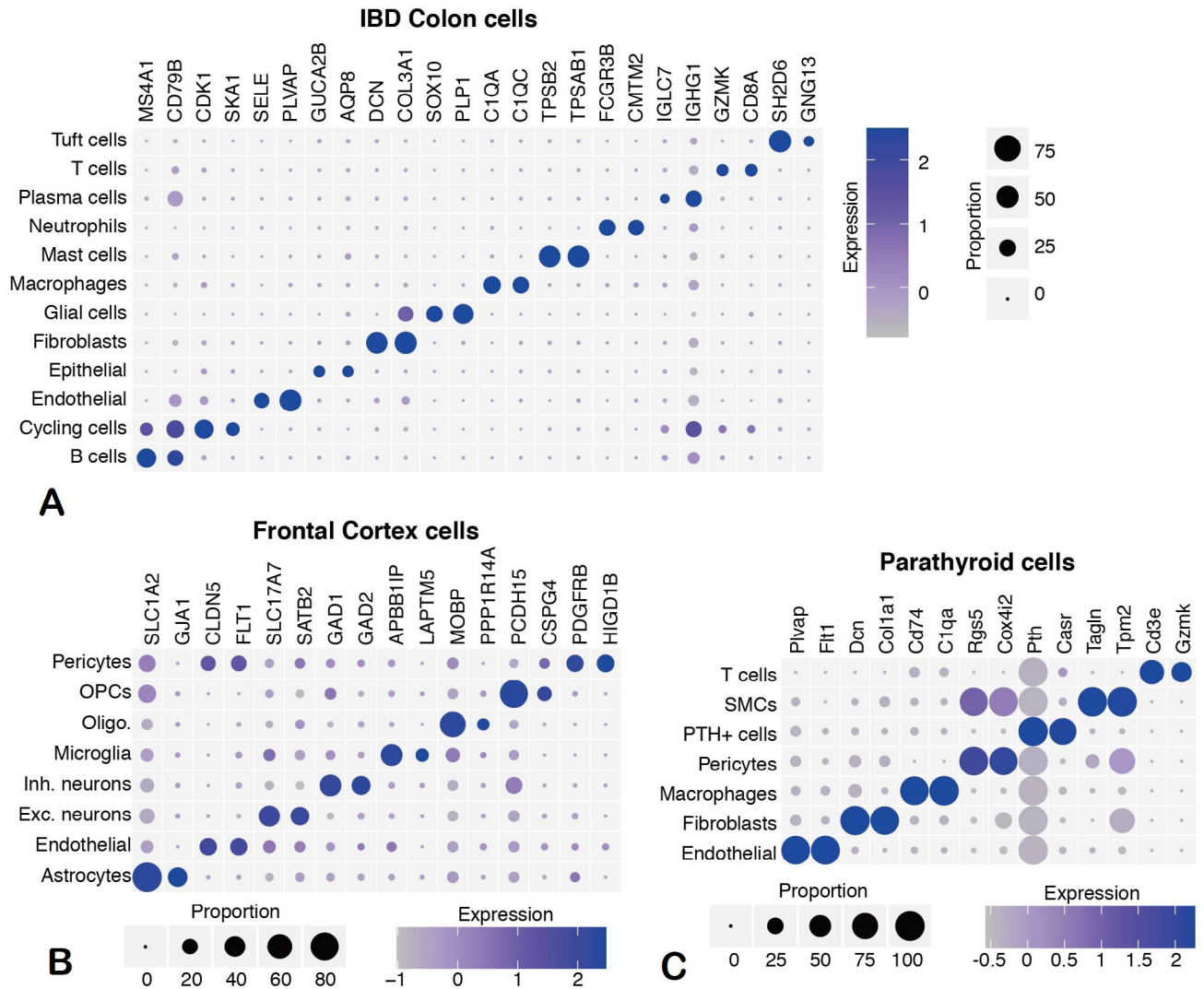
Cell type	CP10K-specific	LC-UMI-specific	Shared
Mouse brain metastatic microenvironment			
ASTRO	CAR8, KCNK1, GABRB1, NAT8, ALDOC, HHATL, TMEM229A, ADRA1A, RFX4, KCNJ16, NR-CAM, PAMR1, BTBD17, FOLH1, SLC14A1, SCG3, ADCY2, PIPOX, NPAS3, USP51	RGS7, SLC30A10, CHIL1, SEMA4A, 6430573F11RIK, SORL1, NRXN2, EFR3B, COLGALT2, GPD2, TMEM136, MAGI3, 2900076A07RIK, MALAT1, ANKS1, MCUR1, DENND4C, NFAT5, DOCK4	C030018K13RIK, SLC6A11, PDZPH1, GPC5, HAPLN1, 2900052N01RIK, GRIN2C, FEZF2, PPP1R3G, IGSF1, CYP4F15, KIRREL2, PRSS50, NR2E1, LHX2, GABBR2, GM266, DIO2, SL-C1A2, GM45650
BAM	H2-EB1, H2-AA, CD74, H2-AB1, CXCL1, TGFB1, ARHGAP19, TRPV4, COL14A1, DSE, HPSE, MSR1, C2, MS4A6B, MDFIC, TMEM8, IFI204, APOBEC1, CLEC4A2, SMAGP	ACP5	LYVE1, MRC1, CD163, FCNA, CCL8, PF4, PLA2G2D, F13A1, CLEC10A, MS4A4A, LILRA5, CBR2, CLEC4N, FOLR2, SAA3, CCL24, MS4A7, SER-PINB8, AOA, CCL7
BC	FCRLA, LY6D, SPIB, P2RY10, CD2, IL12A, GPR171, CIITA, STAP1, B3GNT5, CLEC2I, CHST3, RHOF, IFI203, ARHGAP15, MNDAL, SER-PINB1A, CD69, AC125149.3, KYNU	None	IGLC1, FCMR, MS4A1, IGLC2, FCER2A, IGKC, CD79A, GM43291, PAX5, 4930597A21RIK, CD19, POU2AF1, BLK, CD79B, MZB1, GM8369, IGLC3, CXCR5, VPREB3, IGHD
BG	GM9945, SEPT4, ABLIM2, SOX1, GPM6B, ADRA1B, KLHL13, SCD2, CDK5RAP2, WIP1, NDST1, TPMT, NEK8, IRS1, IQSEC1, ATRN, TSPAN33, IMPAD1, HAUS3, TB-C1D7	KCNC3, BTBD17, EN2, NAT8, FOLH1, GABRB1, ALDH1A1, ALDOC, NEGR1, SLC14A1, RFX4, SCG3, ADCY2, CYP2J9, PHYHIPL, KCNJ16, PFKM, ADRA1A, NRCAM, FXYD7	GM3455, CD70, GDF10, TDO2, GRIA4, DAO, SSTR2, MYBPC1, SYNDIG1, GRIA1, GREB1L, WIF1, MSX2, CPS-F4L, SOWAHA, FRRS1L, A2M, SYT10, CDH22, COX7B2
BMDM	ACE, E2F2, HAAO, OBFC1, ALD-H3B1, LPL, SGK3, VRK1, FEM1C, ACTG1, GM4799, RNH1, SCAND1, SHFM1, TAF10, HSPA8, AKT1, TFEB, EIF3K, UBE2A	SIRPB1C, LILRA6, MEFV, SIRPB1B, PIRA2, FPR2, FLT3, SLPI, GDA, STAP1, DPP4, EMILIN2, CIITA, H2-AA, LSP1, PILRB2, H2-EB1, MCEMP1, H2-AB1, CD300LB	EAR1, EAR2, CHIL3, MS4A8A, CD300E, ADGRE4, TREML4, IFITM6, CD177, GM9733, GPR141, NAPSA, CD209A, RYR1, F10, GPR132, CCR2, LTB4R1, CD300LD, ATP8B4
CPEC	FAM92A, ENY2, CBWD1, ACAT1, POLR2K, PPIA, SUMO3, DNLZ, SEC61G, RARS2, PSMA2, ADSL, BTBD1, NDUFA10, HBS1L, ZRSR2, PSMG1, SLIRP, ANAPC4, FKBP8	ATP2B3, PLK5, MSX1, RBP1, SLC29A4, ACE, CIDEA, CLDN1, 2810433D01RIK, KRT1, CNTNAP2, SEMA3B, FAM161A, IGFBP2, OCIAD2, KCNK1, PRKCQ, FAM160A1, TRPV4, MYRIP	PRR32, KCNE2, SULT1C2, TTR, DEFB9, INS2, DEFB11, COX8B, HTR2C, IGFN1, WDR72, KL, RP24-233B16.2, SOSTDC1, MRAP2, CRB3, WDR86, SLC16A8, 9330185C12RIK, EPN3
EpC	CIDEA, FAM161A, TPPP3, SEM-A3B, CHAF1A, PSMC3IP, CCDC34, B4GALNT4, 2310030G06RIK, EZR, AIF1L, CCDC24, FUOM, ARL6, KCTD15, NRBP2, CEP89, HIST1H4I, BOK, ETFA	FAM216B, HDC, ACBD7, RIT2, WNT4, CCDC114, KCTD8, 2010300C02RIK, LY6H, DZANK1, CYB561, FAM84B, KCNMB4, GM42517, NNAT, TSPYL5, SGSM1, PTPRU, PLTP, WDR34	SNTN, AOC1, STOML3, TMEM212, SLC26A3, KRT15, CCDC153, LYG2, GOT1L1, GM44109, GM20661, GM21149, SLC14A2, FRMPD2, DNAH3, UOX, AKAP14, GM12695, CCDC33, ZFP474
GN	ATP2B3, KCTD8, RIT2, DLX6, EPHA3, DLGAP3, ARFGEF3, GALNT13, KCNH7, EN2, TRNP1, GM42372, DLG2, TMEM108, PNMAL2, 2010300C02RIK, JPH4, CNTNAP2, 2900011008RIK, BEX2	MEG3, ST8SIA3, SYT6, CSRNP3, AUTS2, IPCEF1, DMXL2, RAPGEF2	SCGN, CPA6, RPRML, KCNF1, GM42912, CALB2, GRIN2A, ZMAT4, GABRA5, SYT1, GABRA6, FRMD7, SL-C17A7, NRSN2, GRM2, NPPA, CBLN1, DLGAP2, SNHG11, CBLN3

Table continues on next page.....

Cell type	CP10K-specific	LC-UMI-specific	Shared
HpC	FAM216B, MSX1, ACBD7, CLDN1, TMEM150C, CCDC114, SLC29A4, PLK5, CDON, CYB561, OCIAD2, PTPRU, BEX4, TSPYL5, FAM84B, SERINC2, WDR34, TUBB4B, GALM, KIF21A	3632451O06RIK, DBNDD1, EMB, SGMS2, ARFGEF3, GM42372, PNMAL2, BC067074, A730017C20RIK, SMIM5, BEND7, PNCK, CACNB2, NDN, NRG2, PCGF2, 4430402I18RIK, CCDC24, ANXA2, SDC2	SPDEF, SPP2, MYH2, FERD3L, SSPO, BC048502, RP23-477K6.1, LDLRAD2, D930020B18RIK, RASEF, CAR3, MYH1, WFDC2, NPSR1, SCT, SULT6B2, CREB3L4, MAB21L2, CUZD1, CAR1
MG	SGK1, FMNL1, EVI2A, RBM47, ATP8A2, GARNL3, ELMO1, ZCWPW1, DOCK10, SLC15A4, NRM, FAM134B, KHK, PIK3R6, ARL10, UAP1L1, FKBP5, KLF10, NEK6, SCOC	H2-OB, TRPM2, CSF3R, CD22, SLAMF6, NFKBID, PIL-RA, CCRL2, C5AR1, RAC2, NFAM1, IGSF6, GLIPR1, SNX20, GPR65, SLC16A3, MIR142HG, PTAFR, LILRB4A, LRRC25	CCR6, HSPB3, SIGLECF, ADORA3, SLC2A5, IL1A, 2610528A11RIK, AFP, CXCL13, P2RY12, CST7, CRYBB1, TNF, GPR34, TMEM119, SIGLECH, GPR84, CRYBA4, LAG3, HEXB
NB	SLIT1, ELAVL4, TOX3, FOXO6, LMO3, MYCN, DBNDD1, YPEL1, ST8SIA3, TTC9B, GNG3, MEG3, HMGB3, PAK7, PAK3, DBN1, CDCA7, FOXP2, MFAP2, STMN1	None	IGFBPL1, ISLR2, PROKR2, DCX, SP9, TUBB3, SOX11, TIAM2, SP8, DLX5, DRAXIN, DLX1, RNF165, ELAVL2, DLX2, SCRT2, STMN2, ZFP57, MGAT5B, SRRM4
NEUT	SIRPB1B, LILRA6, MEFV, GDA, SELL, HDC, FPR2, SIRPB1C, LTB, SLPI, CD300LB, PIRA2, IL18RAP, MCEMP1, PILRB2, CD44, EMILIN2, PSTPIP1, FGR, LSP1	None	BC100530, STFA2L1, MRGPRA2B, RETNLG, ASPRV1, GM5483, STFA2, IL1F9, CXCR2, NLRP12, S100A9, DHRS9, S100A8, 4833407H14RIK, WFDC21, TACSTD2, IL1R2, KLRA17, HP, 4732465J04RIK
NIPC	DHPS	SLIT1, TTC9B, MYCN, EPHA3, SAPCD2, HMGB3, FOXO6, CDCA7, DLGAP3, PAK3, DLX6, ELAVL4, KCNH7, B3GNT5, TOX3, MFAP2, LMO3, FRAT2, FOXG1, YPEL1	HELT, GSX2, GSX1, GM29260, NPTX2, INSM1, MXD3, ASCL1, DLL3, PSRC1, FAM64A, LRR1, RRM2, NEIL3, TTK, PBK, SKA1, 2810417H13RIK, HIST-1H2AF, PIF1
NK	FLT3, DPP4, THY1, RUNX3, PRK-CQ, CORO2A, ITGA4, OSBPL3, D16ERTD472E, A530032D15RIK, IFI47, CKM, FAM117A, FOXP4, B4GALNT1, SEPT9, VOPP1, PTGER4, REL, PHF11A	SPIB, LY6D, FCRLA, P2RY10, CD2, GPR171, SELL, LTB, IL18RAP, RHOF, CLEC2I, MNDAL, IFI203, SERPINB6B, KYNU, CD69, DUSP2, ST-8SIA4, HCST, IL2RG	CMA1, KLK1B27, KLK1, D13ERT-D608E, DNNT, PRF1, NCR1, SH2D1A, GZMA, IL2RB, GZMB, CLNK, NKG7, KLRG1, KLRI2, KLRB1C, KLRC1, GATA3, TRBC1, XCL1
OEC	ERBB3, PRIMA1, 2810433D01RIK, APOD, NCALD, SP5, BC067074, FIGN, SLC2A13, UBE2QL1, ZDH-HC2, SCRN1, AHCYL2, SLC25A13, RAP1GAP, PLEKHF1, TSPAN15, FXYD7, NCAM2, PHGDH	HMGCS2, PLSCR2, 9330159F19RIK, GADD45A, CEBPD, ATP8A2, CASKIN2, BCAR3, NECTIN3, STARD13, FGF2, ISG15, OASL2, TMX4, IRGM2, DDX60, DENND1B, SELENOP, PIGH, RNF215	GM45385, NKX3-1, FOXD3, CLCA3A1, GM13499, GM12688, GPD4, OLFR32, FRZB, AQP1, SLC30A2, NPY, PRSS56, GM44691, IL13RA2, PSMB11, MYOZ1, COL28A1, CDH3, FNDC1
OLIG	RNF7, POP4, DCPS, TRIM35, GABARAPL2, MED30, MIEN1, FAM204A, PFDN6, TADA2A, MRFAP1, DNAC19, DCTN2, PCNP, TBCC, CWC15, PPP2CA, TRAPPC2, TMEM60, FXR2	IL12A, POU3F1, PRIMA1, TMEM150C, JPH4, SERPIN-B1A, BFSP2, ELAVL3, SOX10, ERBB3, KLHL4, TPPP3, APOD, FIGN, PAK7, GM9945, NAP1L5, PPP1R14A, ANKS1B, SEPT4	C030029H02RIK, ERMN, KLK6, OPA-LIN, LY6G6F, CYP2J12, TMEM125, GM45624, MOG, GJC2, ITGAD, MOBP, FA2H, HAPLN2, CLDN11, GM4593, NIPAL4, MYRF, NFE2L3, TNNI1

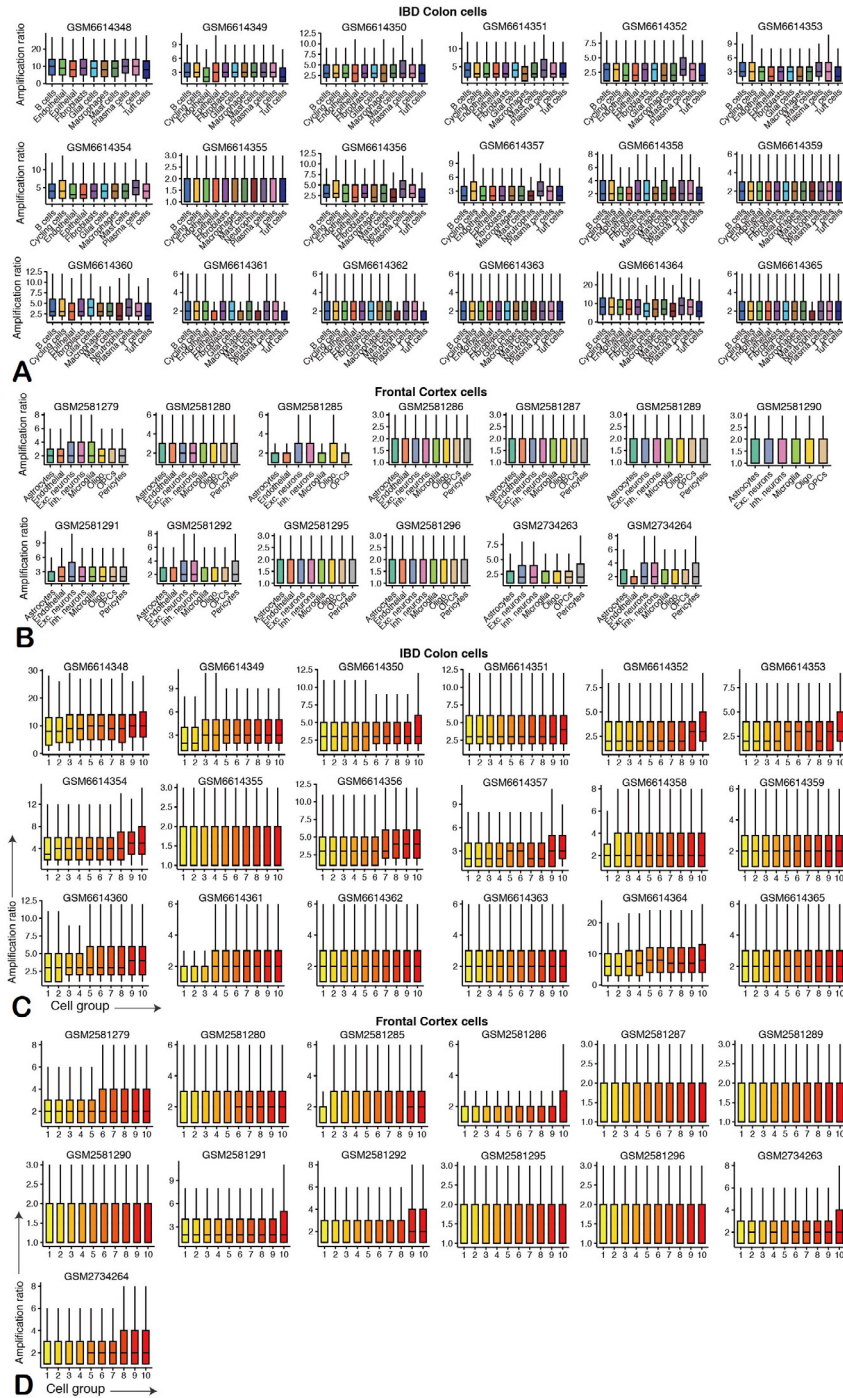
Table continues on next page.....

Cell type	CP10K-specific	LC-UMI-specific	Shared
OPC	3632451O06RIK, SAPCD2, BFSP2, POU3F1, SOX10, KCNK10, ANKS1B, KRT1, RP24-286J21.7, GM42517, EPHB2, GPR45, SH3D19, UNC5C, MYBL1, ZFP462, ZWILCH, MANEAL, DPYD, MAP6	GNG3, GFRA2, CHST3, TMEM108, 2900011O08RIK, CNIH2, NCALD, CAR8, ZDHHC2, NETO2, TMEM35A, REEP2, ASIC1, RAP1GAP, D630045J12RIK, ARXES1, WASF1, PLPPR2, XKR6, NAV1	PYY, PNLIP, GPR17, IQCJ, NEU4, 3110035E14RIK, VGF, SSTR1, ANKRD34C, A930009A15RIK, C1QL2, A930003O13RIK, TNFR, HMGA2, FAM19A2, PRKG2, ZFP488, LHFPL3, CHRNA4, 5730559C18RIK
PER	CASQ2, HIC1, GFRA2, COX4I2, LAMC3, FAP, CHP2, TMEM45A, MND1, NID1, SLC30A10, ARH-GEF17, KCTD1, IMPA2, GRTP1, SL-C19A1, INPP4B, HRCT1, GUCY1A2, HSPA12A	THY1, CXCL1, BMP1, OLFML2B, IL1R1, SERPINE2	ABCC9, KCNJ8, NODAL, ACE2, SERPINA1D, SERPINA1B, ATP13A5, ENPEP, PLA1A, REM1, ART3, GGT1, CYGB, VTN, EPHX3, SLC5A5, SL-C16A12, RGS5, COL5A3, CD248
VEC	STRA6, SERPINB6B, SGMS2, SL-C16A9, MYLPP, CRISPLD1, MYRIP, KLHL4, PLTP, HMGCS2, BEND7, ST8SIA4, SH3BP4, NDRG1, NET1, LIMS2, TANC1, PRRG1, SOCS2, NHSL2	PRKD2, DOCK6, SH2D3C, SMAGP, SWAP70, ETS2	AU021092, GM694, TMEM252, SCGB3A1, SLC38A5, SOX18, MYCT1, ADGRL4, RP23-381H23.2, LY6C1, ROBO4, CYYR1, SOX17, PTPRB, ADGRF5, TEK, SLC01A4, FOXQ1, SHE, FLT1
VLC	RBP1, WNT4, IGFBP2, FAM160A1, ALDH1A1, PARD6G, ITGA3, GSTT1, COPZ2, FDX1, NEDD4, CMC2, CKAP4, AIFM2, ARF2, ARF-GAP3, ACADSB, TBGR1, TBC1D8B, TGDS	GALNT13, FOXP2, CRISPLD1, DLG2, LAMC3, FAP, HIC1, CHP2, SLC16A9, SLC2A13, STRA6, TMEM45A, DDR2, PAMR1, NID1, ANXA1, DACT1, UNC5C, COL14A1, GPX8	PRG4, SLC47A1, PRDM6, PTGDR, LUM, DCN, WNT16, CARD14, ASGR1, GM5860, MPZL2, MFAP5, LYPD2, SLC22A6, ALDH1A2, SLC26A7, VSX1, AOX3, FAM180A, NOV
VSMC	SMIM5, PPP1R14A, DDR2, CACNB2, MPP7, PNCK, MYO1D, TTLL7, RGS7BP, OLFML2B, PCP4L1, SDC2, DACT3, FBXL2, OAZ2, EVA1C, HSPB6, PRKAG2, TUBA1C, FXYD1	COX4I2, CASQ2, EBF1, MND1, CACNA2D1, ARHGEF17, SLC19A1, HRCT1, LIMS2, SCN3A, INPP4B, PEAR1, GUCY1A2, NET1, PRKAR2B, RGS6, UTRN, CAV1, SYDE1, PLP2	RERGL, TAGLN, PLN, MAP3K7CL, MYH11, ACTA2, RP23-37D14.4, LMOD1, XIRP1, GPR20, MYOCD, FBXL22, TPM2, CNN1, NXNL2, MUSTN1, MYL9, OLFRL558, SCN4B, HSPB7

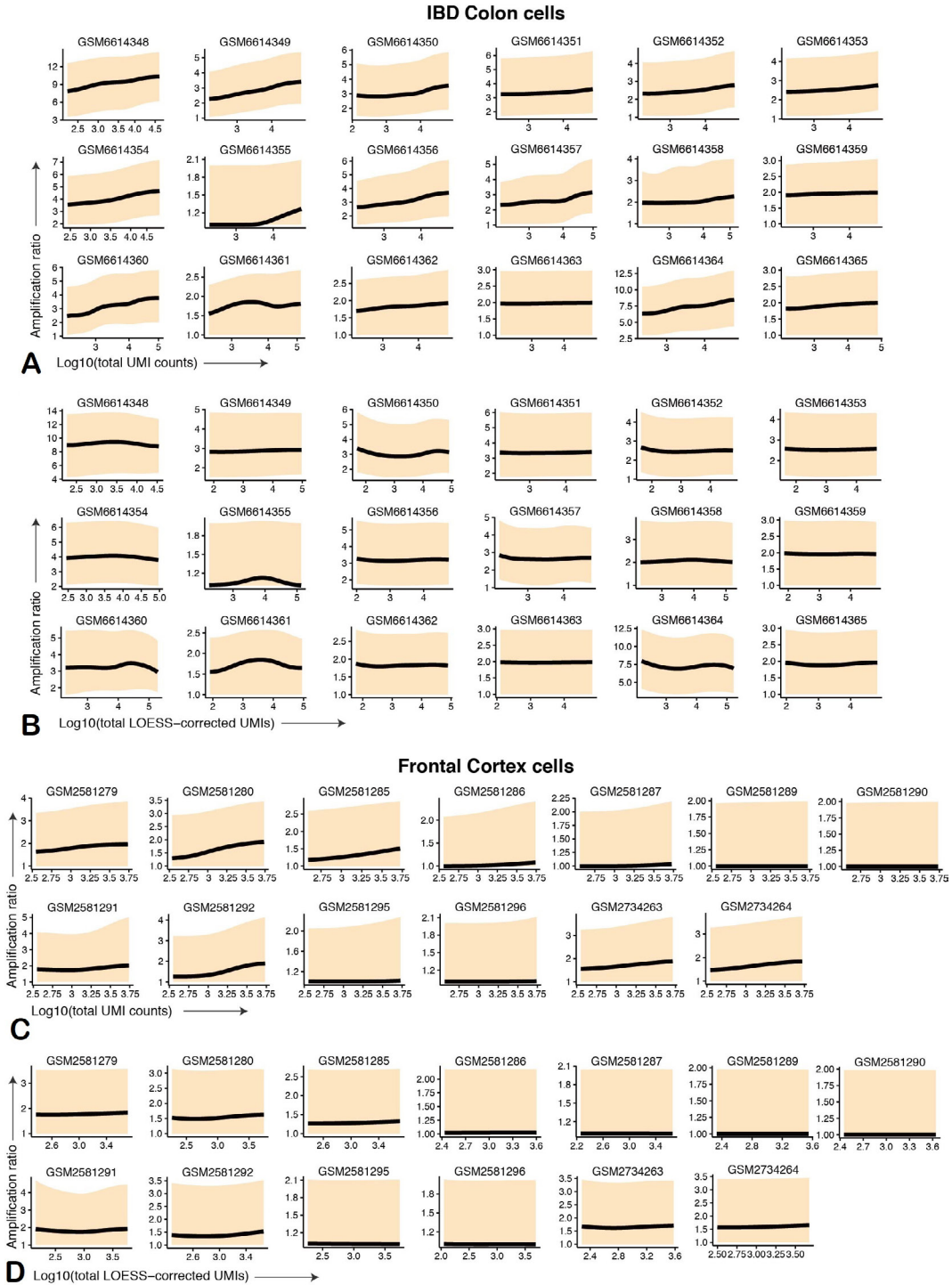


Supplementary Fig. 1. Marker gene expression across cell types in the human IBD colon, human frontal cortex, and rat parathyroid datasets.

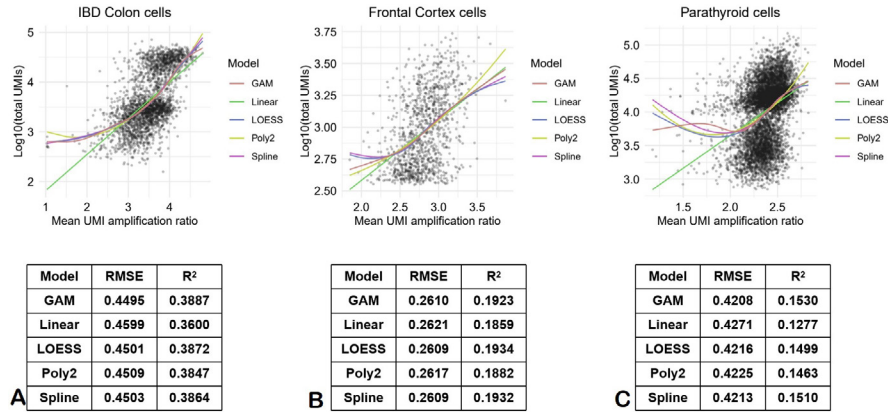
Bubble plots illustrating the expression levels and proportions of canonical marker genes across cell types in the (A) human IBD colon, (B) human frontal cortex, and (C) rat parathyroid datasets. Bubble size denotes the proportion of cells expressing each gene, while color intensity reflects the scaled average expression level.



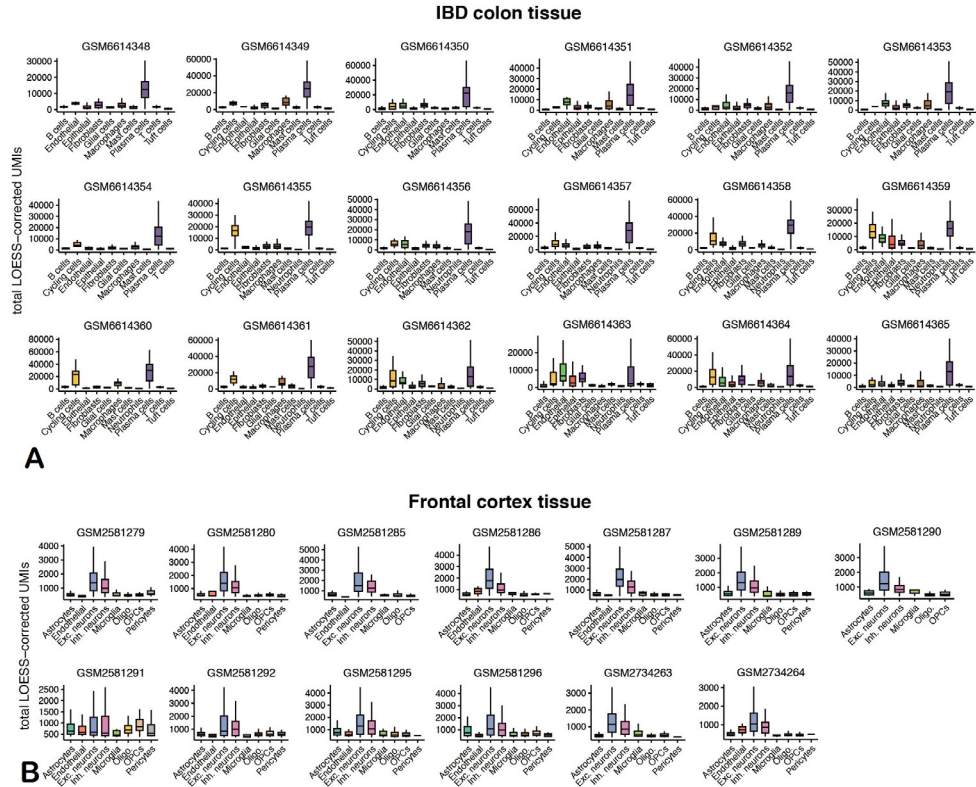
Supplementary Fig. 2. UMI amplification ratio distribution across samples in the human IBD colon and human frontal cortex datasets. (A, B) Distribution of UMI amplification ratio across different cell types in individual samples from the human IBD colon dataset (A) and human frontal cortex dataset (B). Each boxplot represents the UMI amplification ratio distribution within a specific cell type for a given sample. (C, D) UMI amplification distribution across cell groups stratified by total UMI counts in individual samples from the human IBD colon dataset (C) and human frontal cortex dataset (D). Within each sample, cells were divided into ten equal-sized groups based on their total UMI counts, and boxplots depict the distribution of UMI amplification ratio across these groups. The x-axis represents cell groups ordered from lowest to highest total UMI counts, illustrating how UMI amplification ratio varies with total UMI counts.



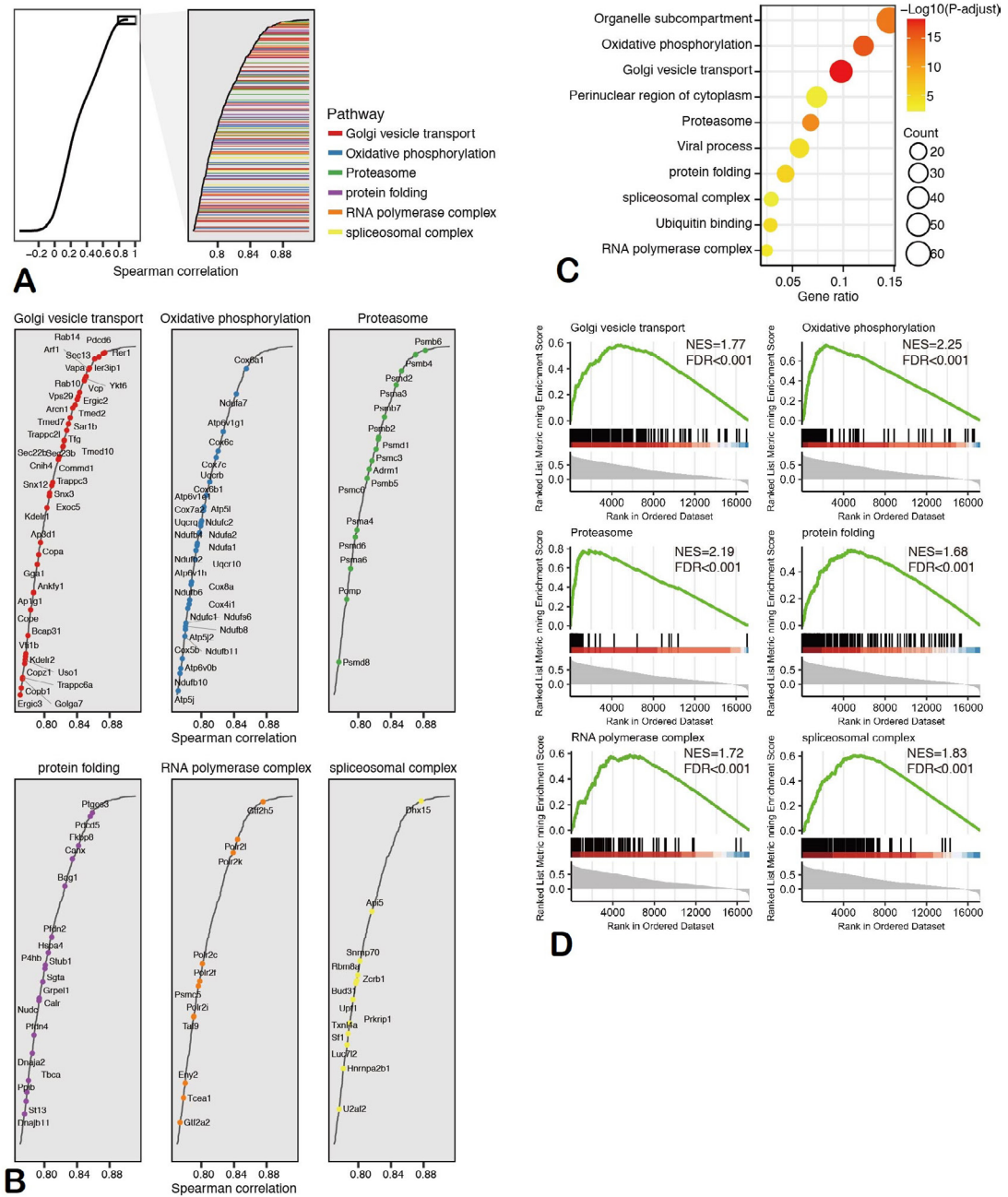
Supplementary Fig. 3. Relationship between cellular total UMI counts (before and after LOESS correction) and UMI amplification ratio. (A) Relationship between $\log_{10}$ -transformed cellular total UMI counts and UMI amplification ratio in individual samples from the human IBD colon dataset. Each panel represents a distinct sample, with a smoothed line indicating the median UMI amplification ratio and the colored region representing the interquartile range. (B) Same as (A), but using total LC-UMIs instead of total raw UMI counts. (C, D) Equivalent analyses performed on individual samples from the human frontal cortex dataset, following the same structure as (A, B).



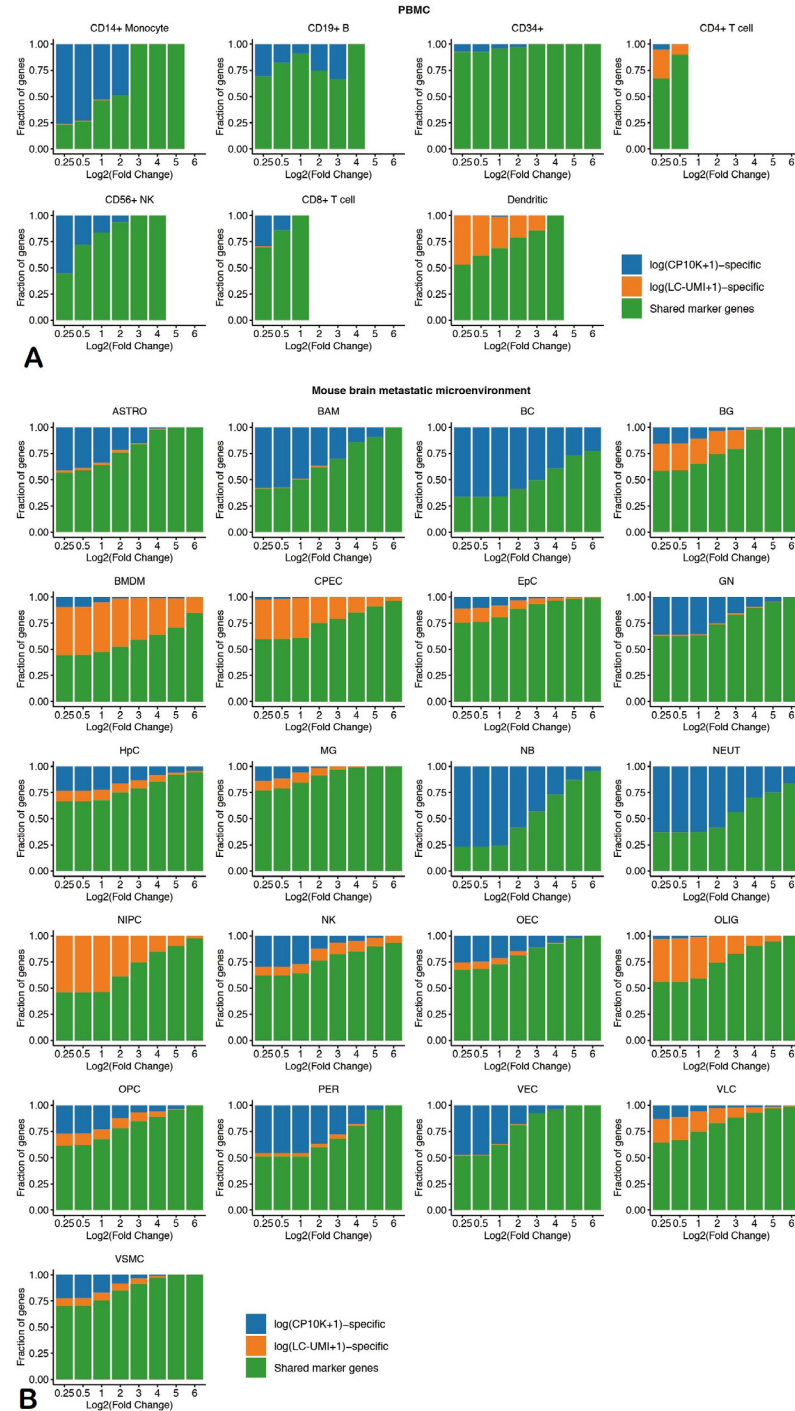
Supplementary Fig. 4. Comparison of regression models for modeling the relationship between UMI amplification ratio and total UMI counts in representative scRNA-seq samples. Scatter plots show data from representative samples of the three datasets presented in **Figure 1**: (A) human IBD colon (GSM6614349), (B) human frontal cortex (GSM2581279), and (C) rat parathyroid. Each point represents a single cell, with the x-axis indicating the mean UMI amplification ratio (reads per UMI) and the y-axis showing the $\log_{10}$ -transformed total UMI count. Five regression models linear, second-degree polynomial (Poly2), natural spline (df = 4), LOESS (span = 0.75), and generalized additive model (GAM) were fitted to each sample. Fitted curves are overlaid in color. Model performance was evaluated by root mean square error (RMSE) and coefficient of determination (R^2), summarized in the tables below each panel. While GAM achieved marginally better fit metrics, LOESS produced smoother and more stable trends without overfitting, supporting its utility for correcting UMI amplification bias.



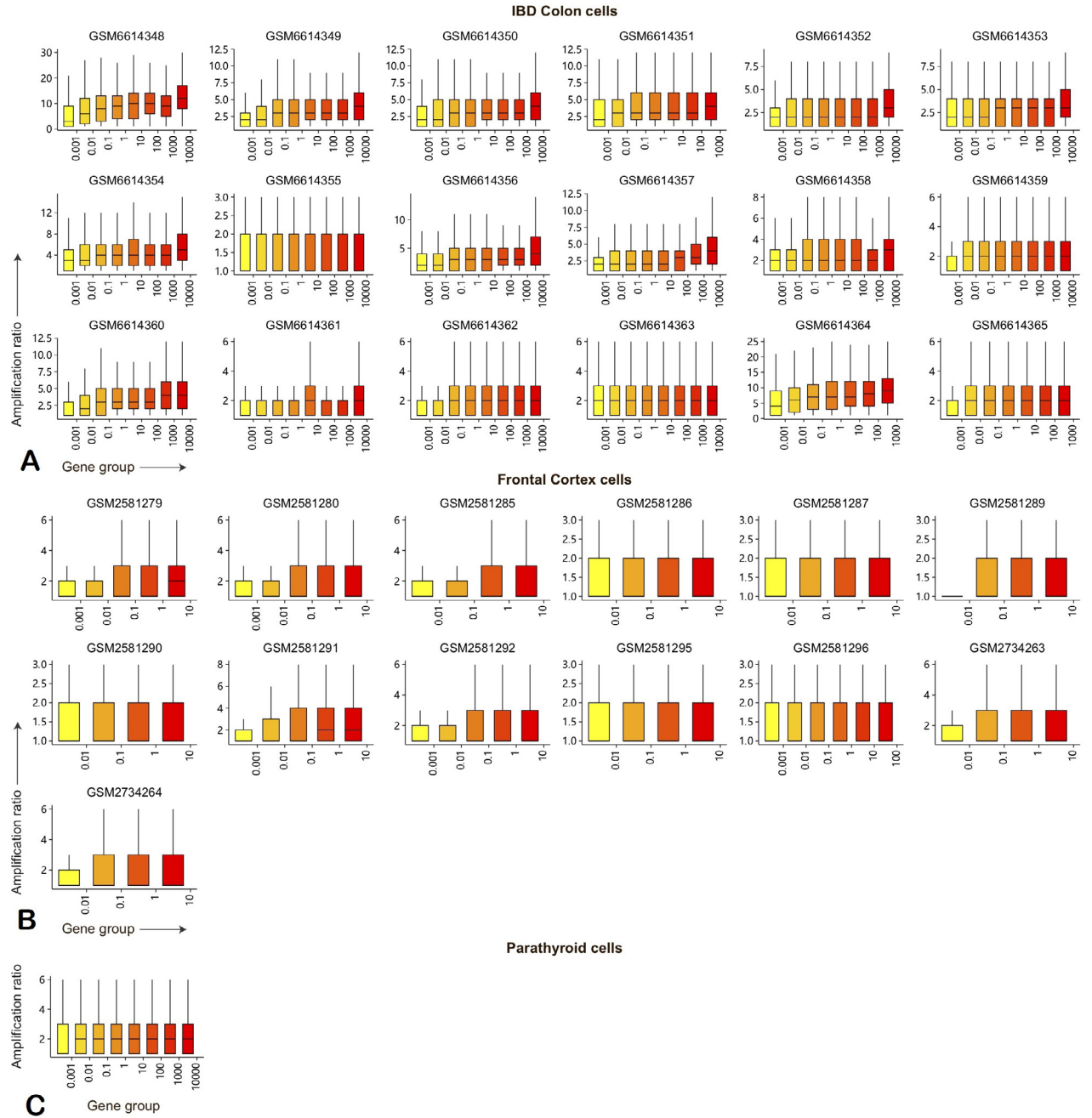
Supplementary Fig. 5. Total LC-UMIs across cell types in the human IBD colon and human frontal cortex datasets. Boxplots showing the distribution of total LC-UMIs across cell types in individual samples from the (A) human IBD colon and (B) human frontal cortex datasets.



Supplementary Fig. 6. Correlation between total LC-UMIs and gene expression, and pathway enrichment in the Tabula Muris dataset. (A) Spearman correlation between total LC-UMIs and gene expression levels across 72 cell types in the Tabula Muris dataset. The 500 genes with the highest correlation coefficients are highlighted on the right, with transcription-related pathway genes color-coded by pathway identity. (B) Detailed breakdown of transcription-related pathways identified in (A), with each pathway displayed separately. Genes within each pathway that rank among the top 500 most correlated genes are highlighted in distinct colors. (C) Top 10 enriched pathways among the top 500 most correlated genes from (A). The x-axis represents the gene ratio, defined as the proportion of pathway-associated genes relative to the total analyzed genes. Bubble size indicates the number of genes per pathway, while color intensity corresponds to statistical significance ($-\log_{10}$ adjusted p-value). (D) GSEA results for transcription-related pathways. Each panel displays the enrichment score (green line) for a given pathway across the ranked gene list ordered by Spearman correlation coefficients from (A). The normalized enrichment score (NES) and false discovery rate (FDR) are shown.



Supplementary Fig. 7. Comparison of marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization in PBMC and mouse brain metastatic microenvironment datasets. (A, B) Stacked bar plots illustrating the proportion of marker genes identified at progressively increasing log₂FC thresholds for each cell type in the (A) PBMC and (B) mouse brain metastatic microenvironment datasets. Marker genes were identified using Wilcoxon tests (adjusted p-value < 0.05), required to have the highest expression in the corresponding cell type compared to all others, and further filtered across a range of log₂FC thresholds. Shared marker genes identified by both normalization methods (log(CP10K+1) and log(LC-UMI+1)) are shown in green, while those uniquely detected by log(CP10K+1) and log(LC-UMI+1) are shown in blue and orange, respectively.



Supplementary Fig. 8. Distribution of UMI amplification ratios across gene expression groups in the human IBD colon, human frontal cortex, and rat parathyroid datasets. (A) Human IBD colon, (B) human frontal cortex, and (C) rat parathyroid datasets. Each panel represents a single sample from the corresponding dataset. Within each sample, genes were grouped by average expression level, defined as the mean UMI count per gene across all cells. Expression bins were determined using log₁₀-scale thresholds. For each group, boxplots show the distribution of UMI amplification ratios, calculated as the number of reads per UMI molecule. Genes with higher expression levels tend to display increased amplification ratios.