The rise of sparser single-cell RNAseq datasets; consequences and opportunities

Gerard A. Bouland[1,2], Ahmed Mahfouz[1,2,3,*], Marcel J.T. Reinders[1,2,3,*]
[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
* Corresponding authors: Ahmed Mahfouz (a.mahfouz@lumc.nl) and Marcel J.T. Reinders (m.j.t.reinders@tudelft.nl)

Abstract

There is an exponential increase in the number of cells measured in single-cell RNA sequencing (scRNAseq) datasets. Concurrently, scRNA-seq datasets become increasingly sparser as more zero counts are measured for many genes. We discuss that with increasing sparsity the binarized representation of gene expression becomes as informative as count-based expression. We show that downstream analyses based on binarized gene expressions give similar results to analyses based on count-based expressions. Moreover, a binarized representation scales to 17-fold more cells that can be analyzed using the same amount of computational resources. Based on these observations, we recommend the development of specialized tools for bit-aware implementations for downstream analyses tasks, creating opportunities to get a more fine-grained resolution of biological heterogeneity.

Introduction

Since its introduction, single-cell RNA sequencing (scRNAseq) 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 being measured in a single experiment. However, at the same time, datasets have become sparser, i.e. no counts measured for an increasing number of genes. This sparsity has generally been seen as a problem, especially since standard count distribution models seem to fail in explaining the excess of zeros(5–8). This zero-inflation has sparked discussions about whether the excess of zeros can be explained by mainly technological or biological factors(5, 8–10). As the field moves towards sparser datasets, it is vital to know what the consequences are of the ever-increasing abundance of zero measurements. In 2020, Qui et al. (11) proposed to embrace zeros as useful signal and developed a clustering algorithm requiring only binarized scRNAseq data (a zero representing a zero count and a one a non-zero count). Although this was the first paper explicitly stating zeros ("dropouts") to be informative, binarization of scRNAseq was already used in practice in 2015 to infer gene regulatory networks(12). Since then, several methods have employed binarized scRNA-seq data. For instance, for dimensionality reduction that improved cell type classification and trajectory inference compared to using counts(13), as well as for differential expression analysis(14).

Using 52 datasets, published between 2015 and 2021, we show that as the sparsity of scRNAseq data increases, the detection pattern of expression (binarized scRNAseq data) becomes as informative as the quantification of expression (counts). We explain this by showing that the driving process behind zero-inflation is mainly explained by biological heterogeneity. Motivated by these findings, we demonstrate that downstream analyses based on solely the detection pattern are in line with those based on quantification. Together with further advantages of binarized scRNAseq data, this opens possibilities in analyzing extremely large datasets in an efficient manner.

More cells, more zeros

Across the 52 datasets the association between the year of publication and the number of cells shows a Pearson's correlation coefficient of r = 0.57 (**Fig. 1a**). For instance, the average dataset in 2015 (n = 7) had 704 cells while the average dataset in 2020 (n = 5) had 68,100 cells. One consequence of having more cells is that datasets are becoming sparser, exemplified by a Pearson's correlation coefficient of r = -0.72 (**Fig. 1b**) for the association between the number of cells and detection rate. It is likely that this trend will continue over the next years as having a larger number of cells has statistical benefits(15, 16).

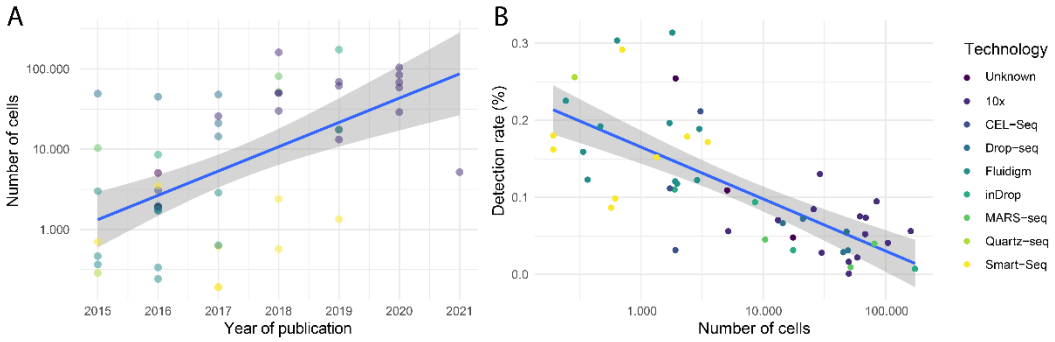

FIGURE 1: Using R-package scRNAseq (v2.8.0), 52 datasets ranging in date of publication from 2015 to 2021 were downloaded. (a) Scatterplot of the number of cells (log scale) against the date of publication. (b) Scatterplot of the detection rate (%) (y-axis) against the number of cells (log scale, x-axis).

More zeros, less signal in expression counts

As zeros become more abundant, the detection pattern might be as informative as the expression counts. Using cells from four datasets, with varying degrees of sparsity, we observed a strong correlation (Pearson's $r \geq 0.73$) between the log-normalized expression counts of a cell and its respective binarized variant (**Fig. 2**). This strong correlation implies that the binarized signal captures most of the signal present in the log normalized count data. Interestingly, there was a strong correlation between this correlation (between counts and binarized profiles) and the detection rate (**Fig. 2**, mean correlation = -0.76). This indicates that as datasets become sparser, the quantification of expression becomes less informative with respect to the detection pattern.

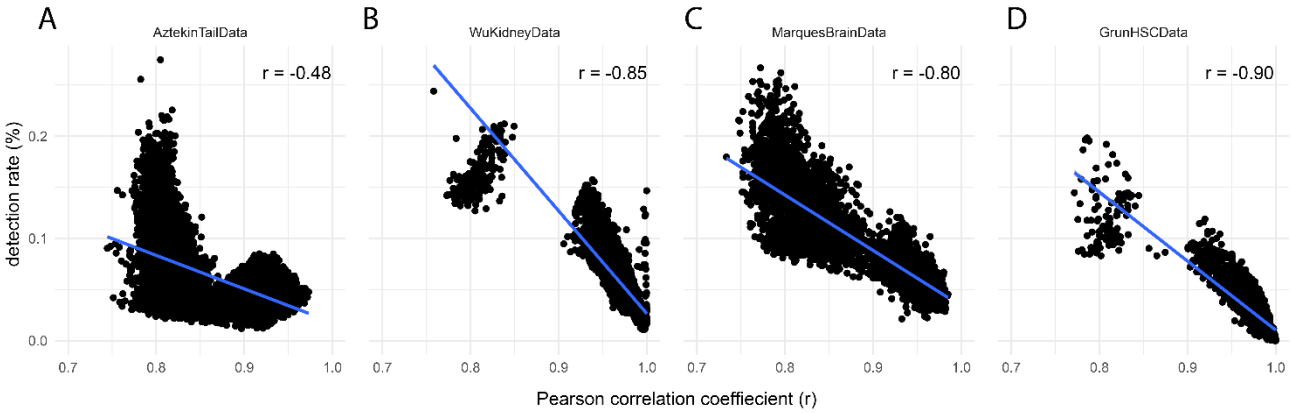

FIGURE 2: Pearson correlations per cell between log normalised expression profile and the binarized profile across four different datasets (between brackets: detection rate% : median, Q1, Q3) : **a)** AztekinTailData(17) (6.3% 4.6%, 8.8%) **b)** WuKidneyData(18) (4.2%, 3.1%, 5.8%) **c)** MarquesBrainData(19) (10.4% 7.8%, 13.6%) and **d)** GrunHSCData(20) (2.2%, 0.5%, 4.7%). X-axis represents the correlation, Y-axis the (detection rate %) in a cell.

Expression counts add little to no information on top of the detection pattern

To test whether counts can actually be discarded in practice, we compared the performance of automatic cell-type identification and differential expression analysis methods using counts and binarized data. First, we used two existing automatic cell-type identification methods, scPred and SingleR(21, 22). The median F1-score as well as the global accuracy were very similar between cell-type identifications based on the binarized data and identifications based on the log-normalized count data (**Fig. 3a, Fig. 3b**). This finding implies that the quantification of expression does not add information for cell-type identification. This conclusion was further supported by randomly shuffling the non-zero counts, which resulted in an almost similar cell-type identification performance (**Fig. 3c, Fig. 3d**).

Next, we evaluated whether counts can also be discarded when pseudobulk data is considered. For each individual, in a dataset containing scRNAseq data of the prefrontal cortex of 34 individuals(23), we generated pseudobulk data by either taking, for each gene, the mean expression across all cells, or the rate of non-zero values across all cells (detection

rate). The Spearman's rank correlation (across all genes) was ≥ 0.99 (**Fig. 3e**) for every individual, indicating that for pseudobulk aggregation, implying that the binarized representation faithfully represents counts.

To quantify this further, we simulated 10 datasets with muscat(15), each with 50,000 cells and 2,000 genes of which 25% were differentially expressed between two groups comprised each of 10 individuals (2,500 cells per individual). Pseudobulk data for each individual was generated. Using Limma trend(24) for mean gene expression and t-test for the detection rate, we identified differential expressed genes and evaluated whether we found the simulated ones back. The F1-score for the count and binarized data representations were similar(**Fig. 3f**), again indicating that the quantification of expression did not add additional information to test for differential expressed genes in pseudobulk data.

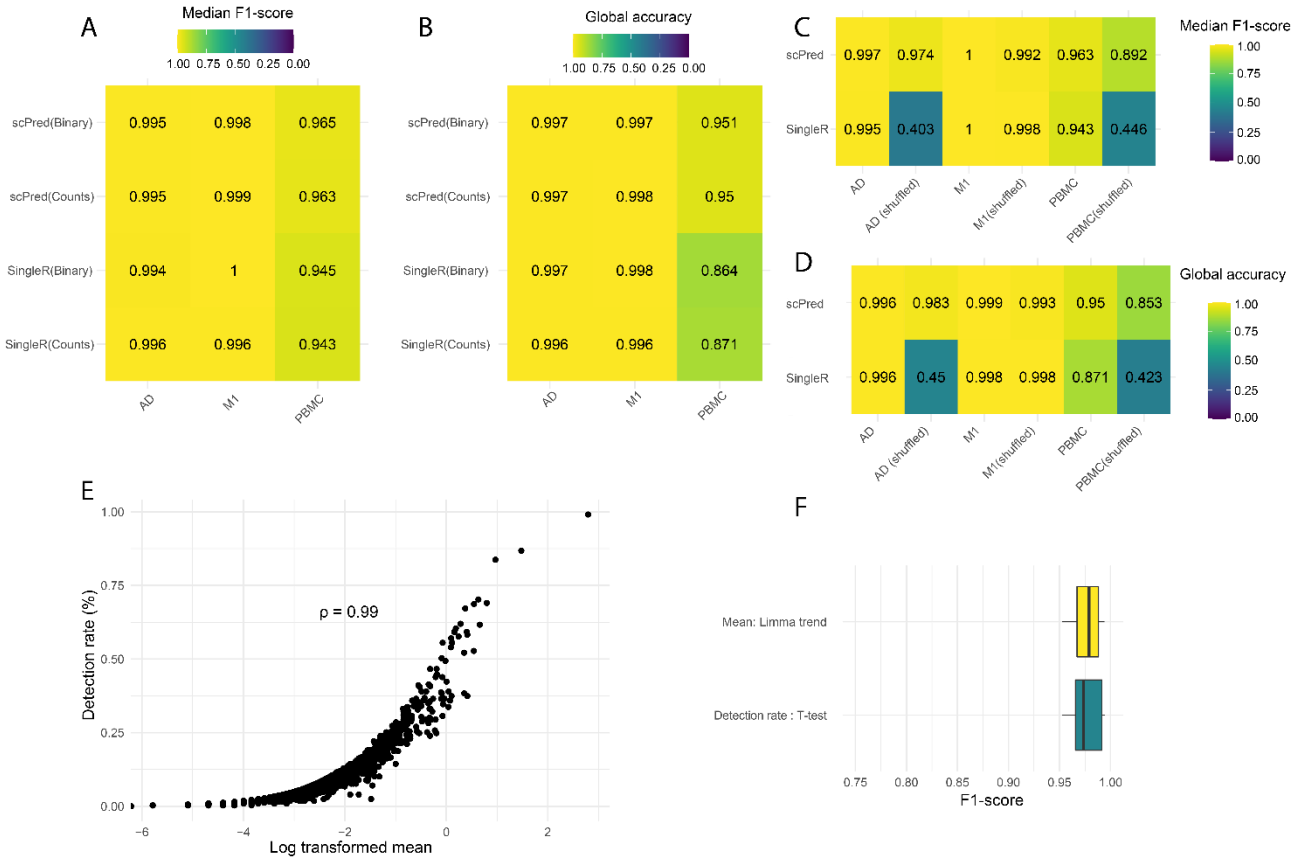

FIGURE 3: (a,b) The performance of cell type identification by SingleR(21) and scPred(22) when applied to binarized data (Binary) or log-normalized expression (Counts) for three relatively recent datasets: AD(25), M1(26), PBMC(27). (a) show the median F1-score, and (b) the global accuracy score. (c,d) Similar as in (a,b) but now with the non-zero expressions being shuffled. (e) Scatterplot of detection rate (y-axis) vs mean expression(x-axis) for all genes (n = 30.062) of one individual. (f) F1-score indicating how well simulated differentially expressed genes in pseudobulk data can be found back when either count data is used or binarized data.

Zero-inflation is driven by biological heterogeneity

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 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(14) 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(26) as well as a Smart-Seq dataset(28). 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). Like earlier work (5), we conclude biological heterogeneity to be the main driver of zero-inflation.

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			

Binarized representation allows for highly scalable analyses

Increasingly larger datasets require increasingly more computational resources. For example, a dataset(29) with 83,262 cells and 20,138 genes after transformation and normalization requires 5.1 Gigabytes of working memory using Seurat(30) and sctransform(31). In contrast, binarizing the same dataset and storing it as bits requires only 300 Megabytes, which is a 17-fold reduction in storage requirements. This potentially boosts scalability of downstream analyses to larger numbers of cells, opening possibilities to get a more fine-grained resolution of biological heterogeneity(32).

Discussion

In a recent article(8), Jian et al. discuss the "zero-inflation controversy". They made a distinction between a biological zero, indicating true absence of a transcript, and a non-biological zero, indicating failure of measuring a transcript while it is present in the cell. Likewise, Sakar and Stephens(33) propose to make a distinction between measurement and expression. Following this reasoning, a Poisson model can be used to explain the observed counts, as well as the frequency of observed zero counts. In this case, a non-biological zero encodes useful information, i.e. the gene is unlikely to be highly expressed. However, this model is not suited for the observed zero-inflation. Instead of resuming to a zero-inflated model, we argue that zero-inflation is primarily the result of biological heterogeneity. In other words, the zero-inflation itself is indicative of the multi modal expression of a gene. As such, the binary representation faithfully captures the underlying biology, as the presence of a zero is mainly dictated by the relative abundance of the respective gene.

We showed that analyses based on a binarized representation of scRNAseq data perform on par with count-based analyses. Working with binarizing scRNAseq data has clear additional advantages. The first is simplicity. Binarization circumvents the need to normalize and transform the scRNAseq data. Hence it avoids making various subjective choices and thus improves reproducibility of the subsequent data analyses. Second, binarization alleviate noise, as it is only subject to detection noise opposed to quantification and detection noise(13). Third, binarization reduces the amount of required storage significantly and allows the analysis of significantly larger datasets. This would, for example, 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(34). We have shown that sparsity is inversely correlated with the amount of additional signal that is captured by the quantification of expression. Consequently, binarization will not be useful for all scRNAseq datasets. Previous work suggested that when the detection rate becomes >90% binarization does not perform on par with count-based representation for the task of dimensionality reduction (13). At what detection rate binarizing is not useful anymore for other analyses should be investigated further.

Concluding, as binarized representations of scRNAseq data capture the relevant biological signal, there are opportunities for binarized data analyses methods so that cells can be put into a context of an extremely large number of other cells and be interrogated accordingly.

Acknowledgements

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.

Author 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.

Competing interests

The authors declare no competing interests.

References

1. Mathys,H., Davila-Velderrain,J., Peng,Z., Gao,F., Mohammadi,S., Young,J.Z., Menon,M., He,L., Abdurrob,F., Jiang,X., *et al.* (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, **570**, 332–337.
2. Van Der Wijst,M.G.P., Brugge,H., De Vries,D.H., Deelen,P., Swertz,M.A. and Franke,L. (2018) Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.*, **50**, 493–497.
3. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A., *et al.* (2018) RNA velocity of single cells. *Nat. 2018 5607719*, **560**, 494–498.
4. Lotfollahi,M., Wolf,F.A. and Theis,F.J. (2019) scGen predicts single-cell perturbation responses. *Nat. Methods 2019 168*, **16**, 715–721.
5. Choi,K., Chen,Y., Skelly,D.A. and Churchill,G.A. (2020) Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.*, **21**, 183.
6. Pierson,E. and Yau,C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 1–10.
7. Kharchenko,P. V., Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods 2014 117*, **11**, 740–742.
8. Jiang,R., Sun,T., Song,D. and Li,J.J. (2022) Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol. 2022 231*, **23**, 1–24.
9. Svensson,V. (2020) Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, **38**, 147–150.
10. Cao,Y., Kitanovski,S., Küppers,R. and Hoffmann,D. (2021) UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nat. Biotechnol. 2021 392*, **39**, 158–159.
11. Qiu,P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1–9.
12. Moignard,V., Woodhouse,S., Haghverdi,L., Lilly,A.J., Tanaka,Y., Wilkinson,A.C., Buettner,F., MacAulay,I.C., Jawaid,W., Diamanti,E., *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol. 2015 333*, **33**, 269–276.
13. Li,R. and Quon,G. (2019) ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol.*, **20**, 1–20.
14. Bouland,G.A., Mahfouz,A. and Reinders,M.J.T. (2021) Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genomics Bioinforma.*, **3**.
15. Crowell,H.L., Soneson,C., Germain,P.L., Calini,D., Collin,L., Raposo,C., Malhotra,D. and Robinson,M.D. (2020) muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.*, **11**, 1–12.
16. Mandric,I., Schwarz,T., Majumdar,A., Hou,K., Briscoe,L., Perez,R., Subramaniam,M., Hafemeister,C., Satija,R., Ye,C.J., *et al.* (2020) Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun. 2020 111*, **11**, 1–9.
17. Aztekin,C., Hiscock,T.W., Marioni,J.C., Gurdon,J.B., Simons,B.D. and Jullien,J. (2019) Identification of a regeneration-organizing cell in the Xenopus tail. *Science (80-.).*, **364**, 653–658.
18. Wu,H., Kirita,Y., Donnelly,E.L. and Humphreys,B.D. (2019) Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: Rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.*, **30**, 23–32.
19. Marques,S., Zeisel,A., Codeluppi,S., Van Bruggen,D., Falcão,A.M., Xiao,L., Li,H., Häring,M., Hochgerner,H., Romanov,R.A., *et al.* (2016) Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, **352**, 1326–1329.
20. Grün,D., Muraro,M.J., Boisset,J.C., Wiebrands,K., Lyubimova,A., Dharmadhikari,G., van den Born,M., van Es,J., Jansen,E., Clevers,H., *et al.* (2016) De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, **19**, 266–277.
21. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R., *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol. 2019 202*, **20**, 163–172.
22. Alquicira-Hernandez,J., Sathe,A., Ji,H.P., Nguyen,Q. and Powell,J.E. (2019) ScPred: Accurate supervised method for

cell-type classification from single-cell RNA-seq data. *Genome Biol.*, **20**, 1–17.

23. Nagy,C., Maitra,M., Tanti,A., Suderman,M., Théroux,J.F., Davoli,M.A., Perlman,K., Yerko,V., Wang,Y.C., Tripathy,S.J., *et al.* (2020) Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.*, **23**, 771–781.
24. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol. 2014 152*, **15**, 1–17.
25. Grubman,A., Chew,G., Ouyang,J.F., Sun,G., Choo,X.Y., McLean,C., Simmons,R.K., Buckberry,S., Vargas-Landin,D.B., Poppe,D., *et al.* (2019) A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.*, **22**, 2087–2097.
26. Bakken,T.E., Jorstad,N.L., Hu,Q., Lake,B.B., Tian,W., Kalmbach,B.E., Crow,M., Hodge,R.D., Krienen,F.M., Sorensen,S.A., *et al.* (2021) Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nat. 2021 5987879*, **598**, 111–119.
27. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun. 2017 81*, **8**, 1–12.
28. Hodge,R.D., Bakken,T.E., Miller,J.A., Smith,K.A., Barkan,E.R., Graybuck,L.T., Close,J.L., Long,B., Johansen,N., Penn,O., *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature*, **573**, 61–68.
29. Almanzar,N., Antony,J., Baghel,A.S., Bakerman,I., Bansal,I., Barres,B.A., Beachy,P.A., Berdnik,D., Bilen,B., Brownfield,D., *et al.* (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, **583**, 590–595.
30. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol. 2015 335*, **33**, 495–502.
31. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol. 2019 201*, **20**, 1–15.
32. Sikkema,L., Strobl,D., Zappia,L., Madissoon,E., Markov,N., Zaragosi,L., Ansari,M., Arguel,M., Apperloo,L., Bécavin,C., *et al.* (2022) An integrated cell atlas of the human lung in health and disease. *bioRxiv*, 10.1101/2022.03.10.483747.
33. Sarkar,A. and Stephens,M. (2021) Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet. 2021 536*, **53**, 770–777.
34. González-Alemán,R., Hernández-Castillo,D., Rodríguez-Serradet,A., Caballero,J., Hernández-Rodríguez,E.W. and Montero-Cabrera,L. (2020) BitClust: Fast Geometrical Clustering of Long Molecular Dynamics Simulations. *J. Chem. Inf. Model.*, **60**, 444–448.