A balanced measure shows superior performance of pseudobulk methods over mixed models and pseudoreplication approaches in single-cell RNA-sequencing analysis

Alan E Murphy[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

Abstract

Sub-heading

Arising From: Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. Nature Communications (2021). https://doi.org/10.1038/s41467-021-21038-1

Summary

Recently, Zimmerman *et al*.,[1] proposed the use of mixed models over pseudobulk aggregation approaches, reporting improved performance on a novel simulation approach of hierarchical single-cell expression data. However, their reported results could not prove the superiority of mixed models as they are based on separate calculations of type 1 (performance of the models on non-differentially expressed genes) and type 2 error (performance on differentially expressed genes). To correctly benchmark the models, a reanalysis using a balanced measure of performance, considering both the type 1 and type 2 errors (both the differentially and non-differentially expressed genes), is necessary.

Contact

Alan Murphy: a.murphy@imperial.ac.uk, Nathan Skene: n.skene@imperial.ac.uk

Code availability

The modified version of hierarchicell which returns the Matthews correlation coefficient performance metric as well as the type 1 error rates, uses the same simulated data across approaches and has checkpointing capabilities (so runs can continue from where they left off if aborted or crashed) is available at: https://github.com/neurogenomics/hierarchicell.

The benchmarking script along with the results are available at: https://github.com/Al-Murphy/reanalysis_scRNA_seq_benchmark.

Introduction

The simulation approach of hierarchical single-cell expression data developed by Zimmerman *et al.,*[1] (hierarchicell) is used to generate non-differentially expressed genes upon which performance is evaluated using the type 1 error rate; the proportion of non-differentially expressed genes indicated as differentially expressed by a model. The authors determined that pseudobulk methods are "overly conservative" relative to mixed models however this can not be determined based on this analysis tool. In single-cell expression analysis, a conservative model does a poor job of capturing truly differentially expressed genes. Thus, such a model would have a high type 2 error rate. Certain methods' 1- type 2 error or power were calculated on specific simulated dataset sizes in Figure 3 and Supplementary Figures 5-11 of the authors' analysis. Importantly though, a systematic analysis of the models' type 2 error rates was not reported, therefore, the authors' statement cannot be concluded.

Considering the systematic analysis of the type 1 error results reported by Zimmerman *et al.,*[1] across the 20,000 iterations of 5 to 40 individuals and 50 to 500 cells at a p-value cut-off of 0.05, it was observed that pseudobulk approaches, in fact, have the lowest type 1 error at every iteration (Supplementary Figure 1). However, as previously outlined, we need a balanced measure of performance that considers both type 1 and type 2 error rate to correctly benchmark the models.

Main

Here, we modified Zimmerman *et al.*'s hierachicell approach to simulate both differentially expressed and non-differentially expressed genes. The differentially expressed genes were randomly simulated with a fold change between 1.1 and 10. We tested the models using the Matthews Correlation Coefficient (MCC) giving a balanced measure of performance as well as the type 1 error. MCC is a well-known and frequently adopted metric in the machine learning

field which offers a more informative and reliable score on binary classification problems[2]. MCC produces scores in [-1,1] and will only assign a high score if a model performs well on both non-differentially and differentially expressed genes. Moreover, MCC scores are proportional to both the size of the differentially and non-differentially expressed genes, so it is robust to imbalanced datasets. Furthermore, hierarchicell uses R's pseudo-random number functionality when generating the single-cell expression data, meaning each iteration will result in a different simulated dataset. However, Zimmerman *et al.'s* approach did not account for this in their benchmarks thus, their comparisons were not based on the same data. We further modified hierachicell to use the same simulated data for all models, enabling a fair comparison.

Our analysis demonstrates that pseudobulk approaches are the best performing across all number of individuals and cells variations (Figure 1). There is one exception for sum pseudobulk which performs worse than Tobit at 5 individuals and 10 cells. Figure 1 also highlights a trend whereby pseudoreplication models; Modified t, Tobit, Two-part hurdle: Default and Two-part hurdle: Corrected (which take cells as independent replicates) show degrading performance as the number of cells increase. This is likely due to the over-estimation of power driven by the dependence between cells from the same individual[3]. On the other hand, both sum and mean pseudobulk approaches (Pseudobulk: Mean and Pseudobulk: Sum) show improved performance as the number of cells increase. This trend is also noted in two of the other models; GEE1 and Tweedie: GLMM.

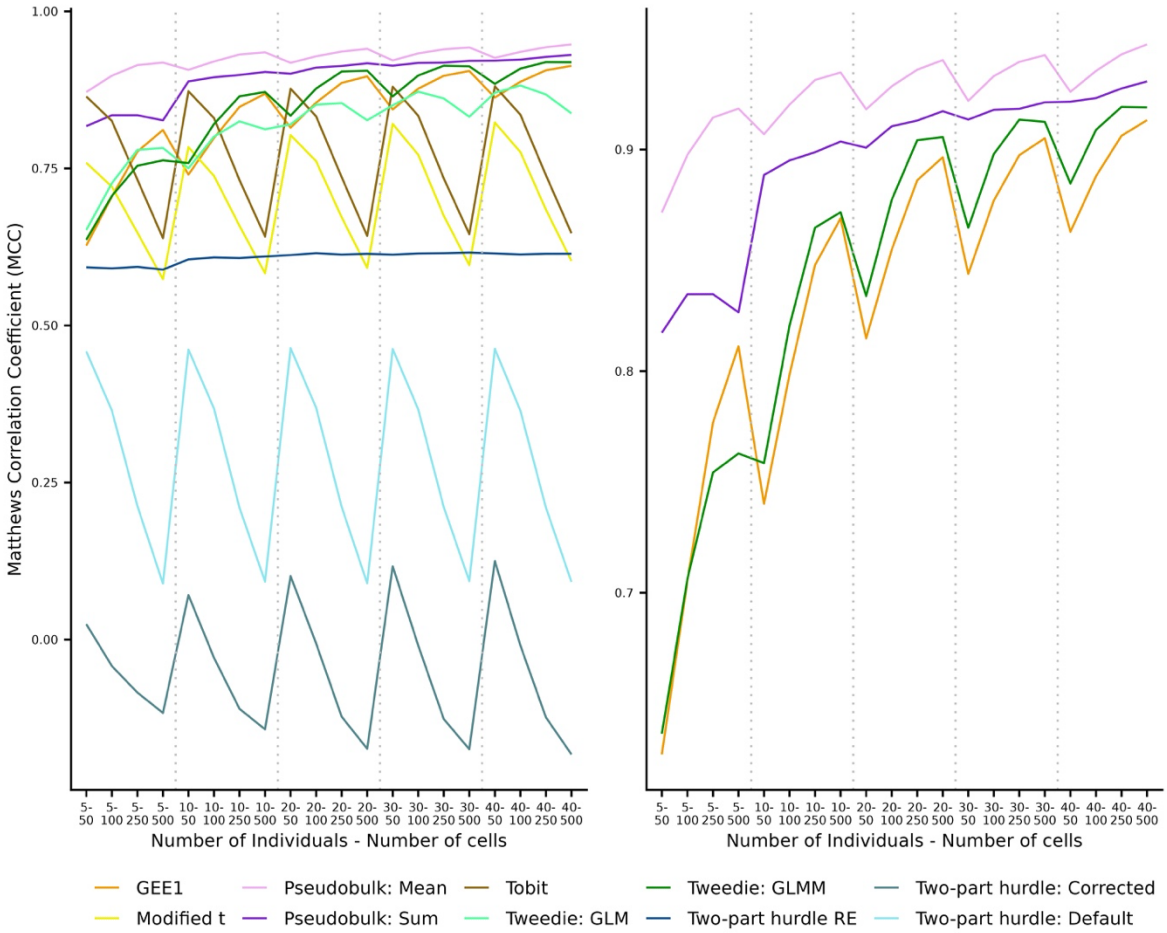


Figure 1*: The average Matthews correlation coefficient from the 20,000 iterations; 50 runs for each of the 5 to 40 individuals and 50 to 500 cells at a p-value cut-off of 0.05 on 10,000 genes. Left shows all benchmarked models whereas right focuses on the top four approaches. The different models are pseudoreplication approaches; Reproducibility-Optimized Statistical Testings - ROTS (Modified t), Monocole (Tobit) and model-based analysis of single-cell transcriptomics – MAST default, corrected, mean and sum pseudobulk approaches; from DESeq2 (Pseudobulk: Mean, Pseudobulk: Sum), generalised linear models: generalized estimating equation (GEE1) and generalized linear model with Tweedie distribution (Tweedie: GLM) and mixed model approaches; generalized linear mixed model with Tweedie distribution (Tweedie: GLMM) and model-based analysis of single-cell transcriptomics – MAST with a random effect for individuals (Two-part hurdle: RE). The performance split by each iteration is given in supplementary table 1.*

In real datasets there would never be equal numbers of cells in each sample. To mirror this, we simulated data with an imbalanced number of cells between case and controls. Pseudobulk mean outperformed all other approaches on this analysis (Supplementary Figure 2). The pseudobulk approach which aggregated by averaging rather than taking the sum appears to be the top performing overall, however, it is worth noting that hierarchicell does not normalise the simulated datasets before passing to the pseudobulk approaches. This is a standard step in such analysis to account for differences in sequencing depth and library sizes[5]. This approach was taken by Zimmerman *et al.* as their data is simulated one independent gene at a time without considering differences in library size. The effect of this step is more apparent on the imbalanced number of cells where pseudobulk sum's performance degraded dramatically. Pseudobulk mean appears invariant to this missing normalisation step because of averaging's own normalisation effect. It should be noted that this is a flaw in the simulation software strategy and does not show an improved performance of pseudobulk mean over sum.

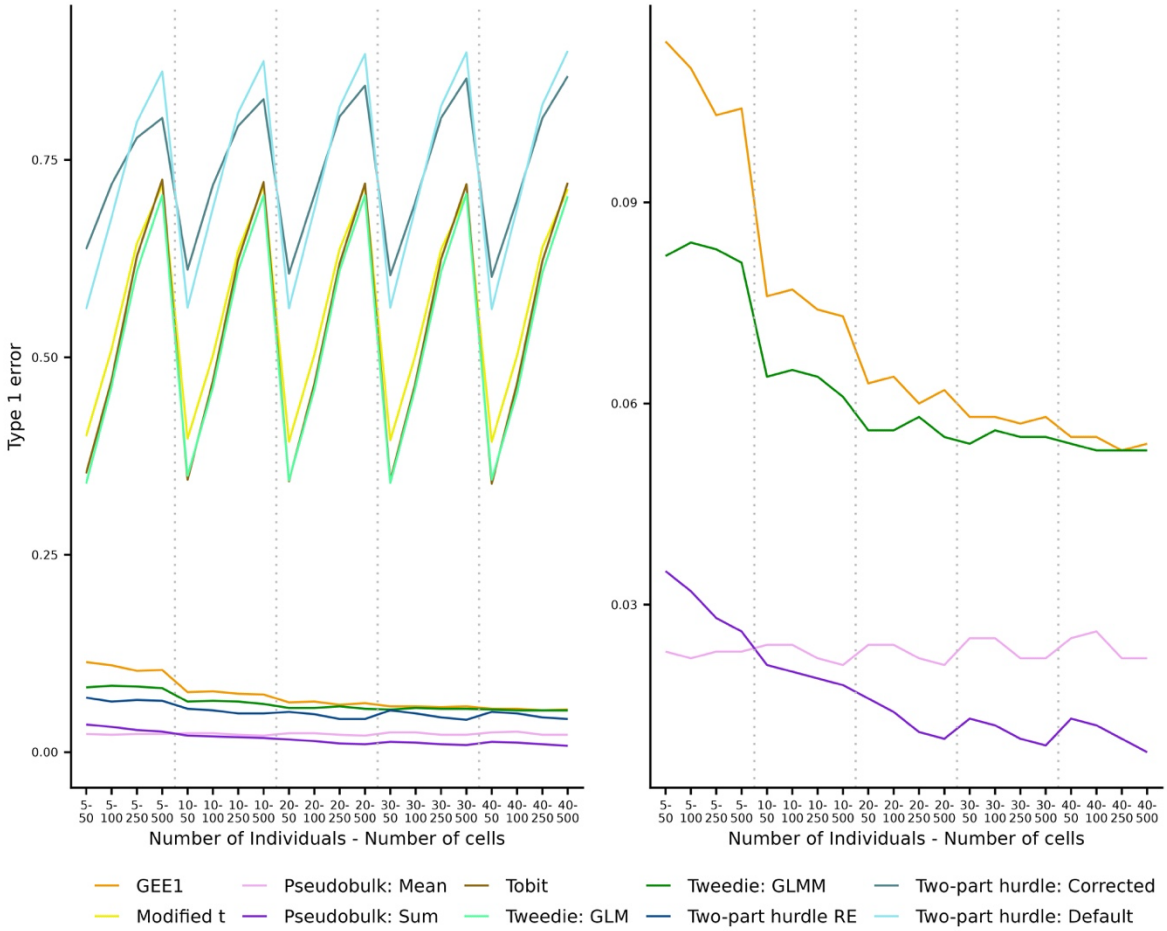
Pseudobulk approaches were also found to be the top performing approaches in a recent review by Squair *et al.,*[4]. Notably, the pseudobulk method used here; DESeq2[5], performed worse than other pseudobulk models in Squair *et al.,*'s analysis and so their adoption may further increase the performance of pseudobulk approaches on our dataset. Conversely, Squair *et al.,* did not consider all models included in our analysis or the different forms of pseudobulk aggregation. Therefore, our results on sum and mean pseudobulk extend their findings and indicate that mean aggregation may be the best performing. However, the reader should be cognisant that the lack of a normalisation step based on the flaw in the simulation software strategy likely causes the increased performance of mean over sum aggregation. Further, the use of simulated datasets in our analysis may not accurately reflect the differences between individuals that are present in biological datasets. Thus, despite both our results and those reported by Squair *et al.,* there is still room for further analysis, benchmarking more models, including different

combinations of pseudobulk aggregation methods and models, on more representative simulated datasets and biological datasets to identify the optimal approach. Specifically, we would expect pseudobulk sum with a normalisation step to outperform pseudobulk mean since it can account for the intra-individual variance which is otherwise lost with pseudobulk mean but this should be tested, including on imbalanced datasets.

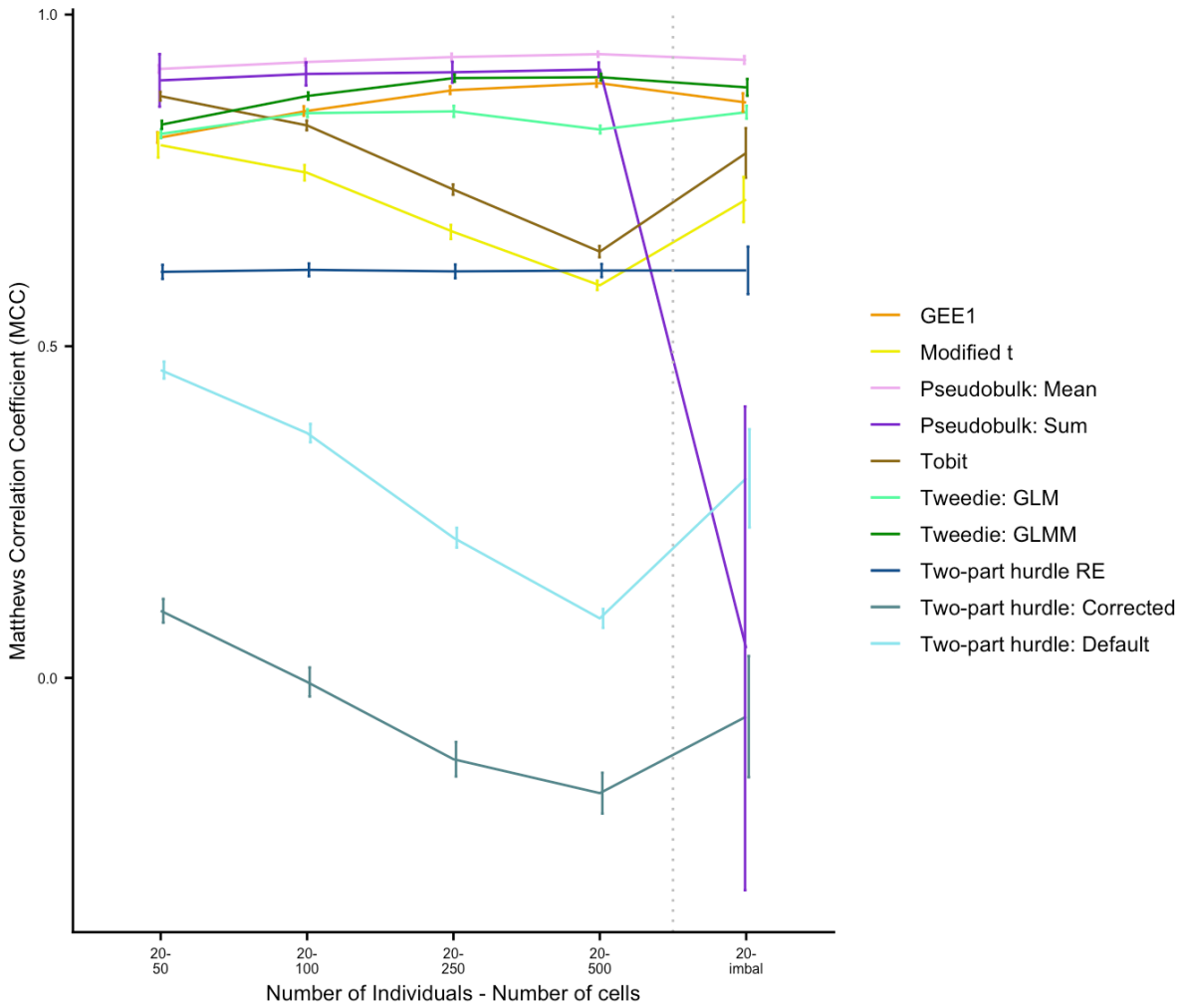
Conclusion

In conclusion, our results demonstrate that pseudobulk approaches are far from being too conservative and are, in fact, the best performing models based on this simulated dataset for the analysis of single-cell expression data.

Supplementary Figures



Supplementary Figure 1*: The average type 1 error from the 20,000 iterations; 50 runs for each of the 5 to 40 individuals and 50 to 500 cells at a p-value cut-off of 0.05 on 5,000 genes reported by Zimmerman et al.[1]. Left shows all benchmarked models whereas right focuses on the top four approaches. The different models are pseudoreplication approaches; Reproducibility-Optimized Statistical Testings - ROTS (Modified t), Monocole (Tobit) and model-based analysis of single-cell transcriptomics – MAST default, corrected, mean and sum pseudobulk approaches; from DESeq2 (Pseudobulk: Mean, Pseudobulk: Sum), generalised linear models: generalized estimating equation (GEE1) and generalized linear model with Tweedie distribution (Tweedie: GLM) and mixed model approaches; generalized linear mixed model with Tweedie distribution (Tweedie: GLMM) and model-based analysis of single-cell transcriptomics – MAST with a random effect for individuals (Two-part hurdle: RE).*



Supplementary Figure 2*: The average Matthews correlation coefficient of all benchmarked models across all balanced number of cells and the imbalanced number of cells for 20 individuals; 50 runs for each at a p-value cut-off of 0.05 on 5,000 genes. The number of cells were randomly chosen using a gamma distribution with shape 4 and scale 45 separately for cases and controls to produce the imbalanced dataset (giving a mean 150-200 cells). The error bars give 1 standard deviation around the mean. The different models are pseudoreplication approaches; Reproducibility-Optimized Statistical Testings - ROTS (Modified t), Monocole (Tobit) and model-based analysis of single-cell transcriptomics – MAST default, corrected, mean and sum pseudobulk approaches; from DESeq2 (Pseudobulk: Mean, Pseudobulk: Sum), generalised linear models: generalized estimating equation (GEE1) and generalized linear model with Tweedie distribution (Tweedie: GLM) and mixed model approaches; generalized linear mixed model with Tweedie distribution (Tweedie: GLMM) and model-based analysis of single-cell transcriptomics – MAST with a random effect for individuals (Two-part hurdle: RE).*

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.

References

1. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
2. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
3. Lazic, S. E. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* **11**, 5 (2010).
4. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
5. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).