Supplementary-File–Chapter-4: 'SPsimSeq: semi-parametric simulation of bulk and single-cell RNA sequencing data'

Alemu Takele Assefa, Jo Vandesompele, Olivier Thas

February 21, 2019

Contents

1 Benchmarking and demonstration of SPsimSeq

1.1 Datasets

For the subsequent demonstrations and benchmarking of the SPsimSeq method, we use bulk and single-cell RNA-seq datasets summarized below,

- Neuroblastoma bulk RNA-seq data retrieved from [1] (GEO accession GSE49711). The data contains 498 neuroblastoma tumors. In short, unstranded poly(A)+ RNA sequencing was performed on the HiSeq 2000 instrument (Illumina). Paired-end reads with a length of 100 nucleotides were obtained. To quantify the full transcriptome, raw fastq files were processed with Kallisto v0.42.4 (index build with GRCh38-Ensembl v85). The pseudo-alignment tool Kallisto [2] was chosen above other quantification methods as it is performing equally good but faster. For this study, a subset of 172 tumors (samples) with high-risk disease were selected, forming two groups: the MYCN amplified ($n_1 = 91$) and MYCN non-amplified ($n_2 = 81$) tumours as used in [3]. Sometimes we refer this dataset to us the Zhang data or the Zhang neuroblastoma data.

- Neuroblastoma NGP cells single-cell RNA-seq data retrieved from [4] (GEO accession GSE119984): This dataset is generated for a cellular perturbation experiment on the C1 instrument (SMARTer protocol) [4]. This total RNA-seq dataset contains 83 NGP neuroblastoma cells, of which 31 were treated with 8μM of nutlin-3 and the other 52 cells were treated with vehicle (controls). In the subsequent sections, this dataset is referred to us the NGP single-cell RNA-seq data.

- Peripheral blood mononuclear cell (PBMC) single-cell RNA-seq data retrieved from https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz: This dataset contains 2700 single-cells sequenced on an Illumina NextSeq 500 using unique molecular identifiers (UMI). The data are generated using the 10x Genomics Chromium V1 protocol (Cell Ranger 1.0.0). In the subsequent sections, this dataset is referred to us the PBMC single-cell RNA-seq data.

1.2 The Splat simulation method

To benchmark the SPsimSeq method, we implemented one of the most commonly used fully parametric simulation method, called the Splat simulation [5]. This simulation algorithm makes use of a gamma-Poisson hierarchical model, which is a particular parametrization of the negative binomial distribution. In particular, the mean expression level of gene g in cell i, λ_{gi}, is sampled from a gamma distribution (i.e. $\lambda_{gi} \sim \Gamma(\alpha_{gi}, \beta_{gi})$). Subsequently, the read counts Y_{gi} for gene g in cell i are sampled from a Poisson distribution (i.e. $Y_{gi}|\lambda_{gi} \sim \text{Poisson}(\lambda_{gi})$). To simulate data with the desired library size and account for the mean-variance trend, Splat uses real RNA-seq data to estimate the hyper-parameters α_{gi} and β_{gi}. Moreover, Splat uses a logistic function for the observed relationship between the mean expression level of a gene and the proportion of zero counts to add excess zeros representing the technical noise (also known as dropouts). To add a set of genes that are DE between, for example, two simulated groups, it multiplies the mean expression of randomly selected genes in one of the groups by a factor, also known as a fold-change. The factor is drawn from a log-normal distribution with user-adjustable location and scale parameters. The Splat simulation is implemented using the `splatter` R Bioconductor package (version 1.6.1) [5]. To simulate bulk RNA-seq data using the Splat procedure, we disabled its feature that adds dropouts (`dropout.type="none"`), which is specifically designed for simulating single-cell RNA-seq data simulation.

1.3 Benchmarking methods

For a particular simulation, both the SPsimSeq and Splat methods started from the same source dataset, and all the user-adjustable simulation parameters (such as the number of genes, number of cells/samples, number of groups, number of batches, fraction of DE genes, and the log-fold change thresholds) set to be the same for both methods. Besides, the source and simulated data have equal number of genes, cells/samples, groups, cells/samples per group, and batches so that the simulated dataset is supposed to be a mirror-image of the source dataset in terms of the various characteristics discussed below.

We compared the simulated datasets (from the SPsimSeq and Splat methods) with their correspond-

ing real source dataset with respect to various gene and cell/sample level metrics used in [5, 6]. In particular, the comparison metrics include

- the distribution of mean, variance and coefficients of variation (CV) of gene expression levels
- the relationship between the mean and variance and the mean and CV of gene expression levels
- the distribution of the fraction of zero counts per gene and its relationship with the mean expression level
- the distribution of the pairwise correlation coefficients between the genes and cells/samples
- the distribution of the library sizes and fraction of zero counts per sample/cell

In addition, the difference in the variance, CV, and fraction of zero counts per gene between the real and the simulated data are computed at a particular range of the mean expression levels (after log-CPM transformation). In particular, the mean log-CPM of genes are first divided into 1000 intervals, and at each interval, the difference in the mean of the variance, CV, and fraction of zero counts are computed between the real and simulated data. The difference is computed as the summary from the simulated data minus that of the real data. Therefore, positive differences indicate over-representation of the characteristic, negative differences indicate under-representation and 0 (or nearly 0) indicates equivalence between the real and the simulated data. The LOESS smoothed regression is also employed for a more precise visualization of the differences across mean log-CPM intervals.

1.4 Simulation of bulk RNA-seq data

We use SPsimSeq to simulate bulk RNA-seq data starting from the Zhang neuroblastoma data. For benchmarking purpose, we also simulate bulk RNA-seq data using the Splat procedures as described above. In particular, we simulate bulk RNA-seq data with the following features

- 5000 genes (`n.genes = 5000`)
- a total of 172 samples (`tot.samples = 172`), which equals with the source data

- the samples are divided into two groups (`group.config = c(0.47, 0.53)`) with 10% of the genes are DE between the groups (`pDE = 0.1`). The group composition is similar to that of the source data.
- the DE genes have a LFC at least 0.5 (`lfc.thrld=0.5`) with t-statistic threshold of 2.5 (`t.thrld=2.5`) and ll.threshold of 5 (`ll.thrld=5`)
- all the samples are generated in a single batch (`batch.config = 1`), similar to the source data
- since zero inflation is not an issue in bulk RNA-seq dataset, we do not model the zeros separately (`model.zero.prob = FALSE`)
- the number of classes to construct the distributions of the gene expression levels is 50% of the sample size (n) (i.e. `w=0.5`). The reason will be discussed later.

The evaluation results for bulk RNA-seq simulation are shown in Figures S1 to S5. The results generally indicate that SPsimSeq generates a realistic bulk RNA-seq dataset in terms of all the comparison metrics. In particular, the mean-variance relationship (Figure S1), the mean-CV relationship (Figure S2), the mean versus the fraction of zero counts (Figure S3) and the dependence between genes (Figure S4) were accurately captured by the SPsimSeq simulations and they were better than that of the Splat simulations. Although both the SPsimSeq and Splat simulations started from the same source dataset and generated an equivalent size of RNA-seq dataset (in terms of the number of genes and library sizes, see Figure S5), a substantial discrepancy was observed between the real data and the Splat simulated data with respect to the majority of the comparison metrics. For example, the Splat method tends to be biassed towards highly expressed genes (Figure S5), under-represents the gene-to-gene and the sample-to-sample correlations (Figure S4 and S5) and under-represents the variance and CV of low-abundance genes (see Figure S1, S2 and S5).

We used the following SPsimSeq code to simulate bulk RNA-seq data based on the features listed above.

```
# load required libraries
library(SPsimSeq)
```

```r
# load the Zhang data (availabl with the SPsimSeq package)
data("zhang.data")

# remove genes with insufficient expression (important step to avoid bugs)
zhang.counts  <- zhang.data$counts[rowSums(zhang.data$counts > 0)>=5, ]
MYCN.status   <- zhang.data$MYCN.status+1

# subset an equivalent size of the source dataset
set.seed(25081988)
zhang.counts2 <- zhang.counts[sample(nrow(zhang.counts), 5000), ]
zhang.counts2 <- zhang.counts$counts
MYCN.status   <- zhang.counts$MYCN.status

# SPsimSeq simulation
SPsimSeq.sim <- SPsimSeq(s.data = zhang.counts2, group = MYCN.status, batch = NULL,
                         n.genes = 5000, group.config = c(0.47, 0.53),
                         batch.config = 1, tot.samples = 172, pDE = 0.1,
                         lfc.thrld=0.5, t.thrld=2.5, ll.thrld=5, w=0.5,
                         model.zero.prob = FALSE,  result.format = "list",
                         seed=25081988)
```

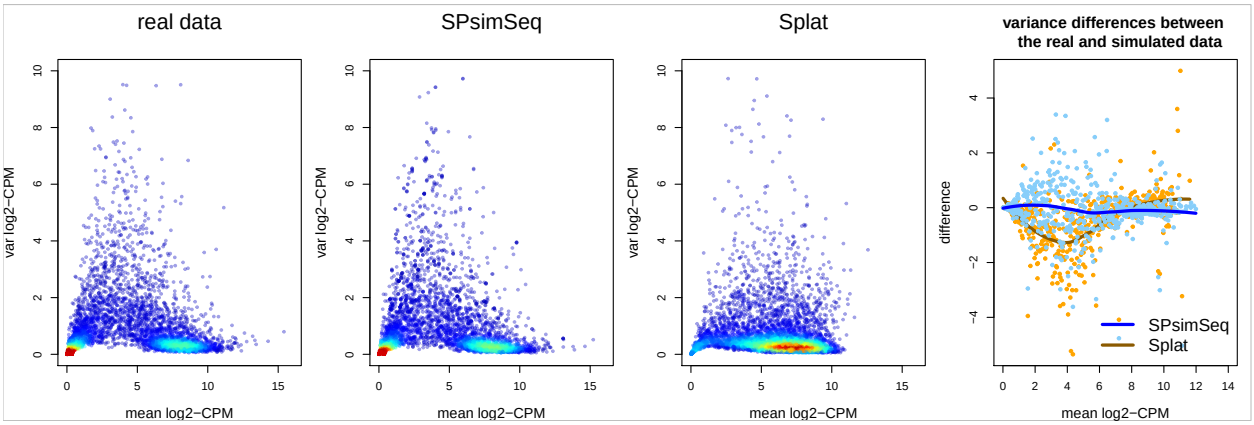

Figure S1: The relationship between the mean and variance of gene expression levels (log-CPM) from the real bulk RNA-seq data (the Zhang data) and the SPsimSeq and Splat simulated datasets.

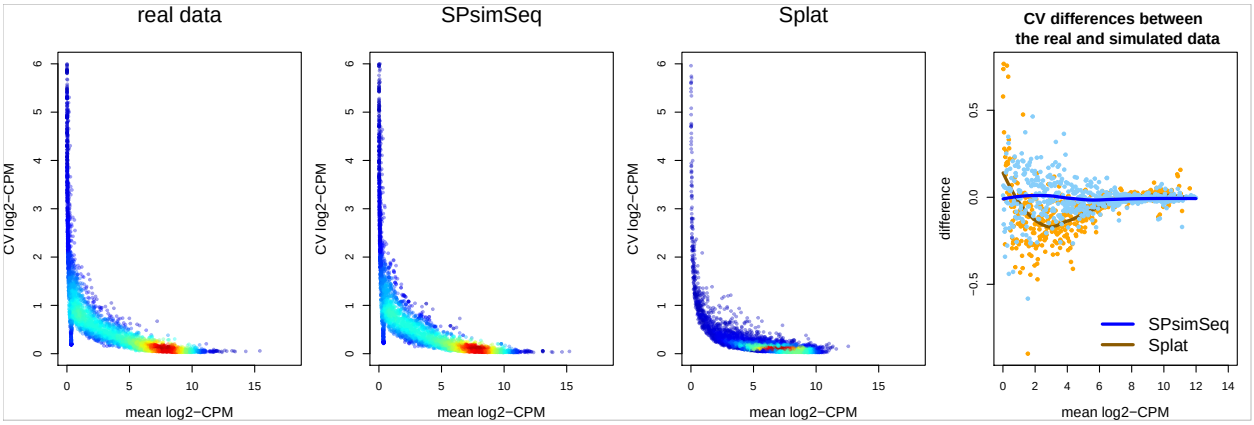

Figure S2: The relationship between the mean and CV of gene expression levels (log-CPM) from the real bulk RNA-seq data (the Zhang data) and the SPsimSeq and Splat simulated datasets.

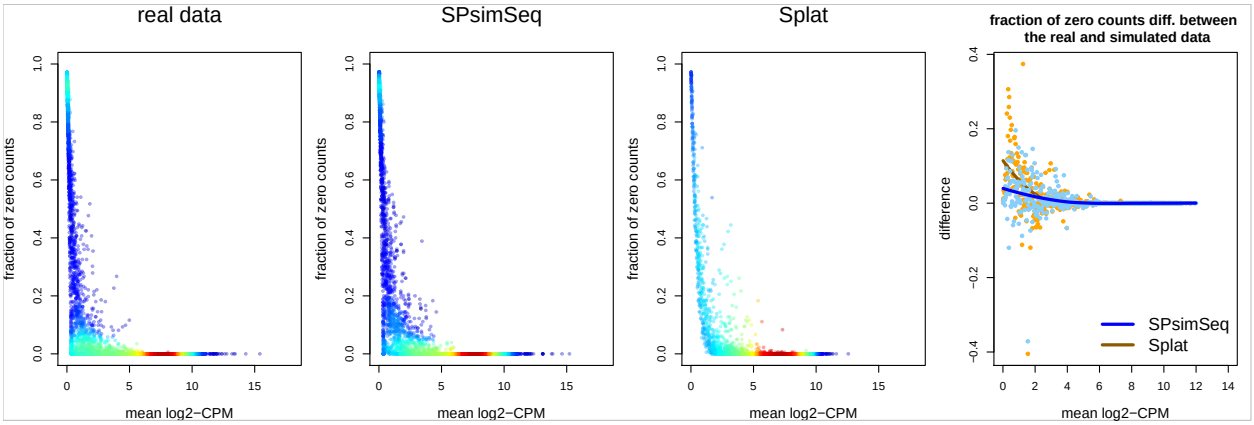

Figure S3: The fraction of zero counts per gene as a function of the mean gene expression levels (log-CPM) from the real bulk RNA-seq data (the Zhang data) and the SPsimSeq and Splat simulated datasets.

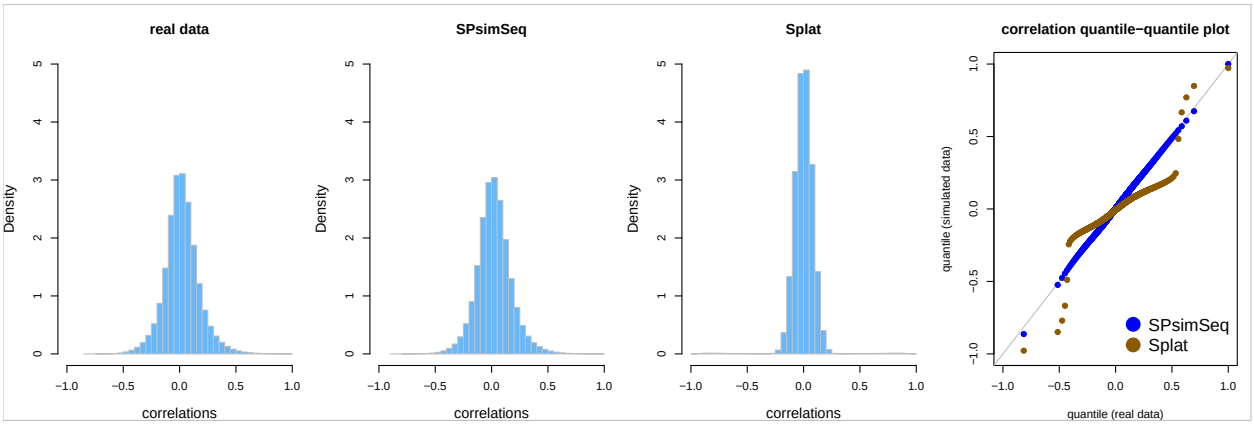

Figure S4: The distributions of the pairwise Pearson correlation-coefficients between genes from the real bulk RNA-seq data (the Zhang data) and the SPsimSeq and Splat simulated datasets.

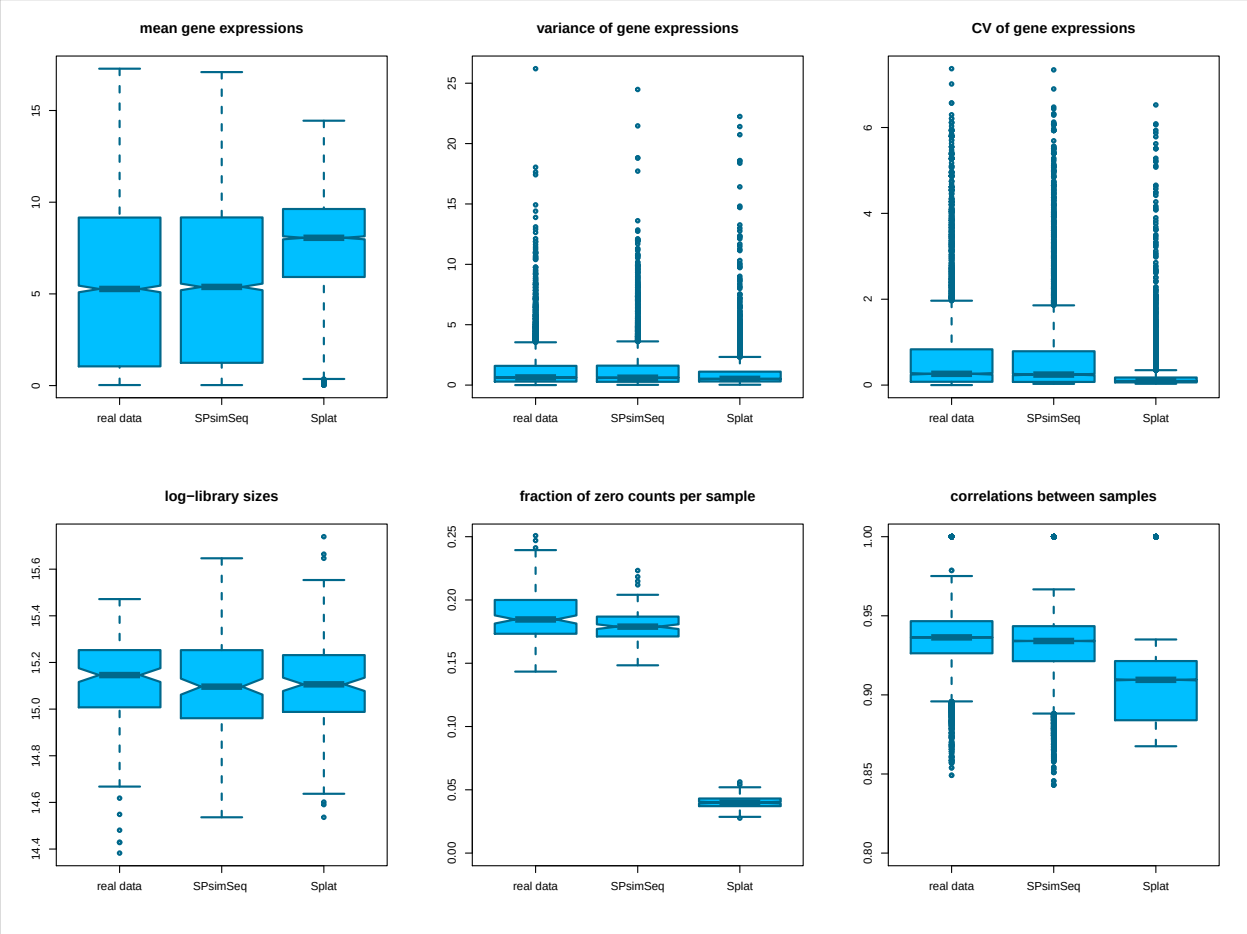

Figure S5: The distributions of the mean, variance and CV of gene expression levels (upper row), and the distributions of the log-library sizes (across samples), fraction of zero counts per sample and the correlations between samples (bottom row) from the real bulk RNA-seq data (the Zhang data) and the SPsimSeq and Splat simulated datasets.

1.5 Simulation of single-cell RNA-seq data (read-count data)

Now we use SPsimSeq to simulate single-cell RNA-seq data starting from the NGP neuroblastoma single-cell RNA-seq data. We also simulate single-cell RNA-seq data using the the Splat procedures with dropouts added. In particular, we simulate a single-cell RNA-seq data with the following features

- 5000 genes (`n.genes = 5000`)
- a total of 83 cells, which is equal to the source data (`tot.samples = 83`)
- cells are divided into two groups according to the composition in the source data (37% nutlin and 63% vehicle)– (`group.config = c(0.37, 0.63)`)
- 10% of the genes are DE between the two treatment groups of cells (`pDE = 0.1`)
- the DE genes have a LFC of at least 0.5 (`lfc.thrld=0.5`) with t-threshold 2.5 (`t.thrld=2.5`) and ll.threshold 5 (`ll.thrld=5`)
- all cells are generated in a single batch (`batch.config = 1`)
- zero counts are modeled separately to account for the zero inflation (`model.zero.prob = TRUE`)
- the number of classes to construct the distributions of the gene expressions is 50% of the sample size (n) (`w=0.5`)

The results in Figures S6 to S10 generally indicate that SPsimSeq succeeds simulating single-cell RNA-seq data (with read-counts) that mimics the real data with respect to all the comparison metrics. In particular, the relationship between the mean and variability of the gene expressions (variance and CV) and the fraction of zero counts are well captured by the SPsimSeq simulations and better than that of the Splat simulations. Both the SPsimSeq and Splat methods showed comparably good performance in terms of capturing the gene-wise correlation (Figure S9) and the distribution of the mean expression level of the genes (Figure S10). However, the Splat simulations over-represented the variance of gene expressions, particularly for highly expressed genes (Figures S6 and S10), and under-represented the similarity between cells (Figure S10).

We used the following SPsimSeq code to simulate single-cell RNA-seq data based on the features listed above.

```
# load required libraries
library(SingleCellExperiment)
library(SPsimSeq)

# load the NGP nutlin data (availabl with the package)
data("scNGP.data")

# filter genes with sufficient expression (important step to avoid bugs)
scNGP.data2 <- scNGP.data[rowSums(counts(scNGP.data) > 0)>=5, ]
treatment   <- ifelse(scNGP.data2$characteristics..treatment=="nutlin",2,1)
scNGP.data2 <- scNGP.data2[sample(nrow(scNGP.data2), 5000), ]

# simulate data (we simulate here only a single data, n.sim = 1)
sim.data.SPsim <- SPsimSeq(n.sim = 1, s.data = scNGP.data2, batch = NULL,
                           roup = treatment, n.genes = 5000, batch.config = 1,
                           group.config = prop.table(table(treatment)),
                           tot.samples = ncol(scNGP.data2), w = 0.5,
                           pDE = 0.1, lfc.thrld=0.5, t.thrld=2.5, ll.thrld=5,
                           model.zero.prob = TRUE, result.format = "SCE",
                           seed=25081988)
```

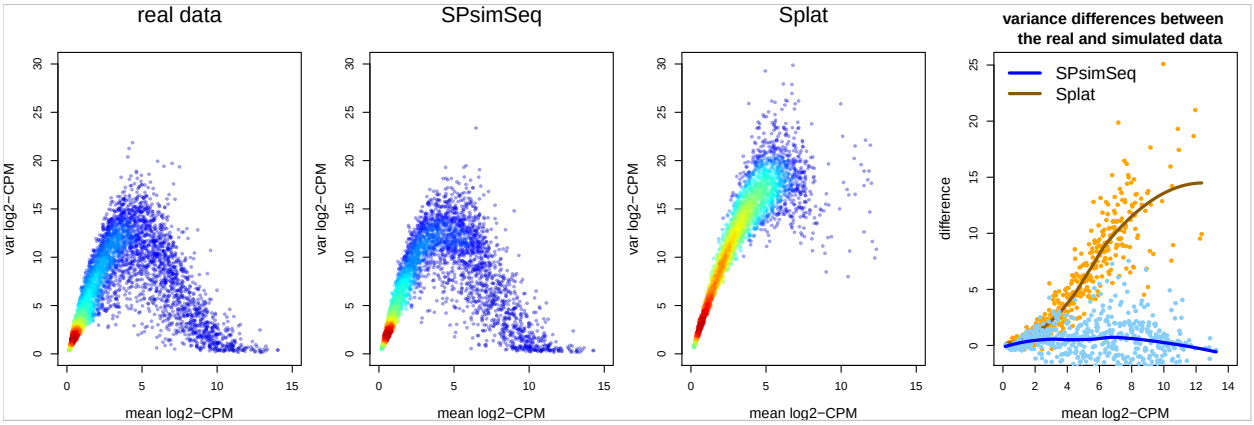

Figure S6: The relationship between the mean and variance of gene expression levels (log-CPM) from the real single-cell RNA-seq data (NGP neuroblastoma) and the SPsimSeq and Splat simulated datasets.

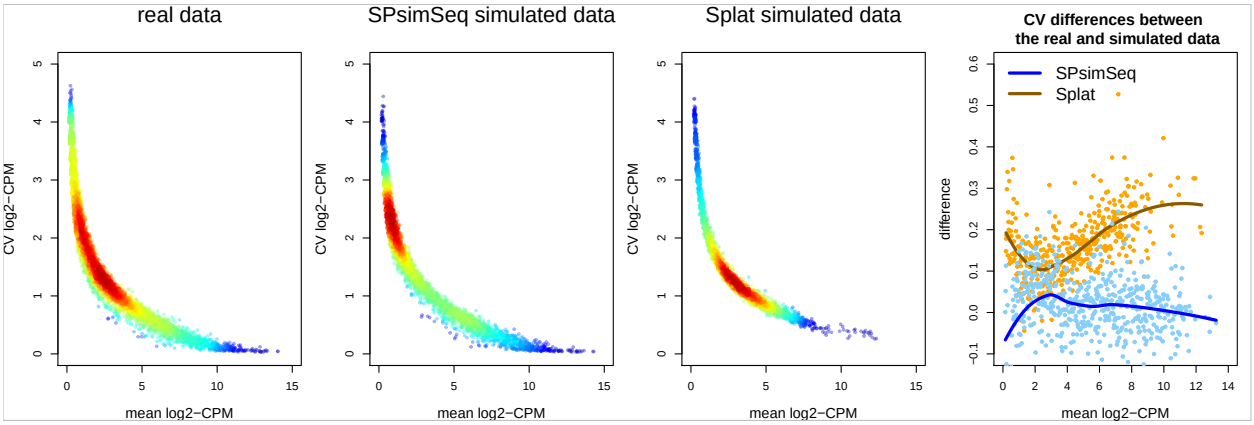

Figure S7: The relationship between the mean and CV of gene expression levels (log-CPM) from the single-cell RNA-seq data (NGP neuroblastoma) and the SPsimSeq and Splat simulated datasets.

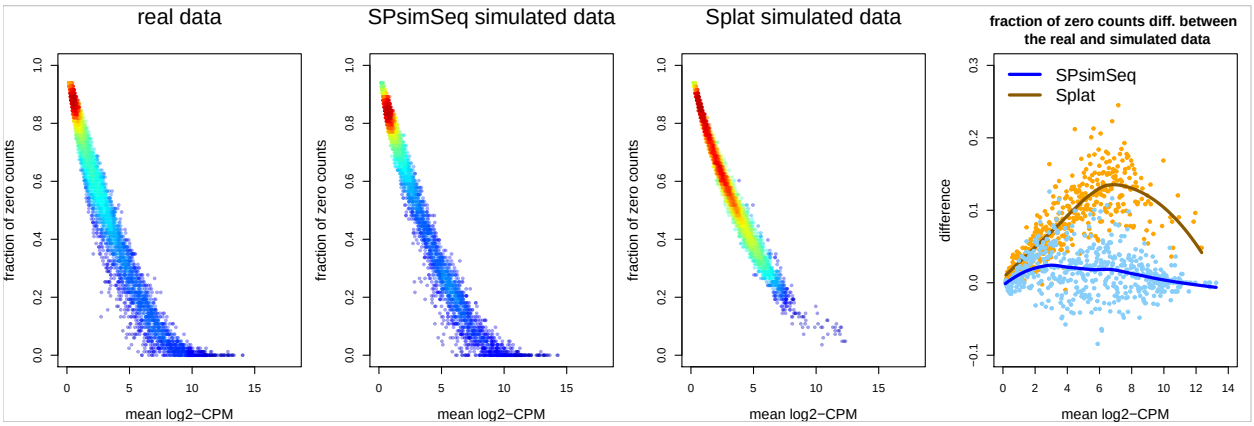

Figure S8: The fraction of zero counts per gene as a function of the mean gene expression levels (log-CPM) from the real single-cell RNA-seq data (NGP neuroblastoma) and the SPsimSeq and Splat simulated datasets.

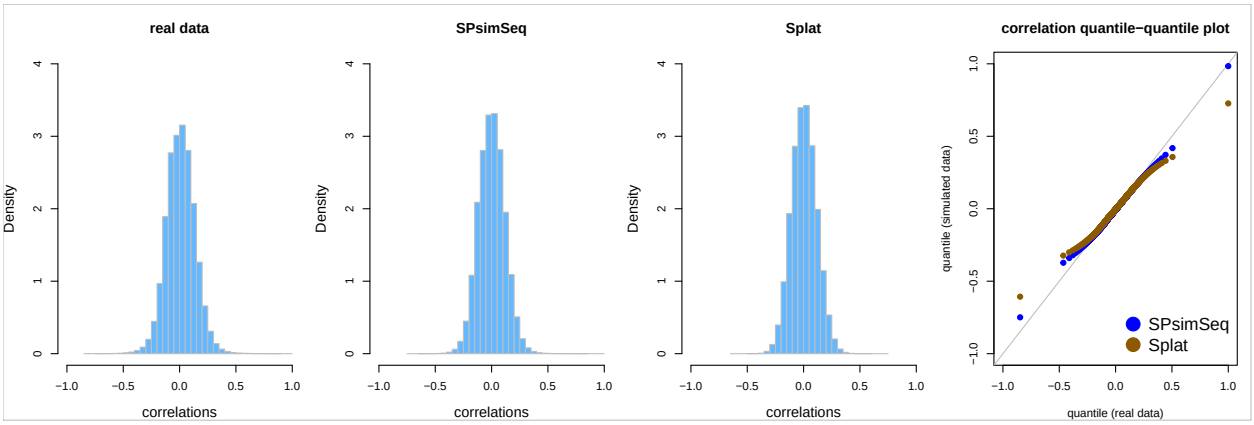

Figure S9: The distributions of the pairwise Pearson correlation-coefficients between genes from the real single-cell RNA-seq data (NGP neuroblastoma) and the SPsimSeq and Splat simulated datasets.

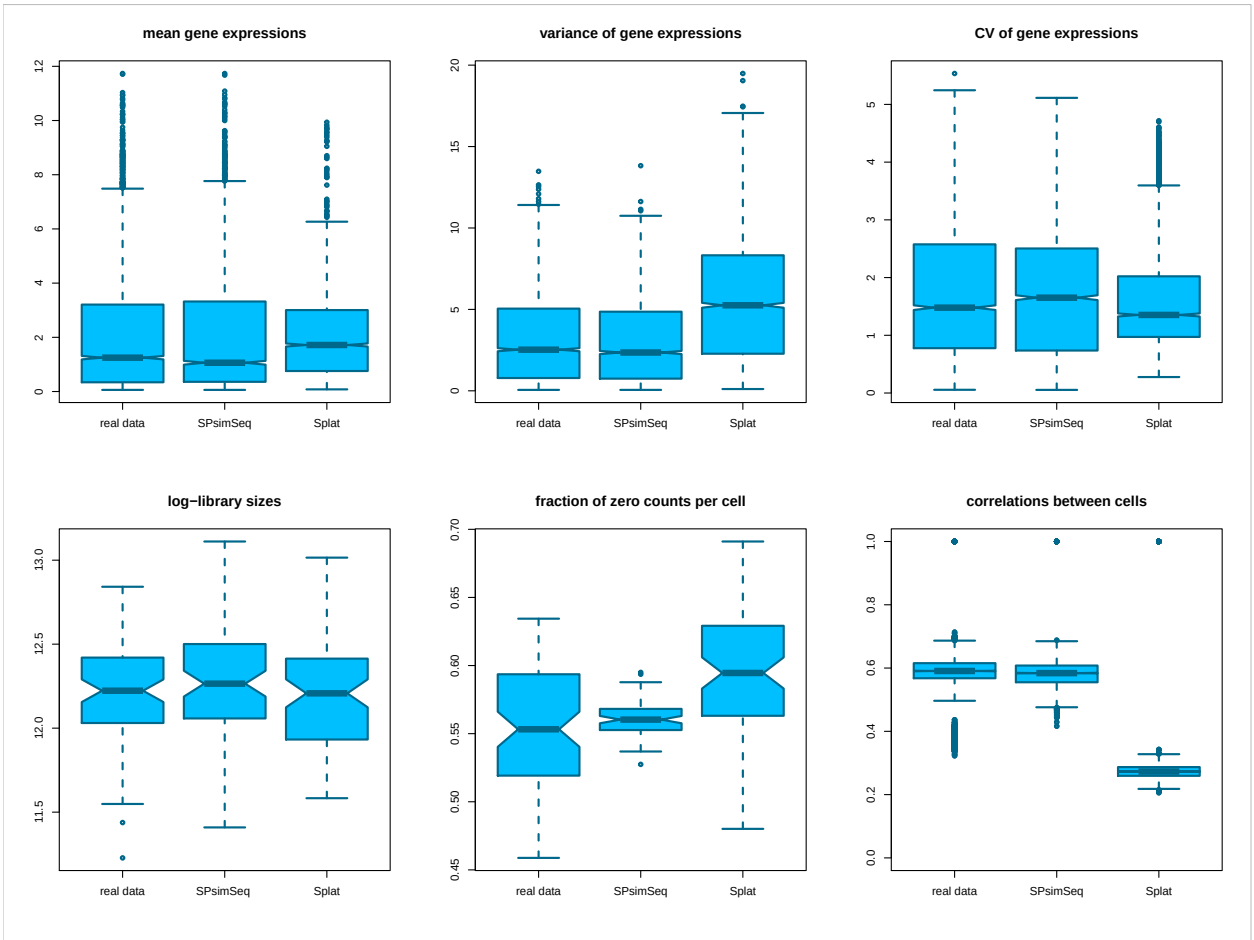

Figure S10: The distributions of the mean, variance and CV of gene expression levels (upper row), and the distributions of the log-library sizes (across cells), fraction of zero counts per cell and the correlations between cell (bottom row) from the real single-cell RNA-seq data (NGP neuroblastoma) and the SPsimSeq and Splat simulated datasets.

1.6 Simulation of single-cell RNA-seq data (UMI-count data)

Unique molecular identifiers (UMI) are highly used in single-cell RNA-seq studies. UMI counts reduce amplification bias and result in a better approximation of gene expression [7]. In this section, we demonstrate how the SPsimSeq method can be used to simulate UMI data for single-cell RNA-seq studies. As a source dataset, we use the PBMC single-cell RNA-seq data generated using Chromium protocol from the 10xGenomics, which contains UMI counts. In particular, we simulate single-cell RNA-seq data with the following features.

- 5000 genes (`n.genes = 5000`)
- a total of 2700 cells, which is equal to the source data (`tot.samples = ncol(PBMCdat2)`)
- all cells are within the same population (`group.config = 1`), and hence no gene is DE (`pDE = 0`)
- all cells are generated in a single batch (`batch.config = 1`)
- zero counts are modeled separately to account for zero inflation (`model.zero.prob = TRUE`)
- the number of classes to construct the distributions of the gene expressions is determined using the Sturges' rule (`w=NULL`)

The evaluation results once more confirm that the SPsimSeq method simulates a realistic single-cell RNA-seq dataset with UMI-counts in terms of all the comparison metrics (Figures S11 to S15). Despite a slight deviation, the Splat simulations also well captured the distribution of the gene expression levels (Figure S15), the relationship between the mean expression level and CV of genes (Figure S12), the relationship between the mean expression level and fraction of zero counts per gene (Figure S12) and the similarity between PBM cells (Figure S15).

We used the following SPsimSeq code to simulate single-cell RNA-seq (UMI-count) data based on the features listed above.

```
# load required libraries
library(SingleCellExperiment)
library(SPsimSeq)
```

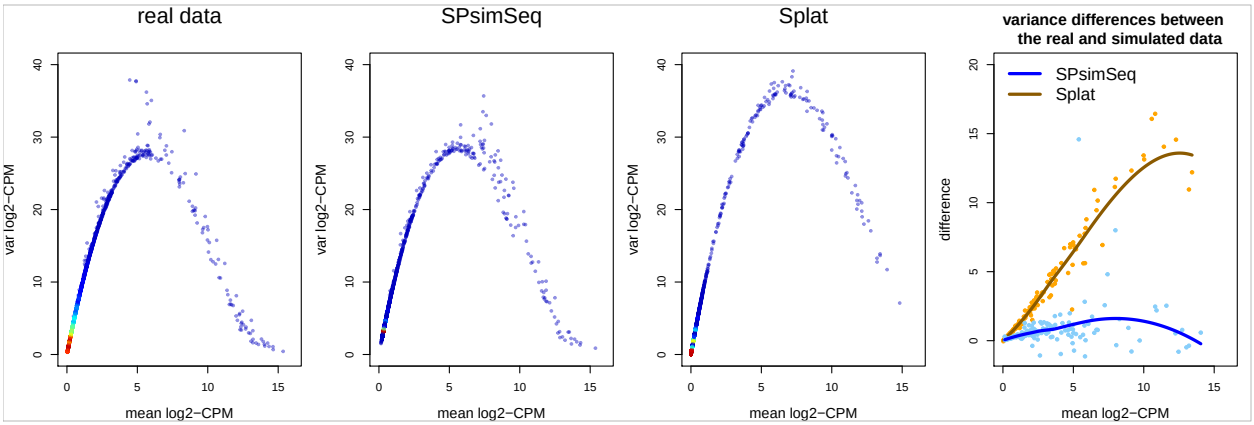

Figure S11: The relationship between the mean and variance of gene expression levels (log-CPM) from the real single-cell RNA-seq data (PBMC data) and the SPsimSeq and Splat simulated datasets.

```
# load the Zhang data (availabl with the package)
data("PBMC.data")

# filter genes with sufficient expression (important step to avoid bugs)
PBMCdat  <- PBMC.10x.data[rowSums(counts(PBMC.10x.data) > 0)>=10, ]
PBMCdat2 <- PBMCdat[sample(nrow(PBMCdat), 5000), ]

# simulate data (we simulate here only a single data, n.sim = 1)
sim.data.SPsim <-  SPsimSeq(n.sim = 1, s.data = PBMCdat2, batch = NULL,
                            group = NULL, n.genes = 5000, batch.config = 1,
                            group.config = 1, tot.samples = ncol(PBMCdat2), pDE = 0,
                            model.zero.prob = TRUE, result.format = "SCE", seed = 25081988)
```

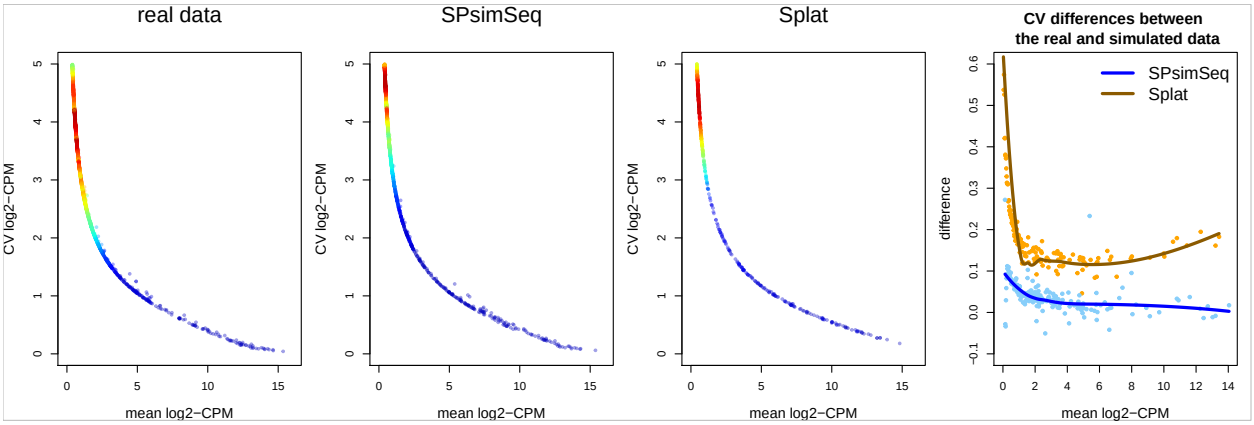

Figure S12: The relationship between the mean and CV of gene expression levels (log-CPM) from the single-cell RNA-seq data (PBMC data) and the SPsimSeq and Splat simulated datasets.

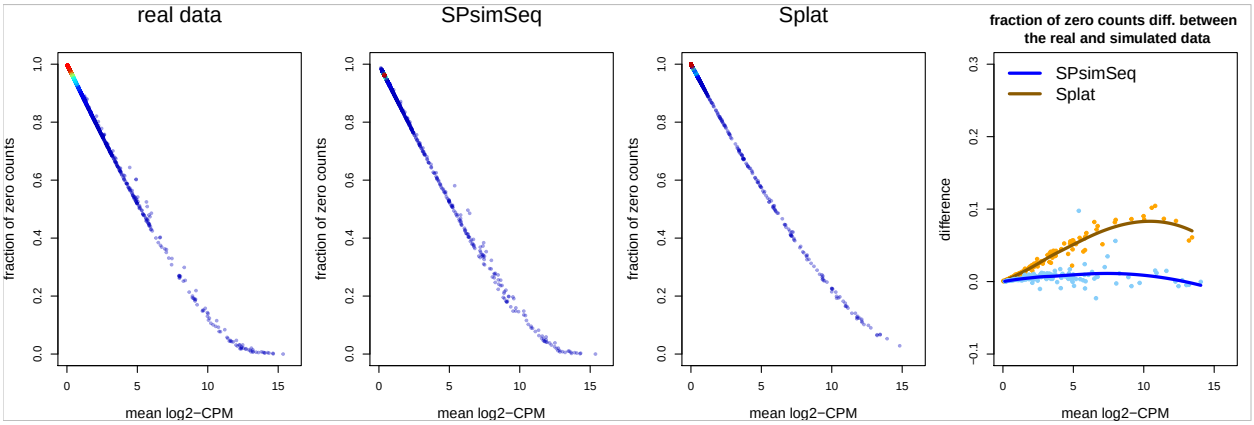

Figure S13: The fraction of zero counts per gene as a function of the mean gene expression levels (log-CPM) from the real single-cell RNA-seq data (PBMC data) and the SPsimSeq and Splat simulated datasets.

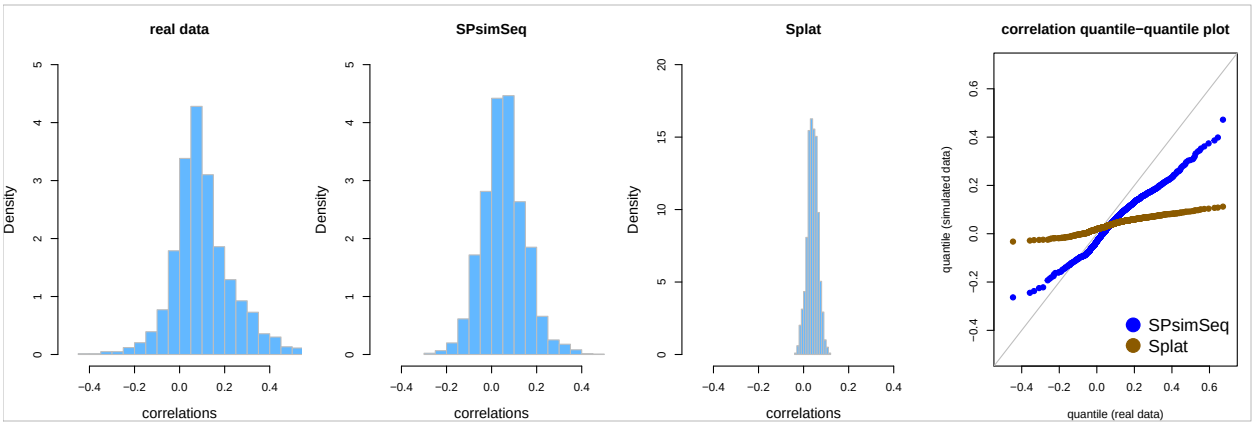

Figure S14: The distributions of the pairwise Pearson correlation-coefficients between genes from the real single-cell RNA-seq data (PBMC data) and the SPsimSeq and Splat simulated datasets.

2 Evauation of the choice of number of classes

Recall that the SPsimSeq procedure involves the partitioning of the log-CPM transformed gene expression levels of the source data into K disjoint classes. In this section, we assess the effect of the choice of K on the characteristics of the simulated dataset from the estimated distributions. In particular, we compare the SPsimSeq simulated datasets generated with different choices of K in terms of the various comparison metric discussed above. K is defined relative to the number of samples/cells in the source data (n). That is, we used $w = K/n$. We simulate new dataset for $w \in \{w^*, 0.1, 0.3, 0.5, 0.7\}$, where $w^* = k^*/n$ with k^* is determined by the Sturges' rule, i.e. $k^* = 1 + 3.322 \log_{10} n$ [8]. Except for the Sturges' rule, w is the same for all genes. We simulate both bulk and single-cell RNA-seq datasets, and we compare them with their corresponding source datasets (real datasets).

The results generally indicate that the different choice of K has a minimal effect on the characteristics of the simulated gene expression levels. However, a very small w may result in less realistic simulated data depending on the type of the source dataset. For the bulk RNA-seq data (Figures S16-S18) and single-cell RNA-seq data with read-counts (Figures S19-S21), a small K relative to n, for instance, $w = 0.1$, resulted in a simulated data that less resembles the source data in terms of the fraction of zero counts per gene and variability of the gene expression levels. For these type of datasets, a sufficiently large K, such as $w \in (0.3, 0.7)$, is sufficient to simulate a new dataset that mirror-images

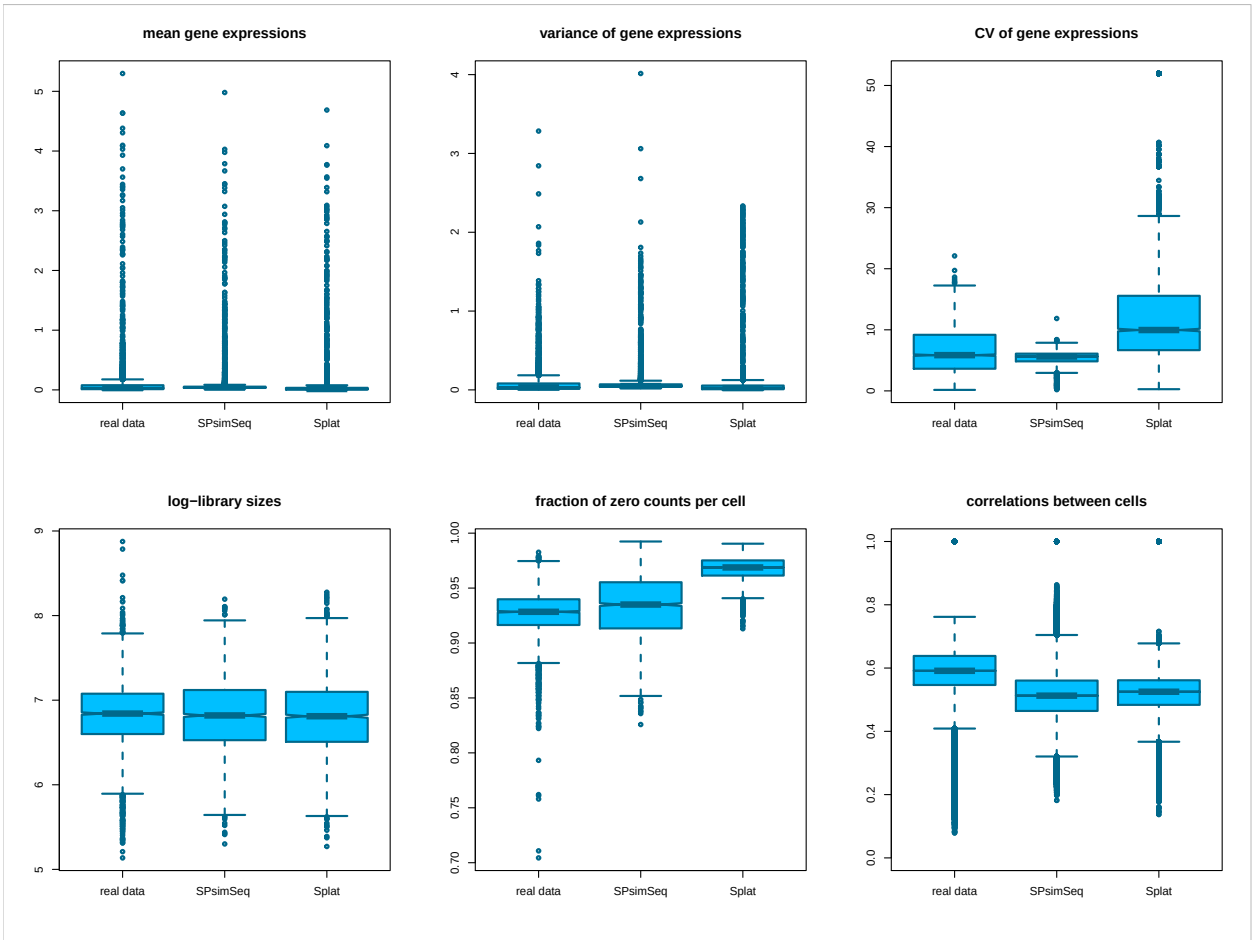

Figure S15: The distributions of the mean, variance and CV of gene expression levels (upper row), and the distributions of the log-library sizes (across cells), fraction of zero counts per cell and the correlations between cell (bottom row) from the real single-cell RNA-seq data (PBMC data) and the SPsimSeq and Splat simulated datasets.

the source dataset. This is because a sufficiently large K is required to accurately estimate of the distributions, especially for datasets that are characterized by multimodal distributions of the gene expression levels, such as single-cell RNA-seq data with read-counts. On the other hand, our results show that for single-cell RNA-seq data with UMI-counts, Sturges' rule works well because of the unimodal characteristics of UMI counts (Figures S22-S24). Generally, a very large or a very small K may result in a failure to estimate the distributions of gene expression levels and a very small K prohibits to accurately capture the shape of the distributions, for example for multimodal distributions.

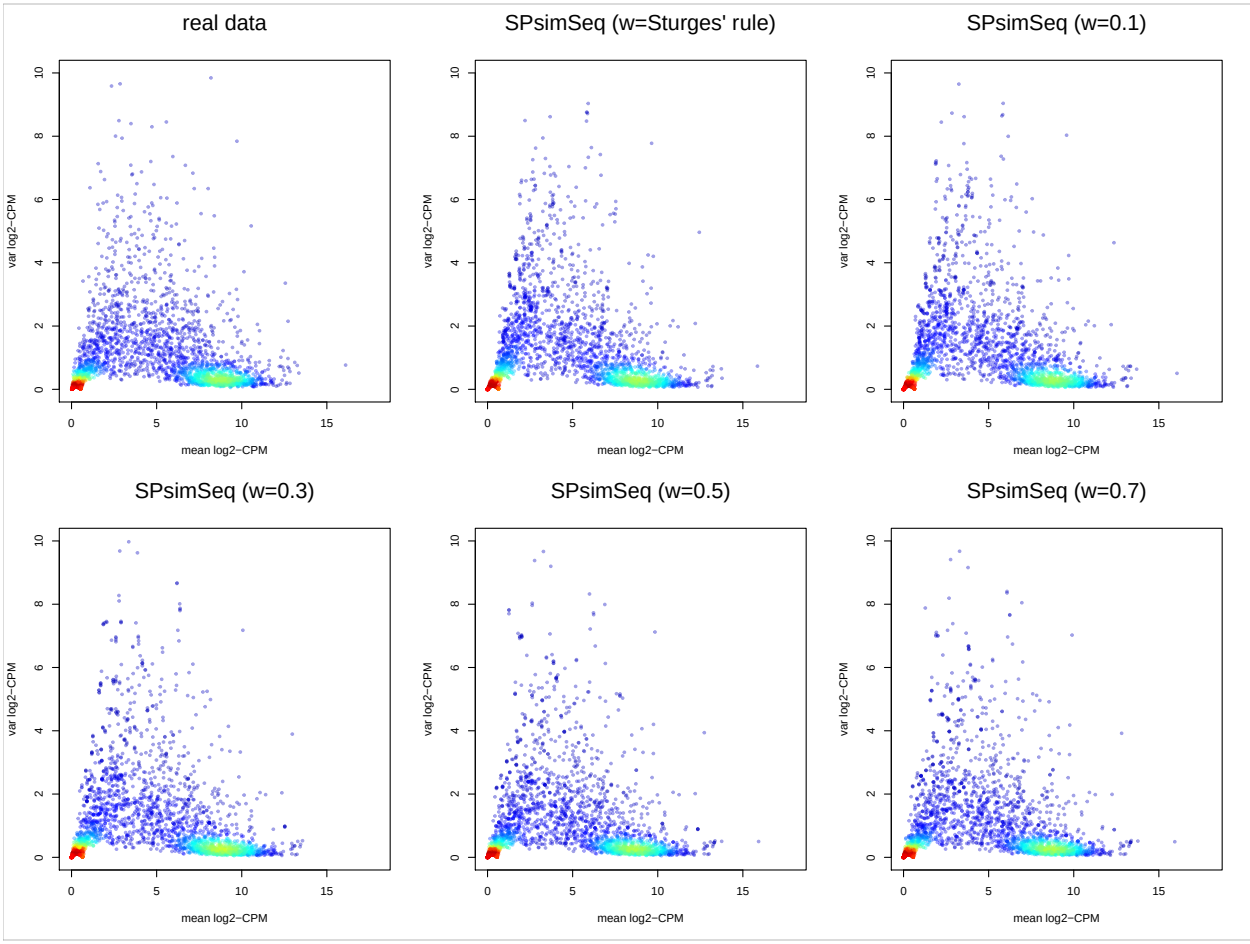

Figure S16: The relationship between the mean and variance of gene expression levels from a real bulk RNA-seq data and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

2.1 Bulk RNA-seq data

Results are shown in Figures S16-S24.

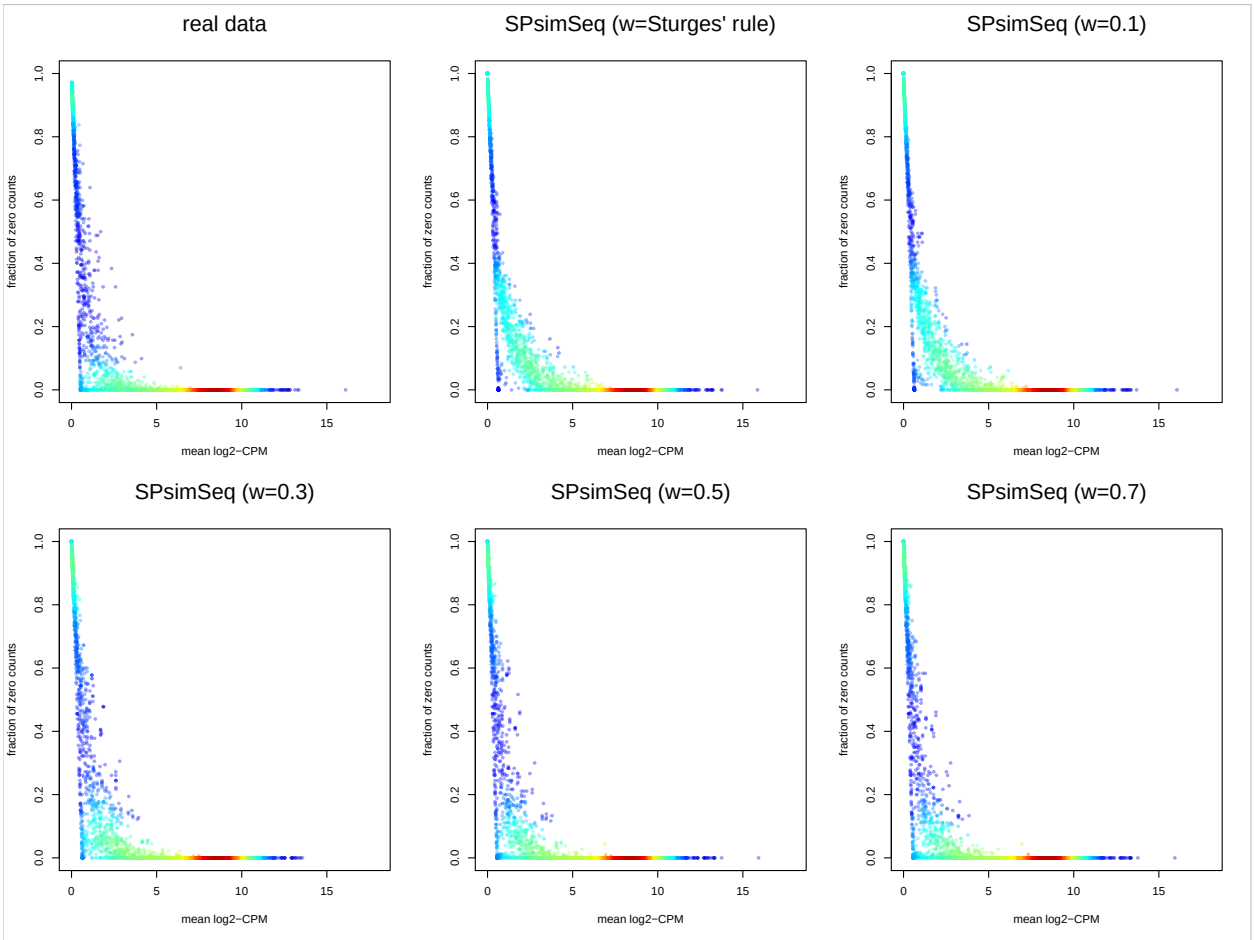

Figure S17: The relationship between the mean gene expression levels and the fraction of zero counts (per gene) from a real bulk RNA-seq data and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

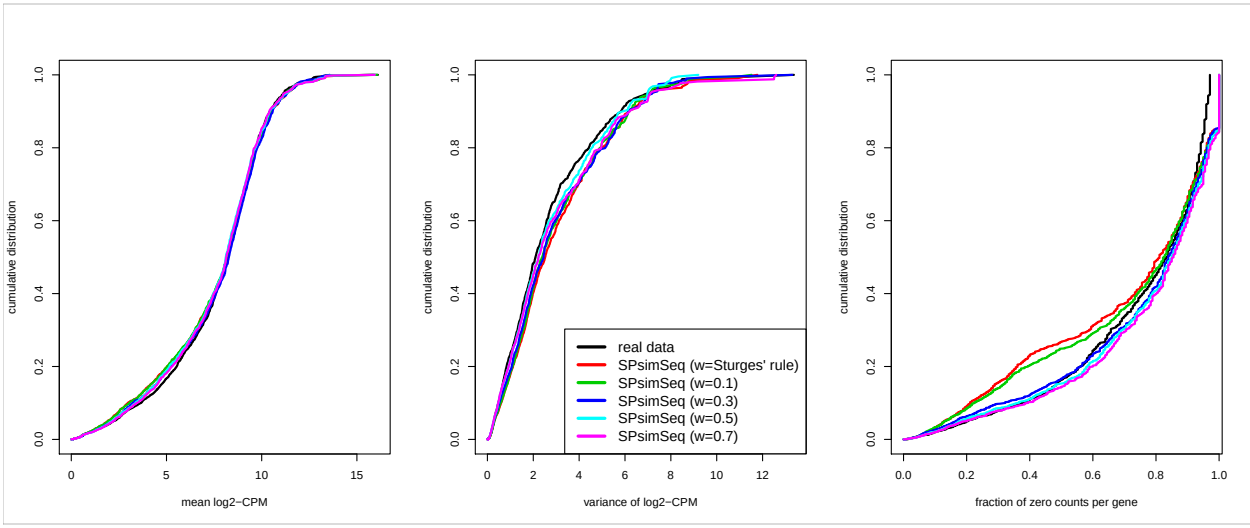

Figure S18: The cumulative distributions of the mean and variance of gene expression level and the fraction of zero counts per gene from a real bulk RNA-seq data and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

2.2 Single-cell RNA-seq data (read-count data)

Results are shown in Figures S19-S21.

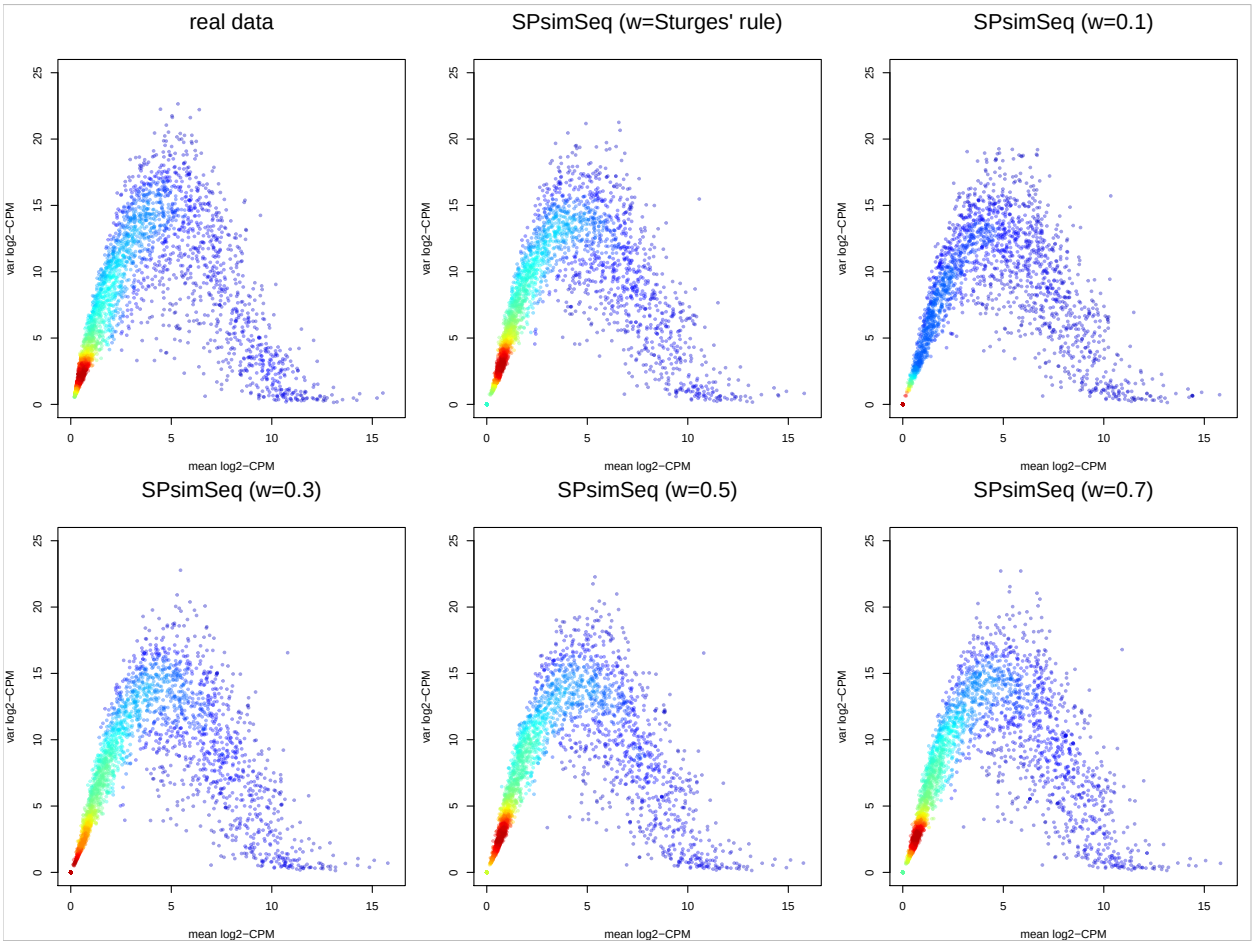

Figure S19: The relationship between the mean and variance of gene expression levels from a real single-cell RNA-seq data (read-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

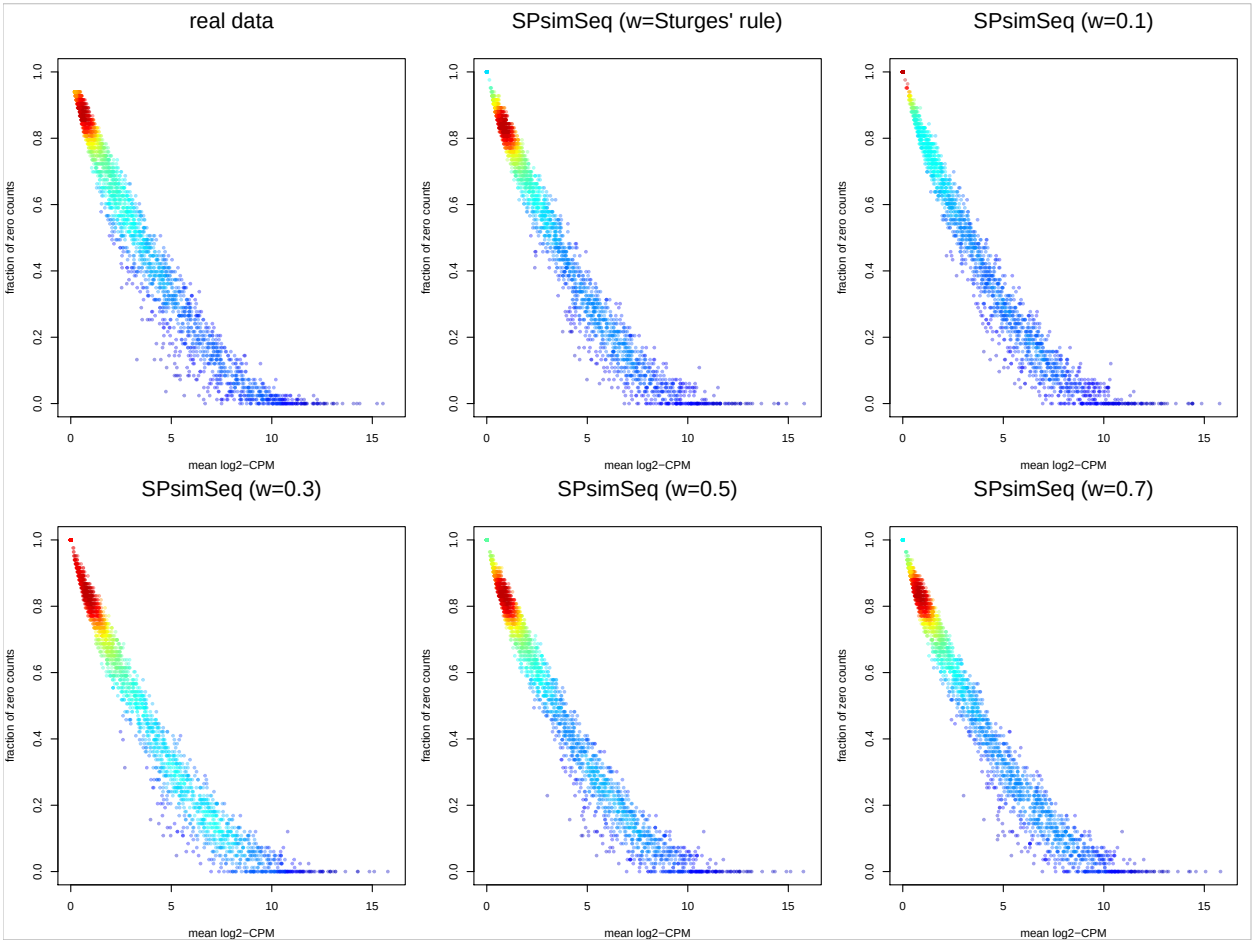

Figure S20: The relationship between the mean gene expression levels and the fraction of zero counts (per gene) from a real single-cell RNA-seq data (read-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

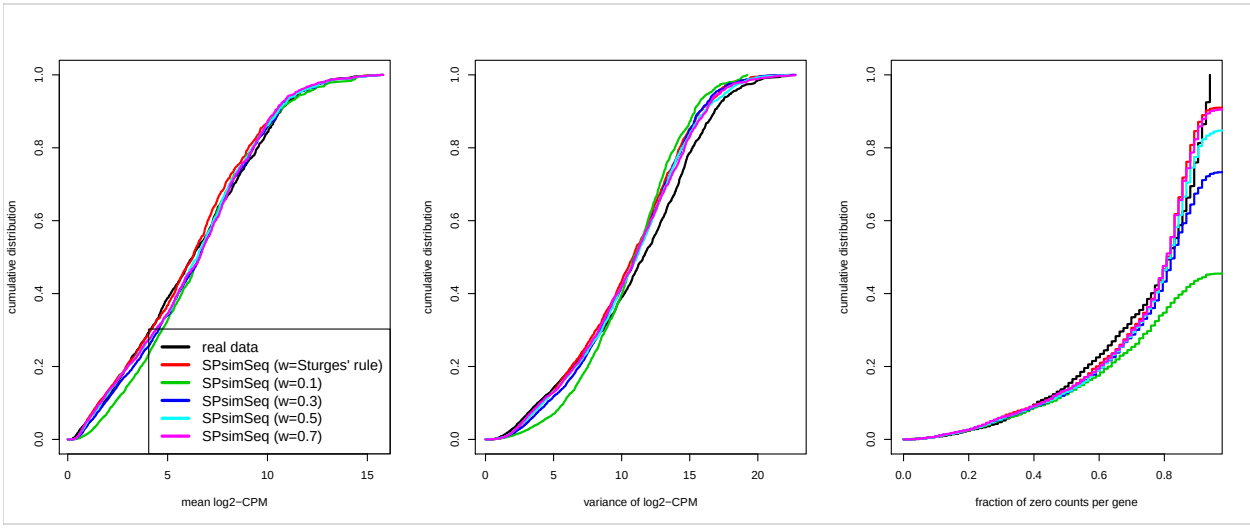

Figure S21: The cumulative distributions of the mean and variance of gene expression level and the fraction of zero counts per gene from a real single-cell RNA-seq data (read-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

2.3 Single-cell RNA-seq data (UMI-count data)

Results are shown in Figures S22-S24.

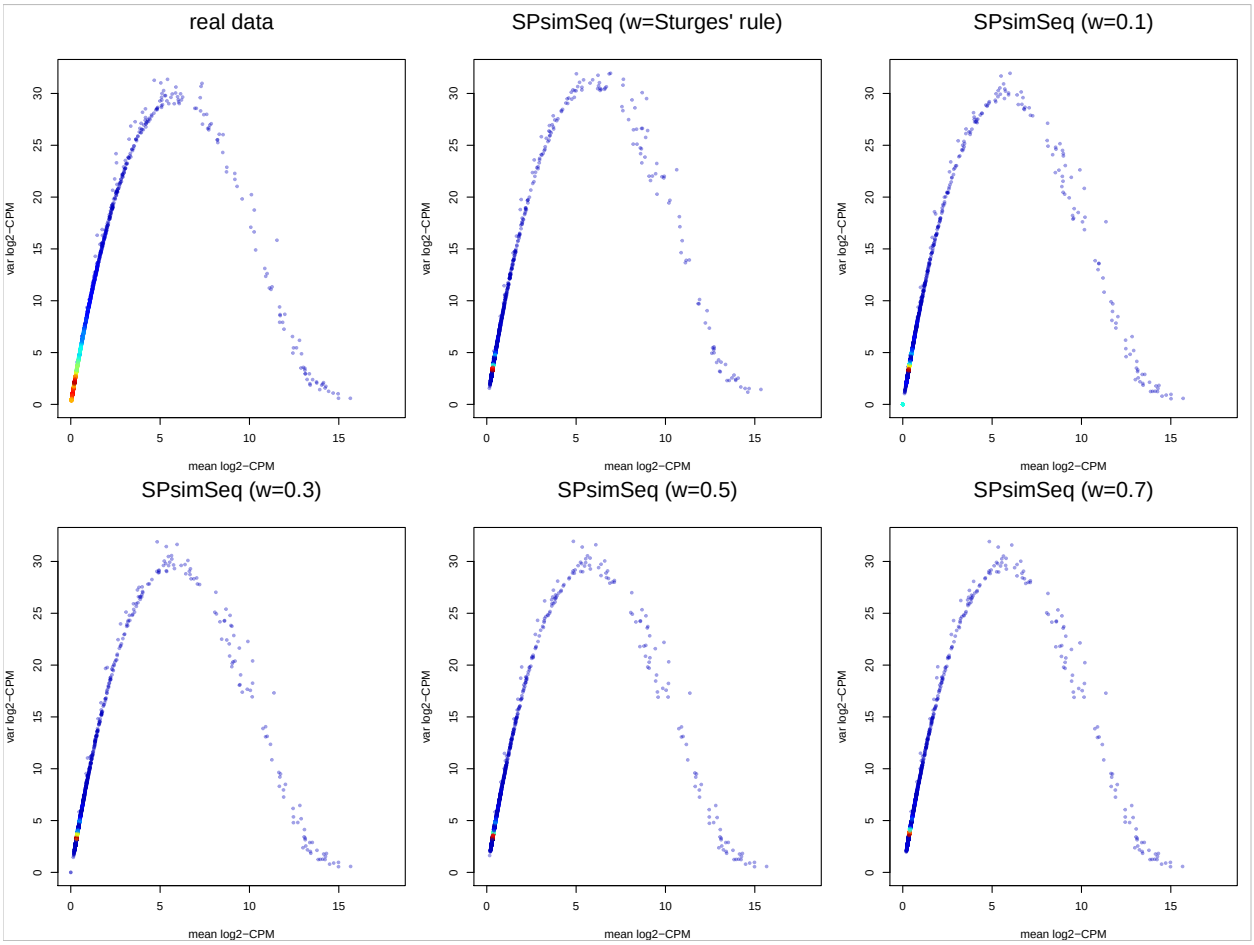

Figure S22: The relationship between the mean and variance of gene expression levels from a real single-cell RNA-seq data (UMI-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

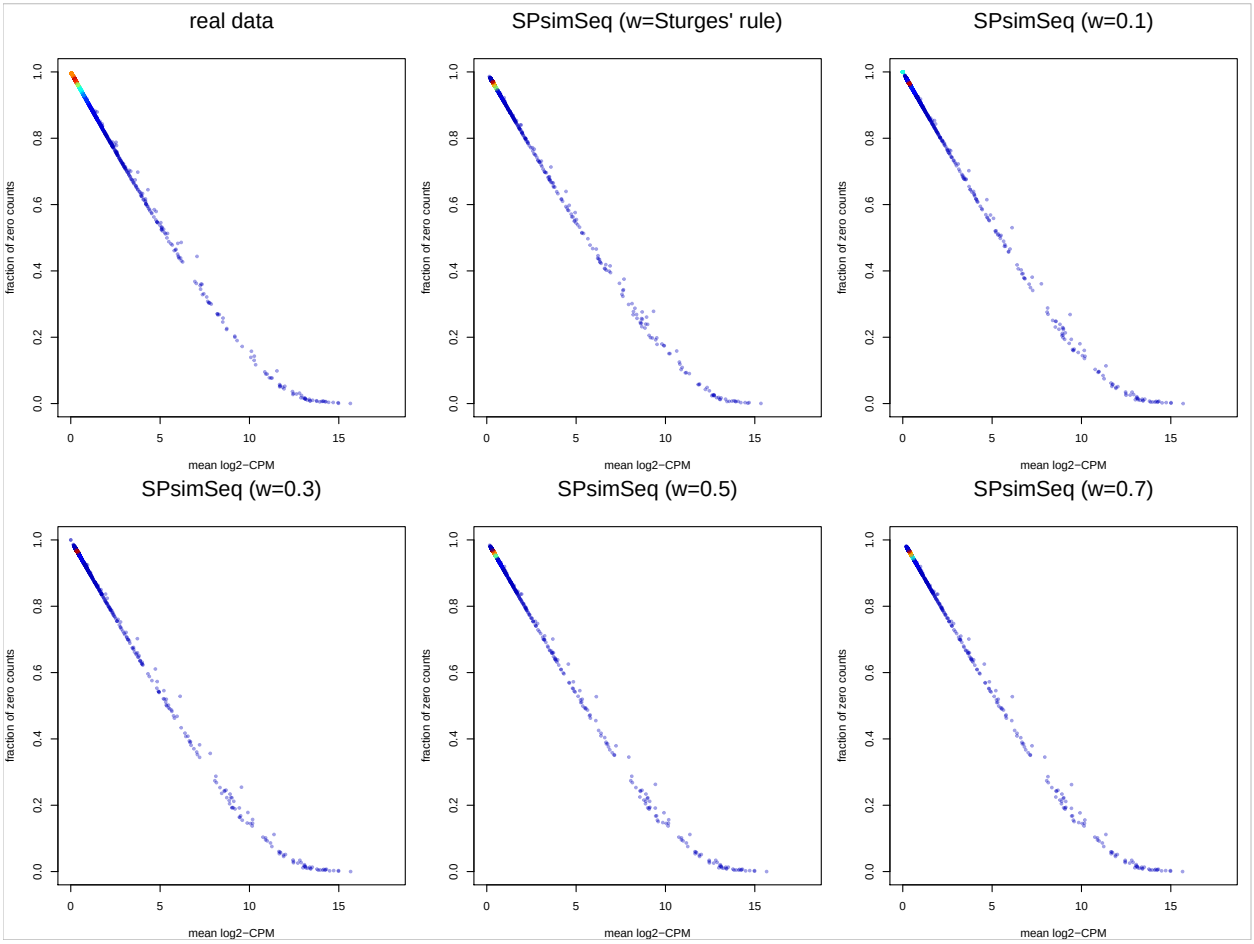

Figure S23: The relationship between the mean gene expression levels and the fraction of zero counts (per gene) from a real single-cell RNA-seq data (UMI-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

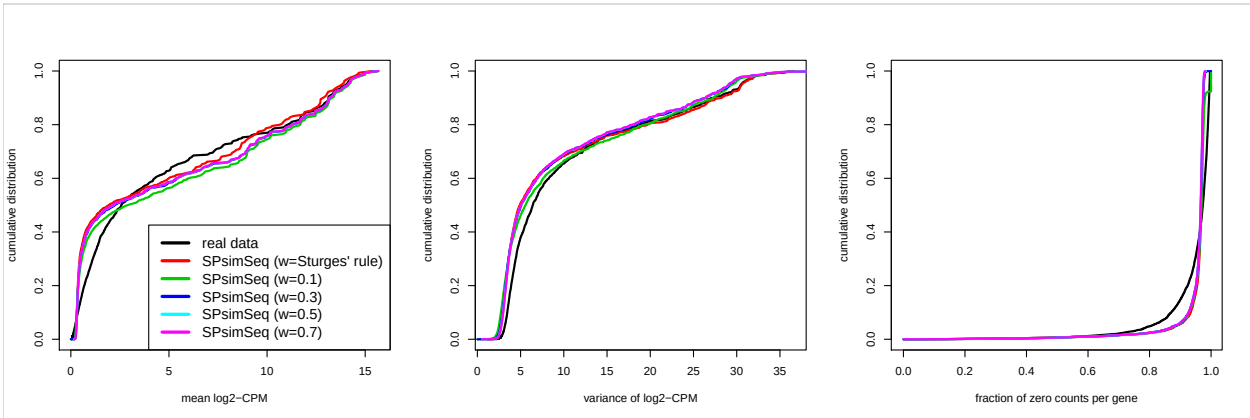

Figure S24: The cumulative distributions of the mean and variance of gene expression levels and the fraction of zero counts per gene from a real single-cell RNA-seq data (UMI-counts) and SPsimSeq simulated datasets with different choices of the number of classes K. In particular, for the SPsimSeq simulation, $w = K/n \in \{0.1, 0.3, 0.5, 0.7\}$ and K is determined using the Sturges' rule are compared.

3 The SPsimSeq R package

The latest version (version 2.0.0) of SPsimSeq package is available in https://github.com/CenterForStatistics-UGent/SPsimSeq. It can be installed using the following code

```
remotes::install_github("CenterForStatistics-UGent/SPsimSeq")
```

The package has the following dependencies (R packages): `stats`, `MASS`, `SingleCellExperiment`, `fitdistrplus`, `graphics`, `edgeR`, `Hmisc`, `WGCNA`, `limma`, and `mvtnorm`.

References

1. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome biology. 2015;16:133.

2. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016;34:525.

3. Assefa AT, De Paepe K, Everaert C, Mestdagh P, Thas O, Vandesompele J. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing

data. Genome Biology. 2018;19.

4. Verboom K, Everaert C, Bolduc N, Livak KJ, Yigit N, Rombaut D, et al. SMARTer single cell total RNA sequencing. Nucleic Acids Research. 2019;47:e93—e93.

5. Zappia L, Phipson B, Oshlack A. Splatter: Simulation of single-cell rna sequencing data. Genome biology. 2017;18:174.

6. Soneson C, Robinson MD. Towards unified quality verification of synthetic count data with countsimQC. Bioinformatics. 2017;34:691—692.

7. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature methods. 2014;11:163.

8. Sturges HA. The choice of a class interval. Journal of the American Statistical Association. 1926;21:65—66. doi:10.1080/01621459.1926.10502161.