Chromosomes and Gene Expression
Neuroscience

Avoiding false discoveries: Revisiting an Alzheimer's disease snRNA-Seq dataset

Alan E Murphy, **Nurun Nahar Fancy**, **Nathan G Skene**

UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK • Department of Brain Sciences, Imperial College London, London W12 0BZ, UK

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 2
November 16, 2023

Reviewed preprint version 1
September 1, 2023 (this version)

Sent for peer review
July 11, 2023

Posted to preprint server
June 14, 2023

Abstract

Arising From

Mathys, H. *et al.* Nature (2019). https://doi.org/10.1038/s41586-019-1195-2 Mathys *et al.*, conducted the first single-nucleus RNA-Seq study (snRNA-Seq) of Alzheimer's disease (AD)[1]. The authors profiled the transcriptomes of approximately 80,000 cells from the prefrontal cortex, collected from 48 individuals – 24 of which presented with varying degrees of AD pathology. With bulk RNA-Seq, changes in gene expression across cell types can be lost, potentially masking the differentially expressed genes (DEGs) across different cell types. Through the use of single-cell techniques, the authors benefitted from increased resolution with the potential to uncover cell type-specific DEGs in AD for the first time[2]. However, there were limitations in both their data processing and quality control and their differential expression analysis. Here, we correct these issues with best-practice approaches to snRNA-Seq processing and differential expression, resulting 892 times fewer differentially expressed genes at a false discovery rate (FDR) of 0.05.

eLife assessment

This work is **useful** as it highlights the importance of data analysis strategies in influencing outcomes during differential gene expression testing. While the manuscript has the potential to enhance awareness regarding data analysis choices in the community, its value could be further enhanced by providing a more comprehensive comparison of alternative methods and discussing the potential differences in preprocessing, such as scFLOW. The current analysis, although insightful, appears **incomplete** in addressing these aspects.

Main

Our questions of Mathys *et al.*,[1] focus around their data processing and their differential expression analysis. Firstly, in relation to their processing approach, the authors discussed the high percentages of mitochondrial reads and low number of reads per cell present in their data. This is indicative of low cell quality[3] however, we believe the authors' quality control (QC) approach did not capture all of these low quality cells. Moreover, the authors did not integrate the cells from different individuals to account for batch effects. As the field has matured since the author's work was published, dataset integration has become a standard step in single-cell RNA-Seq protocols[4]. To address these issues, we used scFlow[4] to reprocess the author's data. This pipeline included the removal of empty droplets, nuclei with low read counts and doublets followed by embedding and integration of cells from separate samples and cell typing. See Supplementary Note 1 for a detailed explanation of these steps. Re-processing resulted in 50,831 cells passing QC, approximately 20,000 less the authors' post-processing set. Additionally, we found different cell type proportions to the authors (**Fig. 1a**) which could be due to accounting for batch effects.

Our second question of Mathys *et al.*,[1] is their differential expression approach. The authors conducted a differential expression analysis between the controls and the patients with AD pathology, concentrating on six neuronal and glial cell types; excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes and oligodendrocyte precursor cells, derived from the Allen Brain Atlas[5]. They performed downstream analysis on their identified DEGs and investigated some of the most compelling genes in more detail. Therefore, all findings put forward by their paper were based upon the validity of their differential expression approach. However, for this approach, the authors conducted a two-part, cell and patient level analysis. The cell-level analysis took each cell as an independent replicate and the results of which were compared for consistency in directionality and rank of their DEGs against their patient level analysis, a Poisson mixed model. The authors identified 1,031 DEGs using this combinatorial approach – DEGs requiring an FDR <0.01 in the cell-level and an FDR<0.05 in the patient level analysis. It is important to note that this cell-level differential expression approach, also known as pseudoreplication, over-estimates the confidence in DEGs due to the statistical dependence between cells from the same patient not being considered[6,7,8]. When we inspect all DEGs identified at an FDR of 0.05 from the authors' cell-level analysis, this number increases to 14,274. Pseudobulk differential expression (DE) analysis has recently been proven to give optimal performance compared to both mixed models and pseudoreplication approaches[6,9]. It aggregates counts to individuals thus accounting for the dependence between an individual's cells.

Here, we apply a pseudobulk DE approach, sum aggregation and edgeR LRT[10], to the reprocessed data. We found 16 unique DEGs when considering the six cell types used by the authors (Supplementary Table 1). This was 892 times fewer DEGs than that reported originally at an FDR of 0.05. When we compare these DEGs, we can see that the absolute log2 fold change (LFC) of our DEGs is over 20 times larger than the authors'; median LFC of 3.33 and 0.16, despite the authors' DEGs having an FDR score 76,000 times smaller; median FDR of 2.89×10^{-7} and 0.02 (**Fig. 1b-c**). We can show that this stark contrast in these DEGs is just an artefact of the authors taking cells as independent replicates by considering the Pearson correlation between the number of DEGs found and the cell counts (**Fig. 1d-f**). There is a near perfect correlation for the authors DEGs, whether we consider the DEGs from the pseudoreplication analysis (**Fig. 1d**) or the 1,031 genes from the authors' combinatorial approach (**Fig. 1e**) which is not present in our re-analysis (**Fig. 1f**).

Although it has been proven that pseudoreplication approaches result in false positives by artificially inflating the confidence from non-independent samples, we wanted to investigate the effect of the approach on the authors' dataset. We ran the same cell-level analysis approach – a Wilcoxon rank-sum test and FDR multiple-testing correction, 100 times whilst randomly

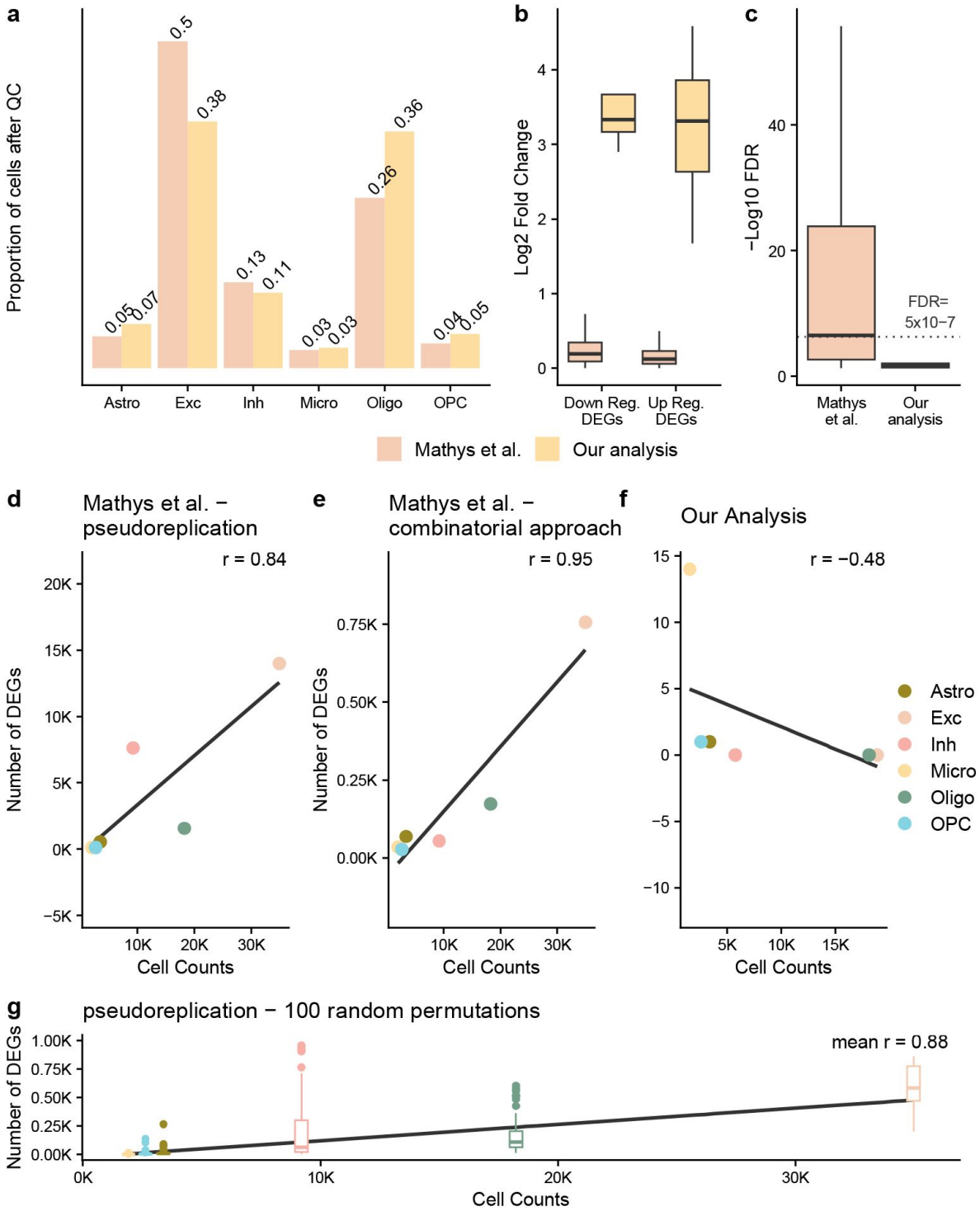

Fig 1

a shows the proportion of cells left after quality control (QC) from the author's original work (Mathys et al.) and our reanalysis (Our analysis). **b, c** highlights the log2 fold change and -log10 false discovery rate (FDR) of the differentially expressed genes. In **c**, we have marked an FDR of 5×10^{-7}, dashed grey line, to highlight how small the p-values from Mathys et al.'s analysis are. **d, e, f, g** show the Pearson correlation between the cell counts after QC and the number of DEGs identified. The different cell types are astrocytes (Astro), excitatory neurons (Exc), inhibitory neurons (Inh), microglia (Micro), oligodendrocytes (Oligo) and oligodendrocyte precursor cells (OPC).

permuting the patient identifiers (**Fig. 1g**). We would expect not to find any differential expressed genes with this approach given the random mixing of case and control patients. However, this pseudoreplication approach consistently found high numbers of DEGs and we observe the same correlation between the number of cells and number of DEGs as with the authors results. As a result, we conclude that integrating this pseudoreplication approach with a mixed model like the authors proposed just artificially inflates the test confidence for a random sample of the genes resulting in more false discoveries in cell types with bigger counts.

Conclusion

In conclusion, the authors' analysis has been highly influential in the field with numerous studies undertaken based on their results, something we show has uncertain foundations. However, we would like to highlight that the use of pseudoreplication in neuroscience research is not isolated to the author's work, others have used this approach[11,12,13] and their results should be similarly scrutinised. Here, we provide our processed count matrix with metadata and also, the DEGs identified using an independently validated, differential expression approach so that other researchers can use this rich dataset without relying on pseudoreplication approaches. While the number of DEGs found here are significantly lower, much greater confidence can be had that these are AD relevant genes. The low number of DEGs found may also cause concern given the sample size and cost of collection and sequencing of such datasets. However, the increasing number of snRNA-Seq studies being conducted for AD, creates the opportunity to conduct differential meta-analyses to increase power. Further work is required in the field to develop methods to conduct such analysis, integrating studies and accounting for their the heterogeneity, similar to that which has been done for bulk RNA-Seq[14]. Some such approaches have already been made in COVID-19 research which could be leveraged for neurodegenerative diseases[15].

Data availability

The differentially expressed genes and processed count matrix from the original study are available with their manuscript. The source data underlying **Figure 1** is provided as a Source Data file. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

The differential expression analysis pipeline is available at: *https://github.com/neurogenomics/reanalysis_Mathys_2019*. This is a general use pipeline which can be run for any single-nucleus or single-cell transcriptomic dataset. The config file containing all the parameters used for the scFlow run is also available in this repository.

Acknowledgements

This work was supported by a UKDRI Future Leaders Fellowship [grant number MR/T04327X/1] and the UK Dementia Research Institute, which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

Authors' Contributions

A.E.M and N.G.S jointly conceived and executed the study. N.N.F conducted the quality control and processing of the data and wrote up these steps. A.E.M wrote the manuscript which was reviewed by N.N.F and N.G.S.

Competing interests

The authors declare no competing interests.

References

1. Mathys H., et al. (2019) **Single-cell transcriptomic analysis of Alzheimer's disease** *Nature* **570**:332–337

2. De Strooper B., Karran E. (2016) **The Cellular Phase of Alzheimer's Disease** *Cell* **164**:603–615

3. Ilicic T., et al. (2016) **Classification of low quality cells from single-cell RNA-seq data** *Genome Biol* **17**

4. Khozoie C., et al. (2021) **Khozoie, C. et al. scFlow: A Scalable and Reproducible Analysis Pipeline for Single-Cell RNA Sequencing Data. 2021.08.16.456499 Preprint at 10.1101/2021.08.16.456499 (2021).** https://doi.org/10.1101/2021.08.16.456499

5. Tasic B., et al. (2018) **Shared and distinct transcriptomic cell types across neocortical areas** *Nature* **563**:72–78

6. Murphy A. E., Skene N. G. (2022) **A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis** *Nat. Commun* **13**

7. Zimmerman K. D., Espeland M. A., Langefeld C. D. (2021) **A practical solution to pseudoreplication bias in single-cell studies** *Nat. Commun* **12**

8. Lazic S. E. (2010) **The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?** *BMC Neurosci* **11**

9. Squair J. W., et al. (2021) **Confronting false discoveries in single-cell differential expression** *Nat. Commun* **12**

10. Chen Y., Lun A. T. L., Smyth G. K. (2016) **From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasilikelihood pipeline** *F1000Research* **5**

11. Fernandes H. J. R., et al. (2020) **Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses** *Cell Rep* **33**

12. Lui J. H., et al. (2021) **Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks** *Cell* **184**:489–506

13. Wakhloo D., et al. (2020) **Functional hypoxia drives neuroplasticity and neurogenesis via brain erythropoietin** *Nat. Commun* **11**

14. Rau A., Marot G., Jaffrézic F. (2014) **Differential meta-analysis of RNA-seq data from multiple studies** *BMC Bioinformatics* **15**

15. Garg M., et al. (2021) **Meta-analysis of COVID-19 single-cell studies confirms eight key immune responses** *Sci. Rep* **11**

Article and author information

Alan E Murphy

UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK, Department of Brain Sciences, Imperial College London, London W12 0BZ, UK
For correspondence: a.murphy@imperial.ac.uk
ORCID iD: 0000-0002-2487-8753

Nurun Nahar Fancy

UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK, Department of Brain Sciences, Imperial College London, London W12 0BZ, UK

Nathan G Skene

UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK, Department of Brain Sciences, Imperial College London, London W12 0BZ, UK
For correspondence: n.skene@imperial.ac.uk
ORCID iD: 0000-0002-6807-3180

Copyright

Editors

Reviewing Editor
Joon-Yong An
Korea University, Seoul, Korea, the Republic of

Senior Editor
Murim Choi
Seoul National University, Seoul, Korea, the Republic of

Reviewer #1 (Public Review):

Murphy, Fancy and Skene performed a reanalysis of snRNA-seq data from Alzheimer Disease (AD) patients and healthy controls published previously by Mathys et al. (2019), arriving at the conclusion that many of the transcriptional differences described in the original publication were false positives. This was achieved by revising the strategy for both quality control and differential expression analysis. I believe the authors' intention was to show the results of their reanalysis not as a criticism of the original paper (which can hardly be faulted for their strategy which was state-of-the-art at the time and indeed they took extra measures attempting to ensure the reliability of their results), but primarily to raise awareness and provide recommendations for rigorous analysis of sc/snRNA-seq data for future studies.

STRENGTHS:

The authors demonstrate that the choice of data analysis strategy can have a vast impact on the results of a study, which in itself may not be obvious to many researchers.

The authors apply a pseudobulk-based differential expression analysis strategy (essentially, adding up counts from all cells per individual and comparing those counts with standard RNA-seq differential expression tests), which is (a) in line with latest community recommendations, (b) different from the "default options" in most popular scRNA-seq analysis suites, and (c) explains the vastly different number of DEGs identified by the authors and the original publication. The recommendation of this approach together with a detailed assessment of the DEGs found by both methodologies could be a useful finding for the research community. Unfortunately, it is currently not fully substantiated and is confounded with concurrent changes in QC measures (see weaknesses).

The authors show a correlation between the number of DEGs and the number of cells assessed, which indicates a methodological shortcoming of the original paper's approach (actually, the authors of the original paper already acknowledged that the lesser number of DEGs for rare cell types was a technical artefact). To be educational for the reader it would be important to provide more information about the DEGs that were "found" and those that were "lost". Given vast inter-individual heterogeneity in humans, it is likely that the study was underpowered to find weaker differences using the pseudobulks (Fig. 1B shows that only genes with more than 4-fold change were found "significant").

All code and data used in this study are publicly available to the readers.

WEAKNESSES:

The authors interpret the fact that they found fewer DEGs with their method than the original paper as a good thing by making the assumption that all genes that were not found were false positives. However, they do not prove this, and it is likely that at least some genes were not found due to a lack of statistical power and not because they were actually "incorrect". The original paper also performed independent validations of some genes that were not found here.

I am concerned that the only DEGs found by the authors are in the rare cell types, foremost the rare microglia (see Fig. 1f). It is unclear to me how many cells the pseudo-bulk counts were based on for these cells types, but it seems that (a) there were few and (b) there were quite few reads per cells. If both are the case, the pseudobulk counts for these cell populations might be rather noisy and the DEG results are liable to outliers with extreme fold changes.

The authors claim they improved the quality control of the dataset. While I do not think they did anything wrong per se, the authors offer no objective metric to assess this putative improvement. This is another major weakness of the paper as it confounds the results of the improved (?) differential analysis strategy and dilutes the results. I detail this weakness in the two following points:

Removing low-quality cells: The authors apply a new QC procedure resulting in the removal of some 20k more cells than in the original publication. They state "we believe the authors' quality control (QC) approach did not capture all of these low quality cells" (l. 26). While all the QC metrics used are very sensible, it is unclear whether they are indeed "better". For instance, removal with a mitochondrial count of <5% seems harsh and might account for a large proportion of additional cells filtered out in comparison to the original analysis. There is no blanket "correct cutoff" for this percentage. For instance, the "classic" Seurat tutorial https://satijalab.org/seurat/articles/pbmc3k_tutorial.html uses the 5% threshold chosen by the authors, an MAD-based selection of cutoff arrived at 8% here https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html, another "best practices" guide choses by default 10% https://bioconductor.org/books/3.17/OSCA.basic/quality-control.html#quality-control-discarded, etc. Generally, the % of mitochondrial reads varies a lot between datasets. As far as I can tell, the original paper did not use a fixed threshold but instead used a

clustering approach to identify cells with an "abnormally high" mitochondrial read fraction. That also seems reasonable. Overall, I cannot assess whether the new QC is really more appropriate than the original analysis and the authors do not provide any evidence in favor of their strategy.

Batch correction: "Dataset integration has become a standard step in single-cell RNA-Seq protocols" (l. 29). While it is true that many authors now choose to perform an integration step as part of their analysis workflow, this is by no means uncontroversial as there is a risk of "over-integration" and loss of true biological differences. Also, there are many different methods for dataset integration out there, which will all have different results. More importantly, the authors go on "we found different cell type proportions to the authors (Fig. 1a) which could be due to accounting for batch effects" but offer no support for the claim that the batch effects are indeed related to the observed differences. An alternative explanation would be a selective loss/gain of certain cell types during quality control. The original paper stated concerns about losing certain cell types (microglia, which do not seem to be differentially abundant in the original paper / new analysis).

Relevant literature is incompletely cited. Instead of referring to reviews of best practices and benchmarks comparing methods for batch correction and or differential analysis, the authors only refer to their own previous work.

Due to a lack of comparison with other methods and due to the fact that the author's methodology was only applied to a single dataset, the paper presents merely a case study, which could be useful but falls short of providing a general recommendation for a best practice workflow.

APPRAISAL:

The manuscript could help to increase awareness of data analysis choices in the community, but only if the superiority of the methodology was clearly demonstrated. The recommended pseudobulk differential expression approach along with the indication of drastic differences that this might have on the results is the main output of the current manuscript, but it is difficult to assess unequivocally how this influenced the results because the differential analysis comes after QC and cell type annotation, which have also been changed in comparison to the original publication. In my opinion, the purpose of the paper might be better served by focusing on the DE strategy without changing QC and instead detailing where/how DEGs were gained/lost and supporting whether these were false positives.

- https://doi.org/10.7554/eLife.90214.1.sa1

Reviewer #2 (Public Review):

Summary: This paper takes on the important topic of preprocessing of single cell/nuclei RNA-seq prior to testing for differential gene expression. However, the manuscript has a number of critical weaknesses.

Strengths: This is an important topic and a key dataset for illustration.

Weaknesses: A major contribution is the use of the authors' own inhouse pipeline for data preparation (scFLOW), but this software is unpublished since 2021 and consequently not yet refereed. It isn't reasonable to take this pipeline as being validated in the field.

The authors claim that Mathys' analysis didn't use batch correction prior to analysis and claim that such processing is routine in the field, but the only citation they give is to the above-mentioned scFLOW. Batch correction for DEG analysis isn't the field standard, for example, Bryois et al. (2022) PMID: 35915177 doesn't perform batch correction. Whether or

not to do such preprocessing is certainly arguable, but the authors need to argue it, not presuppose it.

The authors spend considerable effort in discounting the pseudoreplication analysis of Mathys. It is well understood that this analysis yields a lot of false positives, but Mathys only used this approach for removing genes, not as a valid test in and of itself. They also worry that the significant findings in Mathys' paper are influenced by the number of cells of each type. I'm sure it is since power is a function of sample size, but is this a bad thing? It seems odd that their approach is not influenced by sample size.

- https://doi.org/10.7554/eLife.90214.1.sa0