Methods in Bioinformatics (INFO-F439) Project

Review of "muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data" [1]

Léopold Guyot

May 30, 2025

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionised transcriptomics by enabling gene expression profiling at the resolution of individual cells. Unlike bulk RNA-seq, which averages expression signals across thousands or millions of cells, scRNA-seq reveals cell-to-cell heterogeneity and uncovers distinct cellular subpopulations within complex tissues. This allows to provide crucial insights in biological mechanisms.

In many biological and clinical contexts, it is essential to account for this cellular diversity. For instance, in immunology or oncology, the presence or absence of specific cell subtypes can have significant functional and diagnostic implications. A clear example is acute myeloid leukemia (AML), where the tumor comprises a hierarchy of subpopulations, including leukemia stem cells (LSCs) that are often resistant to treatment and responsible for relapse [2]. Identifying and characterising these rare subpopulations is critical for prognosis and therapeutic targeting. scRNA-seq offers the granularity needed to capture such variation, making it indispensable for studies where subtle but biologically meaningful differences may be masked by population averages.

However, the analytical challenges of scRNA-seq data are considerable. The data are high-dimensional, sparse, and noisy due to dropout events and technical variability. Modelling gene expression at the single-cell level must account for these factors while also respecting the hierarchical structure of the data—cells nested within patients or experimental groups. Furthermore, the design of appropriate statistical models must balance sensitivity, specificity, and computational scalability.

To address these challenges, a variety of statistical models have been developed. The `scDD` method [3] models gene expression as a mixture of distributions, enabling the detection of differential expression patterns beyond mean shifts, such as changes in modality or proportion. Mixed-effect models incorporate both fixed effects (e.g., cell subpopulation) and random effects (e.g., patient) to account for intra-patient correlations and nested data structures.

The reviewed article [1] proposes a pseudobulking approach, where cells are aggregated by the combination of cell subpopulation and patient. This results in a data structure resembling bulk RNA-seq, reducing noise. Once aggregated, well-established bulk RNA-seq methods such as `limma` [4], `DESeq2` [5], and `edgeR` [6] can be applied for differential expression analysis. This strategy leverages the robustness of bulk models while preserving important biological structure related to both cell type and biological replicates.

In this study, we aim to reproduce and evaluate the results presented in the reviewed paper, which investigated differential expression analysis strategies for scRNA-seq data. In that work, the authors used two real scRNA-seq datasets as the basis for simulating artificial datasets with known, controlled variations. These simulated datasets allowed for rigorous benchmarking of modelling approaches.

The original study compared different strategies: Single-cell methods including `scDD`, and mixed-effect models and aggregation-based methods, where cells were grouped by patient and subpopulation to produce pseudobulk profiles, then analysed using established bulk RNA-seq tools such as `limma`, `DESeq2`, and `edgeR`.

In this review, we reproduce the benchmark of the differential expression methods using simulated and real scRNA-seq datasets, compare our findings with existing studies, and explore the literature about pseudobulking.

2 Methods

This project was developed using the R programming language [7] and leveraged packages from the Bioconductor repository [8].

2.1 Datasets

To evaluate the performance of pseudobulking methods, the datasets used had to meet specific criteria. In particular, each dataset needed to include multiple patients and several identifiable cell subpopulations. These datasets were then used to simulate new data that mimic the original distribution patterns.

2.1.1 Kang et al. 2018

The dataset from Kang et al. [9] consists of single-cell RNA-sequencing profiles of peripheral blood mononuclear cells (PBMCs) collected from 8 human donors. It includes both unstimulated and interferon-β–stimulated cells.

The original droplet-based scRNA-seq data is publicly available via the Gene Expression Omnibus (GEO) under the accession `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96583`. It is also distributed in the `SingleCellExperiment` format [10] through the `ExperimentHub` package [11], using the accession code: `EH2259`.

2.1.2 Mouse LPS

The second dataset investigates transcriptomic changes in brain tissue from mice subjected to peripheral lipopolysaccharide (LPS) treatment. It includes samples from 4 vehicle-treated and 4 LPS-treated mice.

This dataset was originally published alongside the reviewed paper [1]. It is accessible via ArrayExpress (accession: E-MTAB-8192) and can also be obtained from the `ExperimentHub` package [11] with the accession code: `EH3297`.

2.1.3 Guo et al. 2018

Guo et al. [12] performed single-cell RNA sequencing on approximately 6,500 testicular cells from 3 healthy adult males, using the 10x Genomics Chromium platform.

This dataset was not included in the original benchmarking paper. It was added here as an independent dataset to assess the generalisability and robustness of the results obtained with other datasets. The aim is to evaluate whether similar conclusions can be drawn using a dataset with different biological context and origin.

The dataset is available through the `CTdata` R package [13], and can be accessed using the `testis_sce` function.

2.2 Simulation Framework

To systematically evaluate differential expression analysis methods, we developed a data-driven simulation framework based on a multi-sample, multi-subpopulation scRNA-seq reference dataset. This framework allows for modulation of key parameters, including the number of cells per subpopulation and sample, as well as the type and magnitude of differential expression patterns. Simulations are based on the negative binomial distribution, which is widely used to model scRNA-seq data. Subpopulation- and sample-specific parameters (means, dispersions, and library sizes) are estimated directly from the reference dataset. Simulated data is then generated by sampling from these empirical distributions, thereby capturing the structure and variability observed in real scRNA-seq data.

To introduce biologically meaningful variation, genes can be designated as subpopulation specific (differential across cell types), condition specific (differential expression between treatment conditions), or non-differential (uniform expression across all variables).

The framework supports a diverse set of expression patterns based on the classification by Korthauer et al. [14]:

- **DE (Differential Expression)**: A change in the mean expression level of a gene between conditions, while the overall distribution shape remains unimodal and similar.
- **DP (Differential Proportion)**: A shift in the proportions of cells in low and high expression states between conditions, without a change in the expression values themselves.

- **DM (Differential Modality)**: A change in the modality of gene expression, such as shifting from an unimodal to a bimodal distribution, indicating that the underlying structure of expression changes.

- **DB (Differential Both)**: A combination of DP and DM, where both the proportions and the modality of expression states differ between conditions.

- **EE (Equally Expressed)**: No differences in expression or distribution between conditions or subpopulations; expression is consistent across all samples.

- **EP (Equal Proportion)**: Genes exhibit bimodal expression, but the proportions of cells in each expression state remain the same across conditions.

This classification enables simulation of complex and realistic gene expression scenarios, making the framework well-suited for benchmarking modelling methods across a variety of biologically plausible settings.

2.3 Pseudobulk Approach

To reduce the complexity of the data to be modeled, an aggregation approach can be applied. This involves summarising the counts of cells belonging to a specific patient and a specific cell group (Fig. 1). Aggregation can be performed using different summary statistics, such as the sum or the mean. By applying this technique, we simplify the data and reduce the noise while preserving meaningful biological variation, specifically the variability between biological replicates and the variability between cell (sub)populations.

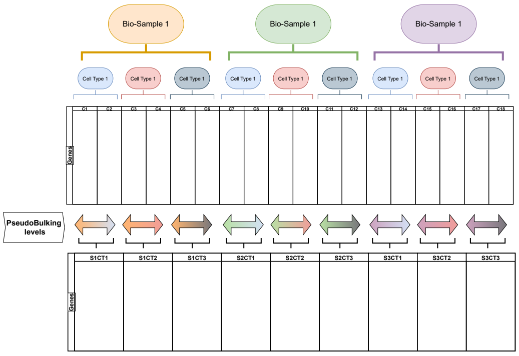


Figure 1: Scheme of the pseudobulk approach. The columns in the first table represent the individual cells. Several cells belonging to the same bio-sample and cell type are aggregated to give the second table in which each column represent the value for a particular bio-sample and cell type.

2.4 Workflow

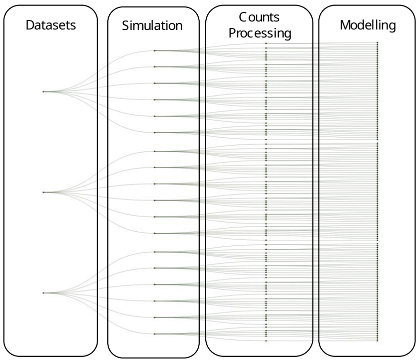


Figure 2: Scheme of the workflow generated with `targets`. The first layer is the datasets retrieval and preprocessing. The second layer is the simulation of new datasets using different parameters. The third layer is the use of different processing method and aggregation techniques on counts. The fourth layer is the modelisation of the data.

Given the complexity and number of steps involved in this workflow, we used the `targets` R package [15] to automate the entire process. Each step is applied to the results produced by all preceding steps, ensuring a consistent and reproducible pipeline. In total, the workflow comprises 291 steps and results in the generation of 144 models (Fig. 2).

2.4.1 Preprocessing

Each dataset underwent a standardised preprocessing procedure. To ensure that only the intended variation was present, we retained a single experimental condition per dataset. Specifically, for the Kang et al. dataset, only the control cells were kept; for the mouse LPS dataset, only vehicle-treated cells were retained; and the testis dataset included only one experimental condition.

We filtered genes to retain those with more than one count in at least 10 cells. Similarly, we filtered cells to keep only those expressing at least 100 genes. Additionally, we retained only those clusters that consisted of at least 100 cells.

2.4.2 Simulation

The datasets were used to generate artificial data according to the simulation framework described earlier. We used functions provided by the `muscat` R package [16] to facilitate the simulation process.

We simulated various scenarios using different parameter combinations. In the first set of simulations, each sample-subpopulation combination contained 400 cells. We considered four cases:

1. 10% of genes were altered in both proportion and modality (DB).
2. 10% of genes were altered in mean expression (DE).
3. 10% of genes were altered in modality (DM).
4. 10% of genes were altered in the proportions of low and high expression-state components (DP).

Additionally, we conducted a series of simulations in which 10% of the genes were altered in mean expression, varying the number of cells per sample-subpopulation combination (20, 100, and 400 cells).

2.4.3 Processing Counts

Each simulated dataset was processed using multiple count normalisation methods. These included log-transformation, residuals, counts-per-million (CPM), and unprocessed raw counts. The transformations were performed using the `calculateCPM`, `normalizeCounts`, and `computeLibraryFactors` functions from the `scuttle` R package [17], as well as the `vst` function from the `sctransform` R package [18].

2.4.4 Pseudobulk

We applied three different pseudobulk aggregation strategies: mean aggregation, sum aggregation, and no aggregation.

2.4.5 Modelisation

We employed a variety of models for differential analysis, using the formula `~ patient + cell_subpopulation`. For non-aggregated data, we applied mixed models and the `scDD` method. The mixed models were implemented using the `lmer` function from the `lme4` package [19], treating the patient effect as a random effect. The `scDD` method was implemented using the `scDD` package [3].

For aggregated (pseudobulk) data, we used established bulk RNA-seq modelling approaches, including `limma` [4], `DESeq2` [5], and `edgeR` [6]. The `limma` package was applied with either the `voom` or the `trend` method.

2.4.6 Process Results

The results were processed using the `iCOBRA` package [20]. Differential expression tables were used to generate FDR–TPR curves at various adjusted p-value thresholds (0.01, 0.05, 0.1, and 0.2) comparing the results between locally adjusted p-value (at the cluster level) and globally adjusted p-value. Additionally, UpSet plots were generated to visualise the intersections of detected gene sets across methods and ground truth. The runtime for each method, comprising both the aggregation and modelling steps, was recorded using the `targets` package [15].

2.5 LPS Downstream Analysis

In the reviewed article, the researchers also applied downstream analysis to pseudobulked LPS dataset. To reproduce the results we applied the same pseudobulk methods as described before with mean and sum aggegation on the LPS mouse data. Then MDS and UMAP plots were drawn.

3 Results

3.1 Effect of the Variation Type Introduced on Model Performance

3.1.1 Kang and LPS datasets

To assess the impact of different expression variation types on method performance, we here show the simulation results based on the Kang dataset (Fig. 4a). Note that the results obtained from the mouse LPS show equivalent results. These simulations introduced distinct categories of gene expression variation. The results are compared based on the true positive rate (TPR) and false discovery rate (FDR) across tested differential state methods.

DB (Differential Both) simulations showed consistently poor TPR across all tested methods. This was expected, as DB genes combine changes in both proportions and modality, presenting a particularly challenging scenario for current models. The inability to detect DB genes highlights the limitations of existing methods in capturing simultaneous changes in expression state prevalence and distribution structure. Even the scDD method which works by comparing the distribution of expression did not successfully model these simulated data.

In contrast, **DE (Differential Expression)** simulations demonstrated that pseudobulk methods outperformed both mixed models (MM) and scDD. Among the pseudobulk approaches, performance was generally strong, although *limma-trend* applied to VST residuals showed a reduced TPR compared to other pseudobulk variants. scDD exhibited very low TPR. Also locally adjusted p-values improved detection rates, this came at the cost of a notable increase in FDR.

DM (Differential Modality) simulation results were similar to those for DE, with aggregation-based methods again showing superior performance relative to MM and scDD. Notably, scDD, while still exhibiting low TPR, showed improved FDR control in this scenario compared to its performance on DE simulations.

DP (Differential Proportion) was the most challenging variation type among DE, DM, and DP. All methods exhibited a small drop in TPR compared to DE and DM. In particular, *limma-trend* using VST residuals failed to model the data effectively, with especially low TPR.

Overall, these results highlight the clear advantage of pseudobulk-based approaches across various expression variation types, particularly DE and DM. Mixed models and scDD lagged behind, especially in more complex scenarios.

3.1.2 Impact of Count Processing Strategies

Across all datasets and variation types, the choice of count processing strategy had minimal impact on model performance for most pseudobulk methods. However, one notable exception was observed for the *limma-trend* method when applied to VST residuals. This count processing method consistently showed reduced TPR and, in some cases, increased FDR, particularly in the DE and DM scenarios.

3.1.3 Validation on an Independent Dataset: testis dataset

To assess the robustness of our observations, we additionally tested all methods with a newly added *testis* dataset (Fig. 5a). Globally, the results obtained on this dataset were consistent with those from the Kang and LPS datasets, reinforcing the validity of our conclusions. For most variation types, the relative performance of the methods remained similar. However, a notable deviation was observed for **DM (Differential Modality)** simulations: the *limma-trend* method applied to VST residuals displayed an elevated FDR, reaching up to 20% false positives. This suggests that this approach may be unreliable for detecting modality-driven changes in certain biological contexts. Despite this exception, the overall agreement in trends between datasets increases our confidence in the comparative evaluation and the general applicability of the simulation framework.

3.2 Effect of sample size on model performance

We assessed the impact of the number of cells per patient and subpopulation on model performance. The results were consistent across all datasets tested: Kang (Fig. 4b), LPS (data not shown), and testis (Fig. 5b). This is confirming the generalisability of the observed trends.

As expected, model performance declined when the number of cells per patient/subpopulation decreased. This effect was evident in both true positive rate (TPR) and false discovery rate (FDR), with fewer cells leading to reduced sensitivity and less stable error control. This is the case for bothe aggregation and non-aggregation methods.

3.3 Runtime Analysis

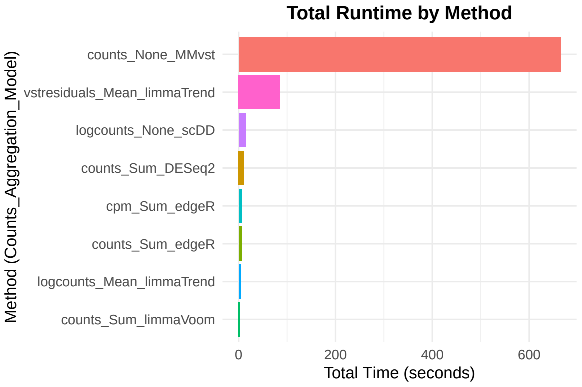


Figure 3: Runtime analysis for the Kang dataset of the counts processing + aggregation and the modelisation. Time in seconds.

To complement our performance evaluation, we also compared the computational efficiency of the tested methods by measuring their total runtime on the Kang dataset (Fig. 3). The results clearly show that aggregation-based pseudobulk methods are consistently the most efficient. This highlight the scalability and practicality of pseudobulk-based workflows, especially for large single-cell datasets.

While most count processing strategies had little impact on computational time, an exception was observed for the *vstresiduals* used in combination with *limma-trend*, which resulted in increased runtimes and poorer performance compared to other aggregation schemes.

Notably, methods based on mixed models (*MMvst*) and scDD were the slowest among all those tested, often requiring several times more computational time than pseudobulk approaches. If we take in consideration our earlier findings, these methods did not yield superior performance. Thus they may not justify their higher computational cost in most applications.

Taken together, these results support the recommendation of pseudobulk approaches not only for their accuracy but also for their computational efficiency, making them well-suited for both small and large scale differential state analyses in scRNA-seq studies.

3.4 LPS Downstream Analysis

To further validate the reproducibility of the findings from the reviewed study, we performed downstream analysis on the LPS dataset using pseudobulked data. Specifically, we reproduced the MDS and UMAP plots as described in the original article. Here, we focus on the MDS results (Fig. 6).

While the original article presented MDS plots based on sum-aggregated counts, we extended this by generating an MDS plot using mean-aggregated (averaged) expression data. We can make the same observations on both pseudobulk methods. The MDS plots revealed clear clustering by cell type, indicating strong subpopulation identity. However, similar to the observations in the review paper, the separation between treatment conditions (LPS vs. Vehicule) was not pronounced. This suggests that while pseudobulking preserves cell-type-specific expression structure well, it may not always capture subtle treatment effects at the global expression level.

4 Discussion

Our benchmark results using the Kang and LPS datasets are in strong agreement with those reported in the reviewed paper. Across both datasets (Kang and LPS), we observed similar trends that reinforce the conclusions from the reviewed article. Notably, pseudobulk approaches consistently outperformed both mixed models (MM) and scDD in terms of true positive rate (TPR) and false discovery rate (FDR). This superior performance was especially clear in the detection of genes with differential expression (DE) and differential modality (DM), where pseudobulk methods yielded both high true positive rates (TPRs) and efficient runtime.

As previously reported, we also observed that scDD exhibited poor performance across most simulated variation types, and that *limma-trend* applied to VST residuals was the least performant among the pseudobulk methods tested. Additionally, our results confirmed the difficulty in detecting DB (differential both) genes, consistent with the original study. These genes, which combine modality and proportion shifts, remain particularly challenging to identify for all evaluated methods. Furthermore, we reproduced the difference between globally and locally adjusted p-values: while local adjustments led to higher sensitivity, they also resulted in inflated FDR.

Importantly, both our analysis and the reviewed article highlighted the critical impact of the number of cells per patient and per subpopulation on model performance. As the number of cells increased, all methods improved in accuracy, but the rate of improvement varied.

To further test the generalisability of our conclusions, we evaluated all methods on an additional dataset derived from testis tissue. Remarkably, the results obtained from this dataset mirrored those from the Kang and LPS datasets, and by extent as those reported in the reviewed article. This consistency across datasets reinforces the robustness of our findings and supports the broader applicability of the conclusions drawn regarding method performance, computational efficiency, and sensitivity to sample size.

Nevertheless, there are important differences between our setup and the one used in the reviewed article that should be acknowledged. For example, we did not replicate each method-simulation pair multiple times as they did, primarily due to computational constraints (our full benchmark already required over eight hours of runtime). Additionally, our benchmark did not cover all models or preprocessing strategies included in their study. In particular, we did not test alternative count processing methods applied to MM or scDD models, nor did we evaluate methods based on Anderson–Darling (AD) tests or MAST, which were used in the original paper.

We also looked at the literature on the pseudobulking subject, we identified an other study by Zimmerman et al. [21] that raised concerns regarding the use of pseudobulk methods for single-cell differential analysis. Their simulations, which accounted for intra-individual correlation structures observed in real data, suggested that while pseudobulk approaches provided good type 1 error control, they were overly conservative—leading to inflated type 2 error rates. This was attributed to the loss of information due to aggregation and imbalanced cell numbers across individuals, which disproportionately impacted statistical power. They recommended generalised linear mixed models with patient random effects as the most effective solution for balancing type 1 and type 2 error control.

However, these findings were critically re-evaluated by Murphy and Skene [22], who highlighted important methodological issues in the original analysis. Notably, they pointed out that Zimmerman et al. did not control for randomness in their simulations, making direct comparisons between methods unreliable. Indeed, different methods were not using the same simulated data. Moreover, they emphasised the limitations of evaluating methods solely based on type 1 or type 2 error rates. By introducing other metrics like the Matthews Correlation Coefficient and ROC curves, Murphy and Skene demonstrated that pseudobulk approaches, particularly aggregation by averaging, outperformed all other tested methods, including GLMMs. Even in the presence of imbalanced datasets, mean-based pseudobulking maintained superior performance, challenging the earlier claims of excessive conservatism. These findings reaffirm our conclusions, strengthening the case for pseudobulk methods as a robust and computationally efficient solution for DS analysis in scRNA-seq data.

Code Availability

All scripts used for this report are available on `https://github.com/leopoldguyot/muscat.Replication`.

References

[1] Helena L Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature communications*, 11(1):6077, 2020.

[2] Peter van Galen, Volker Hovestadt, Marc H Wadsworth II, Travis K Hughes, Gabriel K Griffin, Sofia Battaglia, Julia A Verga, Jason Stephansky, Timothy J Pastika, Jennifer Lombardi Story, et al. Single-cell rna-seq reveals aml hierarchies relevant to disease progression and immunity. *Cell*, 176(6):1265–1281, 2019.

[3] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Li Yuan, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1077-y`.

[4] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi: 10.1093/nar/gkv007.

[5] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.

[6] Yunshun Chen, Lizhong Chen, Aaron T L Lun, Pedro Baldoni, and Gordon K Smyth. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53(2):gkaf018, 2025. doi: 10.1093/nar/gkaf018.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

[8] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. URL `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`.

[9] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, January 2018. ISSN 1546-1696. doi: 10.1038/nbt.4042. URL `https://www.nature.com/articles/nbt.4042`. Publisher: Nature Publishing Group.

[10] Robert Amezquita, Aaron Lun, Etienne Becht, Vince Carey, Lindsay Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Herve Pages, Mike Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie Hicks. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17:137–145, 2020. URL `https://www.nature.com/articles/s41592-019-0654-x`.

[11] Martin Morgan and Lori Shepherd. *ExperimentHub: Client to access ExperimentHub resources*, 2025. URL `https://bioconductor.org/packages/ExperimentHub`. R package version 2.16.0.

[12] Jingtao Guo, Edward J. Grow, Hana Mlcochova, Geoffrey J. Maher, Cecilia Lindskog, Xichen Nie, Yixuan Guo, Yodai Takei, Jina Yun, Long Cai, Robin Kim, Douglas T. Carrell, Anne Goriely, James M. Hotaling, and Bradley R. Cairns. The adult human testis transcriptional cell atlas. *Cell Research*, 28(12):1141–1157, December 2018. ISSN 1748-7838. doi: 10.1038/s41422-018-0099-2. URL `https://www.nature.com/articles/s41422-018-0099-2`. Publisher: Nature Publishing Group.

[13] Axelle Loriot, Julie Devis, and Laurent Gatto. *CTdata: Data companion to CTexploreR*, 2025. URL `https://bioconductor.org/packages/CTdata`. R package version 1.8.0.

[14] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17:1–15, 2016.

[15] William Michael Landau. The targets r package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, 6(57):2959, 2021. URL `https://doi.org/10.21105/joss.02959`.

[16] Helena L. Crowell, Pierre-Luc Germain, Charlotte Soneson, Anthony Sonrel, Jeroen Gilis, Davide Risso, Lieven Clement, and Mark D. Robinson. *muscat: Multi-sample multi-group scRNA-seq data analysis tools*, 2025. URL `https://bioconductor.org/packages/muscat`. R package version 1.22.0.

[17] Davis J. McCarthy, Kieran R. Campbell, Aaron T. L. Lun, and Quin F. Willis. Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *Bioinformatics*, 33:1179–1186, 2017. doi: 10.1093/bioinformatics/btw777.

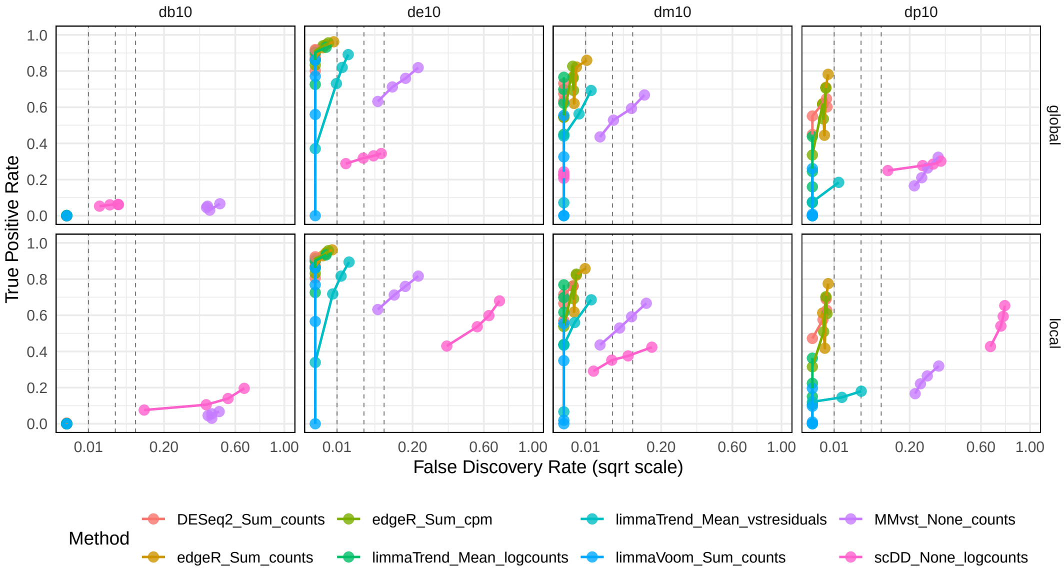
[18] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biology*, 23:20, 2022. doi: 10.1186/s13059-021-02584-9. URL `https://doi.org/10.1186/s13059-021-02584-9`.

[19] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

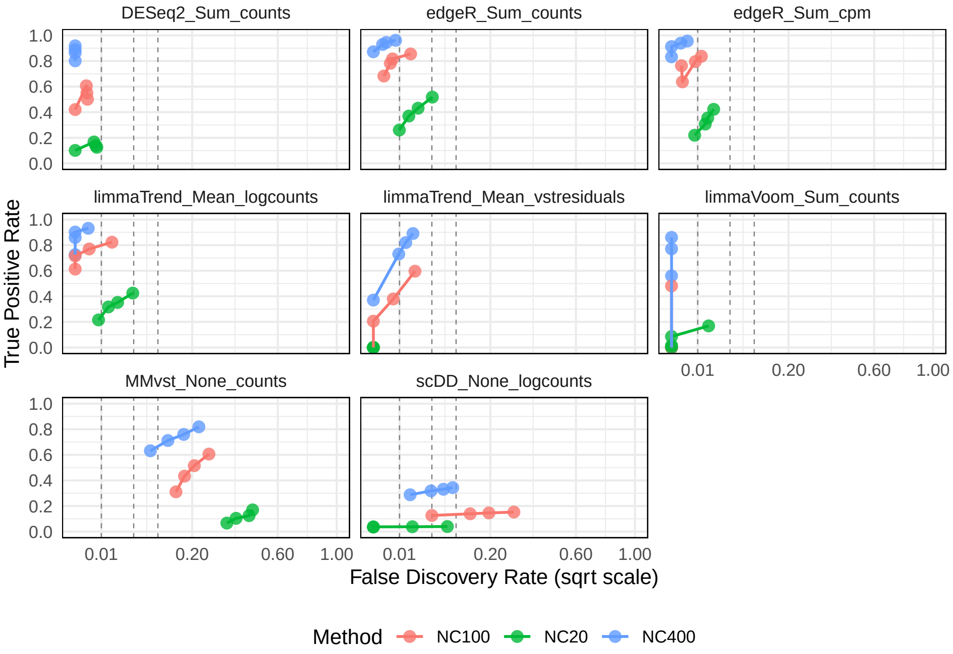
[20] Charlotte Soneson and Mark D Robinson. icobra: open, reproducible, standardized and live method benchmarking. *Nature Methods*, 13(4):283, 2016. URL `http://www.nature.com/nmeth/journal/v13/n4/full/nmeth.3805.html`.

[21] Kip D Zimmerman, Mark A Espeland, and Carl D Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature communications*, 12(1):738, 2021.

[22] Alan E Murphy and Nathan G Skene. A balanced measure shows superior performance of pseudobulk methods over mixed models and pseudoreplication approaches in single-cell rna-sequencing analysis. *bioRxiv*, pages 2022–02, 2022.

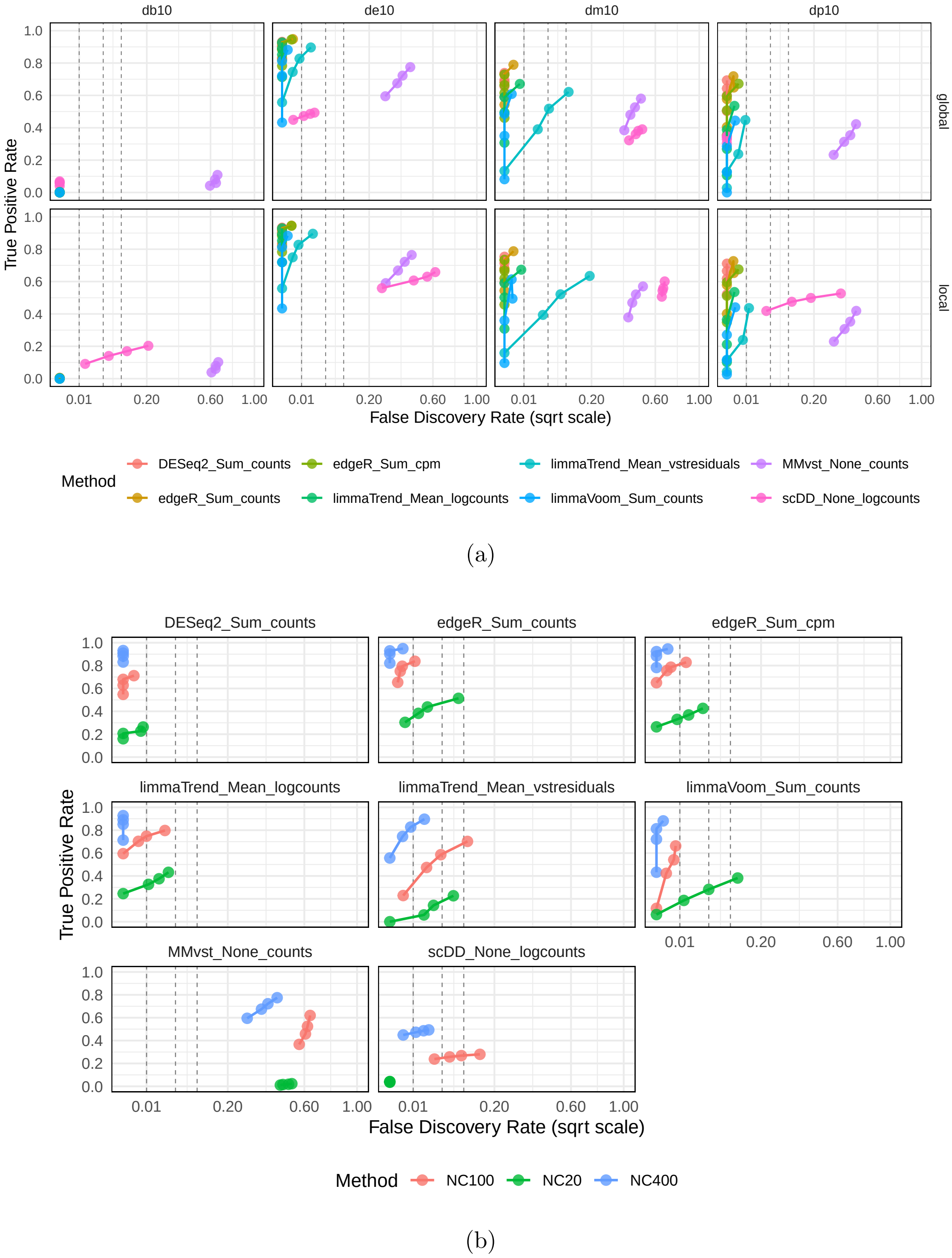


(a)



(b)

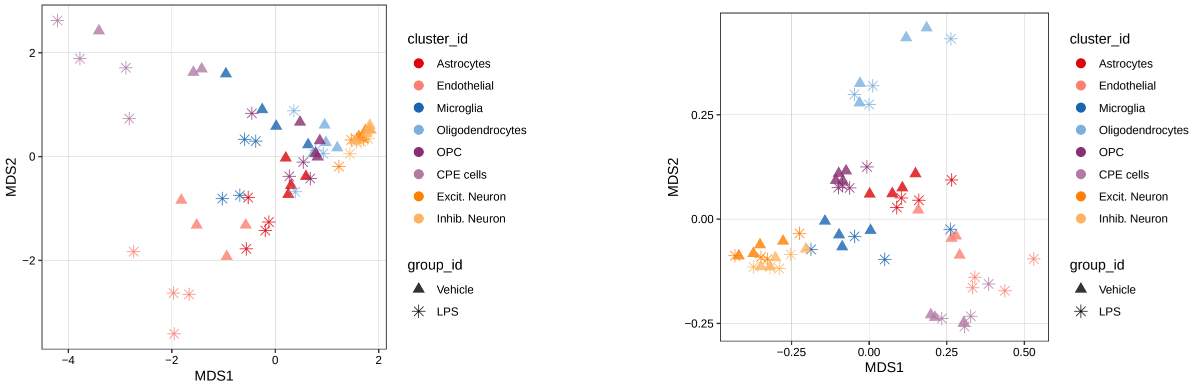
Figure 4: FDR vs TPR of the methods applied on the **Kang dataset** for different adjusted pvalue thresholds (0.01, 0.05, 0.1 and 0.2). Dashed line represent FDR values of 0.01, 0.05 and 0.1 **[a]** Comparison for different types of introduced variation. Each column of plots represent a variation type. db10 = 10% of genes where both the proportions and the modality of expression differ, de10 = 10% of genes altered in their mean expression between condition, dm = 10% of genes with a change in the modality of their expression, dp10 = 10% of genes with a shift in the proportion of cells in low and high expression states between conditions. The first row of plots use globally adjusted pvalues, the second row use locally adjusted pvalues (on each cluster). Colors represent the different methods. **[b]** Comparison of the performance of the methods with different number of cells by patient and clusters. Performance computed on de10 simulated data and with globally adjusted pvalues. NC20 = 20 cells (green), NC100 = 100 cells (red) and NC400 = 400 cells (blue). Each plot stand for a different method.



(a)

(b)

Figure 5: FDR vs TPR of the methods applied on the **LPS mouse dataset** for different adjusted pvalue thresholds (0.01, 0.05, 0.1 and 0.2). Dashed line represent FDR values of 0.01, 0.05 and 0.1 **[a]** Comparison for different types of introduced variation. Each column of plots represent a variation type. db10 = 10% of genes where both the proportions and the modality of expression differ, de10 = 10% of genes altered in their mean expression between condition, dm = 10% of genes with a change in the modality of their expression, dp10 = 10% of genes with a shift in the proportion of cells in low and high expression states between conditions. The first row of plots use globally adjusted pvalues, the second row use locally adjusted pvalues (on each cluster). Colors represent the different methods. **[b]** Comparison of the performance of the methods with different number of cells by patient and clusters. Performance computed on de10 simulated data and with globally adjusted pvalues. NC20 = 20 cells (green), NC100 = 100 cells (red) and NC400 = 400 cells (blue). Each plot stand for a different method.



(a) Sum Aggregation

(b) Mean Aggregation

Figure 6: Multidimensional Scaling (MDS) plot on **pseudobulked LPS mouse dataset**. Each point represents one cluster-patient value, colors represent clusters (cell subpopulations) and shapes represents group ID (treatment).