propeller: testing for differences in cell type proportions in single cell data

Belinda Phipson[1,2,3], Choon Boon Sim[4,5], Enzo Porrello[4,5,6], Alex W Hewitt,[7,8] Joseph Powell[9,10] and Alicia Oshlack[1,3,11,*]

[1] Peter MacCallum Cancer Centre, Melbourne, Victoria, 3000, Australia

[2] Department of Pediatrics, University of Melbourne, Parkville, Victoria, 3010, Australia

[3] Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, 3010, Australia

[4] Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne 3052, VIC, Australia.

[5] Melbourne Centre for Cardiovascular Genomics and Regenerative Medicine, The Royal Children's Hospital, Melbourne 3052, VIC, Australia.

[6] Department of Anatomy and Physiology, School of Biomedical Sciences, The University of Melbourne, Melbourne 3010, VIC, Australia.

[7] Menzies Institute for Medical Research, School of Medicine, University of Tasmania, Tasmania, Australia

[8] Centre for Eye Research Australia, The University of Melbourne, Victoria, Australia

[9] Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, 2010, Australia;

[10] UNSW Cellular Genomics Futures Institute, University of New Souith Wales, Kingston, NSW, 2052, Australia

[11] School of Biosciences, University of Melbourne, Parkville, Victoria, 3010, Australia

[*] corresponding author

Email addresses:

Alicia Oshlack: Alicia.Oshlack@petermac.org

Belinda Phipson: phipson.b@wehi.edu.au

Abstract:

Single cell RNA Sequencing (scRNA-seq) has rapidly gained popularity over the last few years for profiling the transcriptomes of thousands to millions of single cells. To date, there are more than a thousand software packages that have been developed to analyse scRNA-seq data. These focus predominantly on visualization, dimensionality reduction and cell type identification. Single cell technology is now being used to analyse experiments with complex designs including biological replication. One question that can be asked from single cell experiments which has not been possible to address with bulk RNA-seq data is whether the cell type proportions are different between two or more experimental conditions. As well as gene expression changes, the relative depletion or enrichment of a particular cell type can be the functional consequence of disease or treatment. However, cell type proportions estimates from scRNA-seq data are variable and statistical methods that can correctly account for different sources of variability are needed to confidently identify statistically significant shifts in cell type composition between experimental conditions. We present *propeller*, a robust and flexible method that leverages biological replication to find statistically significant differences in cell type proportions between groups. The *propeller* method is publicly available in the open source `speckle` R package (https://github.com/Oshlack/speckle).

1. Introduction

Single cell RNA-sequencing (scRNA-seq) technology has led to breakthroughs in the discovery of novel cell types and enhanced our understanding of the development of complex tissues. As the technology has matured it has become relatively straightforward to profile the transcriptomes of hundreds of thousands of cells, resulting in valuable insight into the composition of tissues.

While many of the first published single cell papers focused on defining the resident cell types in complex tissues[1–4], the field is now using this technology for complex experimental comparisons with biological replication[5–8]. While experiments with different conditions and multiple biological samples can be costly, substantial savings can be made by pooling cells from multiple samples. If samples are genetically diverse, they can be demultiplexed using genetic information[9,10]. An alternative approach is to use molecular cell multiplexing protocols, some of which are now commercially available, e.g. CellPlex from 10x Genomics. Collectively, cell multiplexing makes designing larger scRNA-seq experiments more feasible.

While the first step in analysis for a scRNA-seq experiment with multiple experimental conditions and biological replicates is to identify the cell types present in each sample, downstream analysis requires sophisticated tools to address specific hypotheses about how a perturbation affects the biological system. Two analysis tasks are commonly performed following cell type identification in order to understand the effect of the condition. One task is to find genes that are differentially expressed between groups of samples, for every cell type observed in the experiment, similar to the analysis of bulk RNA-seq experiments[11]. However, a benefit of scRNA-seq data is that we have additional information on the composition of the samples. The relative change in abundance of a cell type can be a consequence of disease or treatment. Due to technical as well as biological sources, the cell type proportions estimates can be quite variable. The focus of this work is to find statistically significant differences in cell type proportions between groups of samples that appropriately takes into account sample-to-sample variability. Here we present *propeller*, a robust and flexible linear modelling based solution to test for differences in cell type proportions between experimental conditions. Our *propeller* method is publicly available in the `speckle` R package (https://github.com/Oshlack/speckle).

2. Methods

Propeller is a function in the `speckle` R package that uses cell level annotation information to calculate sample level cell type proportions, followed by data transformation and statistical testing for each cell type. *Propeller* leverages biological replication to estimate the high sample-to-sample variability in cell type counts often observed in real single cell data (Figure 1a). While the variability in cell type proportions estimates between samples can be due to technical sources such as variation in dissociation protocols, there may be valid biological reasons for variations. For example, blood cell type composition is known to change with age[12]. Taking into account sample-to-sample variability when analysing differences in cell type proportions is critical as observed cell type variances are far greater than variances estimated under a binomial or Poisson distribution, which can only account for sampling variation (Figure 1b).

The first step of *propeller* is to calculate the cell type proportions for each sample which can be derived from a Seurat or SingleCellExperiment object. This results in a matrix of proportions where the rows are the cell types and the columns are the samples. The binomial distribution has the property that proportions close to 0 and 1 have small variance, and values close to 0.5 have large variance i.e. the variances are heteroskedastic. To overcome this, we have implemented two transformations in *propeller*: (1) arcsin square root transformation, and (2) logit transformation. While the arcsin square root transformation is not as effective at stabilising variances as the logit transformation, it will always produce a real value. If the logit transformation is selected an offset of 0.5 is added to the raw cell type counts matrix prior to transformation.

Next we test whether the transformed proportions for every cell type are significantly different between two or more experimental conditions. If there are exactly two groups, we perform moderated t-tests; if there are more than two groups, we perform moderated ANOVA tests[13]. These tests are moderated using an empirical Bayes framework, allowing information to be borrowed across cell types to stabilise the cell type specific variance estimates. This is particularly effective when the number of biological replicates is small, currently a common situation in scRNA-seq experiments. The final step in *propeller* is to calculate false discovery rates[14] to account for testing across multiple cell types. The output of *propeller* consists of

condition specific proportions estimates, *p*-values and false discovery rates for every cell type observed in the experiment.

The minimal annotation information that *propeller* requires for each cell is cluster/cell type, sample and group/condition, which can be automatically extracted from Seurat and SingleCellExperiment class objects. More complex experimental designs can be accommodated using the *propeller.ttest* and *propeller.anova* functions.

3. Simulating cell type counts data

We simulated cell type counts from a beta-binomial distribution. The parameters α and β of the beta distribution were estimated from real data (Figure 1a, heart single nuclei RNA-seq) using the *estimateBetaParamsFromCounts* function available in `speckle`. We compared the performance of *propeller* to other commonly used statistical models for analysing differences in proportions that can take into account biological variability[15–19]. A dataset with two groups of five samples and seven cell types of varying abundance was simulated resembling the heart single nuclei RNA-seq (snRNA-seq) dataset. Three cell types change proportions between the two groups by 2-3 fold, while the remaining four cell types do not. Across 1,000 simulated datasets, we found that the power to detect true differences in proportions was strongly influenced by cell type abundance (Figure 1c). Methods based on the negative binomial distribution had reduced power to detect changes in cell types with larger abundances, while *propeller* with arcsin square root transform had reduced power to detect differences in cell types that were relatively rare. *Propeller* with logit transform and the beta-binomial model performed comparably well across the range of cell type abundances.

4. Application to real single cell datasets

Complex experimental designs can be modelled using the *propeller* functions. In order to demonstrate the types of experimental designs that can be accommodated, we applied *propeller* to three different scRNA-seq datasets:

1. 20 PBMC samples across young and old male and female samples[6]. We modelled age and sex as categorical variables.

2. 9 human heart biopsy samples across development (fetal, young, adult)[5]. We modelled development as a continuous variable and sex as a categorical variable.
3. 13 bronchoalveolar lavage fluid immune cell samples across three groups (healthy controls, moderate and severe COVID-19 infection)[20]. We modelled disease status as a categorical variable and performed an ANOVA to find cell type differences between the three groups.

Figure 1d-f shows examples of statistically significant cell types for the three different datasets and models using *propeller* with a logit transformation. In the study of the immune system landscape, CD8 naive cells are depleted in old samples compared to young, taking sex into account (Figure 1d). Across healthy human heart development, the relative proportion of cardiomyocytes decline with age (Figure 1e). In a study of moderate and severe COVID-19, we find neutrophils are significantly different between healthy controls and moderate and severe bronchoalveolar lavage samples from COVID-19 patients (Figure 1f). These studies highlight the different types of analysis that can be performed using the *propeller* functions.

5. Conclusions

Propeller is a method for testing for differences in cell type proportions from single cell data. It takes account of sample-to-sample variability, which is large due to both technical and biological sources. Features of *propeller* include:

- Easy implementation suitable for novices and experienced users
- Data extraction from Seurat and SingleCellExperiment class objects
- Options for different transformations of proportions
- Ability to model complex experimental designs, including mixed effects models
- Empirical Bayes variance estimation
- Interpretable output
- Plotting functions for cell type counts and cell type proportion dispersion estimates
- Extensive user guide in the form of a vignette

Funding

BP is funded by an NHMRC Investigator grant GNT1175653, AO is funded by an NHMRC Investigator grant GNT1196256, JP is funded by an NHMRC Investigator grant GNT1175781, and ERP is funded by an NHMRC Investigator grant GNT2008376. The work was supported by NHMRC Grant GNT1187748. The snRNA-seq dataset for heart was funded by Royal Children's Hospital Foundation and NHMRC Project grant GNT1160257.

References

1. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
2. Bornstein, C. *et al.* Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells. *Nature* **559**, 622–626 (2018).
3. Liu, Y. *et al.* Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res.* **28**, 819–832 (2018).
4. Combes, A. N. *et al.* Single cell analysis of the developing mouse kidney provides deeper insight into marker gene expression and ligand-receptor crosstalk. *Development* **146**, (2019).
5. Sim, C. B. *et al.* Sex-Specific Control of Human Heart Maturation by the Progesterone Receptor. *Circulation* **143**, 1614–1628 (2021).
6. Huang, Z. *et al.* Effects of sex and aging on the immune cell landscape as assessed by single-cell transcriptomic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
7. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913.e19 (2021).
8. Bunis, D. G. *et al.* Single-Cell Mapping of Progressive Fetal-to-Adult Transition in Human Naive T Cells. *Cell Rep.* **34**, 108573 (2021).

9. Xu, J. *et al.* Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.* **20**, 290 (2019).

10. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).

11. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).

12. Tan, Q. *et al.* Handling blood cell composition in epigenetic studies on ageing. *International journal of epidemiology* vol. 46 1717–1718 (2017).

13. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).

14. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

15. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

16. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

17. Chen, Y., Pal, B., Visvader, J. E. & Smyth, G. K. Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. *F1000Res.* **6**, 2055 (2017).

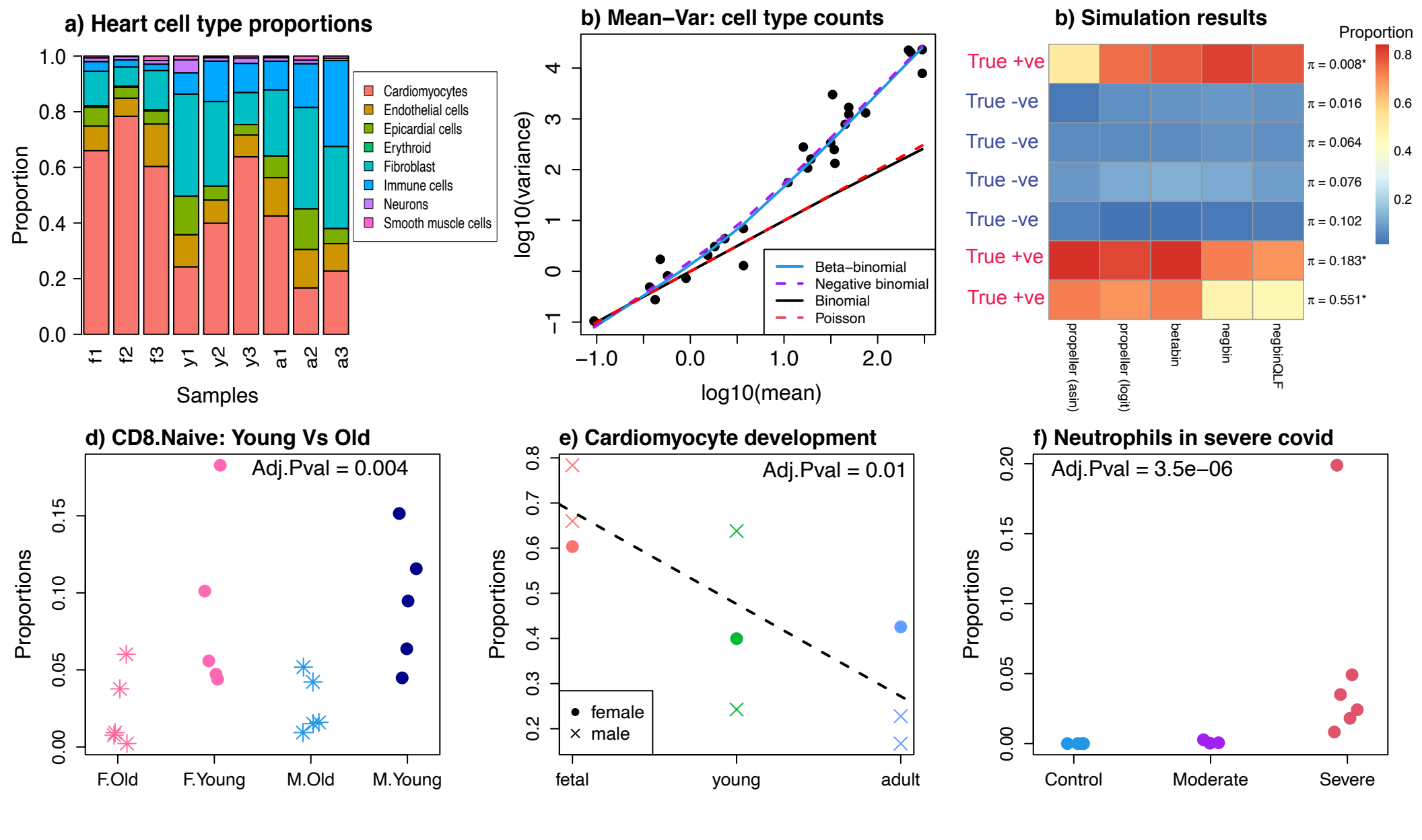
18. Martin, B. D., Witten, D. & Willis, A. D. MODELING MICROBIAL ABUNDANCES AND DYSBIOSIS WITH BETA-BINOMIAL REGRESSION. *Ann. Appl. Stat.* **14**, 94–115

(2020).

19. Orchestrating Single-Cell Analysis with Bioconductor. https://bioconductor.org/books/release/OSCA/.

20. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).

Figure 1. Analysing cell type proportions from single cell RNA-seq data with *propeller*. a. Barplot showing high levels of variability of cell type proportion estimates between nine samples in a human heart development snRNA-seq dataset. b. Mean-variance relationship for 27 cell types in 12 healthy PBMC scRNA-seq samples showing that cell type counts are over-dispersed. The plot is produced using the *plotCellTypeMeanVar* function in the `speckle` package. c. Heatmap showing the proportion of simulated datasets with p-value < 0.05 for each of the seven cell types for each of the 5 methods. For the true positives, dark red indicates greater power to detect significant cell type differences between two groups (proportion is high). For the true negatives, dark blue indicates good false discovery rate control (proportion is low). The methods tested are *propeller* with arcsin square root transform, *propeller* with logit transform, beta-binomial regression, negative binomial regression and quasi-likelihood negative binomial regression. d. There is a statistically significant difference in the proportions of CD8 naive cells between young and old PBMC samples, taking sex into account. e. Treating developmental stage as a continuous variable, the cardiomyocyte populations show a relative decline across development in human heart samples. f. Neutrophils are statistically significantly different between healthy control, moderate and severe COVID-19 bronchoalveolar lavage samples.


a) Heart cell type proportions
Proportion
Samples
f1
f2
f3
y1
y2
y3
a1
a2
a3
Cardiomyocytes
Endothelial cells
Epicardial cells
Erythroid
Fibroblast
Immune cells
Neurons
Smooth muscle cells
b) Mean−Var: cell type counts
log10(variance)
log10(mean)
Beta−binomial
Negative binomial
Binomial
Poisson
b) Simulation results
Proportion
True +ve
True -ve
True -ve
True -ve
True -ve
True +ve
True +ve
π = 0.008*
π = 0.016
π = 0.064
π = 0.076
π = 0.102
π = 0.183*
π = 0.551*
propeller (asin)
propeller (logit)
betabin
negbin
negbinQLF
d) CD8.Naive: Young Vs Old
Adj.Pval = 0.004
Proportions
F.Old
F.Young
M.Old
M.Young
e) Cardiomyocyte development
Adj.Pval = 0.01
Proportions
female
male
fetal
young
adult
f) Neutrophils in severe covid
Adj.Pval = 3.5e−06
Proportions
Control
Moderate
Severe