Supplementary Materials to the paper by Wang M. et al. (Guidelines for Bioinformatics of Single-Cell Sequencing Data Analysis in Alzheimer's Disease: Review, Recommendation, Implementation and Application. (2022) *Molecular Neurodegeneration*)

Correspondence:
Bin Zhang, PhD
Professor
Department of Genetics & Genomic Sciences
Department of Pharmacological Sciences
Mount Sinai Center for Transformative Disease Modeling
Icahn School of Medicine at Mount Sinai
1425 Madison Avenue, New York, NY 10029
(O) 212-659-1726 (x81726)(F) 212-659-5507
Email: bin.zhang@mssm.edu

Power Analysis for scRNA-seq Study Design

[Authors: Won-min Song, Minghui Wang, Xian Xiao Zhou and Bin Zhang]

Power analysis is a critical step to rationalize scRNA-seq experimental design to ensure robustness and reproducibility of scientific findings such as cell type proportion changes or rare cell types (cell type discovery), differential expression, and expression quantitative loci (eQTLs). However, the established power calculation approaches for bulk RNA-seq studies cannot be directly adapted to scRNA-seq data due to the sparsity in count matrices by dropout events and low cell-wise sequencing depths[1,2], cell type-specific expressions, and inherently different data distributions[3,4]. Several key parameters need to be determined to ensure statistically powerful experimental design, such as the number of cells per sample, cell-wise sequencing depth, and sample size, given the budget constraint[5]. These parameters include scRNA-seq experimental protocols, primarily tag-based protocols that leverage unique molecular identifiers (UMIs) to sequence large pool of cells with shallow read depths per cell[6] and full-length protocols aimed at smaller pool of cells with deep cell-wise sequencing.

Assuming that cells in different subpopulations or cell types are sampled according to a multinomial distribution, Single-Cell One-sided Probability Interactive Tool (SCOPIT) provides an R package *pmultinom* along with an interactive web application to estimate the number of cells required for sequencing, conditioned on the frequency of rarest subpopulation[7]. SCOPIT is similar to a previously unpublished web application at https://satijalab.org/howmanycells. However, these methods do not compute the power of distinguishing cell subpopulations, a process affected by the level of gene expression difference between subpopulations and other scRNA-seq data parameters such as dropout rate. To help determine the power of identifying cell subpopulations in scRNA-seq experiments, Kim e*t al.*[8] proposed in their R package *ncells* to model the separation of cell subpopulations by measuring the distance between the cell subpopulation centers projected onto the first principal component of the sample data. *ncells* uses simulation to estimate the statistical power of detecting subpopulation separation at a given number of sequenced cells from multiple user-defined parameters, including the proportion of a cell subpopulation of interest, the proportion of up-regulated marker genes of a subpopulation, fold change of subpopulation specific marker genes' expression , and overall dropout rates. A number of simulation-based approaches such as powsimR[9], scDesign[10], Hierarchicell[11] and muscat[12] have been developed to estimate the statistical power of identifying cell-type specific DEGs conditioned on effect sizes (e.g. fold changes in differential expression detection). These methods first infer the key parameters of generative models of single-cell transcriptome, accounting for sparsity and dropout events. The fitted models enable scRNA-seq data simulation with control over the key aspects of experimental design such as sample size, the number of cells per sample, and library size. Then, established differential expression analysis approaches (e.g. voom-limma[13], MAST[14], edgeR[15]) are applied to the simulated data at the cell level[9,10] or pseudo-bulk level by combining cells from a sample[12], to

evaluate the accuracy of differential expression under a desired experimental scenario. The results determine the required number of cells per cell type/state to robustly identify differential expression signals with a desired power. However, the simulation-based approaches are computationally expensive and are prohibitive for studies involving a large number of cells in data simulation[16].

Alternatively, analytical approaches, via fitting the model parameters for differential expression detection rates from adequate training data, can substantially reduce the computational burden and extrapolate the power estimation to large-scale single-cell studies. To this end, Schmid *et al.* developed an analytical solution to estimate the power by using DESeq approach to detect differential expression [17] from cell type-wise 'pseudo-bulk' data[16], leading to an R software package scPower. The key design parameters such as cell-wise read depth, sample size, number of cells per sample, number of cells per cell type and dropout rates are parametrized in scPower, conditioned on priors learned from pilot datasets and experimental design parameters. Especially, the cost-related parameters in generating scRNA-seq data (read depth per cell, number of cells per sample, sample sizes) can optimize the experimental design under a given budget. However, the efficiency of analytical solutions is achieved at the expense of learning priors from appropriate pilot data, which should be context-specific (e.g., tissue type and disease type)[16].

Overall, the power calculation need take into account scRNA-seq specific noises, different sequencing protocols, and study objectives etc. For example, Li and Li considered robust detection of glial subpopulations, astrocytes and oligodendrocytes, and applied scDesign to evaluate the optimal number of cells to identify differential expression between the cell types[10]. They estimated that 512 cells per cell type is required in the Fluidigm C1 protocol, and 4,096 cells when the inDrop protocol is used[10]. The larger optimal cell number within the inDrop protocol originates from the

shallow read depth per cell, thus requiring more cells to reach the same level of differential expression detection accuracy as the Fluidigm C1 protocol[10].

Recommended workflow and applications to AD

As AD is a very heterogeneous disease[18] with varying presence of key cell types in different brain regions at different disease stages[19] [20], the key study parameters such as number of cells per sample and cell-wise read depth should be optimized on the basis of the key cell populations of interest. We recommend *ncells* for estimating the cell number required to achieve a given power to identify a cell subpopulation from single cell RNA-seq experiments, as it accounts for scRNA-eq specific noises such as high dropout rate.

Then, the statistical analyses of interest can also substantially affect the key sequencing parameters. For example, eQTL studies require greater number of reads to reliable detect expression changes and respective allele-specific expressions, contingent on genetic variants of interest[16]. On contrary, differential expression analysis within each cell type will affect the study design by requiring the presence of certain cell types per experimental condition. To this end, we recommend scPower which provides a very comprehensive set of toolboxes to estimate sequencing parameters, contingent on the budget constraints and statistical analyses of interest[16]. As an example, we applied scPower to 4 samples (2 ADs and 2 controls) from the ROSMAP cohort[21] to estimate the parameters required for power analysis, including the negative binomial distribution parameters for each gene, the gamma mixed distribution parameters for each sample, the parameters of the gamma fits by the mean UMI counts per cell, and finally the median dispersion function for each cell type. Based on these parameter estimates, we calculated the power under a range of sample sizes, cell numbers per sample, cell type frequencies and read depths to

accommodate various experimental scenarios and provided the workflow as tutorial (https://songw01.github.io/AD_scRNAseq_companion/vignettes/scPower.ROSMAP.html).

References

1 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).

2 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

3 Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17**, 75, doi:10.1186/s13059-016-0947-7 (2016).

4 Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* **15**, 539-542, doi:10.1038/s41592-018-0033-z (2018).

5 Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nature Communications* **11**, 774, doi:10.1038/s41467-020-14482-y (2020).

6 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74, doi:10.1038/nmeth.1778 (2011).

7 Davis, A., Gao, R. & Navin, N. E. SCOPIT: sample size calculations for single-cell sequencing experiments. *BMC Bioinformatics* **20**, 566, doi:10.1186/s12859-019-3167-9 (2019).

8 Kim, K. I., Youn, A., Bolisetty, M., Palucka, A. K. & George, J. Calculating sample size for identifying cell subpopulation in single-cell RNA-seq experiments. *bioRxiv*, 706481, doi:10.1101/706481 (2019).

9 Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486-3488, doi:10.1093/bioinformatics/btx435 (2017).

10 Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41-i50, doi:10.1093/bioinformatics/btz321 (2019).

11 Zimmerman, K. D. & Langefeld, C. D. Hierarchicell: an R-package for estimating power for tests of differential expression with single-cell data. *BMC Genomics* **22**, 319, doi:10.1186/s12864-021-07635-w (2021).

12 Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun* **11**, 6077, doi:10.1038/s41467-020-19894-4 (2020).

13 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

14 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).

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, doi:10.1093/bioinformatics/btp616 (2010).
16 Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat Commun* **12**, 6625, doi:10.1038/s41467-021-26779-7 (2021).
17 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
18 Neff, R. A. *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* **7**, doi:10.1126/sciadv.abb5398 (2021).
19 Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290 e1217, doi:10.1016/j.cell.2017.05.018 (2017).
20 Batiuk, M. Y. *et al.* Identification of region-specific astrocyte subtypes at single cell resolution. *Nat Commun* **11**, 1220, doi:10.1038/s41467-019-14198-8 (2020).
21 Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337, doi:10.1038/s41586-019-1195-2 (2019).