Supplementary File - Chapter 5: 'Probabilistic index models for testing differential expression in single cell RNA sequencing data'

Alemu Takele Assefa, Jo Vandesompele, Olivier Thas

April 10, 2019

Contents

1 Additional simulation results

1.1 Simulations

In this section we show results of an empirical comparison between the real data and simulated data (Splat (Zappia, Phipson, and Oshlack 2017) and SPsimSeq (Assefa, Vandesompele, and Thas 2019) methods). In particular, we look at the distribution of the fraction of zero counts per gene, cell to cell similarity, the relationship between log mean expression and fraction of zero counts, and the relationship between log mean expression and coefficients of variation.

In this supplementary file, we show only a subset of the results. More results are available at (Assefa, Vandesompele, and Thas 2019). On the following pages, we show the comparison for the two simulations: scRNA-seq data with read-counts (source data processed with SMARTer/C1 protocol = dataset A), and UMI-counts (generated using Chromium protocol = dataset B).

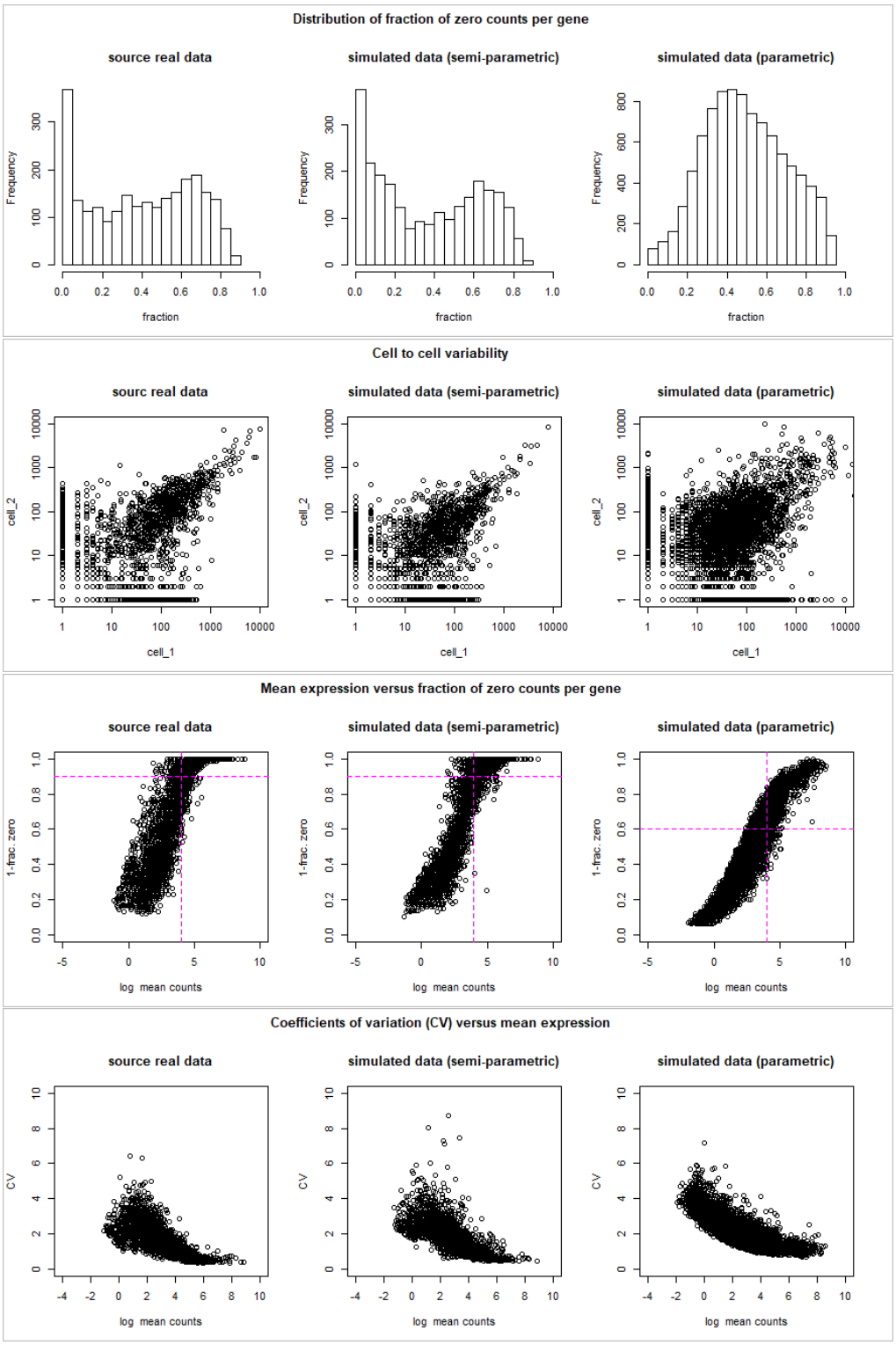

Figure S1: Comparison of the real data (source dataset A) and simulated data starting from this source data. The semi-parametric simulation implemented using SPsimSeq (Assefa et al 2019) R software package and the parametric simulation is using the splatter R bioconductor package (Zappia et al 2017).

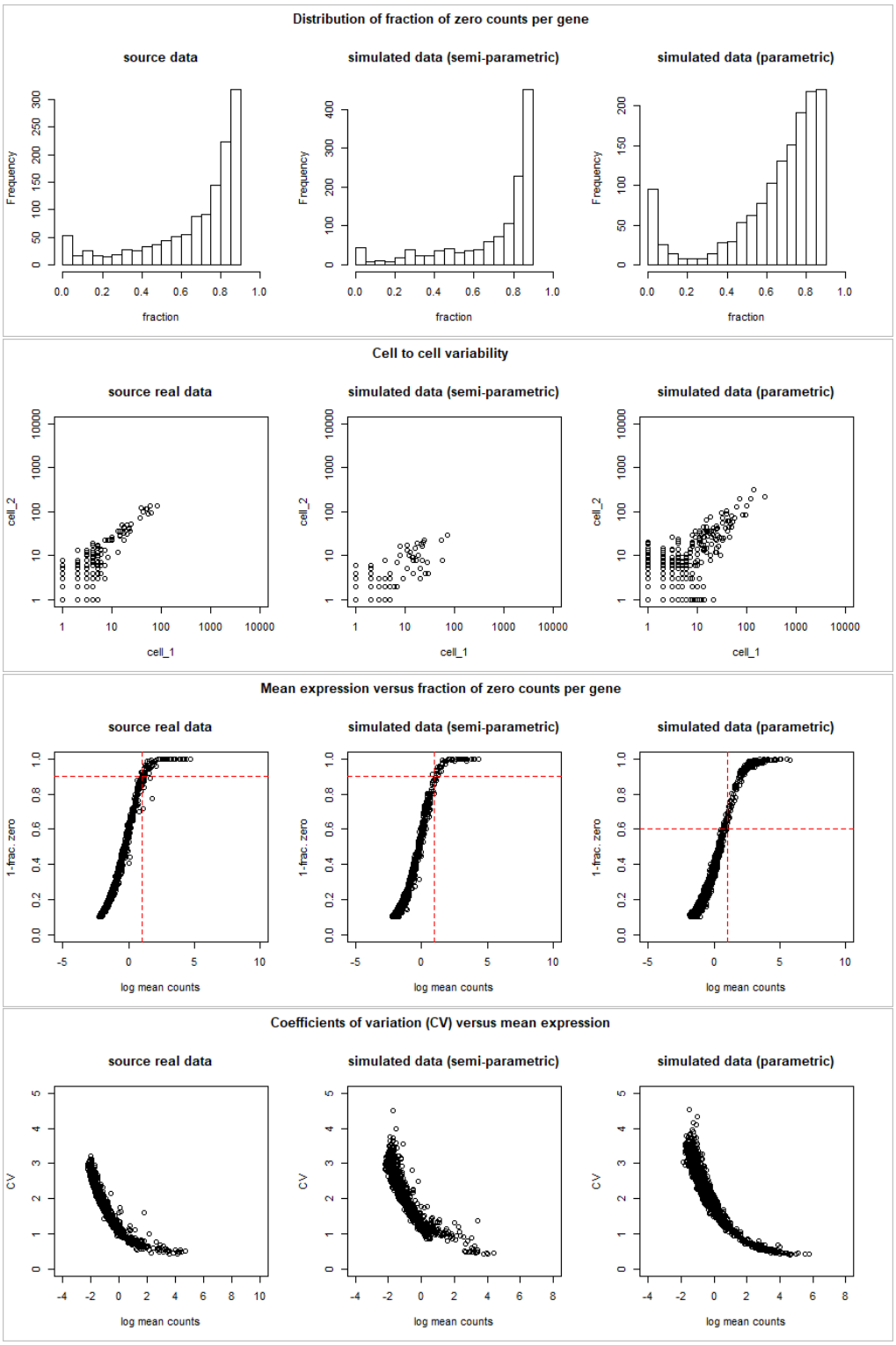

Figure S2: Comparison of the real data (source dataset B) and simulated data starting from this source data. The semi-parametric simulation implemented using SPsimSeq (Assefa et al 2019) R software package and the parametric simulation is using the splatter R bioconductor package (Zappia et al 2017).

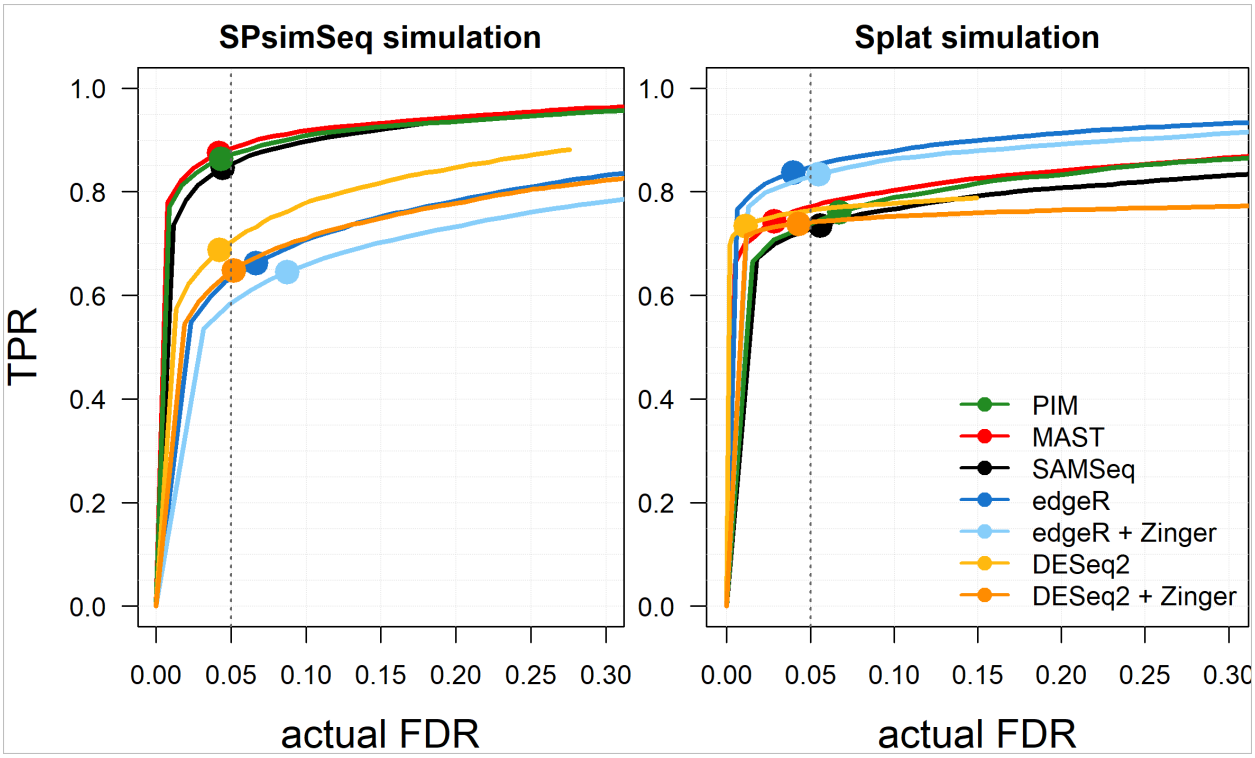

Figure S3: **Results from the simulation study starting from source dataset A.** Dataset A is neuroblastoma scRNA-seq data generated with SMARTer/C1 protocol. Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 100 cells. For the SPsimSeq simulation, the DE genes have LFC ≥ 1 in source data (data A), whereas for the Splat simulation the FC for DE genes is sampled from a log-Normal(location=1.5, scale=0.4), such that more than 97.5% of the DE genes have a LFC of at least 1. The performance measures (actual FDR and TPR) are averaged over a total of 50 independent simulation runs. The curves represent actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.3, and the solid dots show the performances at the 5% nominal FDR level.

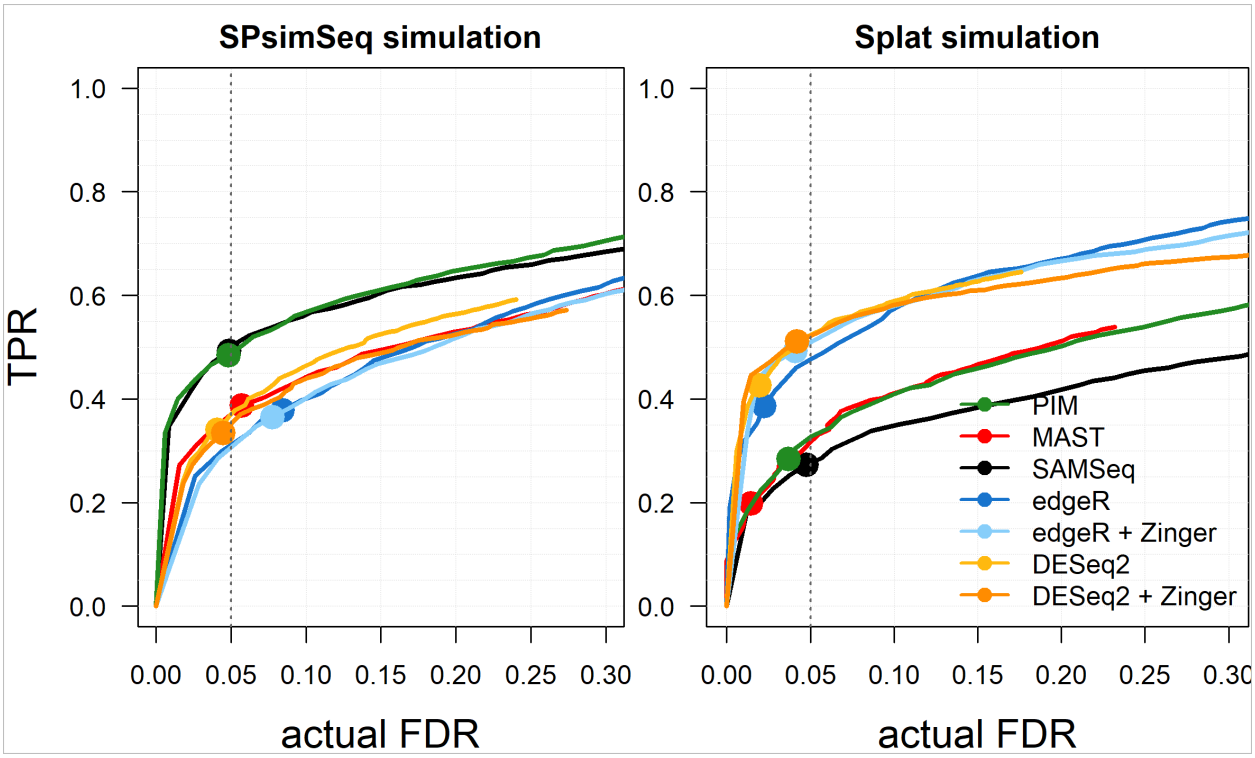

Figure S4: **Results from the simulation study starting from source dataset A.** Dataset A is neuroblastoma scRNA-seq data generated with SMARTer/C1 protocol. Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 100 cells. For the SPsimSeq simulation, the DE genes have LFC ≥ 0.5 in source data (data A), whereas for the Splat simulation the FC for DE genes is sampled from a log-Normal(location=1.25, scale=0.4), such that more than 97.5% of the DE genes have a LFC of at least 0.5. The performance measures (actual FDR and TPR) are averaged over a total of 50 independent simulation runs. The curves represent actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.3, and the solid dots show the performances at the 5% nominal FDR level.

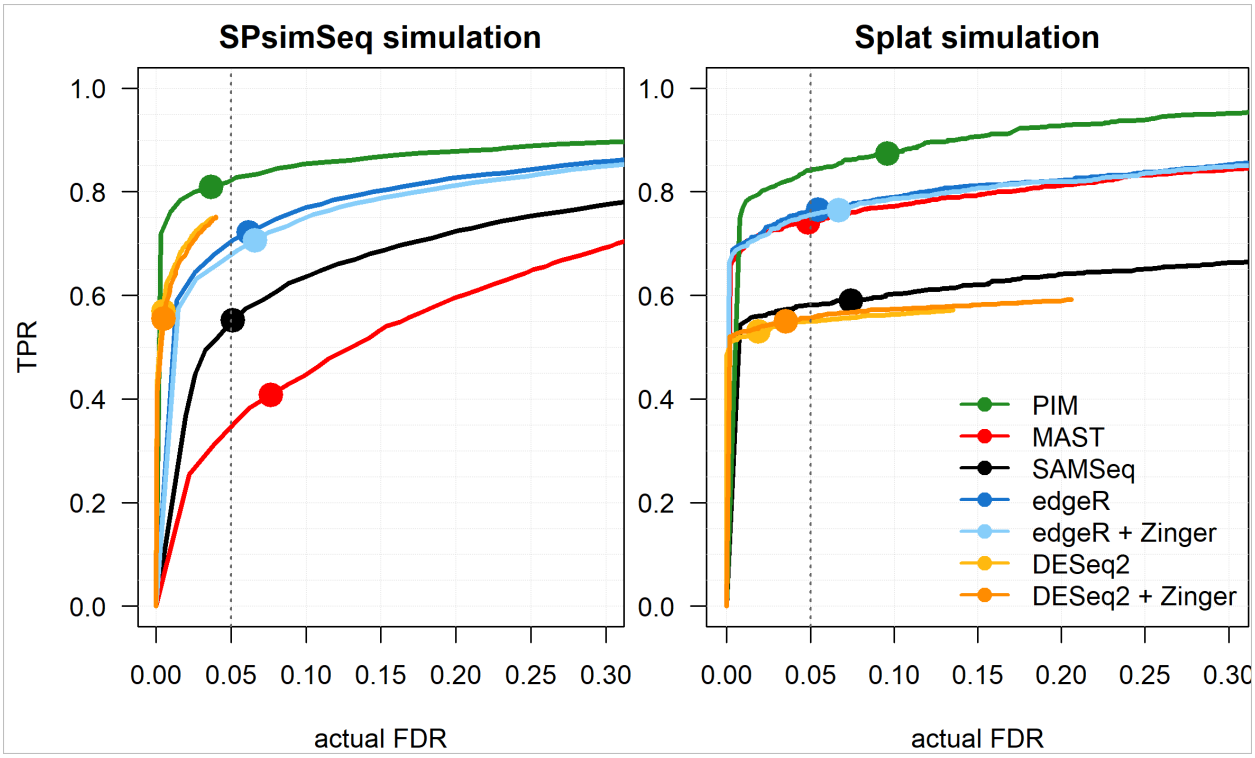

Figure S5: **Results from the simulation study starting from source dataset B.** Dataset B is neuroblastoma scRNA-seq data generated with Chromium protocol (UMI counts). Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 100 cells. For the SPsimSeq simulation, the DE genes have LFC ≥ 1 in source data (data B), whereas for the Splat simulation the FC for DE genes is sampled from a log-Normal(location=1.5, scale=0.4), such that more than 97.5% of the DE genes have a LFC of at least 1. The performance measures (actual FDR and TPR) are averaged over a total of 50 independent simulation runs. The curves represents actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.3, and the solid dots show the performances at the 5% nominal FDR level.

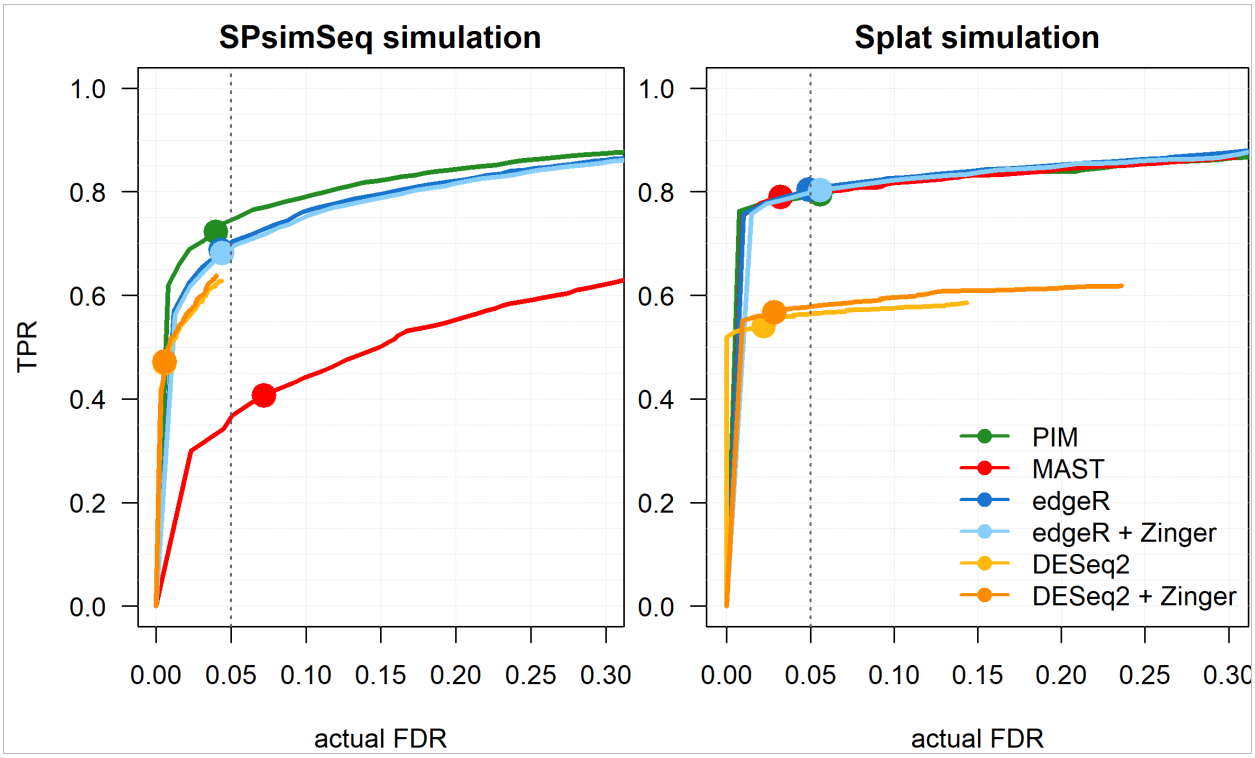

Figure S6: **Results from the simulation study starting from source dataset B.** Dataset B is neuroblastoma scRNA-seq data generated with Chromium protocol (UMI counts). Each simulated dataset includes 2500 genes among which 10% DE, and 2 experimental groups with each 200 cells. For the SPsimSeq simulation, the DE genes have LFC ≥ 0.5 in source data (data B), whereas for the Splat simulation the FC for DE genes is sampled from a log-Normal(location=1.25, scale=0.4), such that more than 97.5% of the DE genes have a LFC of at least 0.5. The performance measures (actual FDR and TPR) are averaged over a total of 50 independent simulation runs. The curves represents actual FDR and TPR evaluated at nominal FDR levels ranging from 0 to 0.3, and the solid dots show the performances at the 5% nominal FDR level. SAMSeq failed for datasets simulated in this simulation setting and exluded from the result.

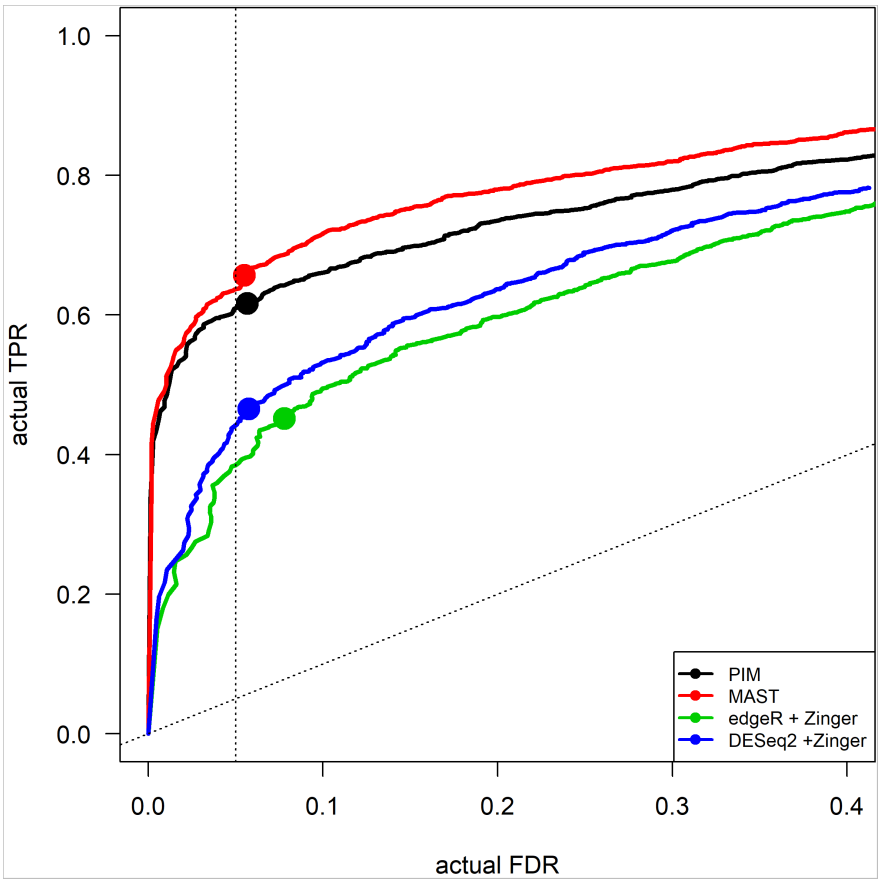

Figure S7: Performance of DE tools for testing DGE across three independent groups of cells based on the SPsimSeq simulation starting from dataset A . In particular, three groups of cells were generated by partitioning the simulated control group (vehicle) into two mock groups. The objective is to demonstrate the applicability of PIM in multiple group designs and evaluate its performance in comparison to the other regression based parametric tools. Each simulated dataset includes 2500 genes, among which 10% DE, and 3 experimental groups, each with 50 cells. The DE genes have LFC ≥ 0.5 between the control and the treatment group. The actual FDR and TPR are calculated by averaging over a total of 30 independent simulation runs. The curve represents actual FDR and TPR evaluated at different nominal FDR levels, ranging from 0 to 0.5, and the solid dots show the performances at the 5% nominal FDR level.

1.2 Mock comparison

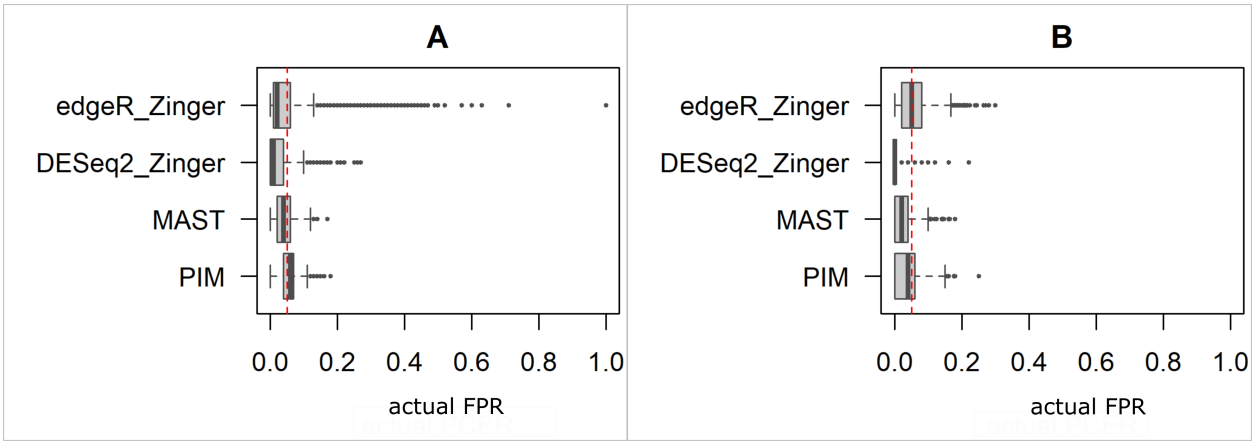

Figure S8: The actual false positive rate (FPR), which is defined for each gene as the fraction of simulations with unadjusted p-value less than 5% (nominal PCER=5% and indicated by red dashed vertical line), as calculated from the mock simulations, for datasets A (panel A) and B (panel B). The box plot shows the distribution across genes.

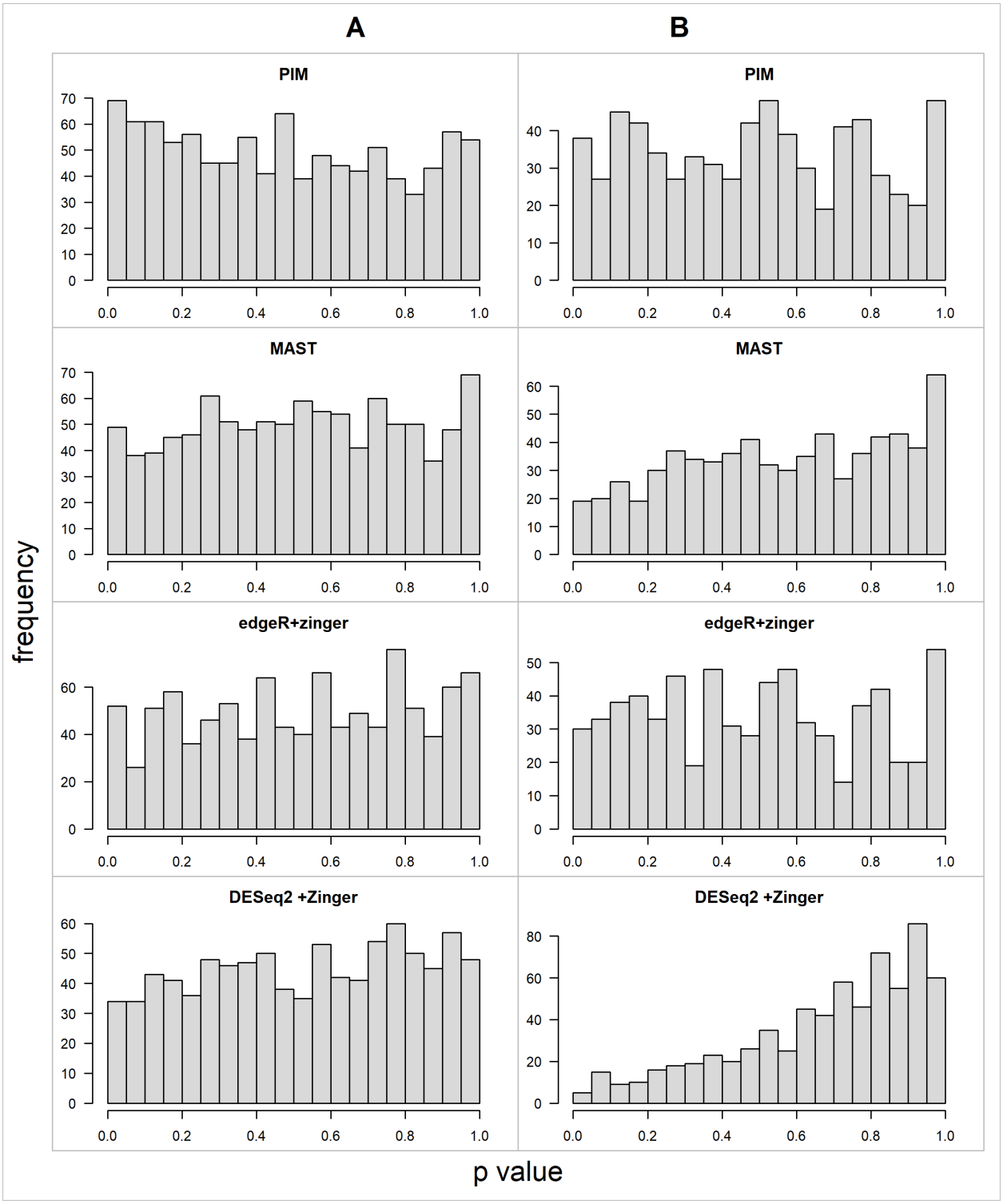

Figure S9: Distribution of unadjusted p-values from mock simulations, starting from source datasets A (A) and B (B).

2 Additional results from analysis of real scRNA-seq datasets

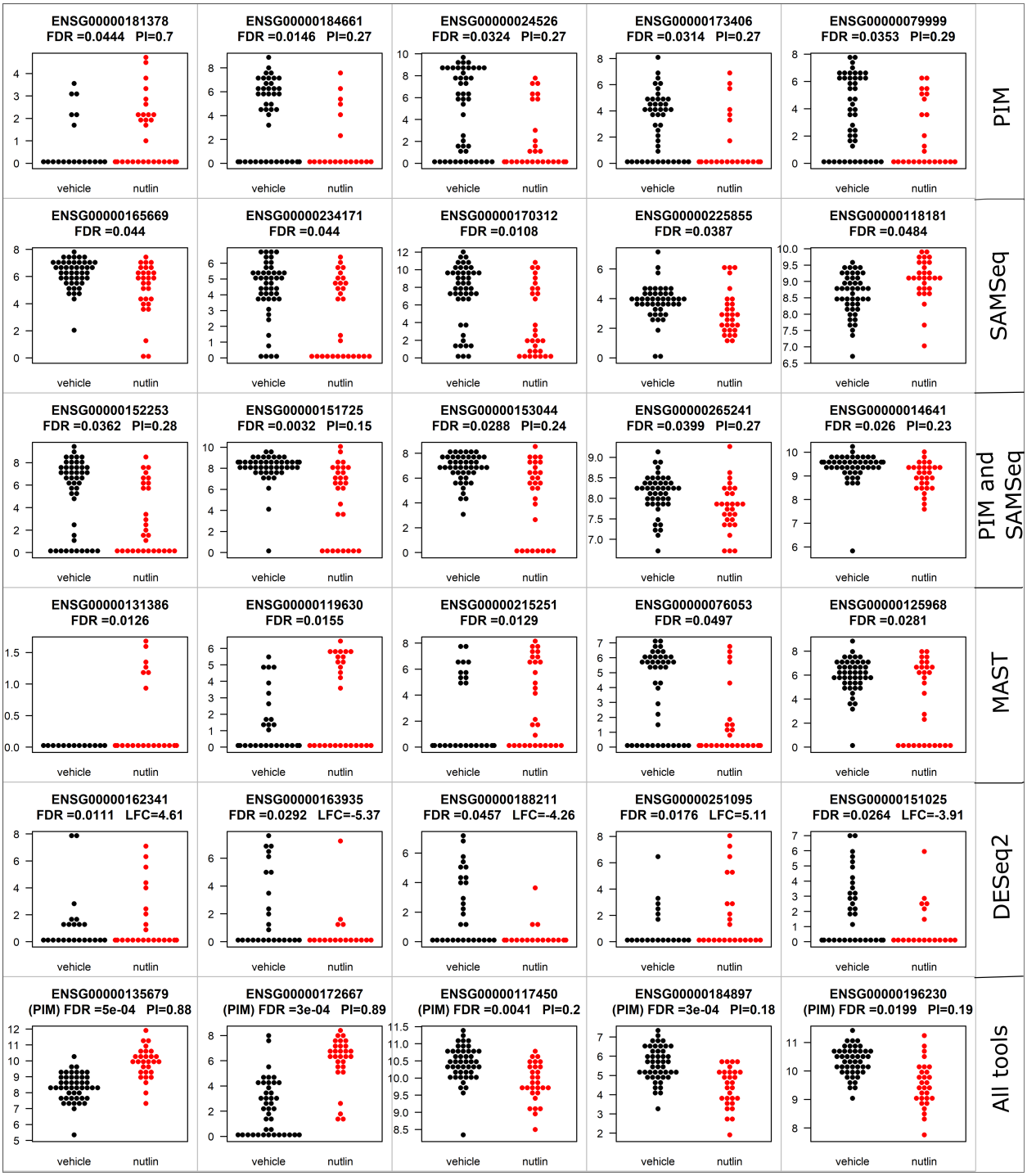

Figure S10: Beeswarm plot to demonstrate the distribution of log-CPM for 5 randomly selected genes among those that are uniquely called DE by PIM, SAMSeq, PIM and SAMSeq, MAST, and DESeq2 at the 5% FDR level and from commonly detected DE genes (bottom panel). The distributions are separately shown for each group (nutlin and vehicle)

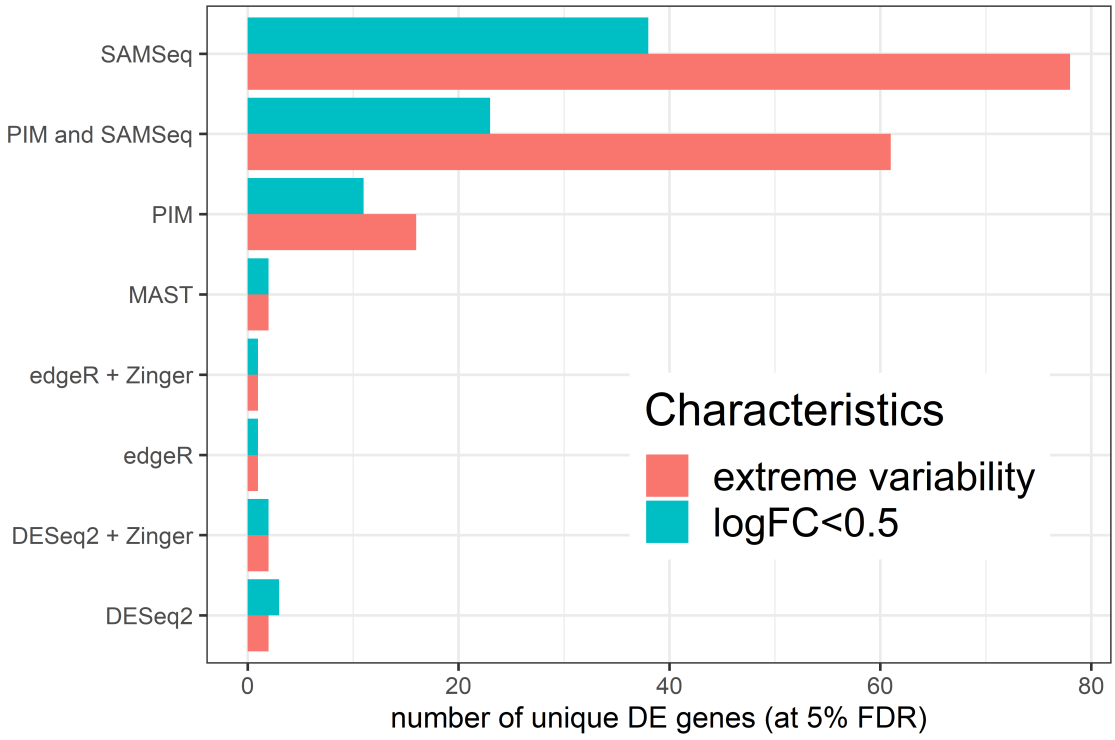

Figure S11: The characteristics of genes uniquely called as DE (at 5% FDR) by each tool. The characteristics include variability CPM across cells and the log fold change estimates. For each tool we counted the number of uniquely identified DE genes that have extreme variability (the upper 10%), and those with estimated log-fold-change below 0.5.

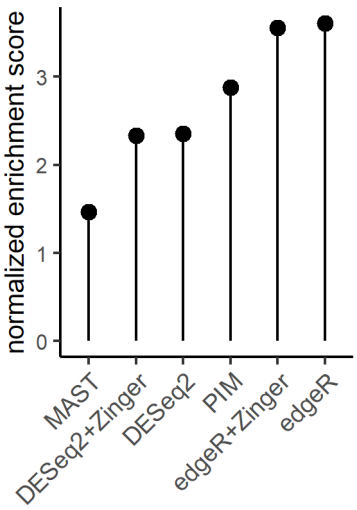

Figure S12: GSEA results for dataset B.

References

Assefa, Alemu Takele, Jo Vandesompele, and Olivier Thas. 2019. "SPsimSeq: Semi-Parametric Simulation of Bulk and Single Cell Rna Sequencing Data." *bioRxiv*. https://doi.org/10.1101/677740.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack. 2017. "Splatter: Simulation of Single-Cell Rna Sequencing Data." *bioRxiv*, 133173.