Consequences and opportunities arising due to sparser single-cell RNA-seq datasets

Bouland, Gerard A.; Mahfouz, Ahmed; Reinders, Marcel J.T.

DOI
10.1186/s13059-023-02933-w

Publication date
2023

Document Version
Final published version

Published in
Genome biology

Genome Biology

SHORT REPORT

Open Access

Consequences and opportunities arising due to sparser single-cell RNA-seq datasets

Gerard A. Bouland[1,2], Ahmed Mahfouz[1,2,3*] and Marcel J. T. Reinders[1,2,3*]

*Correspondence:
a.mahfouz@lumc.nl;
m.j.t.reinders@tudelft.nl

[1] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
[2] Department of Human Genetics, Leiden University Medical Center, Leiden 2333ZC, The Netherlands
[3] Leiden Computational Biology Center, Leiden University Medical Center, Leiden 2333ZC, The Netherlands

Abstract

With the number of cells measured in single-cell RNA sequencing (scRNA-seq) datasets increasing exponentially and concurrent increased sparsity due to more zero counts being measured for many genes, we demonstrate here that downstream analyses on binary-based gene expression give similar results as count-based analyses. Moreover, a binary representation scales up to ~50-fold more cells that can be analyzed using the same computational resources. We also highlight the possibilities provided by binarized scRNA-seq data. Development of specialized tools for bit-aware implementations of downstream analytical tasks will enable a more fine-grained resolution of biological heterogeneity.

Background

Since its introduction, single-cell RNA sequencing (scRNA-seq) has been vital in investigating biological questions that were previously impossible to answer [1–4]. Continuous technological innovations are resulting in a consistent increase in the number of cells and molecules being measured in a single experiment. However, at the same time, datasets appear to become sparser, i.e., more zero measurements across the whole dataset. The sparsity has generally been seen as a problem, especially since standard count distribution models (e.g., Poisson) do not account for the excess of zeros [5–8]. This sparked discussions about whether the excess of zeros can be explained by mainly technological or biological factors [5, 8–10]. Jiang et al. [8] discuss the "zero-inflation controversy," in which a distinction is made between a biological zero, indicating the true absence of a transcript, and a non-biological zero, indicating failure of measuring a transcript that was present in the cell. Similarly, Sarkar and Stephens [11] make a distinction between measurement and expression. They proposed a model that is a combination of an expression model that encodes the true absence of a transcript, i.e., a (biological) zero, with a measurement model, for which they use a Poisson model (which can result in non-biological zeros due to limited sequencing depth). Consequently, even non-biological zeros encode useful biological information as then the gene is unlikely to be highly expressed.

Or, in other words: all zeros in scRNA-seq datasets have biological significance. Aligned with this, Qui et al. [12] proposed to "embrace" all zeros as useful signal and developed a clustering algorithm requiring only binarized scRNA-seq data (a zero representing a zero count and a one for non-zero counts). Using binarized scRNA-seq data, Qui et al. identified clusters similar to clusters identified using a count-based approach. Although this was the first paper explicitly embracing zeros as useful signal, binarization of scRNA-seq was already used to infer gene regulatory networks [13]. Since then, several methods have employed binarized scRNA-seq data. For instance, scBFA [14], a dimensionality reduction method for binarized scRNA-seq data, showed improved visualization and classification of cell identity and trajectory inference when compared to methods that use count data. Likewise, we introduced binary differential analysis (BDA) [15], a differential expression analysis method relying on binarized scRNA-seq data. We showed that differential expression analysis on binary representations of scRNA-seq data faithfully captures biological variation across cell types and conditions.

Provided that a binarized data representation has the potential to reduce required computational resources considerably, and as scRNA-seq datasets are becoming increasingly bigger and sparser, we wondered if binary should be the preferred data representation for other tasks. In this work, we explore the consequences of sparser datasets and the applicability of binarized scRNA-seq data for various single-cell analysis tasks.

Results and discussion

We downloaded 56 datasets published between 2015 and 2021. Based on these datasets, a clear association between the year of publication and the number of cells can be observed (Pearson's correlation coefficient of $r = 0.46$, Fig. 1a). For instance, the average dataset in 2015 ($n = 7$) had 704 cells while the average dataset in 2020 ($n = 7$) had 58,654 cells. Another clear trend that can be seen is that an increasing number of cells is highly correlated with decreasing detection rates (fraction of non-zero values) (Pearson's correlation coefficient of $r = -0.47$, Fig. 1b). Note that this trend of measuring more cells per dataset outweighs improved chemistry over time and thus still results in sparser datasets. It is likely that this trend will continue over the next years as, for many biological questions, shallow sequencing of many cells is more cost effective than deep sequencing of a few cells [16]. Moreover, by measuring more cells, we can better estimate the probability whether a gene is expressed, and the overall power to detect differentially expressed genes in a given dataset increases [17]. This trend will be amplified, as more population scale and multi-condition scRNA-seq datasets are emerging [17, 18], for which a low coverage sequencing is sufficient to capture cell type specific gene expression (given enough cells are measured per individual and per cell type) [19]. Altogether, these developments will result in sparser scRNA-seq datasets with larger numbers of cells.

As zeros become more abundant, a binarized expression might be as informative as counts. Using ~1.5 million cells from 56 datasets, we observed on average a strong point-biserial correlation (Pearson correlation coefficient $p = 0.93$) between the normalized expression counts of a cell and its respective binarized variant, although differences between datasets exist (Additional file 1: Fig. S1). This strong correlation implies that the binarized signal already captures most of the signal present in the normalized

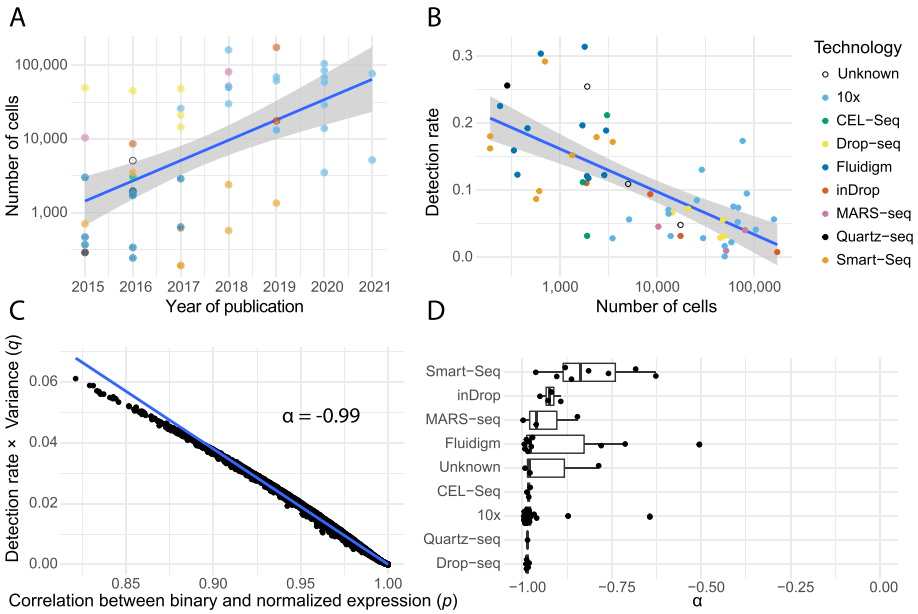

Fig. 1 More cells, more zeros. Binarized scRNA-seq datasets were generated by binarizing the raw count matrix, where zero remains zero and every non-zero value is assigned a one. **A** Association between year of publication, total number of cells. Scatterplot of the number of cells (log scale) against the date of publication. **B** Scatterplot of the detection rate (*y*-axis) against the number of cells (log scale, *x*-axis). **C** On the *x*-axis the Pearson's correlation coefficient (*p*) of every cell from the PaulHSC dataset between the binarized and normalized expressions. On the *y*-axis the product of the detection rate and the variance of the non-zero values (*q*). *a* is the Pearson's correlation coefficient between these values *p* and *q* across all cells. **D** Boxplots of the *α*-values for all 56 datasets grouped by technology. One dataset (LawlorPancreasData) was excluded as *α*-value ($\alpha = 0.42$) for this dataset was a clear outlier

count data. This strong correlation is primarily explained by the detection rate (Additional file 1: Fig. S2a) and the variance of the non-zero counts of a cell (Additional file 1: Fig. S2b). In cells where the detection rate is low (many zeros) and the variance of the non-zero counts is small, the correlation between the normalized expression values and their binary representation is high (Fig. 1c). Across all datasets, the detection rate and variance of measured expressions were good predictors for the correlation between the binary representation and the normalized representation, although differences between technologies exist (Fig. 1d). This indicates that as datasets become sparser, counts become less informative with respect to binarized expression.

To assess whether counts can actually be discarded in practice, we assessed whether binarized data can give comparable results to counts in four common single-cell analysis tasks: (1) dimensionality reduction for visualization, (2) data integration, (3) cell type identification, and (4) differential expression analysis using pseudobulk. First, for dimensionality reduction, we used three different dimensionality reduction approaches on binarized scRNA-seq data; (i) scBFA [14], (ii) PCA (Fig. 2a), and (iii) eigenvectors of the Jaccard cell–cell similarity matrix (see Additional file 2). All three approaches were compared to the standard approach of applying PCA to the normalized counts (Fig. 2b, Additional file 1: Fig. S3). Further, for all four methods, the first ten components were used to generate a non-linear embedding using UMAP (Additional file 1: Fig. S4). Qualitatively, we observed that the results of binary-based dimensionality reduction are

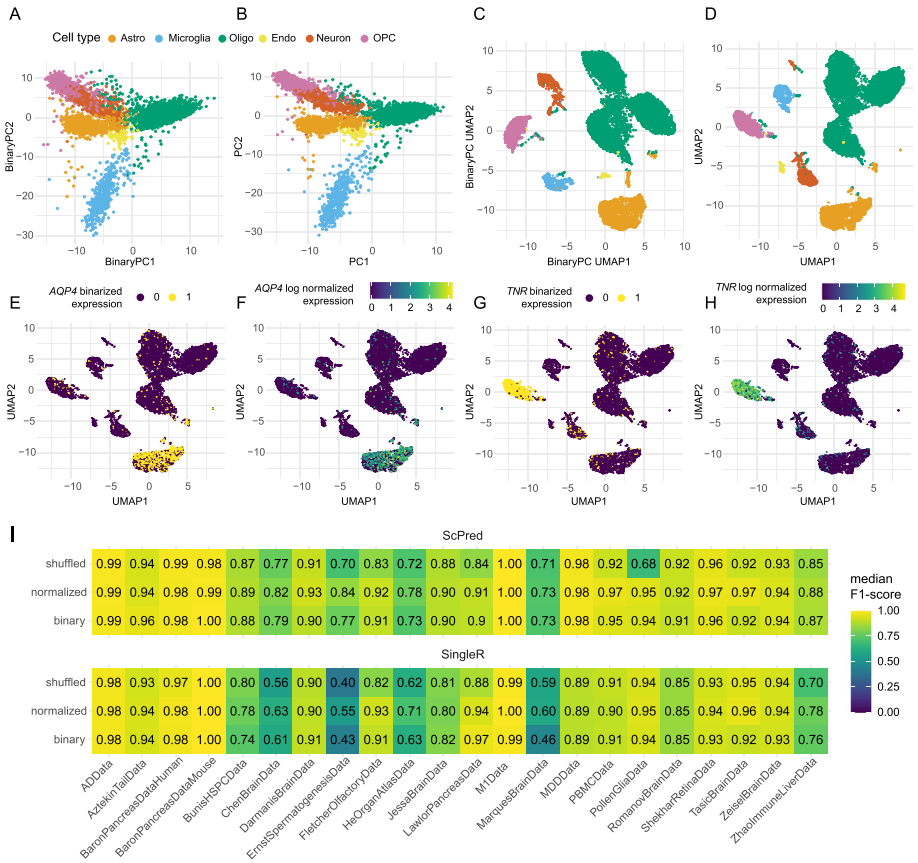

Fig. 2 **A**, **B** Cells plotted against the first two principle components of the AD dataset [20]. **A** PCA based on binary representation, and **B** PCA based on count representation. UMAP generated from data presented with **C** the binary-based PCs and **D** the count-based PCs. Colors indicate annotated cell type. **E**, **H** UMAP based on the count based PCs, in which cells are colored according to the binary representation of the marker genes *AQP4* (**E**) and *TNR* (**H**) which are known markers for astrocytes and OPCs respectively [21]. **F**, **G** Similar as **E** and **H** but showing the normalized expression of the marker gene. **I** The performance (median F1-score) of cell type identification by SingleR [22] and scPred [23] when applied to binary (binarized data), normalized (normalized expression), and shuffled (shuffled normalized expression) for 22 datasets

comparable to standard count-based methods. This was confirmed quantitatively, as the pairwise distances between cells based on the binary-based UMAPs were highly correlated with the pairwise distances from the count-based UMAP ($r \geq 0.73$, Additional file 1: Fig. S5). Especially the UMAP generated with the binary-based PCs was visually very similar to the UMAP generated with the count-based PCs (Fig. 2c, d). Calculating the silhouette score (SS) for each cell type with the reduced dimensions ($n = 10$) resulted in slightly lower scores for scBFA (SS = 0.32) and binary-based PCA (SS = 0.39) compared to the count-based PCA (SS = 0.44) (Additional file 1: Fig. S6). However, in the UMAP space (2-dimensional), silhouette scores for scBFA (SS = 0.43) and binary-based PCA (SS = 0.42) were higher than count-based PCA (SS = 0.35).

Second, we integrated three scRNA-seq datasets [20, 24, 25] with Harmony [26], using count- and binary-based PCA. Both, visually and quantitatively, we observed an improved mixing of cells for the binary representation (LISI = 1.18) as compared to counts (LISI = 1.12) (Additional file 1: Figs. S7 and S8). Third, we evaluated the effect

of binarization on cell annotation using (i) marker genes and (ii) classification methods. Using a set of known brain cell type markers [21], we annotated the binarized AD dataset [20] based on solely the detection of respective cell type markers (see Additional file 2). The annotations were compared to cell type labels that were originally assigned based on the markers' expression level (i.e., counts). We observed a high level of concordance between annotations as quantified by a median F1-score of 0.93. (Additional file 1: Fig. S9). Additionally, we found that the visualization of the binarized expression of cell type markers to be highly similar to the visualization of their normalized expression in UMAP plots (Fig. 2e–h, Additional file 1: Fig. S10). Next, we compared the performance of automatic cell type identification using scPred and SingleR [22, 23] on 22 datasets for which cell type annotations were available. The median F1-scores were highly similar between cell type identifications based on the binarized and the normalized count data, despite large variation of sparseness between these datasets. This finding implies that counts do not add information for cell type identification. This conclusion was further supported by randomly shuffling the non-zero counts, which resulted in a comparable performance (Fig. 2I, Additional file 1: Fig. S11).

Forth, we evaluated whether counts can also be discarded when pseudobulk data is used for differential expression analysis [18]. In a dataset containing scRNA-seq data of the prefrontal cortex of 34 individuals [27], we generated pseudobulk data by either taking the mean expression of each gene across all cells, or the fraction of non-zero values across all cells (detection rate), per individual. Spearman's rank correlation between the binarized profile and the mean counts (across all genes) was ≥ 0.99 (Additional file 1: Fig. S12) for every individual, implying that pseudobulk aggregation with binarized expression faithfully represents counts. To quantify this further, we generated 960 datasets using muscat [18] with 96 unique simulation settings (see Additional file 2). In each dataset, pseudobulk data for each individual was generated and we identified differential expressed genes using Limma trend [28] for the mean gene expression and a *t*-test for the detection rate. In general, the F1-scores for the count and binary representations were very similar across the different settings; however, with small sample sizes and fewer cells, analyses based on a count representation performed better, while analyses based on a binarized expression performed better with larger sample sizes and more cells (Additional file 1: Fig. S13). Additionally, count-based analyses resulted in more false positives (Additional file 1: Fig. S14), while binarized-based analyses resulted in more false negatives (Additional file 1: Fig. S15). The false negatives were primarily due to highly expressed genes that show no differences in the detection rate. At larger sample sizes and with more cells, the false negatives diminished (Additional file 1: Fig. S16). All together, these result show that most of the information is indeed captured in the binary representation, only when genes have a high detection rate (>0.9), or when the number of cells per sample becomes low, then, changes in expression are not reflected in the binary representation and, consequently, information from counts is needed.

Whether zero-inflation associates with technical or biological origins is heavily debated [8]. One compelling reason for this debate is the fact that within a single dataset some genes are zero-inflated, while others are not [5, 8]. We argue that this observation is mostly related to whether a gene is only expressed in a subpopulation of cells (e.g., marker genes) or whether a gene has a stable expression (e.g., housekeeping genes).

To substantiate our claim, we used BDA [15] to identify the top 100 most differentially expressed genes between two cell populations and the top 100 most stable expressed genes in a 10X dataset [24] as well as a Smart-Seq dataset [29]. Next, we applied scRATE [5] to identify the best distribution model for the observed expression of the identified genes, being either a Poisson, a negative binomial, or their zero-inflated counterparts. A Fisher exact test showed that a zero-inflated model was enriched in the top 100 differentially expressed genes, and a non-zero inflated model was enriched in the top 100 stable expressed genes (Table 1). Hence, like earlier work [5], we conclude biological heterogeneity to be the main driver of zero-inflation.

Increasingly larger datasets require increasingly more computational resources. The storage required for all 56 datasets used in this study was 764 gigabytes after normalization using sctransform [30] or 276 gigabytes when log-normalized and stored as sparse matrices. In contrast, binarizing the same datasets and storing them as bits required only 73 gigabytes, which is an ~11-fold and ~fourfold reduction in storage requirements, respectively (Additional file 1: Fig. S17). Yet, there are big differences across datasets. For example, a reduction of ~50-fold and ~20-fold, respectively, was acquired for the BuettnerESC dataset [31]. The amount of storage that can be saved is highly correlated with the detection rate (Additional file 1: Fig. S18), with the highest gain for datasets with a high detection rate. The considerable storage reduction of the binary representation gives the potential to boost downstream analyses to larger numbers of cells, opening possibilities to get a more fine-grained resolution of biological heterogeneity [32].

We showed that analyses based on a binary representation of scRNA-seq data perform on par with count-based analyses. Working with binarized scRNA-seq data has clear additional advantages. The first is simplicity. For the various tasks that we explored, such as dimensionality reduction, data integration, cell type prediction, differential expression analysis [15], and clustering [12], the binary representations required no normalization. Hence, various subjective choices on the normalization could be avoided, which improves reproducibility of these tasks. However, as sequencing depth has an effect on the detection rate of a cell, it is likely this is not the case for all downstream tasks. Second, binarization reduces the amount of required storage significantly and allows the analysis of significantly larger datasets. For example, binary-based data allow for a bit implementation of clustering as has been done before in the field of molecular dynamics resulting in a significant reduction of run time and peak memory usage compared to existing methods [33]. It has also been suggested that binarization alleviates noise [14] as it is insensitive to count errors. However, binarization remains sensitive to detection errors caused by, e.g., the presence of ambient RNA. Consequently, detection of ambient

Table 1 Enrichment of zero-inflated distributions for the top100 differential expressed genes and the enrichment of non-zero inflated distributions for the top100 stable genes

Platform	Top 100	Zero-inflated	Not zero-inflated	logOR	95% CI	*P*-value
10x	Differentially expressed genes	99	1	5.19	3.36, 8.87	3.03×10^{-25}
	Stable genes	35	65			
Smart-seq	Differentially expressed genes	97	3	3.70	2.50, 5.36	5.46×10^{-18}
	Stable genes	44	56			

RNA [34] poses a challenge for binary representations when studying individual cells and thus might require specialized methods to be developed.

At first glance, binarizing scRNA-seq data seems to remove signal. However, genes that are highly expressed across cells will not have a lot of zeros, whereas genes that are lowly expressed across cells will have many. This implies we might be able to infer the relative expression of a gene within an individual cell by exploiting the detection pattern of similar other cells. Using this reasoning, we indeed were able to reconstruct the expression levels of genes from the detection pattern using neighboring cells (Additional file 1: Fig. S19, Additional file 2). Hence, we conclude that the detection rate of a gene in a group of cells, such as a cell type, do faithfully represents the (mean) expression levels of that gene in that group of cells, underpinning why binarization for most of the downstream tasks apparently does not have lost signal.

We have shown that sparsity is inversely correlated with the amount of additional signal that is captured with counts. Consequently, binarization will not be useful for all scRNA-seq datasets. Previous work suggested that when the detection rate is >90%, visualizations based on the binary representation do not perform on par with count-based representation [14]. With our simulation experiments, we have shown a similar trend when considering the task of detecting differential expressed genes based on pseudobulk values.

Conclusions

Concluding, our results support existing literature in showing that binarized scRNA-seq data can be used for the following: dimensionality reduction, data integration, visualization, clustering, trajectory inference, batch correction, differential expression analysis, and cell type prediction. We believe scRNA-seq tool developers should be aware of the possibility of using a binary representation of the scRNA-seq data instead of count-based data, as it gives opportunities to develop computational- and time-efficient tools.

Methods

Detailed methods are available in Additional file 2.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02933-w.

Additional file 1: Fig. S1. The distributions of correlation coefficients between the binarized and count-based expressions of every cell. Fig. S2. Dot plots of association between the correlation coefficient between the binarized and count-based representation and detection rate and variance. Fig. S3. Comparison of binary-based dimensionality reductions. Fig. S4. Comparison of binary-based UMAPs. Fig. S5. Association of pairwise Euclidean distances between cells from count based UMAP and binary based UMAP. Fig. S6. Silhouette scores of count- and binary-based dimensionality reduction. Fig. S7. Count- and binary-based UMAP plots of three brain datasets, not integrated. Fig. S8. Count- and binary-based UMAP plots of three brain datasets, integrated. Fig. S9. Heatmap of concordance between binary-based cell type annotations using markers and counts-based cell type annotations using markers. Fig. S10. UMAP plot with expressions of marker genes, using binarized and normalized representations. Fig. S11. Boxplots of the median F1-score of the automatic cell type prediction with different data representations. Fig. S12. Association of detection rate vs mean expression for all genes of one individual. Fig. S13. F1-score on 960 simulated datasets identifying differentially expressed genes in pseudobulk data with either count data or binarized data. Fig. S14. Number of false positives on 960 simulated datasets identifying differentially expressed genes in pseudobulk data with either count data or binarized data. Fig. S15. Number of false negatives on 960 simulated datasets identifying differentially expressed genes in pseudobulk data with either count data or binarized data. Fig. S16. Number of false negatives binned on the detection rate on 960 simulated datasets identifying differentially expressed genes in pseudobulk data with either count data or binarized data. Fig. S17. Storage requirements for the different data representations.

Fig. S18. Association of detection rate with fold reduction. Fig. S19. Association between recovered expression and normalized expression.

Additional file 2. Methods.

Additional file 3. Review history.

Review history
The review history is available as Additional file 3.

Peer review information
Barbara Cheifet and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
GAB, AM, and MJTR conceived the study designed the experiments. GAB performed all experiments and drafted the manuscript. GAB, AM, and MJTR reviewed and approved the manuscript.

Funding
This research was supported by an NWO Gravitation project: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012) and the European Union's Horizon 2020 research.

Availability of data and materials
All codes, processed data, and analysis results in this paper are publicly available at GitHub [35] and Zenodo [36]. Code used for the analyses and a vignette describing a complete analysis workflow on binarized scRNA-seq data are available on GitHub. The source code is released under the MIT license. The ADData dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138852 [37]. The M1Data dataset is available at https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x. The smart-Seq brain dataset available at https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq. The MDDData dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144136 [38]. The PBMCData is available at https://doi.org/10.5281/zenodo.3357167 [39]. All other datasets used in this study are described in the Additional file 2: Table S1 and are available for download from the R-package scRNAseq (v2.8.0) [40].

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 June 2022 Accepted: 10 April 2023
Published online: 21 April 2023

References

1. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;2019(570):332–7. https://doi.org/10.1038/s41586-019-1195-2. Nature Publishing Group. Cited 2020 Sep 10.
2. Van Der Wijst MGP, Brugge H, De Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50:493–7. Available from: https://www.nature.com/naturegenetics493. Cited 2021 Jun 11.
3. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494–8. Nature Publishing Group. Available from: https://www.nature.com/articles/s41586-018-0414-6. Cited 2022 May 11.
4. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715–21. Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-019-0494-8. Cited 2022 May 11.
5. Choi K, Chen Y, Skelly DA, Churchill GA. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. Genome Biol. 2020;21:183. BioMed Central Ltd. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02103-2. Cited 2020 Oct 29.
6. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:1–10. BioMed Central Ltd. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0805-z. Cited 2022 Feb 4.
7. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. Nature Publishing Group. Available from: https://www.nature.com/articles/nmeth.2967. Cited 2022 Feb 4.

8. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 2022;23(1):1–24. BioMed Central. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02601-5. Cited 2022 Feb 1.
9. Svensson V. Droplet scRNA-seq is not zero-inflated. Nat Biotechnol. 2020;38(2):147–50. Nature Research. Available from: https://www.nature.com/naturebiotechnology. Cited 2020 Oct 29.
10. Cao Y, Kitanovski S, Küppers R, Hoffmann D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. Nat Biotechnol. 2021;39(2):158–9. Nature Publishing Group. Available from: https://www.nature.com/articles/s41587-020-00810-6. Cited 2022 Mar 28.
11. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet. 2021;53(6):770–7. Nature Publishing Group. Available from: https://www.nature.com/articles/s41588-021-00873-4. Cited 2021 Oct 25.
12. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nat Commun. 2020;11:1–9. https://doi.org/10.1038/s41467-020-14976-9. Nature Research. Cited 2020 Oct 29.
13. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015;33(3):269–76. Nature Publishing Group. Available from: https://www.nature.com/articles/nbt.3154. Cited 2021 Nov 29.
14. Li R, Quon G. ScBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. Genome Biol. 2019;20:1–20. BioMed Central Ltd. Available from: https://link.springer.com/articles/10.1186/s13059-019-1806-0. Cited 2021 Nov 29.
15. Bouland GA, Mahfouz A, Reinders MJT. Differential analysis of binarized single-cell RNA sequencing data captures biological variation. NAR Genom Bioinform. 2021;3(4):lqab118. Oxford Academic. Available from: https://academic.oup.com/nargab/article/3/4/lqab118/6478878. Cited 2022 Jan 18.
16. Zhang MJ, Ntranos V, Tse D. Determining sequencing depth in a single-cell RNA-seq experiment. Nat Commun. 2020;11(1):1–11. Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-020-14482-y. Cited 2022 Oct 4.
17. Schmid KT, Höllbacher B, Cruceanu C, Böttcher A, Lickert H, Binder EB, et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. Nat Commun. 2021;12(1):1–18. Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-021-26779-7. Cited 2022 Nov 7.
18. Crowell HL, Soneson C, Germain PL, Calini D, Collin L, Raposo C, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11:1–12. https://doi.org/10.1038/s41467-020-19894-4. Nature Research. Cited 2021 Apr 14.
19. Mandric I, Schwarz T, Majumdar A, Hou K, Briscoe L, Perez R, et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. Nat Commun. 2020;11(1):1–9. Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-020-19365-w. Cited 2021 Nov 30.
20. Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, 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. https://doi.org/10.1038/s41593-019-0539-4. Nature Research. Cited 2020 Oct 27.
21. McKenzie AT, Wang M, Hauberg ME, Fullard JF, Kozlenkov A, Keenan A, et al. Brain cell type specific gene expression and co-expression network architectures. Sci Rep. 2018;8(1):1–19. Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-018-27293-5. Cited 2022 Nov 4.
22. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72. Nature Publishing Group. Available from: https://www.nature.com/articles/s41590-018-0276-y. Cited 2022 Feb 9.
23. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. ScPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019;20:1–17. BioMed Central Ltd. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1862-5. Cited 2022 Feb 9.
24. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598(7879):111–9. Nature Publishing Group. Available from: https://www.nature.com/articles/s41586-021-03465-8. Cited 2022 Feb 15.
25. Nagy C, Maitra M, Tanti A, Suderman M, Théroux JF, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat Neurosci. 2020;23(6):771–81. Nature Publishing Group. Available from: https://www.nature.com/articles/s41593-020-0621-y. Cited 2021 Nov 30.
26. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96. Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-019-0619-0. Cited 2022 Sep 15.
27. Nagy C, Maitra M, Tanti A, Suderman M, Théroux JF, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat Neurosci. 2020;23:771–81. https://doi.org/10.1038/s41593-020-0621-y. Nature Research. Cited 2020 Oct 27.
28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. Oxford University Press. Available from: https://academic.oup.com/nar/article/43/7/e47/2414268. Cited 2021 Jan 18.
29. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573:61–8. https://doi.org/10.1038/s41586-019-1506-7.
30. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):1–15. BioMed Central. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1. Cited 2021 Oct 26.
31. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33:155–60. Available from: https://pubmed.ncbi.nlm.nih.gov/25599176/. Cited 2022 Sep 15.

32. Sikkema L, Strobl D, Zappia L, Madissoon E, Markov N, Zaragosi L, et al. An integrated cell atlas of the human lung in health and disease. bioRxiv. 2022;2022.03.10.483747. Cold Spring Harbor Laboratory. Available from: https://www.biorxiv.org/content/10.1101/2022.03.10.483747v1. Cited 2022 May 11.
33. González-Alemán R, Hernández-Castillo D, Rodríguez-Serradet A, Caballero J, Hernández-Rodríguez EW, Montero-Cabrera L. BitClust: fast geometrical clustering of long molecular dynamics simulations. J Chem Inf Model. 2020;60:444–8. American Chemical Society.
34. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21:1–15. BioMed Central Ltd. Available from: https://link.springer.com/articles/10.1186/s13059-020-1950-6. Cited 2022 Nov 7.
35. Bouland GA, Mahfouz A, Reinders MJT. Arising_sparsity_scRNAseq. Github; 2023. https://github.com/gbouland/Arising_sparsity_scRNAseq.
36. Bouland GA, Mahfouz A, Reinders MJT. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. Zenodo; 2023. https://doi.org/10.5281/zenodo.7732380.
37. Chew G, Grubman A, Ouyang JF, Rackham O, Polo J, Petretto E. A single-cell atlas of the human cortex reveals drivers of transcriptional changes in Alzheimer's disease in specific cell subpopulations. Gene Expression Omnibus; 2019. https://identifiers.org/geo:GSE138852.
38. Turecki G. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Gene Expression Omnibus; 2020. https://identifiers.org/geo:GSE144136.
39. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Me H, Reinders MJT, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA-sequencing data. Zenodo; 2019. https://doi.org/10.5281/zenodo.3357167.
40. Risso D, Cole M. scRNAseq: collection of public single-cell RNA-Seq Datasets. R package version 2.8.0. 2022.

Publisher's Note