Pakistan J. Zool., vol. 58(4), pp 1835-1851, 2026 **DOI**: https://dx.doi.org/10.17582/journal.pjz/20250606044826

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

Rulin Hua[1], Wenjing Yan[2], Xiaoyan Liu[3], Suyang Xuan[4], Xuejuan Li[5], Shafiq Ur Rahman[1], Dekang Lv[4]* and Feng Zheng[1,5]*

[1]*Advanced Institute for Medical Sciences, Dalian Medical University, Dalian, 116044, China*
[2]*Dalian Center for Disease Control and Prevention, Dalian, 116021, China*
[3]*Department of Nephrology, the Second Hospital of Dalian Medical University, Dalian, 116023, China*
[4]*Institute of Cancer Stem Cell, Dalian Medical University, Dalian, 116044, China*
[5]*Wuhu Hospital and Health Science Center, East China Normal University, Shanghai, 200241, China*

Rulin Hua and Wenjing Yan contributed equally.

Article Information
Received 06 June 2025
Revised 25 July 2025
Accepted 08 August 2025
Available online 29 January 2026 (early access)
Published 18 June 2026

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

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.

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-Eltze and Huber, 2023; Hafemeister and Satija,

* Corresponding author: dekanglv@126.com, Fzheng@hsc.ecnu.edu.cn
0030-9923/2026/0004-1835 $ 9.00/0

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 NH_4Cl, 0.1 mM Na_2EDTA, 10 mM $KHCO_3$) 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 (25^{th}–75^{th} 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(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 cluster Profiler (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 (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(CP10K+1), sctransform, and log(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(CP10K+1) matrix and the log(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 log2FC 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 log2FC 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(CP10K+1), sctransform, and log(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-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.

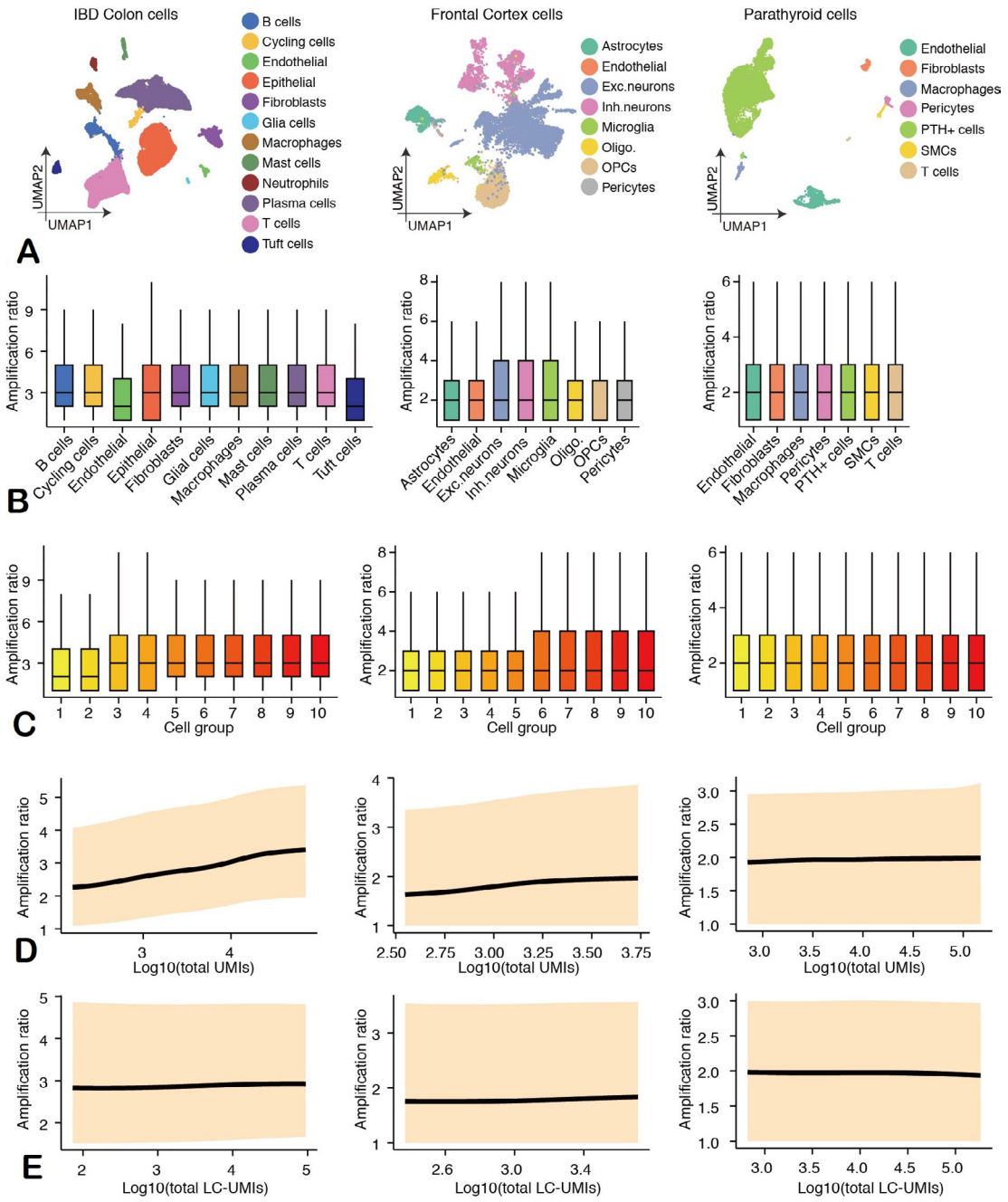

Fig. 1. Distribution of UMI amplification ratio across the human IBD colon, human frontal cortex, and rat parathyroid datasets. Each column represents data from one of the three datasets. (A) UMAP projections of cells from three datasets, shown on the left (human IBD colon, 46,698 cells), middle (human 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.

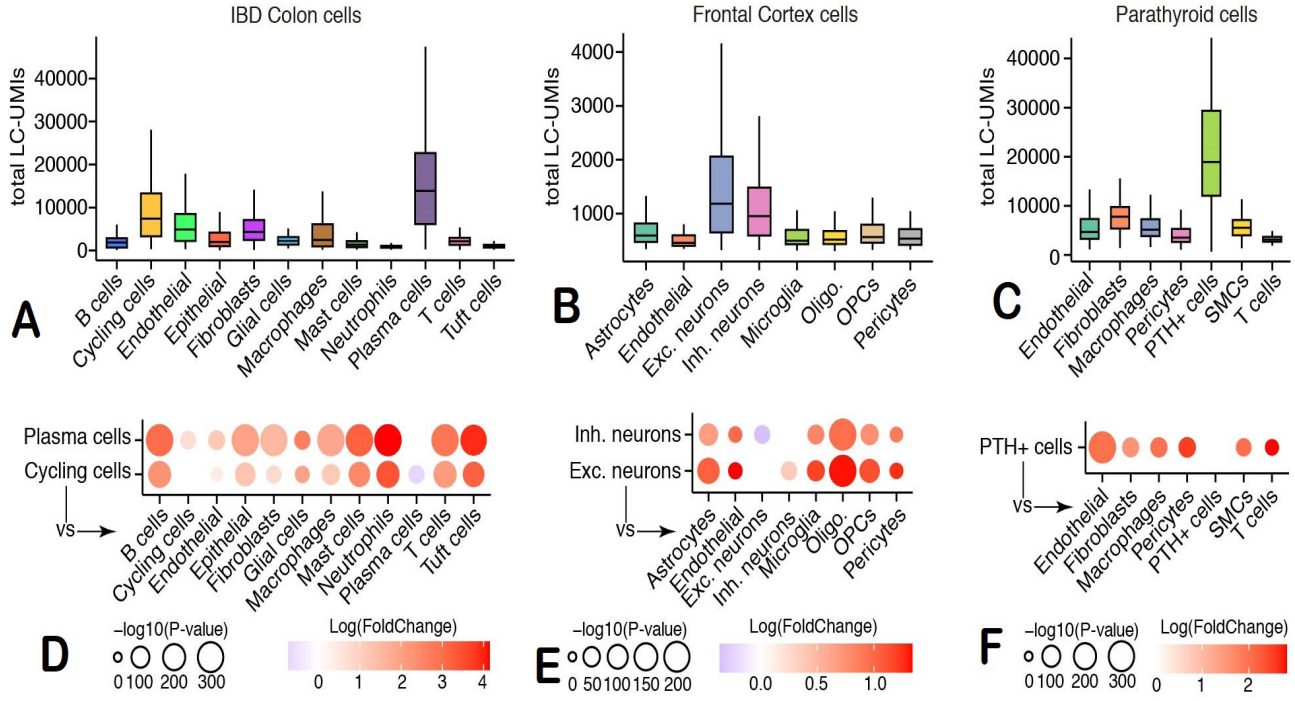

Fig. 2. LC-UMI provides an accurate measure of RNA content across cell types.
Each column corresponds to one of the three datasets, consistent with **Figure 1**. (A–C) Distribution of total LC-UMIs across different cell types in the human IBD colon dataset (A), human frontal cortex dataset (B), and rat 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 -log10(p-value) from the Wilcoxon test, while color gradient indicates the log2(fold change) between transcriptionally active and other cell types.

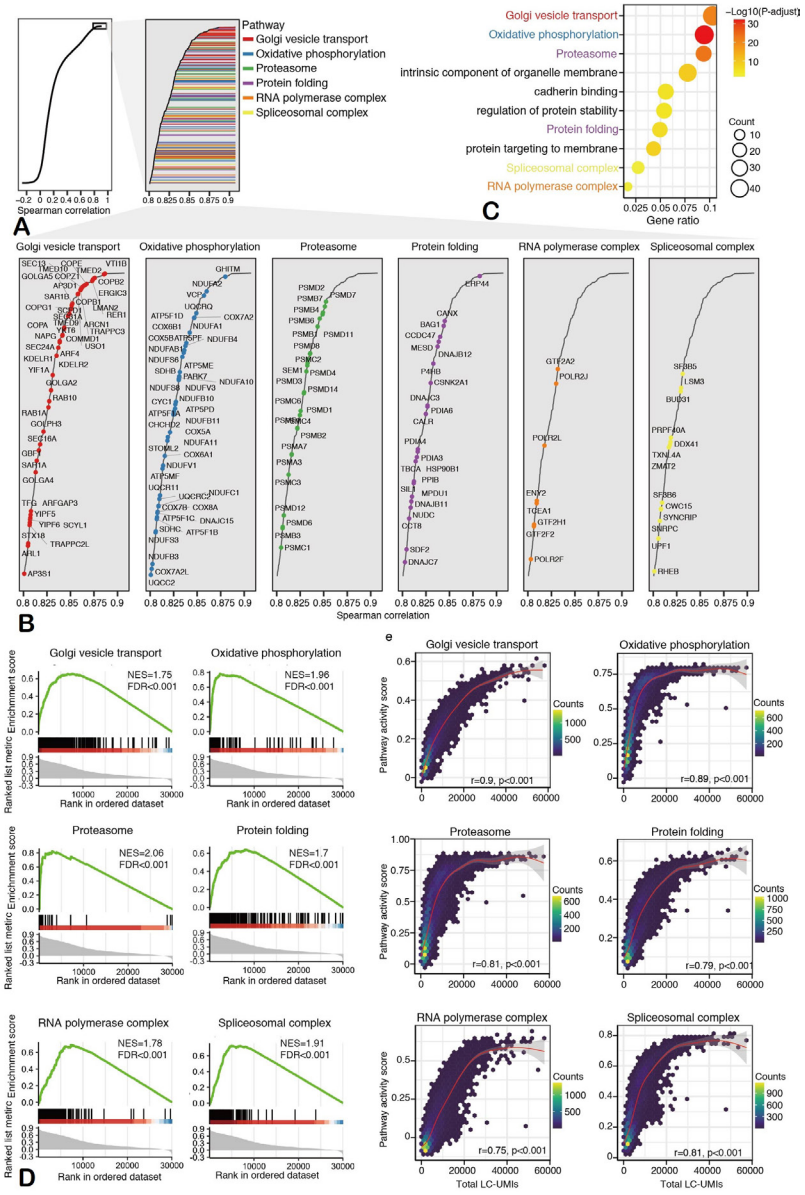

Fig. 3. Correlation analysis between total LC-UMIs and gene expression levels.
(A) Spearman correlation analysis between total LC-UMIs and gene expression levels across 341 cell types in the Tabula Sapiens dataset. The top 500 genes with the highest correlation coefficients are highlighted on the right, 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 -log10(adjusted p-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(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(LC-UMI+1)) to approximate a normal distribution. We then compared its performance against two widely used Seurat normalization methods: log (CP10K+1) (Butler *et al*., 2018) and sctrans 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(LC-UMI+1) was 0.789, which was nearly identical to log(CP10K+1) (0.794) and sctransform (0.788) (Fig. 4B). Similarly, in the mouse brain metastatic microenvironment dataset, log(LC-UMI+1) achieved an F1 score of 0.973, comparable to log(CP10K+1) (0.975) and sctrans 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

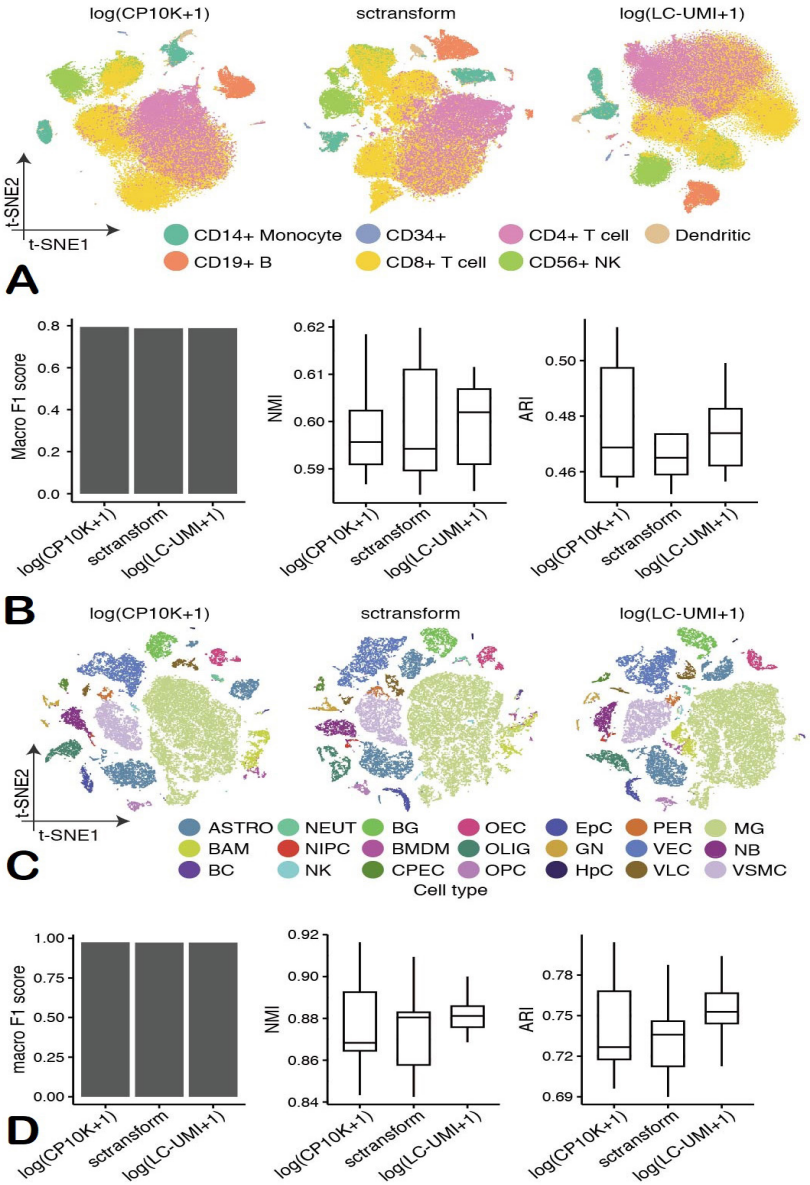

Fig. 4. Benchmarking the effect of different normalization methods on cell type classification in reduced dimensionality. (A) t-SNE embeddings of the PBMC dataset after PCA reduction, comparing three normalization methods: log(CP10K+1) (left), sctransform (middle), and log(LC-UMI+1) (right). Cells are 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 under 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(LC-UMI+1)-normalized clustering results exhibited slightly better concordance with known cell type labels than those obtained with log(CP10K+1) and sctransform (Fig. 4B, D). Next, we evaluated the computational efficiency of the three normalization methods. Log(CP10K+1) was the most computationally efficient, while log(LC-UMI+1) was consistently faster than sctransform across both datasets (Supplementary Table SII). Importantly, similar to log(CP10K+1), the log(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(LC-UMI+1) normalization achieves superior or comparable performance to log(CP10K+1) and sctransform in preserving biological variability and distinguishing cell types in scRNA-seq data. Given its simplicity and computational efficiency, log(LC-UMI+1) serves as a robust alternative for single-cell normalization.

Comparison of marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization

To evaluate the impact of log(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(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 log2 fold change (log2FC) 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., log2FC < 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(CP10K+1) and log(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(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(CP10K+1) normalization, but as a $CD14^+$ monocyte marker under log(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(LC-UMI+1) as a reliable normalization method for differential gene expression analysis in single-cell RNA-seq data.

Log(LC-UMI+1) supports robust cell differentiation trajectory analysis

To assess the effectiveness of log(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(CP10K+1), sctransform, and log(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(LC-UMI+1) preserves the underlying differentiation structure. In the trajectory inferred using log(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.

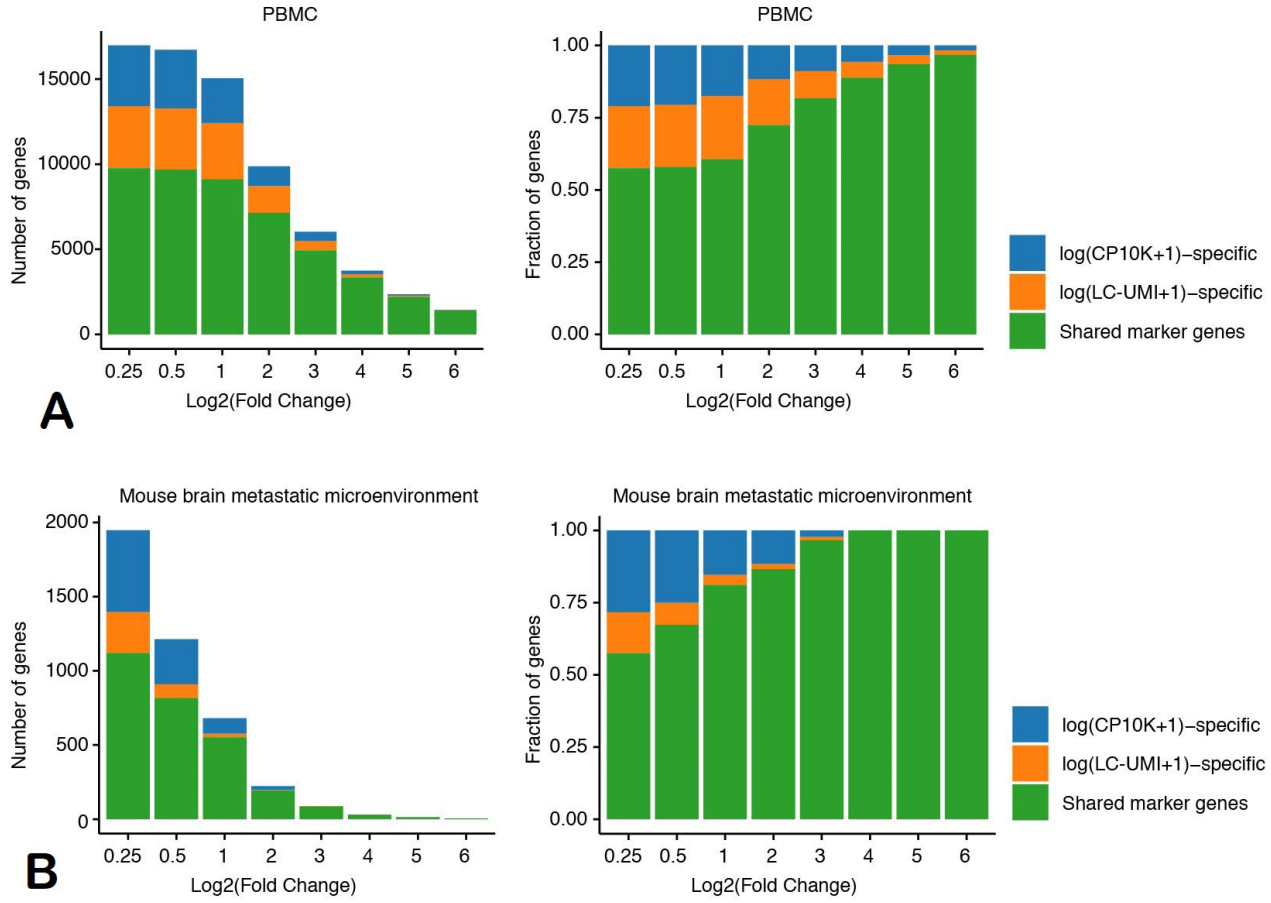

Fig. 5. Comparison of marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization across different fold change thresholds.
Stacked bar plots comparing marker gene identification using log(CP10K+1) and log(LC-UMI+1) normalization in the PBMC dataset (A) and the mouse brain metastatic 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

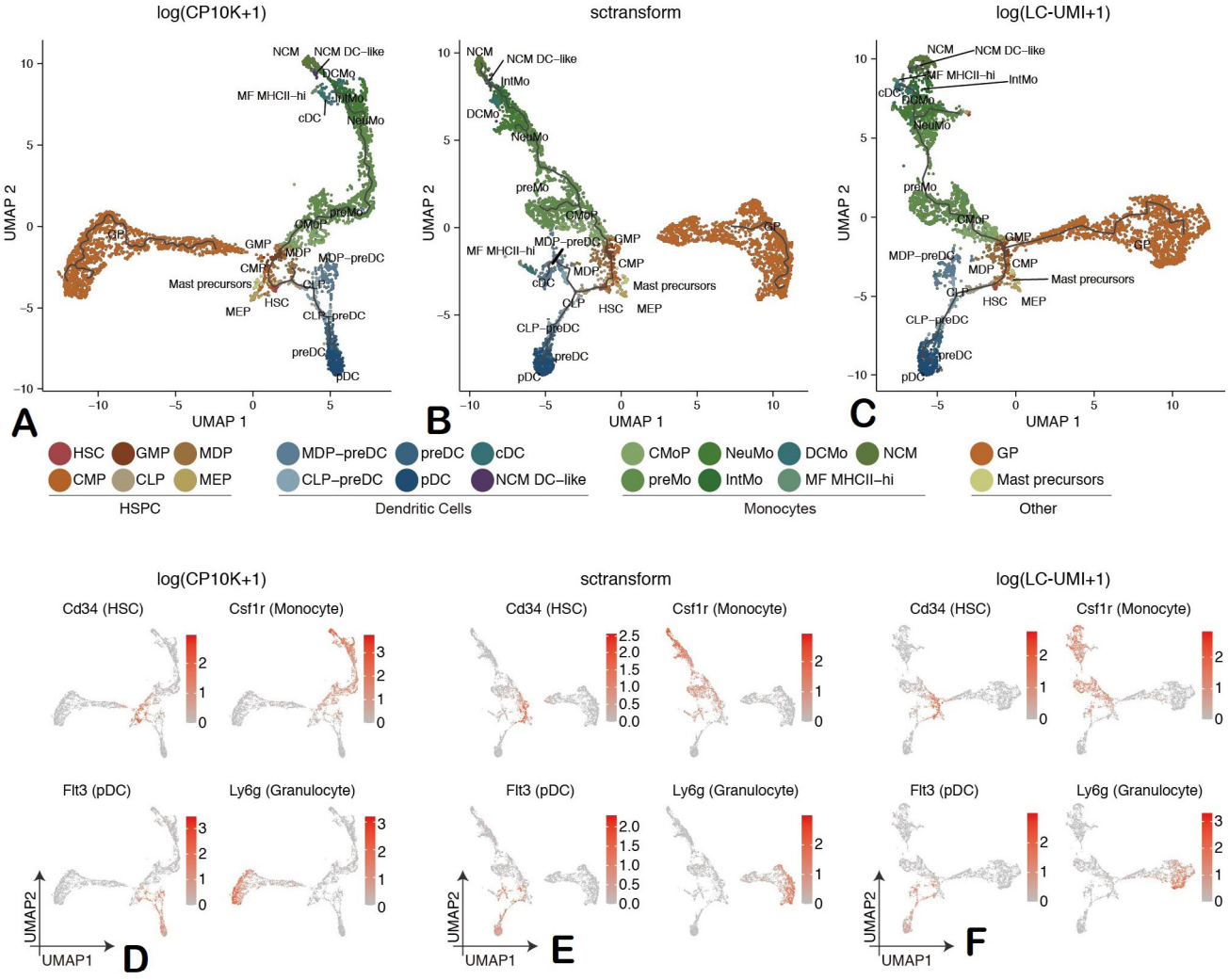

Fig. 6. Differentiation trajectories of the mouse bone marrow myeloid dataset using different normalization methods. (A–C) UMAP projections of differentiation trajectories inferred using Monocle3 with expression matrices normalized by (A) log(CP10K+1), (B) sctransform, and (C) log(LC-UMI+1). Cells are 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 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 (Homberger *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., Martin-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