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

Alan E Murphy[1,2], Nurun Nahar Fancy[1,2], Nathan G Skene[1,2]

[1] UK Dementia Research Institute at Imperial College London, London W12 0BZ, UK

[2] Department of Brain Sciences, Imperial College London, London W12 0BZ, UK

Correspondence should be addressed to A.M. (email: Alan Murphy: a.murphy@imperial.ac.uk) and N.S. (email: Nathan Skene: n.skene@imperial.ac.uk).

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.

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.

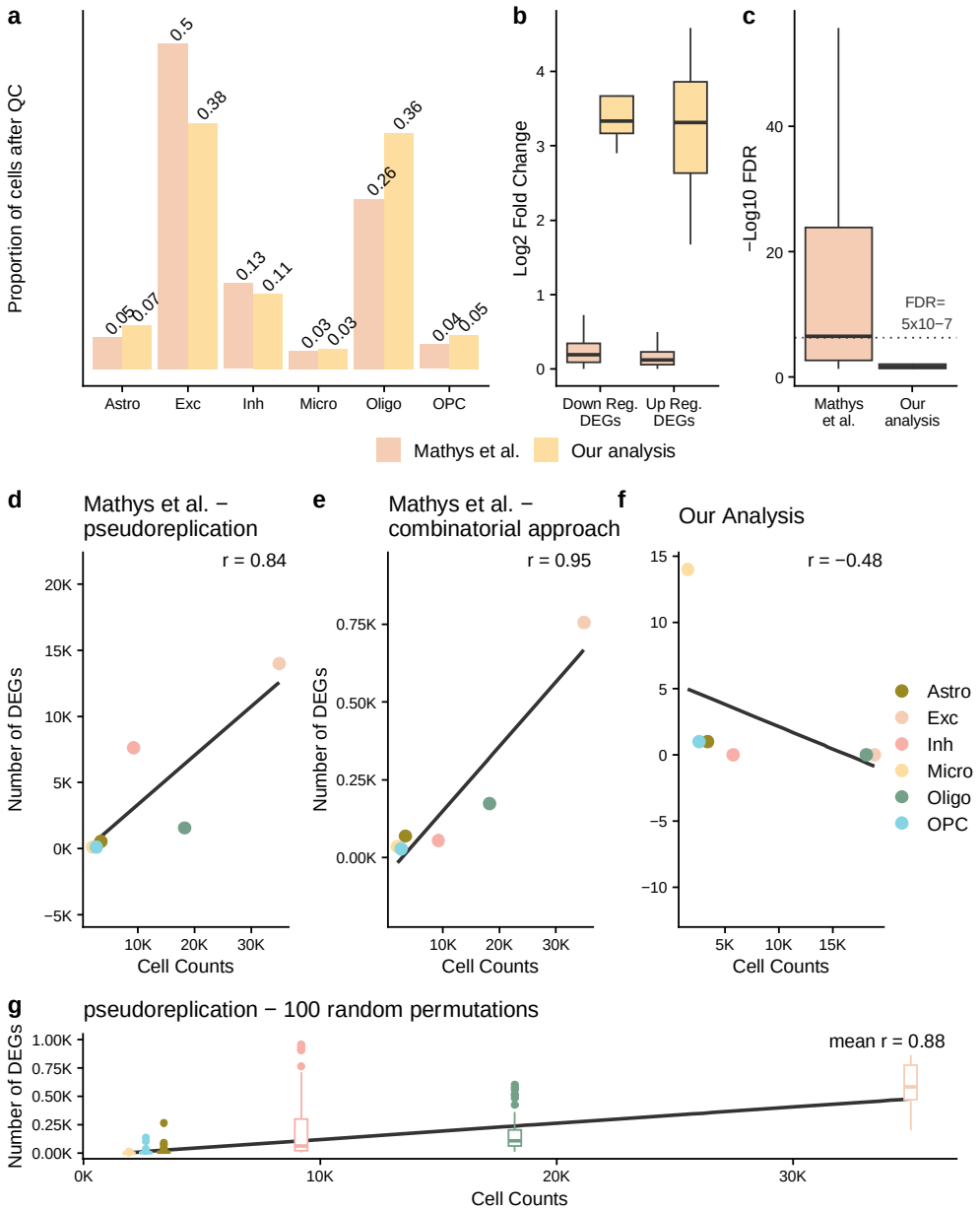

***Fig 1**: **a** shows the proportion of cells left after quality control (QC) from the author's original work (Mathys et al.) and our reanalysis (Our analysis). **b, c** highlights the log2 fold change and -log10 false discovery rate (FDR) of the differentially expressed genes. In **c**, we have marked an FDR of $5x10^{-7}$, 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).*

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 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.

References

1. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
2. De Strooper, B. & Karran, E. The Cellular Phase of Alzheimer's Disease. *Cell* **164**, 603–615 (2016).
3. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
4. Khozoie, C. *et al.* scFlow: A Scalable and Reproducible Analysis Pipeline for Single-Cell RNA Sequencing Data. 2021.08.16.456499 Preprint at https://doi.org/10.1101/2021.08.16.456499 (2021).
5. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
6. Murphy, A. E. & Skene, N. G. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* **13**, 7851 (2022).
7. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
8. Lazic, S. E. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* **11**, 5 (2010).
9. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
10. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* **5**, 1438 (2016).

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

12. Lui, J. H. *et al*. Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks. *Cell* **184**, 489-506.e26 (2021).

13. Wakhloo, D. *et al.* Functional hypoxia drives neuroplasticity and neurogenesis via brain erythropoietin. *Nat. Commun.* **11**, 1313 (2020).

14. Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* **15**, 91 (2014).

15. Garg, M. *et al.* Meta-analysis of COVID-19 single-cell studies confirms eight key immune responses. *Sci. Rep.* **11**, 20833 (2021).

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.