Wu *et al. Genome Biology* (2025) 26:58
https://doi.org/10.1186/s13059-025-03525-6

Genome Biology

RESEARCH Open Access

Exploring and mitigating shortcomings in single-cell differential expression analysis with a new statistical paradigm

Chih-Hsuan Wu[1], Xiang Zhou[2] and Mengjie Chen[3*]

*Correspondence: mengjiechen@uchicago.edu

[1] Department of Statistics, University of Chicago, Chicago, USA
[2] Department of Biostatistics, University of Michigan, Ann Arbor, USA
[3] Department of Human Genetics and Department of Medicine, University of Chicago, Chicago, USA

Abstract

Background: Differential expression analysis is pivotal in single-cell transcriptomics for unraveling cell-type–specific responses to stimuli. While numerous methods are available to identify differentially expressed genes in single-cell data, recent evaluations of both single-cell–specific methods and methods adapted from bulk studies have revealed significant shortcomings in performance. In this paper, we dissect the four major challenges in single-cell differential expression analysis: excessive zeros, normalization, donor effects, and cumulative biases. These "curses" underscore the limitations and conceptual pitfalls in existing workflows.

Results: To address the limitations of current single-cell differential expression analysis methods, we propose GLIMES, a statistical framework that leverages UMI counts and zero proportions within a generalized Poisson/Binomial mixed-effects model to account for batch effects and within-sample variation. We rigorously benchmarked GLIMES against six existing differential expression methods using three case studies and simulations across different experimental scenarios, including comparisons across cell types, tissue regions, and cell states. Our results demonstrate that GLIMES is more adaptable to diverse experimental designs in single-cell studies and effectively mitigates key shortcomings of current approaches, particularly those related to normalization procedures. By preserving biologically meaningful signals, GLIMES offers improved performance in detecting differentially expressed genes.

Conclusions: By using absolute RNA expression rather than relative abundance, GLIMES improves sensitivity, reduces false discoveries, and enhances biological interpretability. This paradigm shift challenges existing workflows and highlights the need for careful consideration of normalization strategies, ultimately paving the way for more accurate and robust single-cell transcriptomic analyses.

Background

Differential expression (DE) analysis in single-cell transcriptomics provides essential insights into cell-type–specific responses to internal and external stimuli [1–4]. While many methods are available to identify differentially expressed genes from single-cell transcriptomics, recent studies raise important concerns about the performance of state-of-the-art methods, including both methods tailored to single-cell data and techniques that work well in bulk [5–7]. As population-level single-cell studies rapidly become more feasible, powerful and accurate analytical methods will be essential for obtaining meaningful results. In this context, we discuss the four "curses" that currently plague the DE analysis of single-cell data: excessive zeros, normalization, donor effects, and cumulative biases, highlighting various limitations and conceptual flaws in the current workflows. We demonstrate these limitations using a few datasets from 10X single-cell RNA-seq (sRNA-seq) protocols [8]. Finally, we present a new paradigm as a potential solution to some of these issues and illustrate its performance using three case studies.

The curse of zeros

Bulk RNA-seq provides the average transcriptional output of each gene expressed within a population of heterogenous cell types [9, 10]. Due to the sample characteristics, even a moderate sequencing depth can yield information about many thousands of different genes. In comparison, scRNA-seq data is much sparser, with fewer genes expressed per sample and a high proportion of genes with zero UMI counts. Zero UMI counts for a gene can arise from any one of three scenarios: a genuine zero, indicating that the gene is not expressed; a sampled zero, indicating that the gene is expressed at a low level; or a technical zero, indicating that the gene is expressed at a high level, but not captured by the assay. Despite an increasing body of evidence suggesting that cell-type heterogeneity is the major driver of zeros observed in 10X UMI data [11–13], the prevailing notion within the single-cell community is that zeros are largely uninformative technical artifacts caused by "drop-out" genes (i.e., technical zeros).

Accordingly, many single-cell DE studies include pre-processing steps aimed at removing so-called zero inflation. Several popular pre-processing methods include (1) performing feature selection by aggressively removing genes based on their zero detection rates, such as requiring non-zero values in at least 10% of total cells and restricting DE analysis to a smaller gene set; (2) imputing zeros and performing DE on imputed values [14–17]; and (3) modeling zeros explicitly as an extra component and essentially performing DE on non-zero values only [18, 19].

However, if zeros are genuine biological zeros due to no or very low expression, dismissing or correcting for zeros in scRNA-seq discards a significant portion of information in the dataset before any analysis. By failing to account for cell-type heterogeneity, zero-inflation pre-processing steps such as normalization and imputation can introduce unwanted noise into downstream analyses, including DE. Ironically, the most desired markers in single-cell DE analysis—e.g., genes that are exclusively expressed in a rare cell type—may be obscured by current pre-processing steps for handling zeros.

The curse of normalization

The term "normalization" has been used to denote multiple distinct approaches in genomics [20, 21]. For example, it can refer to the process of correcting PCR amplification biases introduced during sequencing library preparation (i.e., library size normalization) [22–24], the process of harmonizing data across different experimental batches (i.e., batch normalization) [25–29], or to the process of transforming the data to adhere to a normal distribution (i.e., data distribution normalization) [30]. All three normalization approaches have been applied to both bulk and single- cell RNA-seq data, aiming to minimize unwanted technical variations. Choosing appropriate normalization techniques for DE analysis of scRNA-seq data is clearly important to maintain the integrity of the data, but the field has yet to establish a definitive gold standard outlining the circumstances in which different normalizations should be performed.

Library-size normalization is critical in bulk RNA-seq analysis, as it is impossible to track the absolute abundance of RNA molecules in typical bulk RNA-seq protocols because the level of amplification introduced by PCR during library construction is unknown. In this instance, normalization focuses on estimating and subsequently correcting for a sample-specific size factor. This process allows bulk RNA-seq to estimate relative RNA abundances. Post-normalization, samples are calibrated against a common reference, resulting in most genes displaying similar expression levels across samples. When performing DE analysis with bulk RNA-seq data, genes are classified as either up-regulated or down-regulated, based on the assumption that the majority remain unchanged across groups. While this size-factor-based normalization technique is suitable for bulk RNA-seq, it does not translate effectively to scRNA-seq. Protocols in scRNA-seq, such as the 10X, employ unique molecular identifiers (UMIs), which discern between genuine RNA molecules and those generated via PCR. This enables the absolute quantification of RNA levels. Unfortunately, size-factor-based normalization methods (e.g., counts per million reads mapped, or CPM), convert data into relative abundances, erasing useful data provided by the UMIs. Furthermore, because the uniform number of molecules found in CPM-normalized data does not accurately represent true expression levels, CPM-normalized data does not account for competition among genes for cellular resources ultimately leading to suboptimal DE analysis results.

In batch effect normalization, dimension reduction methods pinpoint genes with consistent expression patterns across various batches; these genes act as anchors, guiding the alignment and integration of data [31]. However, in scRNA-seq analysis, only highly expressed or highly variable genes are retained for estimating batch effects and subsequent integration. As a result, gene numbers in integrated scRNA-seq datasets are noticeably reduced compared to the raw UMI data.

For data distribution normalization, the field offers both straightforward (e.g., log-transformation) and advanced strategies (e.g., variance stabilizing transformation, or VST). A notable implementation of VST for scRNA-seq is sctransform [32], which employs a regularized negative binomial regression model, preserving the Pearson residuals for future analytical steps, including DE analysis [33]. However, if the underlying data distribution deviates significantly from the assumed model, the application of VST may introduce bias into the analysis.

To demonstrate the effects of various normalization methods on single-cell data, we compared the raw UMI counts of 10 × scRNA-seq data obtained from post-menopausal fallopian tubes (see Methods) with data normalized using one of three methods: 1) CPM; 2) integrated log-normalized counts after removing batch effects using the Seurat CCA model [34]; and 3) VST using sctransform [32].

We found substantial variation in library sizes across different cell types using total UMI counts; notably macrophages (MP) and secretory epithelial (SE) cells exhibited significantly higher RNA content than other cell types (Fig. 1a). Furthermore, SE cells exhibited larger mean library sizes than mast (MA) cells across all donors. These findings align with the understanding that the main active cell types in post-menopausal fallopian tubes are MP and SE cells, with other cell types remaining dormant post-menopause. However, in the integrated data, the disparities in library size distribution are mitigated, even within cell types (Fig. 1a). While integration reduced differences across donors, it masks variation across cell types. It is worth mentioning that CPM normalization equalizes library sizes across all cell types; such normalizations may potentially obscure differences between cell types that are vital for understanding their unique biological functions. As discussed in the previous section, samples with cell-type heterogeneity may be adversely impacted by normalization methods. In the fallopian tube data set, we observed variation in expression patterns across cell types for UMI counts; however, these differences are less apparent in imputed or certain transformed datasets (Fig. 1a). Gene expression frequency also differs across cell types (Fig. 1b). However, normalization processes can substantially alter the distribution of both non-zero UMI (Fig. 1c) and zero UMI counts (Fig. 1d) counts. For example, while the frequency of genes exponentially declines as raw UMI counts increase, VST data forms a more bell-shaped curve with a mode around 1.5 for non-zero raw UMI counts. Non-zero CPM-normalized data, (scaled by 1000) peaks near 0.2 and is more right-skewed than the VST data. Following batch integration, UMI counts primarily fall below 5 and are not as strongly right-skewed. It is noteworthy that zero UMI counts can be given non-zero values via normalization (except with CPM normalization); for example, zeros in VST data are adjusted to negative values and are left-skewed (Fig. 1d). Conversely, the integration process transforms the original zeros to values clustered closely around zero. We further examined the distributions of gene expression from one gene. Using the gene *RUNX3* as an example (Fig. 1e), the distributions in raw UMI counts and CPM data remain right-skewed. In contrast, the VST and integrated data showcase broader, bell-shaped distributions. The handling of zeros in these latter datasets (i.e., VST and integrated) intrinsically sets them apart from the former. This variability, combined with shifts in distribution skewness, raises concerns about performing DE analysis with normalized values.

The curse of donor effects

Recent reviews have highlighted that many single-cell DE analysis methods are susceptible to generating false discoveries [5]. This is mainly due to failing to account for variations between biological replicates, commonly referred to as "donor effects." In single-cell studies, donor effects are confounded with batch effects since cells from one biological sample are typically processed in the same experimental batch. While single-cell

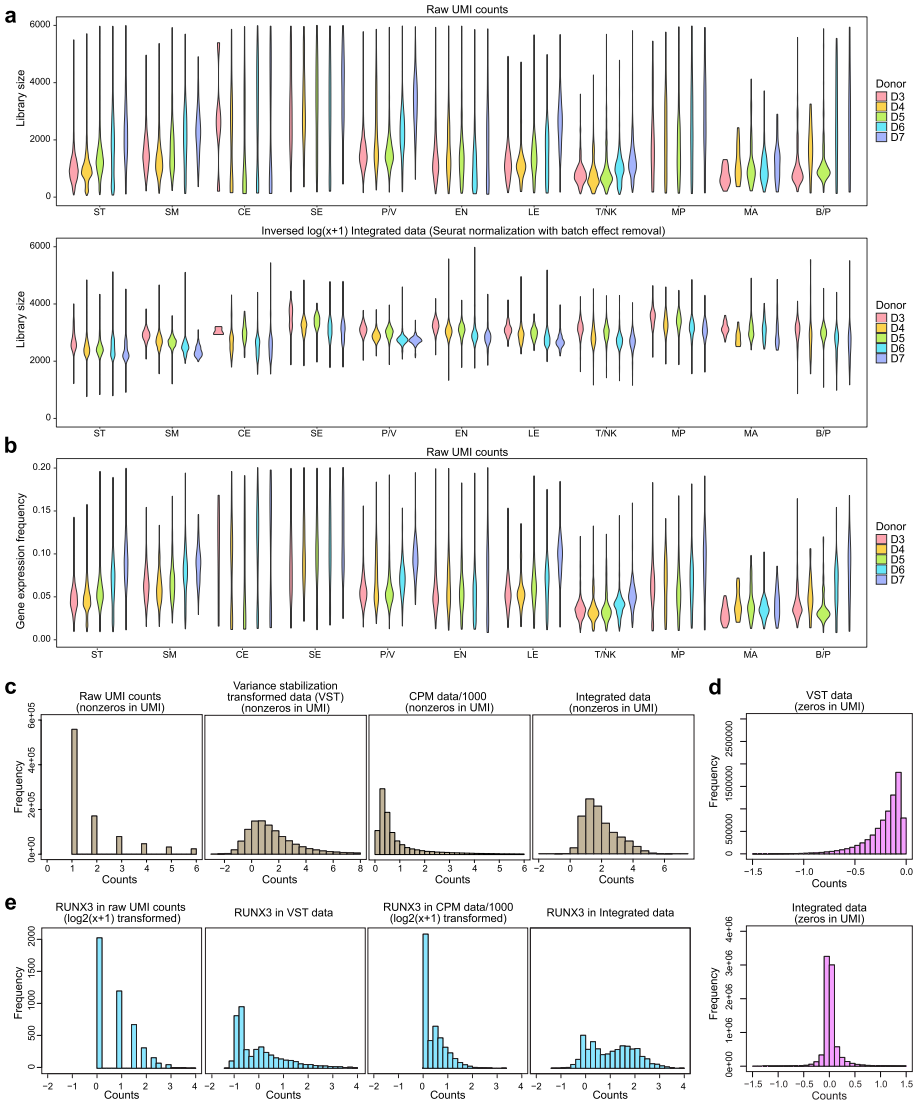

Fig. 1 Effects of normalization on library size, zero frequency, and gene count distributions. **a** Violin plots display library sizes based on raw UMI counts (top) and after data integration (bottom), categorized by cell types and donors. Cell type abbreviation: stromal (ST), smooth muscle (SM), ciliated epithelial (CE), secretory epithelial (SE), parietal/vascular (P/V), endothelial (EN), lymphatic endothelial (LE), T cells/natural killer (T/NK), macrophages (MP), mast (MA), B cells/plasma (B/P). **b** Violin plot illustrating the frequency of gene expression (non-zero counts) in raw UMI data. **c** Histograms representing the distribution of non-zero counts in raw UMI data across various data transformations. **d** Histograms detailing the zero counts in raw UMI data, comparing VST with integrated data where zeros are imputed or converted to non-zeros. **e** Histograms showing the distribution of gene *RUNX3* across different data transformations

studies that contain multiple samples will perform batch correction as pre-processing, they usually do not correct for donor effects when performing DE tests in the downstream analysis [35, 36].

However, it is unclear if batch effect correction alone suffices to eliminate donor-related effects. To address this, we investigated the contributions of variation from different sources before and after batch correction. Using the same fallopian tube dataset described above, we further separated 4553 T/NK cells into 20 subtypes using HIPPO

[37] (Fig. 2a, Additional file 1: Fig. S1). With the aid of canonical markers, we identified specific subtypes, including NK, CD4 + T, CD8 + T, and mature naive T cells. We then focused on subtypes that were observed in all donors (Fig. 2b–c).

To quantify the proportion of variation originating from different sources, we fit a linear model for each gene in several subtype pairs, using cell types and donors as covariates. Through all pairs, the integration reduces donor variation (Fig. 2d, Additional file 1: Fig. S2). However, in comparisons of two subtypes of the same cell type (12 vs.13) and two subtypes of different cell types (13 vs. 19), we observed a decrease in the proportion of cell-type–related variation. This underscores that integration not only mitigates batch effects but also impacts the phenotypes of interest. Importantly, our analysis indicated that even after implementing batch correction, a notable percentage of genes still exhibited donor-related effects (Fig. 2e). As batch effects are often estimated from leading principal components, representing a consensus from a subset of genes, it is possible

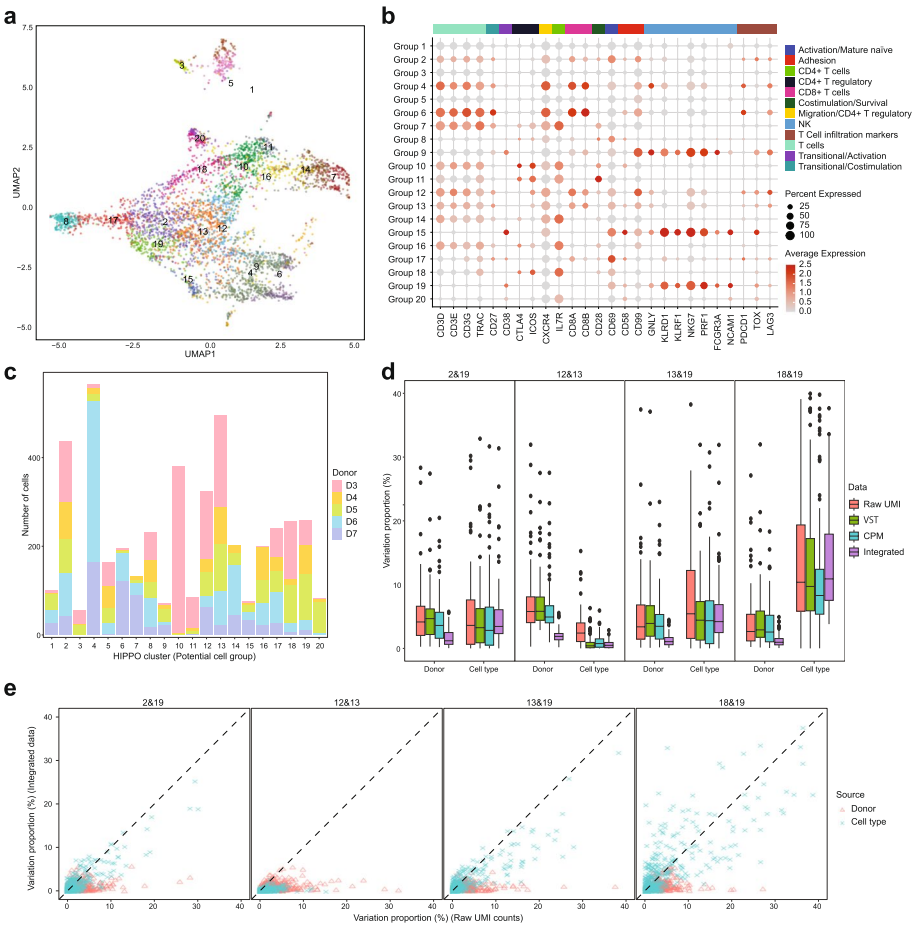

Fig. 2 Cluster and variation analysis of single-cell data from the fallopian tube. **a** UMAP visualizing 20 clusters identified by HIPPO in case study 1. **b** Canonical markers delineate specific cell subtypes: clusters 9, 15, and 19 as NK cells; clusters 7, 10, 11, 14, 16, 18, and 20 as CD4 + T cells; clusters 4, 6, 12, and 13 as CD8 + T cells; clusters 8 and 17 as mature naive T cells. **c** Distribution of donors across the 20 identified clusters. **d** Comparative analysis of variation proportions attributable to donor and cell type effects across different pairs and datasets. **e** Scatter plots comparing variation proportions due to donor and cell type effects across various pairings and data sources

that residual donor effects persist on some, if not all, genes. Therefore, it is crucial to account for donor effects when performing DE tests to avoid false discoveries and obtain accurate results, even after removing batch effects.

A widely used approach to address donor effects in single-cell studies is pseudo-bulk analysis. This involves aggregating cells from the same donor and conducting DE analysis using tools such as DESeq2 [38] or edgeR [39] on the combined data. While effective in some cases, this framework has notable limitations. By treating donor effects as fixed and assuming uniform influence across all cells, it overlooks within-sample heterogeneity. This oversimplification can make the analysis overly conservative, potentially missing significant findings [5]. Moreover, bulk RNA-seq DE tools perform normalization by default, which may have the same drawbacks mentioned earlier in the context of single-cell studies. Therefore, while pseudo-bulk analysis can be useful, it requires careful consideration as it may not always provide a fully accurate resolution to the challenges posed by donor effects in single-cell research.

The curse of cumulative biases

In scRNA-seq analysis, it is common to follow a hierarchical, sequential workflow for clustering and DE analysis. This approach can carry biases from one step to the next; such cumulative biases can ultimately diminish the power to detect differentially expressed genes [40].

Unsupervised learning, especially clustering analysis, is essential in single-cell studies. It groups cells based on gene expression patterns, facilitating the cell-type annotation. While clustering is effective with normalized values like CPMs, it essentially reweights gene features based on their relative contributions. As a result, clustering provides a generalized perspective of variation in gene expression across cell types. The reliance on relative expression also makes clustering fairly resilient to errors and biases introduced by the pre-processing steps.

On the other hand, DE analysis operates at the gene level, using group labels from the clustering process. The effects of biases, whether from donors or batch processing, can vary for each gene. Although DE analysis technically follows clustering—given its reliance on group labels—the metrics used do not need to be identical for both. As we show later in the case studies with data that complete clustering and annotation successfully, if DE analysis is performed using processed expression levels, the cumulative biases can still lead to false discoveries and/or missed differentially expressed genes (DEGs).

An alternative paradigm: *G*eneralized *LI*near *M*ixed-*E*ffects model for *S*ingle-cell expression studies (GLIMES)

To minimize the pre-processing biases discussed above, we propose an approach that conducts DE analysis on raw UMI counts or zero proportion prior to implementing batch correction, normalization, imputation, or feature selection. This approach, which uses generalized linear mixed models (GLMMs) [41], preserves sample-specific structures and biological signals in the data. Furthermore, our proposed approach can adjust for any potential confounding factors, such as age, sex, or ancestry, by incorporating them as covariates with fixed effects. This framework enables us to explicitly account for the variation among biological replicates in comparison to other effects (Fig. 3). We

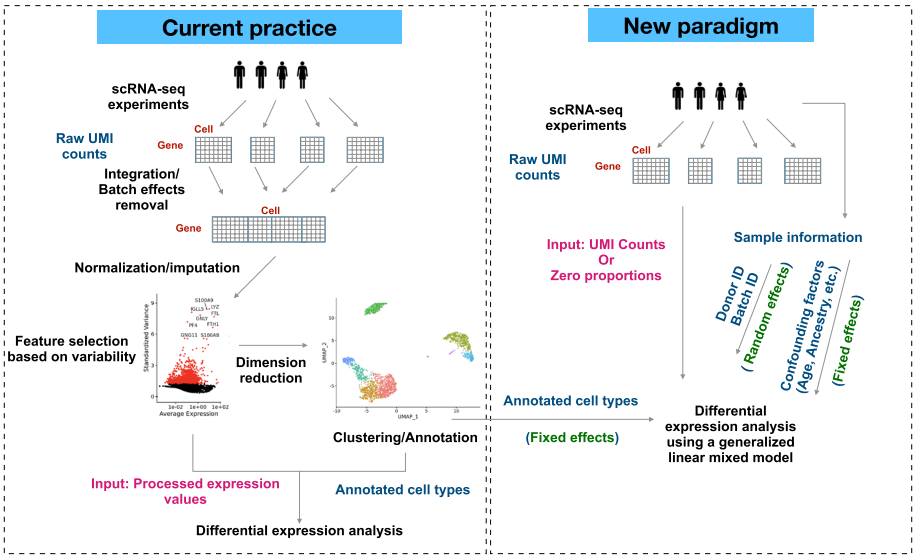

Fig. 3 Comparison of established workflows and proposed paradigm for single-cell analysis. Left: Under the current single-cell analysis pipeline, the raw UMI counts collected from multiple donors are integrated to remove the batch effects and normalized for further analysis. It is common to perform DE analysis on processed data. Right: Our new paradigm directly performs a generalized linear mixed model on raw UMI counts. The random effect can account for the batch effect due to samples. The fixed effect contains the group of interest (e.g., cell types, control/treatment) and other adjustments (e.g., age, sex). The annotated cell types can be obtained from the existing pipeline or HIPPO algorithm which clusters cells based on the zero proportions of UMI counts

provide two GLMM models, Poisson-glmm and Binomial-glmm, which model UMI counts and zero proportion respectively. The proposed procedures have been implemented in software GLIMES (https://github.com/C-HW/GLIMES).

Unlike existing GLMM-based packages such as Muscat [42], GLIMES models the group of interest as a fixed effect while accounting for donor-specific variations and other batch effects as random effects. This distinction is significant because Muscat is primarily designed for DE analysis across different states or conditions, under the assumption that groups of interest are sequenced in separate experimental batches. In Muscat's implementation, the experimental unit—defined as the combination of donor and group of interest—is treated as a random effect in such scenarios.

However, this design may not be suitable for other analyses, such as comparisons across cell types, where the random effect does not align with the study's structure. Additionally, Muscat normalizes counts by incorporating library size factors as an offset in its GLMMs, which emphasizes relative abundance rather than raw counts. This approach mirrors the behavior of pseudo-bulk methods, as Muscat groups counts within the same donor prior to normalization. Consequently, its performance is often comparable to pseudo-bulk methods, as shown in subsequent examples.

To benchmark the performance of our new paradigm, we rigorously evaluated eight distinct methods (Table 1) for DE analysis: two new methods from GLIMES, Poisson-glmm and Binomial-glmm; two traditional pseudo-bulk methods, DESeq2 and edgeR; and four existing single-cell–specific methods, MAST [19], Wilcox in Seurat, and two Muscat GLMMs (MMvst and MMpoisson). Each method was applied in three case

Table 1 Comparison of DE methods used in this paper

	Poisson-glmm	Binomial-glmm	Pb-DESeq2	Pb-edgeR	MAST	Wilcox	MMvst	MMpoisson
Package	GLIMES	GLIMES	Muscat	Muscat	MAST	Seurat	Muscat	Muscat
Input	UMI	Zero counts	UMI	CPM	CPM	Integrated (v4)/ Log normalized (v5)	VST	UMI
Model base	Poisson glmm	Binomial glmm	Negative binomial model	Negative binomial model	Zero-inflated model	Rank-sum test	LMM	Poisson glmm
Normalization	X	X	V	V	V	V	V	V
Normalization method			1. M median of ratio size factor and variance stabilizing transformation in the method	1. CPM normalization 2. Trimmed mean of M values (TMM) in the model	1. CPM normalization	1. Integration applied on log normalized data by Seurat in case study 1 2. Log2-transformed normalized data by Muscat in case studies 2 and 3	1. VST normalization	1. Library size factor as offset in the model

studies (i.e., across cell types, tissue regions, and cell states). We evaluated various scenarios, such as variations in library size between groups, pronounced heterogeneity within groups, and confounding batch effects.

As mentioned, Poisson-glmm fits a GLMM model on UMI counts, while Binomial-glmm fits a GLMM model on the zero proportion of each gene, adding donors as random effect. Pseudo-bulk DESeq2 applies both VST and library size normalization. Pseudo-bulk edgeR applies library size normalization. MAST adopts a zero-inflated negative binomial model, using log-transformed CPM counts and incorporating the cellular detection rate as covariates. The Wilcox test is non-parametric, using integrated normalized counts (Seurat v4) or normalized counts (Seurat v5). The two Muscat models, MMvst with VST counts and MMpoisson with raw UMI counts, account for library size. Both Muscat models consider donor–group combinations as random effects. See the "Methods" section for more details.

Results

Case study 1: DE analysis on different immune cell types in the fallopian tube

Using the fallopian tube data set described above, we examined the efficacy of various methods across three distinct scenarios: homogeneous groups with highly variable library sizes, homogeneous groups with similar library sizes, and heterogenous groups.

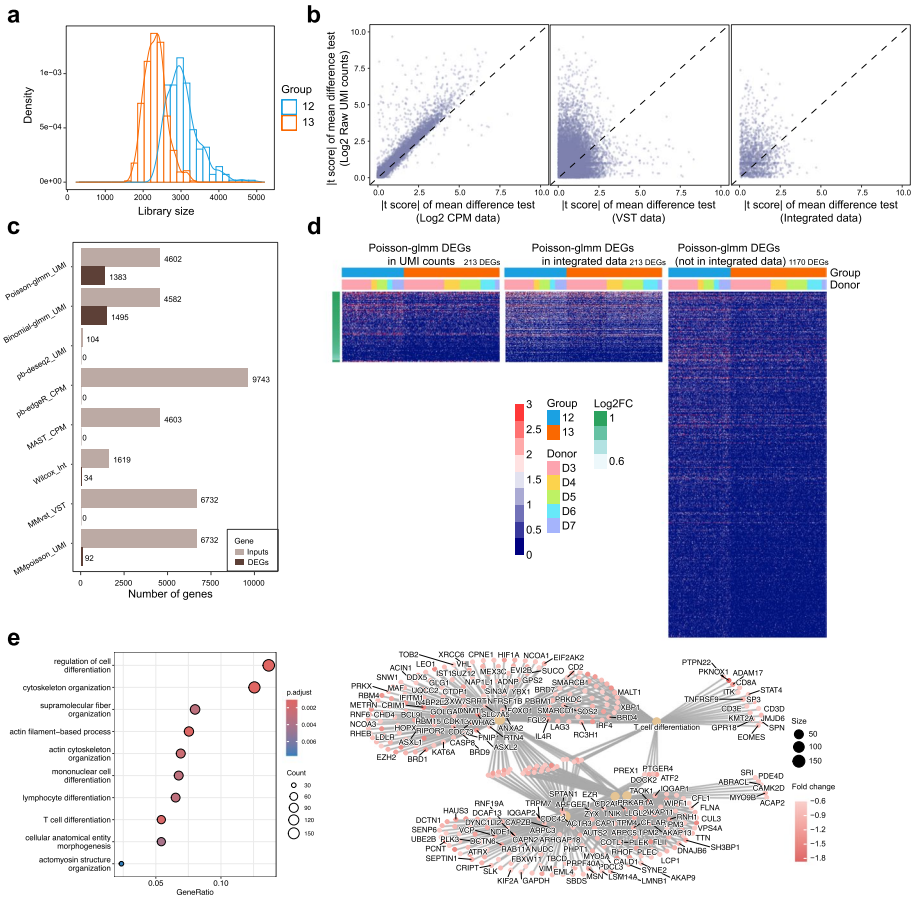

Fig. 4 DE analyses on homogeneous CD8 + T cell subgroups. **a** Density plot of the library size for groups 12 and 13. **b** Scatterplot comparisons of *t*-scores from mean difference tests between raw UMI counts and other transformed data. Each gene's expression in two different groups is compared, showcasing the pairwise absolute *t*-scores from various data sources. **c** Counts of input genes and DEGs in different DE methods. **d** Heatmaps visualize Poisson-glmm DEGs. Order: UMI counts (left), integrated data (middle), and genes not included in the integrated data but shown in UMI counts (right). Heatmaps arrange genes by descending Poisson-glmm fold change estimates and columns group cells by cell clusters and donors. **e** GO analysis of the DEGs identified by Poisson-glmm

For each scenario, we illustrate the overarching gene expression profile, describe the DE results using diagnostic plots, and conduct a gene ontology (GO) analysis to investigate the biological foundations of our DE findings.

Contrasting CD8 + T cell subgroups with marked library size differences

We compared groups of CD8 + T cells (clusters 12 and 13), where there are notable differences in library sizes (*t*-statistics 26.5 on 513 df, p-value < 2.2e − 16) (Fig. 4a) to investigate the impact of library-size–based normalization on single-cell data. Using a two-sample *t*-test, we compared gene expression means between these groups with raw UMI counts and three normalization methods (Fig. 4b) using absolute t-scores. While t-scores from CPM mirror those from UMI counts, albeit with minor shrinkage, both VST and integration show substantial shrinkage, showing that normalization dampens gene expression differences between the groups prior to DE analysis (Fig. 4b). Each

tested method employs its unique filtering approach within the implemented function, resulting in varying numbers of input genes. Specifically, the GLIMES methods (i.e., Poisson-glmm, Binomial-glmm) and MAST each utilize nearly 4600 genes as input (Fig. 4c). In contrast, pseudo-bulk DESeq2 applies default quality control criteria to both genes and cells, resulting in only 104 genes being retained. Pseudo-bulk edgeR retains 9743 input genes in the CPM data. Wilcox in Seurat retains 1619 genes using a prefilter on average log2 fold change, while Muscat mixed models utilize 6732 genes.

Most methods have evenly distributed fold changes and have predominately large adjusted p-values (Additional file 1:Fig. S3b). In particular, pseudo-bulk and mixed models from the Muscat package have p-values clustered around one. In contrast, GLIMES volcano plots show heavily imbalanced expression patterns, that align with the correlating density plots (Additional file 1: Fig. S3c) and identify a substantial number of DEGs (Fig. 4c). The heatmaps of DEGs further emphasize that raw counts can better capture the differences between groups compared to integrated counts (Fig. 4d). In fact, 1170 DEGs were excluded from the integrated data before testing.

Our GLIMES GLMM methods directly use UMI counts or zero proportions, which allows the detection of DEGs by preserving the contribution of DEGs to differences in expression. The DEG contributions are masked by normalization in the other methods because genes that are not differentially expressed have a much larger effect on the difference in library size between groups 12 and 13 than DEGs have.

The DEGs prominently feature GO terms associated with cytoskeleton reorganization and cell differentiation (Fig. 4e). As T cells detect antigens on an antigen-presenting cell, they establish an immunological synapse, necessitating substantial actin filament restructuring. Actin polymerization within this synapse aids the transit of receptors and signaling molecules, crucial for T cell activation [43]. In addition, cytoskeleton reorganization is essential for CD8 + T cell differentiation by enabling migration, signal transduction, and the establishment of effector and memory cell states. Our results hint that among the two CD8 + T cell groups (i.e., groups 12 and 13), group 12 cells are likely activated T cells.

A glimpse at CD4 + T cells vs. NK cells: smaller library size differences

To determine if normalization also affects clusters with smaller library-size differences, we analyzed CD4 + T and NK cells (clusters 2 and 19) using all eight DE methods. The CD4 + T and NK cell clusters have more similar library sizes than the two CD8 + T cells discussed above (t-statistics − 8.08 on 543 df, p-value 4e − 15). All donors, except donor 7, have similar library-size distributions (Additional file 1: Fig. S4a). The zero proportions of genes in these two clusters fit a Poisson distribution well, indicating relative homogeneity within each cell cluster (Additional file 1: Fig. S4a).

In this comparison, GLIMES GLMMs and MAST each use nearly 4000 genes as input (Additional file 1: Fig. S4b). Methods implemented in the Muscat package, including pseudo-bulk methods DESeq2 and edgeR and mixed models MMvst and MMpoisson, input 1384, 9960, 5694, and 5693 genes, respectively, in accordance with their filtering procedures. The Wilcox method from the Seurat package includes 1,740 input genes due to the filtering based on the log2 fold change between two groups of interest. MAST, pseudo-bulk methods, and MMvst each identify fewer than 100 DEGs, while Wilcox

identifies 293 DEGs. In contrast, the methods that use UMI counts, Poisson-glmm, Binomial-glmm, and MMpoisson, identified 608, 639, and 317 DEGs, respectively (Additional file 1: Fig. S4b). GLIMES methods provide at least 338 unique DEGs not identified by other methods (Additional file 1: Fig. S4b).

GLIMES GLMMs show positive log2 fold changes, signifying that genes are upregulated (Additional file 1: Fig. S4c). From the pairwise comparisons of log2 fold change (Additional file 1: Fig. S4d), MAST, Wilcox, and MMvst exhibit smaller log2 fold change estimates than our methods, likely due to normalization processes that shrink the values. Pseudo-bulk methods tend to yield more conservative p-values (Additional file 1: Fig. S4e). While the log2 fold change estimates are consistent across GLIMES GLMMs, pseudo-bulk methods, and MMpoisson, the presence of deviant p-values leads to significant disparities in the identification of DEGs. Our GLMMs identified many more DEG candidates that surpass the thresholds of adjusted p-value and fold change than any other method.

In Additional file 1: Fig. S4f, we display gene expression from DEGs identified by Poisson-glmm alongside heatmaps for VST, CPM, and integrated data. In this comparison, the distinctions between groups are visible in all data sources, but integrated data only contains 153 DEGs, missing 455 DEGs. The heatmaps of DEGs from GLIMES GLMMs show the validity of DEGs (Additional file 1: Fig. S4g), while most of the DEGs identified by MMpoisson and Wilcox do not display any differential expression pattern in UMI counts (Additional file 1: Fig. S4h, i). In addition, our methods provide 440 DEGs not in MMpoisson DEGs (Additional file 1: Fig. S4i).

We performed GO enrichment analysis on DEGs from Poisson-glmm. The DEGs are enriched for GO terms related to leukocyte activation, cell activation, and lymphocyte activation (Additional file 1: Fig. S4j), suggesting NK cells represented by cluster 19 are more active than the CD4 + T cells in the fallopian tube samples.

In summary, even with smaller library-size differences between clusters, normalization still obscures information and hinders the identification of potential DEGs.

Deciphering the complexities of heterogeneous groups: mature T cells vs. CD4 + T cells

To explore each method's effectiveness in identifying DEGs in heterogenous groups-of-interest, we merged clusters of mature T cells (i.e., 8 and 17) and CD4 + T cells (i.e., 2 and 19). The distributions of library sizes for these clusters exhibit noticeable differences (Fig. 5a), and the zero proportions deviate from a Poisson distribution (Additional file 1: Fig. S5a), confirming their heterogeneity.

In this comparison, the two GLIMES GLMMs and MAST each use ~ 3480 genes as input. Pseudo-bulk DEseq2, edgeR, Wilcox, and mixed models utilize 1937, 10,483, 1774, and 7099 genes, respectively. Our GLMM methods show predominantly positive estimates of fold change in volcano plots (Fig. 5b), suggesting higher expression of abundant genes in CD4 + T cells (i.e., groups 2 and 19). MAST and MMvst show a similar tendency, but the imbalance is less pronounced. In contrast, the two pseudo-bulk methods, Wilcox, and MMpoisson provide more evenly distributed estimates in both directions, with Wilcox and pseudo-bulk DESeq2 identifying an abundance of negative DEGs.

The estimates of log2 fold change are not quite identical across different methods (Additional file 1: Fig. S5c). Both pseudo-bulk methods exhibit a negative shift compared

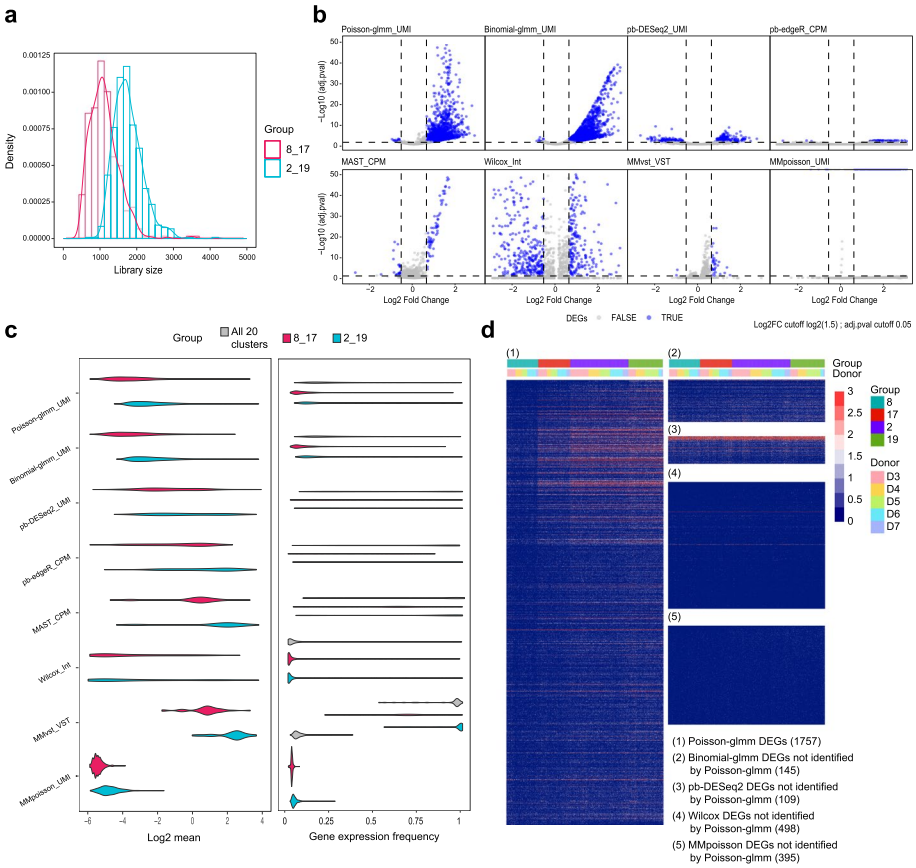

Fig. 5 DE analyses on heterogeneous groups: Mature T Cells vs. CD4 + T Cells. **a** Density plots comparing library sizes for combined groups 8 and 17 and 2 and 19. **b** Volcano plots for each method. The signs of log2 fold change are adjusted such that positive signs represent higher expressions in group 2_19. **c** Left: Violin plot of log2 gene mean for DEGs from different methods. Right: Comparisons of the gene expression frequency of the DEGs from different methods. **d** Heatmaps of DEGs from different methods

to Poisson-glmm, while MMpoisson had a positive shift. MAST, Wilcox, and MMvst once again show shrinkage. This observation sheds light on how normalization and logarithmic transformation during pre-processing influences the estimation of differences in gene expression.

When we examine the violin plots of gene expression frequency and log2 mean for the DEGs identified by each method, it becomes apparent that MAST and MMvst capture fewer DEGs with lower gene expression frequency and smaller gene means than the remaining methods (Fig. 5c). It is worth noting that MAST is a zero-inflated model, which incorporates excessive zeros as an additional component. However, MAST might not effectively characterize the zeros, as demonstrated in previous studies on UMI counts [26, 40]. Consequently, potential DEGs that are lowly expressed may be masked by the model. MMvst, despite having a considerable number of input genes ($n = 7099$), only identified 35 DEGs.

The heatmap of DEGs in Poisson-glmm reveals distinct expression patterns between the two groups (Fig. 5d (1)). However, in this example, the inherent heterogeneity within each group impacts the fitness of the Poisson model, potentially leading to false discoveries. We further examined additional DEGs identified by other methods, but not

by Poisson-glmm (Fig. 5d (2)–(5)). The heatmaps make it evident the DEGs that differentiate between the two groups are largely identified by Poisson-glmm only; the other methods did not contribute additional valid DEGs that differentiate the two groups. Conversely, most of the DEGs detected by Poisson-glmm exhibit differential expression despite the heterogeneity within each group.

Notably, MMpoisson mainly detected DEGs with small means (Fig. 5c), not showing clear differences between different groups (Fig. 5d (5)). And the DEGs are mutually exclusive to those identified by Poisson-glmm. Although Poisson-glmm and MMpoisson both use UMI counts, MMpoisson includes group information as a random variable and involves library size as an offset; our result underscores the significance of using an appropriate random effect in a mixed model and suggests that the cell group information should be excluded from the random component.

The DEGs are enriched for GO terms related to peptide metabolic process and cytoplasmic translation, indicating lower ribosomal RNA activities in mature T cells (Additional file 1: Fig. S5e). Indeed, mature T cells exhibit lower levels of ribosomal RNA activity compared to their immature counterparts, mainly due to the state of activation and the metabolic requirements of the cells. On the other hand, mature T cells, which are not rapidly proliferating, have less need for protein synthesis and thus exhibit lower levels of rRNA activity. However, upon antigen recognition and activation, mature T cells can rapidly upregulate rRNA activity and protein synthesis to support clonal expansion and effector function. This differential regulation of rRNA activity is one of the many ways in which cells regulate their metabolic activities to adapt to different physiological conditions.

In this example, Poisson-glmm detected more valid DEGs for heterogenous cell populations than other methods. Normalization still diminished measurable differences between groups. We also raise concerns about the masking of genes with low expression by the improper treatment of zeros, as seen in the MAST method and VST data.

Case study 2: DE analysis on different regions of human spinal cord cells

In this case study, we analyzed a human spinal cord dataset containing 48,644 cells and 8092 common genes across nine patients [44]. The samples were distributed across 16 slides, meaning that some donor samples were processed in different experimental batches. Cells were annotated with anatomical regions (Additional file 1: Fig. S6a). UMAP plots show that both donor (i.e., patient identity) and batch (i.e., slide) effects have a stronger influence on cluster formation than the anatomical regions themselves (Additional file 1: Fig. S6b), underscoring the importance of correcting for donor and batch effects. Violin plots for the gene *PLP1* further highlight significant variation across donors and batches (Additional file 1: Fig. S6c).

For clustering, we utilized the Seurat (v5) integration workflow, which produces integrated dimensional reduction embeddings, but does not provide integrated gene-level counts (Additional file 1: Fig. S6d). As a result, using normalized counts in the Wilcox method does not effectively account for batch effects, as batch-specific variations may still be present in the normalized counts.

We show the raw counts retain more variation among different regions than the normalized counts in library-size distribution (Additional file 1: Fig. S6e). Additionally, gene

expression in cells from the central canal (Cent_Can) shows higher raw counts compared to other regions, but the difference in expression is diminished after normalization. This discrepancy could lead to completely opposite outcomes in DE analysis, depending on the type of data used as input. It is crucial in DE analysis to carefully consider the choice of input data type, as improper handling of batch effects or library size differences could mask biologically relevant signals, especially in region-specific analyses.

To examine the effectiveness of DE analysis methods across regions, we compared gene expression between cells from the dorsal horn and ventral lateral white regions. For both regions, the zero-proportion plot deviates from the expected Poisson curve, likely reflecting the presence of multiple cell types (Additional file 1: Fig. S7a). Furthermore, histograms of library size reveal differences between the two regions (t-statistics 17.06 on 7783.8 df, p-value $< 2.2e-16$), as well as variations across donors and batches, which could impact the DE results (Additional file 1: Fig. S7a).

In this case study, we excluded MMpoisson from the analysis due to execution issues. Among the other methods, GLIMES GLMMs, MAST, and Wilcox use between 5600 and 6000 genes as input, while the pseudobulk and mixed models in Muscat use 7500 to 8000 genes. Poisson-glmm, Binomial-glmm, pb-DESeq2, pb-edgeR, and Wilcox identified 288, 190, 518, 361, and 209 DEGs, respectively (Additional file 1: Fig. S7b). The upset plot shows that most DEGs identified by our methods overlap with those identified by other methods, although pb-DESeq2 uniquely identified 121 DEGs. In the volcano plot, foldchange estimates from GLMM methods are predominantly negative, while pseudobulk methods and Wilcox yield a more balanced distribution, with some upregulated DEGs (Additional file 1: Fig. S7c). MAST and MMvst failed to identify DEGs, likely due to overall fold-change shrinkage (Additional file 1: Fig. S7d). Additionally, the p-values from our methods differ significantly from those produced by other methods, likely due to differences in normalization approaches (Discussion) (Additional file 1: Fig. S7d). Finally, the Poisson-glmm DEGs heatmap shows distinct expression patterns between the two regions, while Wilcox identifies additional low-expression DEGs, and pb-DESeq2 DEGs display less distinct patterns (Additional file 1: Fig. S7e).

GO analysis indicates that the gene expression differences between the dorsal horn and the ventral lateral white matter regions, especially in pathways such as synaptic signaling and anterograde trans-synaptic signaling. Dorsal horn is primarily associated with sensory input, particularly the processing of pain and tactile sensations. Upregulated synaptic signaling pathways here could indicate its specialized role in receiving and modulating sensory information through neuronal communication. Genes involved in trans-synaptic signaling in the dorsal horn could contribute to the plasticity and relay of sensory signals to higher brain centers. Ventral lateral white matter primarily contains descending motor pathways that are involved in transmitting motor commands from the brain to the spinal cord and subsequently to muscles. Differences in expression related to anterograde signaling could reflect its role in propagating motor signals efficiently. The enrichment of synaptic and trans-synaptic signaling genes in these regions underscores their respective roles in sensory (dorsal horn) and motor (ventral lateral white) processing.

In this example, we observed that donor effects and additional batch effects are unavoidable and must be accounted for in DE analysis. Additionally, normalizing data using

library-size factors can skew gene expression in specific groups, potentially reversing expression patterns and leading to misleading conclusions. Proper handling of these factors is essential to preserve the biological relevance of DE results.

Case study 3: DE analysis on different states of B cells

In our final case study, investigated how different methods affect the DE analysis of different cellular states. We applied our proposed DE framework to data collected by Kang et al. [41], which consists of 29,065 cells and 7661 genes from eight distinct cell types, collected from peripheral blood mononuclear cells of eight lupus patients. Within each cell type, the cells are evenly split into two groups for perturbation: unstimulated control and IFN-β stimulated (Additional file 1: Fig. S8a). UMAP plots (Fig. 6a) highlight

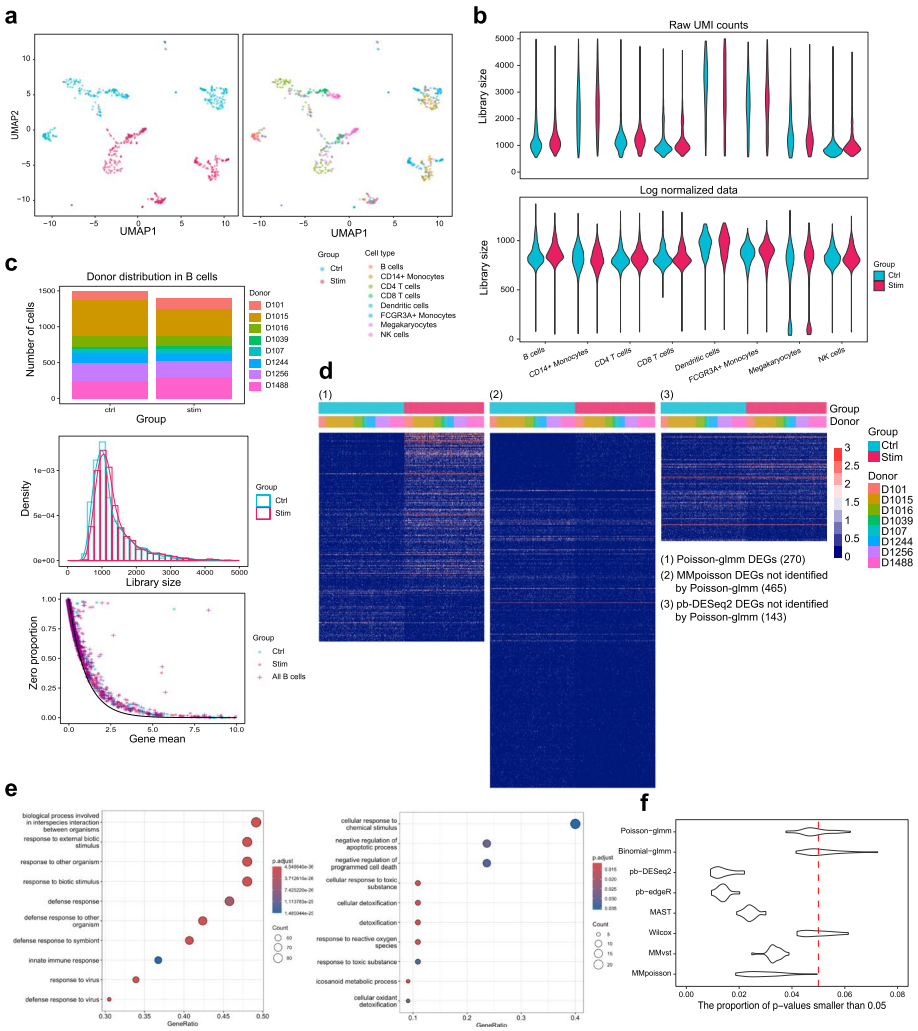

Fig. 6 DE analyses on different states in B cells. **a** UMAP showing groups and cell types for case study 2. **b** Library size comparisons before (raw UMI counts) and after normalization (log-normalized data) by cell type. **c** Top: Donor distribution among B cells. Middle: Density plot of library size in different states. Bottom: Zero proportion plots for different states and combined states. **d** Heatmaps of DEGs identified by different methods. **e** GO analysis for up-regulated (left) and down-regulated genes (right). **f** Violin plots depicting the proportion of *p*-values below 0.05 for each method

that gene expression patterns are more differentiated between stimulation states than between cell types. The zero-proportion plots fit better to Poisson distribution when separated by stimulation states than only by cell types (Additional file 1: Fig. S8b). This observation motivated us to focus on DEGs between the cell states rather than between the cell types. In this dataset, different cell states were processed in separate batches, complicating the analysis, as it becomes challenging to disentangle true biological differences from batch-induced variations. However, it is essential to account for these effects to avoid biased results in DE analysis.

Similar to case study 1, we found that the distribution of library sizes is significantly different after normalization (Fig. 6b). Raw UMI counts show that each cell type has a unique library size distribution. However, these differences are less pronounced following normalization, while library sizes remain relatively consistent between states within a single cell type. However, normalization seems to predominantly affect differences across cell types rather than between states.

For the remainder of our case study, we focused on B cells. The cells from each donor were divided approximately equally between the control and stimulated groups (Fig. 6c top), and the library size distribution in these two groups is similar (Fig. 6c middle). The data does not perfectly fit the Poisson distribution in the zero-proportion plot, indicating a mixture of subtypes within B cells (Fig. 6c bottom).

In our analysis of unstimulated and stimulated B cells, GLIMES GLMM methods, pbDESeq2, and MAST use 2170 to 2650 genes as inputs (Additional file 1: Fig. S9a). Wilcox approach within Seurat uses 5545 genes, and the other methods use around 6600 genes. The upset plot shows that most DEGs identified by our GLMM methods overlap with those identified by other methods, while MMpoisson uniquely identifies 334 DEGs. The estimates of fold change for the two states in B cells exhibit an even spread across all methods, as depicted in the volcano plots (Additional file 1: Fig. S9b). MAST and MMvst struggle to identify differential patterns. In contrast to the previous case studies, GLIMES GLMM approaches flag fewer DEGs than both pseudo-bulk techniques and MMpoisson. Notably, the DEGs that are not shared between pseudo-bulk DESeq2 or MMpoisson and Poisson-glmm predominantly have extremely low expression levels (Fig. 6d (2), (3)).

We hypothesized that this result could be explained by using fold change as a DEG criterion. In bulk RNA-seq, a gene is typically labeled as a DEG if its adjusted p-value is below a certain threshold, often 0.05, and the fold-change estimate exceeds a predetermined value, typically 1.5 or 2 (Additional file 1: Fig. S10a). Most single-cell DE methods use the same criteria as bulk methods; however, in single-cell datasets, the mean counts for many genes are exceedingly close to zero, meaning fold change may not be a reliable metric to differentiate nuances in read counts. For instance, if gene means are 2^{-3} for one group and 3^{-3} for another, the fold-change threshold of 1.5 is met, but the actual difference is a mere 0.0625, which does not convey a significant disparity in expression, especially when juxtaposed with genes having larger means. Moreover, near-zero values can result in computational inaccuracies, causing a fold-change ratio that deviates from the underlying true value.

To overcome the limitation of using fold-change ratios on small counts, we established a new criterion for DEGs based on absolute differences. Specifically, we mandated that

the mean difference between two groups exceeds a set threshold, such as − 1. In the volcano plot, numerous genes would be designated as DEGs when relying on ratio-defined fold change. Yet, the mean vs. mean difference plot shows that many genes that meet the *p*-value criteria showcase only modest changes in absolute mean differences (Additional file 1: Fig. S10a). This approach emphasizes genes with significant absolute differences, yielding more biologically pertinent results. We applied the new criterion on several comparisons in case study 1, leading to a substantial reduction in the number of identified DEGs and a clearer distinction between the compared groups (Additional file 1: Fig. S10d).

We performed GO enrichment analysis on up-regulated and down-regulated genes separately (Fig. 6e). We found IFN-β stimulated B cells have increased activities in the interaction between organisms, defense response, defense response to virus, and defense response to symbiont, while their activities in the cellular detoxification process are decreased (Additional file 1: Fig. S11a). Pseudo-bulk techniques detected similar GO terms (Additional file 1: Fig. S11b), while MMpoisson failed to detect down-regulated GOs (Additional file 1: Fig. S11c).

In this example, we demonstrated that conventional metrics to detect DEGs, especially fold-change ratios, are ill-suited for low-count data where the large fold changes reported by current methods may be attributed to the ratio of two very small gene means. Careful post-processing is needed to prioritize signals and manage false discoveries.

False discovery rates and power assessment

To evaluate p-value calibration in empirical data, we performed a permutation analysis on a null dataset for a specified group of interest. This analysis was conducted on three datasets: CD4 + T cells (group 2), CD8 + T cells (group 13) from case study 1, and B cells (control) from case study 3.

Violin plots (Fig. 6f, Additional file 1: Fig. S12) show that our GLMM methods and the Wilcox method consistently produce well-calibrated *p*-values across different null datasets. In contrast, pseudo-bulk methods and Muscat's mixed models are overly conservative, with *p*-value proportions significantly below 0.05. MAST exhibited conservative behavior for B cells but not for case study 1.

The histograms of *p*-values across the 20 runs demonstrate a consistently flat distribution for our GLMM methods and the Wilcox method, indicative of adherence to the null hypothesis (Additional file 1: Fig. S12). Conversely, other methods display overestimated *p*-values, yielding conservative outcomes. Despite Wilcox's strong performance in the permutation analysis, its power to detect true DEGs remains limited, as demonstrated in earlier case studies. Under both the existing criteria and our newly established criteria for determining DEGs, each method detected no more than one false discovery in each run.

We assessed the power of each method on synthetic datasets simulated with Splatter under four configurations, varying donor and DE effects: (donor effect, DE effect) = (0.5, 0.5), (0,5, 1), (1, 0.5), (1, 1), where the parameter controls the location and scale factor for the effects. The principal component analysis (PCA) plots (Fig. 7a, Additional file 1: Fig. S13a) indicate that the donor effect remains more pronounced than the group of interest, consistent with previous case studies.

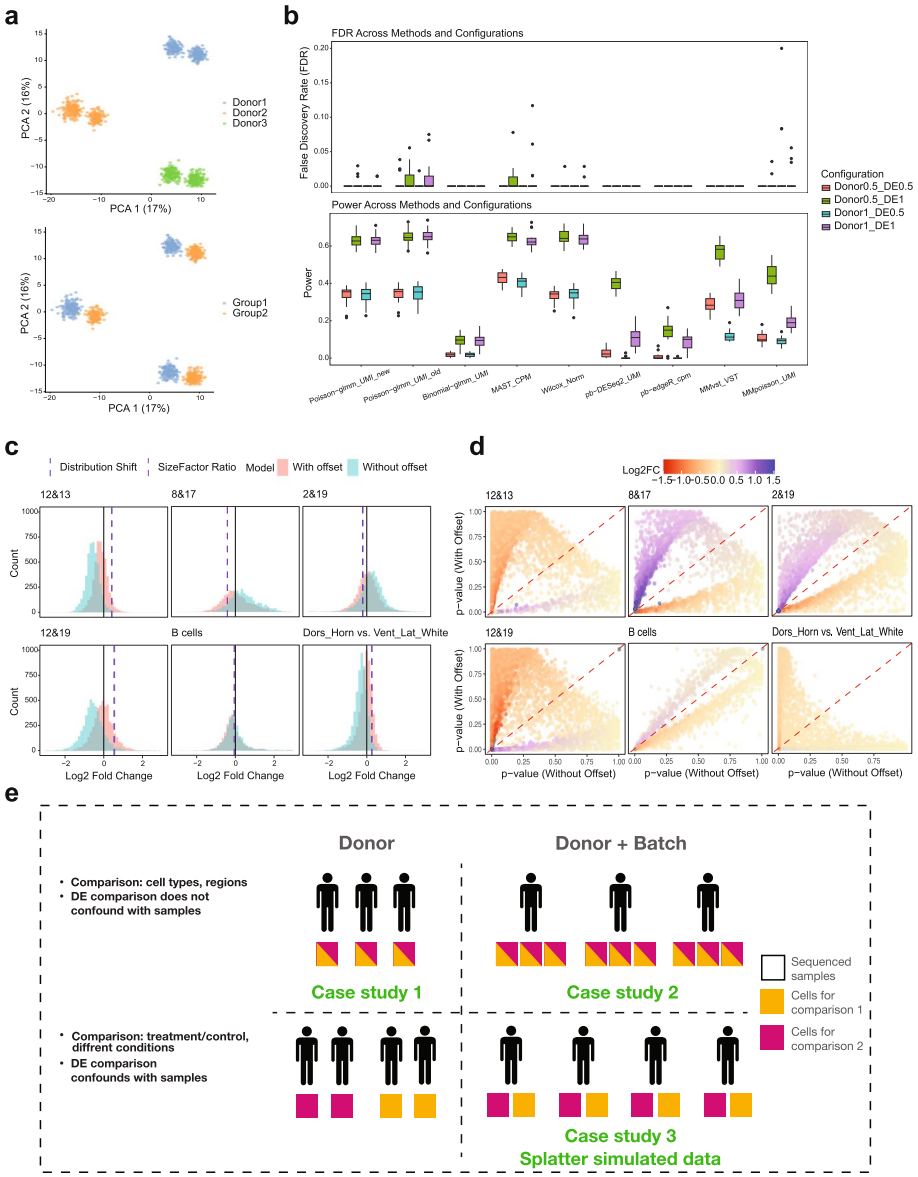

Fig. 7 Simulation results, analysis on Possion-glmm with and without offset, and common experimental design for scRNA-seq DE analysis. **a** PCA of synthetic data simulated by Splatter with donor effect 0.5 and DE effect 1. **b** Top: FDR across all DE methods and configuration. Bottom: power across all DE methods and configurations. **c** Histogram of log2 fold change estimated by Poisson-glmm with and without library-size factor as an offset (pink and green respectively). The blue dashed line indicates the distribution shift after including offset. The purple dashed line indicates the average library-size factors across all cells. Zero is shown in a dark solid line. The dashed line indicates the distribution shift after including offset. The purple dashed line indicates the average library size factors across all cells. Zero is shown in a dark solid line. **d** The *p*-value comparison for Poisson-glmm with and without library-size factor as offset. The red dashed line is $x = y$. Each dot represents the *p*-value of a gene and it is colored by log2 fold change (red for − 1.5, blue for 1.5). Library size factor as an offset. The red dashed line is $x = y$. Each dot represents the *p*-value of a gene and it is colored by log2 fold change (red for − 1.5, blue for 1.5). **e** Common experimental design for scRNA-seq DE analysis. In scRNA-seq, cells are sourced from multiple donors resulting donor effect (left). The experiment may contain other batch effects when multiple samples are from the same donor (right). Some comparisons in DE analysis do not confound with samples when the compared cells appear in the same sample (top), while other comparisons may confound with samples as the compared groups are in different samples (bottom)

We calculated the false discovery rate (FDR) and power for each method and configuration (Fig. 7b). Generally, the FDR is well-controlled for Poisson-glmm (using the new DEG criteria), Binomial-glmm, Wilcox, pseudobulk methods, and MMvst. However, Poisson-glmm (with conventional DEG criteria), MAST, and MMpoisson exhibit higher FDRs when the donor effect is large. In terms of power, Poisson-glmm, MAST, and Wilcox perform well across all configurations. Pb-DESeq2, MMvst, and MMpoisson show moderate power in the (0.5, 1) setting but perform poorly in other settings. Binomial-glmm and pb-edgeR demonstrate consistently lower power across all settings.

Although Splatter provides ground truth for identifying DE genes, enabling straightforward FDR and power calculations, it has notable limitations. First, the library size cannot be customized for different groups (Additional file 1: Fig. S13b), which may prevent it from accurately representing the differences observed in real data. Additionally, we are unable to define more complex batch effect structures, limiting the simulation's ability to capture the intricacies of batch-related variations. Finally, the comparison groups are randomly assigned, which does not reflect scenarios where groups are processed independently, such as distinct cell states.

In terms of computation time (Additional file 1: Fig. S13c), Wilcox and pseudobulk methods are the most efficient. MAST requires more time but remains reasonably fast. Mixed models—including our methods, MMvst, and MMpoisson—demand significantly more processing time, taking over 30 times longer than the Wilcox method, which could become a concern as the number of cells and genes increases.

Poisson-glmm with and without library-size factor

In Muscat's mixed models, library-size factors are included as offsets to perform normalization.

However, as discussed previously, this normalization can distort DE results. To evaluate its impact, we compared Poisson-glmm results with and without the library-size factor across previous case studies. The fold-change estimates shifted by approximately the same magnitude as the average library-size factor ratio between groups (Fig. 7c). Scatter plots for p-value comparisons reveal a significant discrepancy between models, as most p-values diverge sharply from the diagonal line, indicating opposite significance outcomes between the two models (Fig. 7d). From this analysis, we conclude that including library-size normalization in DE analysis can lead to biased results. The decision to use library-size normalization should be carefully considered based on the specific study design and objectives.

Discussion

In this manuscript, we examined existing DE approaches, focusing on pre-processing, input values, test statistics, and fold-change definitions in the context of single-cell DE analysis. Through extensive real-data examples, we demonstrated the limitations of current methods. Specifically, we showed that conventional normalization and preprocessing techniques, which rely heavily on relative RNA abundances, can obscure differentially expressed genes (DEGs) by ignoring or overcorrecting for biological zeros. Additionally, we highlighted how commonly used volcano plots, which depend on

relative RNA abundances, can lead to false discoveries, particularly for low-expression genes, by prioritizing fold changes over absolute expression changes.

A critical issue in single-cell DE analysis is the inappropriate handling of donor effects, which can inflate false discovery rates. Biases also accumulate due to the sequential nature of standard workflows, further compromising accuracy. We emphasize that the decision to apply library-size normalization should be carefully evaluated based on the study design, as its appropriateness depends on the specific research question and data structure.

We outlined four typical DE analysis scenarios (Fig. 7e), each requiring different considerations for handling donor and batch effects. For instance, comparisons between distinct cell types or tissue regions (case studies 1 and 2) are generally less affected by batch effects, whereas DE analyses comparing treatment versus control groups (case study 3) are more susceptible to confounding. Additionally, when multiple batches originate from the same donor (case studies 2 and 3), more complex batch effects must be addressed.

The impact of library-size normalization varies across these scenarios. While normalization can mitigate biases in datasets with large library-size variations, it may also obscure biologically relevant signals or introduce artifacts, particularly when group composition or cell states differ significantly. Thus, careful evaluation of study design, data structure, and confounding factors is essential to determine the most appropriate normalization strategy for accurate and biologically meaningful results.

Conclusion

To address the shortcomings of existing methods, we propose GLIMES, a new statistical paradigm that leverages UMI counts and zero proportions as input while modeling batch effects and within-sample variation using generalized Poisson and Binomial mixed-effects models. By using absolute RNA expression rather than relative abundance, GLIMES enhances sensitivity and robustness, reducing model misspecification and improving biological insight extraction from single-cell data.

The adoption of UMI counts for DE analysis in scRNA-seq has the potential to significantly improve current practices, potentially rendering some existing approaches—such as volcano plots as a diagnostic tool—obsolete. However, this shift also requires careful consideration. Since UMI counts assume strict single-cell measurements, rigorous removal of doublets and triplets is necessary before DE analysis. Furthermore, integrating this new paradigm into widely used analysis pipelines presents a challenge.

A successful transition to this framework will require ongoing efforts to educate and train researchers, update computational tools, and reshape standard practices. By embracing this new approach, the field can move toward more accurate and biologically meaningful single-cell transcriptomic analyses.

Methods and materials

Datasets and pre-processing

In case study 1, a 10X scRNA-seq dataset of post-menopausal fallopian tubes, with 57,182 cells sourced from five donors, covering 29,382 genes was analyzed. We obtained 20 clusters via the HIPPO algorithm. We did not apply a pre-filtering procedure on this dataset, except for built-in filtering steps in each method. We used sctransform to get the VST

data and the integration workflow provided by Seurat (v4) (https://satijalab.org/seurat/reference/integratedata) to obtain the integrated data. In cross-batch integration, only the top 2000 highly expressed genes were retained, which significantly reduced the number of genes for downstream analysis. The dataset had already been fully analyzed and annotated with cell types. We utilized the annotations to examine the effects of normalization/integration on distributions of library sizes across cells.

In case study 2, a 10X scRNA-seq dataset of human spinal cord cells, with 48,644 cells sourced from 9 patients was analyzed. After filtering out ambiguous genes and retaining only genes common across all samples, the dataset included 8092 genes. The cells were annotated for 11 regions and undefined regions. The integrated data was replaced by log2-transformed normalized expression values, which were obtained via computeLibraryFactors and logNormCounts functions in Scater.

In case study 3, the dataset comprised 10X droplet-based scRNA-seq PBCM data from eight Lupus patients obtained before and after 6 h-treatment with IFN-β. After removing undetected and low-expression genes (less than 10 cells expressing more than 1), the dataset consisted of 29,065 cells and 7661 genes. The integrated data was replaced by log2-transformed normalized expression values obtained via computeLibraryFactors and logNormCounts functions in Scater.

All integration or normalization processes were performed on the entire dataset, since cell types are typically unknown during the pre-processing stage.

Variation analysis

To gain a deeper understanding of the donor effect and cell type effect concerning various types of counts, we conducted a variation analysis across multiple group comparisons. To ensure the consistency of our results, we restricted our analysis to genes presented in all datasets. For each gene, we employed linear models (lm (count ~ donor + group)) and used the sum of squares attributed to three components (donor, group, and the residual from the ANOVA table) to obtain the proportion of variation. Logarithm transformation was applied to UMI counts and CPM data to address skewness. The outcomes of this analysis were then presented and compared based on the proportion of variation explained by the first two components across different count types and various pairs. The results of the top 500 genes with the lowest residual variations were exhibited.

Poisson-glmm and Binomial-glmm

By default, we excluded any genes detected in fewer than 5% of cells in the compared groups from differential testing. The GLMMs were implemented with the glmmPQL function of the MASS package. We calculated adjusted p-values with Benjamini–Hochberg correction. Each model fitting was applied to one gene and the two compared groups.

We fit Poisson-glmm on UMI counts. Each count X_{cgk} sampled from cell c, donor k, and gene g, was modeled by

$$X_{cgk}|\lambda_{cgk} \sim Poisson(\lambda_{cgk})$$

$$log\left(\lambda_{cgk}\right) = \mu_g + X_c\beta_g + \epsilon_{gk} + \delta_{gk}.$$

We fit Binomial-glmm on the zero proportions. Each count X_{cgk} was modeled by

$$1[X_{cgk} = 0]|p_{cgk} \sim Bernoulli(p_{cgk})$$

$$log\left(\frac{p_{cgk}}{1-p_{cgk}}\right) = \mu_g + X_c\beta_g + \epsilon_{gk} + \delta_{gk},$$

Where X_c is the indicator for compared groups (e.g., cell types in case study 1, regions in case study 2, control/stimulus in case study 3), and $\epsilon_{gk} \sim N(0, {\sigma_g}^2)$ represents the random effects for donor k, and $\delta_{gk} \sim N\left(0, {\phi_g}^2\right)$ for other batch effects if applicable (e.g., slides in case study 2 and experimental batch in case study 3, but no such term in case study 1). Our goal is to test $H_0 : \beta_g = 0$.

For both methods, we provided "log2 fold change" computed by $\log_2\left(\exp\left(\beta_g\right)\right)$. In Poisson-glmm, this estimate indicates the increment of $\log_2(\lambda_2)$ against $\log_2(\lambda_1)$, which is the conventional log2 fold change. However, this term in Binomial-glmm does not represent the same meaning. It is the difference between logit(p_2) and logit(p_1). The p-value and BH-adjusted p-value are provided.

Benchmarked methods

Pseudo-bulk DESeq2 and pseudo-bulk edgeR are aggregation-based methods used in our comparison. The input counts were summed up for a given gene over all cells in each group and by the donor. The pseudo-bulk data matrix has dimensions GxS, where S denotes the number of interactions of donors and groups. For example, if there are two groups and "a" and "b" donors in each group, then "S" is equal to 2(a + b). We used raw counts as the input for DESeq2, while CPM counts were used for edgeR. The log fold change was converted to log2 fold change in all the comparisons. We implemented these two pseudo-bulk methods following the guided tutorial in the Muscat package; https://www.bioconductor.org/packages/devel/bioc/vignettes/muscat/inst/doc/analysis.html.

For MAST, we fitted a zero-inflated regression model (function zlm) for each gene and applied a likelihood ratio test (function lrTest) to test for between-group differences in gene expression. Besides the labels of groups and the cellular detection rate, we also included donor labels in the covariates. This method was run on log(CPM + 1) counts. We followed the tutorial https://github.com/RGLab/MAST.

Wilcox, a rank sum test, is the default DE method in the FindMarkers function in the Seurat package. For case study 1, we used v4 integration workflow (https://satijalab.org/seurat/archive/v4.3/integration_introduction), which provided integrated counts for 2000 genes. For the other two case studies, we used v5 workflow (https://satijalab.org/seurat/articles/seurat5_integration), generating an integrated dimensional reduction embedding which can be used as input for clustering. In this workflow, the input for DE analysis is simply normalized counts. The log2 fold change in the package is calculated using the formula $\log_2((1 + \text{totalcount1})/\text{n1}) - \log_2((1 + \text{totalcount2})/\text{n2})$ (n1, n2 stand for the number of cells) on the input data, which can be normalized/integrated data by Seurat or other packages. We applied the default filter in FindMarkers to only test genes with a log fold change greater than 0.1. The adjusted p-value provided from the function is based on Bonferroni correction. We followed the guided tutorial found here: https://satijalab.org/seurat/articles/de_vignette.

MMvst and MMpoisson are mixed models implemented in the Muscat package for different states. MMvst fits linear mixed models on variance-stabilizing transformation

data. MMpoisson fits Poisson generalized linear mixed models with an offset equal to the library size factors. In both models, they fit a ~ 1 + group + (1|sample) model for each gene, where "sample" denotes the experimental units (the interaction of donors and compared groups). We followed the tutorial found at: https://www.bioconductor.org/packages/devel/bioc/vignettes/muscat/inst/doc/analysis.html.

GO enrichment analysis

GO over-representation analyses were performed using the enrichGO function in the R package clusterProfiler with default parameters and the functional category for enrichment analysis set to the GO 'Biological Processes' category.

The criteria to determine DEGs

For the benchmarked methods, we adhered to conventional criteria for the identification of DEGs. Specifically, a gene was classified as a DEG if its absolute log2 fold change exceeded a predefined threshold and the adjusted p-value was below a specified cutoff. Typically, DEGs are visually represented in volcano plots. In the first dataset, the log2 fold change threshold was set at log2(1.5), whereas in the second dataset, it was set at 1. The adjusted p-value threshold for both datasets was 0.05.

We proposed new DEG criteria based on the convention plus the gene mean and the difference in mean. If the log2 gene mean in two groups is lower than a certain value (− 2.25 in case study 1) and the log2 mean difference is smaller than a threshold (− 1 in case study 1), the gene would not be considered as a DEG. These can also be used as a filter before any DE analysis to speed up the computation. Both criteria are adjustable, depending on the dataset's performance and characteristics. An examination of heatmaps and mean difference against mean plot in advanced can be helpful to determine the thresholds when analyzing a new dataset (Additional file 1: Fig. S10b, c).

False discovery rate and power

The permutation analysis was conducted within a null dataset focusing on a group of interest. We specifically conducted the analysis on three datasets: B cells (control) from case study 3, CD4 + T cells (group 2), and CD8 + T cells (group 13) from case study 1. Each underwent random assignment to either the control or stimulus group. Subsequently, *p*-values for each gene were computed employing various methods, with the gene set confined to those input into the Poisson-glmm model. To mitigate potential gene filtering, the threshold for the Wilcox method was relaxed. This process was iterated 20 times, and on each iteration, the proportion of *p*-values below 0.05 was calculated along with the corresponding false discovery of differentially expressed genes method was relaxed. This process was iterated 20 times, and on each iteration, the proportion of *p*-values below 0.05 was calculated along with the corresponding false discovery of differentially expressed genes.

We used a Splatter simulator to generate synthetic data for power assessment. Each simulated data contains 1000 genes, and 900 cells evenly distributed to 3 donors. The cells are assigned to two cell types with a 50% probability, and the genes are DE on the second cell type with a 10% probability. The donor effect and DE effect parameters

control the location and scale factor for the effects. The average power and FDR were computed based on 20 replications.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03525-6.

Additional file 1. Supplementary Figures: Figs. S1-S13.

Additional file 2. Review history.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 2.

Authors' contributions
M.C. conceived and led this work. C.W. and M.C. developed the methods and performed the analyses. C.W. implemented the software. X.Z. participated in critically revising the draft. C.W. and M.C. wrote the paper with feedback from X.Z. All authors read and approved the final manuscript.

Funding
The work was supported by National Institutes of Health grant R01 GM126553, R01 HG011883, and HG012927, and additional grant no. NSF 2016307 and Sloan Research Fellowship to M.C.

Data availability
All scRNA-seq datasets used in this study are publicly available. Processed and de-identified human single-cell RNA sequencing data scRNA-seq dataset of post-menopausal fallopian tubes has been deposited at Cellxgene under the following URL: https://cellxgene.cziscience.com/collections/d36ca85c-3e8b-444c-ba3e-a645040c6185. The raw human spinal cord scRNA-seq dataset used in case study 2 is available in the following URL:https://als-st.nygenome.org/. The droplet scRNA-seq data used in case study 3 is deposited under the Gene Expression Omnibus under the accession number GSE96583. The dataset is also available in R through the Bioconductor ExperimentHub package. We provide an R package, GLIMES, implementing Poisson-glmm and Binomial-glmm methods for DE analysis discussed in this study. The GLIMES package is available from GitHub (https://github.com/C-HW/GLIMES) under the BSD 3-Clause license. In addition, the R source code to reproduce all data analysis in the study is available on Zenodo at DOI: 10.5281/zenodo.14279028 [45].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 December 2023 Accepted: 5 March 2025
Published online: 17 March 2025

References

1. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42:8845–60.
2. Greenwald WW, et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun. 2019;10:2078.
3. Grubman A, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci. 2019;22:2087–97.
4. Lawlor N, et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. Genome Res. 2017;27:208–22.
5. Squair JW, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12:1–15.
6. Das S, Rai A, Merchant ML, Cave MC, Rai SN. A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. Genes. 2021;12:1947.
7. Das S, Rai A, Rai SN. Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges. Entropy. 2022;24:995.

8. Lengyel E, et al. A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing. Cell Rep. 2022;41(12):111838.
9. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10:380.
10. Yang Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell rep. 2021;36(4):109442.
11. Kim TH, Zhou X, Chen M. Demystifying, "drop-outs" in single-cell UMI data. Genome Biol. 2020;21:196.
12. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nat Commun. 2020;11:1169.
13. Svensson V. Droplet scRNA-seq is not zero-inflated. Nat Biotechnol. 2020;38:147–50.
14. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19:1–10.
15. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9:997.
16. Tracy S, Yuan G-C, Dries R. RESCUE: imputing dropout events in single-cell RNA-sequencing data. BMC Bioinformatics. 2019;20:1–11.
17. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol. 2018;19:196. https://doi.org/10.1186/s13059-018-1575-1.
18. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:1–10.
19. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:1–13.
20. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. 2015;16:1–9.
21. Zyprych-Walczak J, et al. The impact of normalization methods on RNA-Seq data analysis. BioMed res int. 2015;2015:621690.
22. Dillies M-A, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14:671–83.
23. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:1–9.
24. Lytal N, Ran D, An L. Normalization methods on single-cell RNA-seq data: an empirical survey. Front Genet. 2020;11:41.
25. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
26. Korsunsky I, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
27. Chen M, Zhou X. Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. Sci Rep. 2017;7:13587. https://doi.org/10.1038/s41598-017-13665-w.
28. Chen M, et al. Alignment of single-cell RNA-seq samples without overcorrection using kernel density matching. Genome Res. 2021;31:698–712.
29. Hu J, Chen M, Zhou X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab1147.
30. Schmid R, et al. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. BMC Genomics. 2010;11:1–17.
31. Tran HTN, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:1–32.
32. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:1–15.
33. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021;22:1–20.
34. Argelaguet R, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:1–17.
35. Tian L, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16:479–87.
36. Stuart, T. et al. Comprehensive integration of single-cell data. Cell. 2019:177:1888–1902. e1821.
37. Kim TH, Zhou X, Chen M. Demystifying, "drop-outs" in single-cell UMI data. Genome Biol. 2020;21:196. https://doi.org/10.1186/s13059-020-02096-y.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
39. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformat. 2010:26:139–140.
40. Nguyen HC, Baik B, Yoon S, Park T, Nam D. Benchmarking integration of single-cell differential expression. Nat Commun. 2023;14:1570.
41. Clayton DG. Generalized linear mixed models. Markov chain Monte Carlo in practice. 1996;1:275–302.
42. Crowell HL, et al. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11:6077.
43. Verdon DJ, Mulazzani M, Jenkins MR. Cellular and molecular mechanisms of CD8+ T cell differentiation, dysfunction and exhaustion. Int J Mol Sci. 2020;21:7357.
44. Maniatis S, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science. 2019;364:89–93.
45. Wu, C-H. Code and data for: A new statistical paradigm for single-cell differential expression using generalized linear mixed models. Zenodo; 2024. https://doi.org/10.5281/zenodo.14279028.

Publisher's Note