Internship report

Alternative PCA algorithms analysis with missing values

Léopold Guyot

May 28, 2025

Abstract

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized 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.

FIND REAL EXAMPLE OF CELL HETERO IMPORTANCE 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. 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. Modeling 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 (Korthauer et al., 2016) 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 (Crowell et al., 2025) 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` (Ritchie et al., 2015), `DESeq2` (Love et al., 2014), and `edgeR` (Chen et al., 2025) 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 modeling 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`.

PREVIEW OF RESULTS

2 Methods

This project was developed using the R programming language (R Core Team, 2025) and leveraged packages from the Bioconductor repository (Huber et al., 2015).

2.1 Datasets

To evaluate the performance of pseudobulking methods, the datasets used had to satisfy 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. (Kang et al., 2018) 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 (Amezquita et al., 2020) through the `ExperimentHub` package (Morgan and Shepherd, 2025), 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 (Crowell et al., 2025). It is accessible via ArrayExpress (accession: E-MTAB-8192) and can also be obtained from the `ExperimentHub` package (Morgan and Shepherd, 2025) with the accession code: `EH3297`.

2.1.3 Guo et al. 2018

Guo et al. (Guo et al., 2018) 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 generalizability 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 (Loriot et al., 2025), and can be accessed using the `testis_sce` function.

2.2 Simulation Framework

2.3 Pseudobulk Approach

scheme

2.4 Workflow

ADD woRKFLOW FIGURE

Given the complexity and number of steps involved in this workflow, we used the `targets` R package (Landau, 2021) 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 X models.

2.4.1 Preprocessing

Each dataset underwent a standardized 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 synthetic data according to the simulation framework described earlier. We used functions provided by the `muscat` R package (Crowell et al., 2025) 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.
2. 10% of genes were altered in mean expression.
3. 10% of genes were altered in modality.
4. 10% of genes were altered in the proportions of low and high expression-state components.

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 normalization 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 (McCarthy et al., 2017), as well as the `vst` function from the `sctransform` R package (Choudhary and Satija, 2022).

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 (Bates et al., 2015), treating the patient effect as a random effect. The `scDD` method was implemented using the `scDD` package (Korthauer et al., 2016).

For aggregated (pseudobulk) data, we used established bulk RNA-seq modeling approaches, including `limma` (Ritchie et al., 2015), `DESeq2` (Love et al., 2014), and `edgeR` (Chen et al., 2025). 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 (Soneson and Robinson, 2016). 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 visualize the intersections of detected gene sets across methods and ground truth. The runtime for each method, comprising both the aggregation and modeling steps, was recorded using the `targets` package (Landau, 2021).

3 Results

– Describe your results • This can be in relation to: – Different input/validation data – Different method parameterisations, . . . – Explain what the results mean – Use tables for numbers (do not list in the text) – Figures for distributions, relationships, . . . (easier to understand than text) – 2-3 pages

4 Discussion

Which issues did you identify, and which problems did you encounter? – What is different about your approach (and the results you get) in comparison to the original method? Why? – What are advantages/disadvantage of each method? – 1 page

5 Conclusion

References

R. Amezquita, A. Lun, E. Becht, V. Carey, L. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Soneson, L. Waldron, H. Pages, M. Smith, W. Huber, M. Morgan, R. Gottardo, and S. Hicks. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17:137–145, 2020. URL `https://www.nature.com/articles/s41592-019-0654-x`.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Y. Chen, L. Chen, A. T. L. Lun, P. Baldoni, and G. 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.

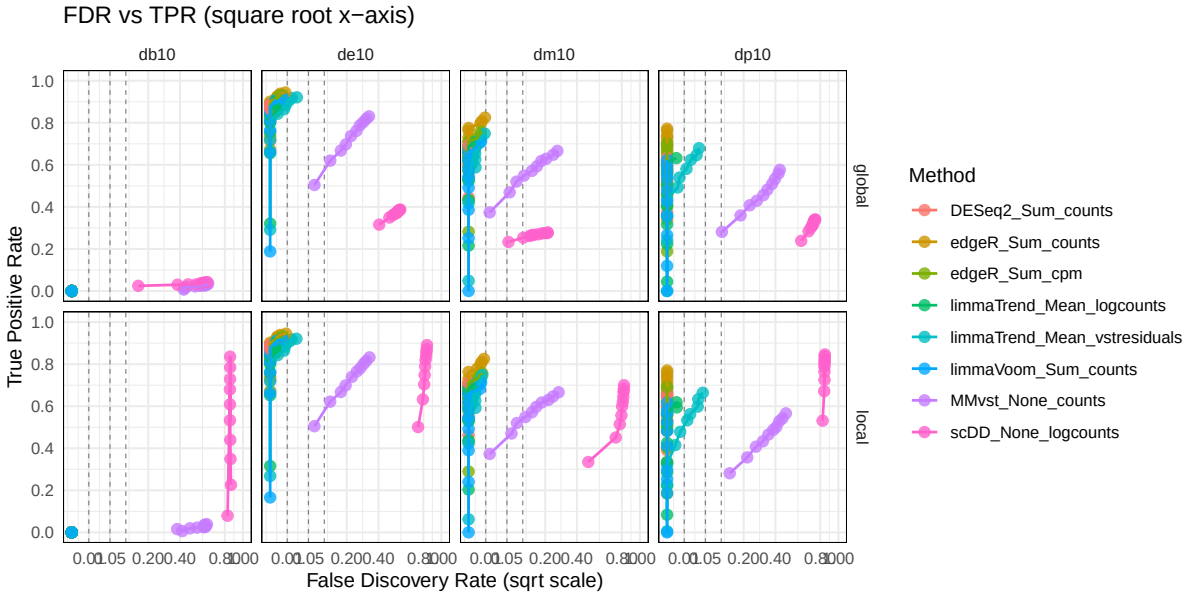

Figure 1:

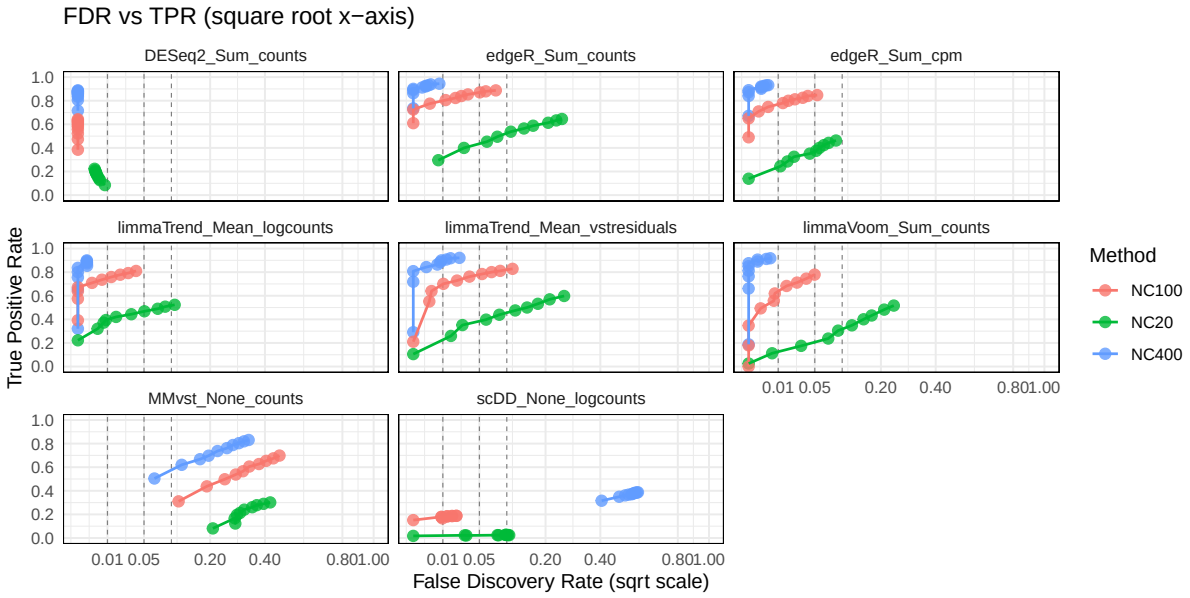

Figure 2:

S. Choudhary and R. 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`.

H. L. Crowell, P.-L. Germain, C. Soneson, A. Sonrel, J. Gilis, D. Risso, L. Clement, and M. 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.

J. Guo, E. J. Grow, H. Mlcochova, G. J. Maher, C. Lindskog, X. Nie, Y. Guo, Y. Takei, J. Yun, L. Cai, R. Kim, D. T. Carrell, A. Goriely, J. M. Hotaling, and B. R. Cairns. The adult human testis transcriptional cell atlas. *Cell Research*, 28 (12):1141–1157, Dec. 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.

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`.

H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell, and C. J. Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, Jan. 2018. ISSN 1546-1696. doi: 10.1038/nbt.4042. URL `https://www.nature.com/articles/nbt.4042`. Publisher: Nature Publishing Group.

K. D. Korthauer, L.-F. Chu, M. A. Newton, L. Yuan, J. Thomson, R. Stewart, and C. 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`.

W. M. 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`.

A. Loriot, J. Devis, and L. Gatto. *CTdata: Data companion to CTexploreR*, 2025. URL `https://bioconductor.org/packages/CTdata`. R package version 1.8.0.

M. I. Love, W. Huber, and S. 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.

D. J. McCarthy, K. R. Campbell, A. T. L. Lun, and Q. 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.

M. Morgan and L. Shepherd. *ExperimentHub: Client to access ExperimentHub resources*, 2025. URL `https://bioconductor.org/packages/ExperimentHub`. R package version 2.16.0.

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/`.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. 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.

C. Soneson and M. 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`.