SPsimSeq: semi-parametric simulation of bulk and single cell RNA sequencing data

June 20, 2019

Alemu Takele Assefa [1], Jo Vandesompele [2,3,4] and Olivier Thas [1,3,5,6,]

[1]Data analysis and mathematical modeling, Ghent University, Belgium,
[2]Biomolecular Medicine, Ghent University, Belgium,
[3]Cancer Research Institute Ghent, Ghent University, Belgium,
[4]Center for Medical Genetics, Ghent University, Belgium,
[5]National Institute for Applied Statistics Research, University of Wollongong, Australia, and
[6]I-BioStat, Hasselt University, Belgium

Summary: SPsimSeq is a semi-parametric simulation method for bulk and single cell RNA sequencing data. It simulates data from a good estimate of the actual distribution of a given real RNA-seq dataset. In contrast to existing approaches that assume a particular data distribution, our method constructs an empirical distribution of gene expression data from a given source RNA-seq experiment to faithfully capture the data characteristics of real data. Importantly, our method can be used to simulate a wide range of scenarios, such as single or multiple biological groups, systematic variations (e.g. confounding batch effects), and different sample sizes. It can also be used to simulate different gene expression units resulting from different library preparation protocols, such as read counts or UMI counts.
Availability and implementation: The R package and associated documentation is available from `https://github.com/CenterForStatistics-UGent/SPsimSeq`.
Supplementary information: Supplementary data are available at *bioRχiv* online.

Introduction

The number of computational tools for the analysis of bulk and single cell RNA sequencing (scRNA-seq) data is growing rapidly [8]. Several methods have been introduced for a single task, e.g. testing for differential gene expression (DGE). These tools typically pass through an evaluation process, often focusing on false discovery rate control and sensitivity. While such an evaluation often relies on simulated data with a built-in truth, to realistically assess the performance of these data analysis tools, the simulated data must faithfully recapitulate the data characteristics of real data [6, 5].

Various methods have been proposed for simulating either bulk or single cell RNA-seq data. The starting point is typically a distributional assumption of the gene expression data, for example the (zero inflated) negative binomial distribution [7]. While these parametric simulation methods are flexible and allow simulating various scenarios by generating synthetic data with good fit to the real data [5], such strong distributional assumptions do not hold in general. Due to the intrinsic biological variability and technical noise, scRNA-seq data sometimes show multimodal distributions [2]. There are also fully non-parametric approaches that employ subsampling from real data [3]. Although non-parametric simulators generate realistic synthetic data, they have limited flexibility and require a large source dataset to subsample from [1].

Here, we present a new simulation procedure for simulating bulk and single cell RNA-seq data. It is designed to maximally retain the characteristics of real RNA sequencing data with reasonable flexibility to simulate a wide range of scenarios. In a first step, the logarithmic counts per million (log-CPM) values from a given real data set are used for semi-parametrically estimating gene-wise distributions. This method is based on a fast log-linear model estimation approach developed by [4]. Arbitrarily large datasets, with realistically varying library sizes, can be sampled from these distributions. Our method has an additional step to explicitly account for the high abundance of zero counts, typical for scRNA-seq data. This step models the probability of zero counts as a function of the mean expression of the gene and the library size (read depth) of the cell (both in log scale). Zero counts are then added to the simulated data such that the observed relationship (zero probability to mean expression and library size) is maintained. In addition, our method simulates DGE by separately estimating the distributions of the gene expression from the different populations (for example treatment groups) in the source data, and subsequently sampling a new dataset from each group.

Our simulation procedure enables benchmarking of statistical and bio-informatics tools with realistic simulated data. In the result section, we demonstrate that the simulated data from our method retains the characteristics of the source data in terms of variability, distribution of mean expression, fraction of zero counts, and the relationship to each other (Figure 1). The details of the procedures and implementations can be found in the supplementary file. Data simulated with our procedure are compared with the original real source data and with data simulated with the parametric Splat procedure [7], which uses a gamma-Poisson hierarchical model (*splatter* R Bioconductor package, version 1.6.1, [7]).

Results

Using three different source RNA-seq datasets (one bulk and two single cell) we benchmarked the novel SPsimSeq simulation method. In particular, we compared the simulated data (using SPsimSeq and Splat) with the real data with respect to various gene and sample (cell) level characteristics as used by [5] and [7]. To simulate bulk RNA-seq data using Splat, we disabled its feature for adding dropouts (*dropout.type="none"*), which is specifically designed for scRNA-seq data simulation. The results generally show that our simulation procedure sufficiently captured the properties of the real data both for bulk and single cell RNA-seq (Figure 1 and supplementary file). The coefficients of variation, variability, distribution of mean expression and fraction of zero counts (per gene and sample/cells) in SPsimSeq simulated data resemble that of the real datasets. Compared with Splat, SPsimSeq generates more realistic data with respect to the majority of the considered metrics. In

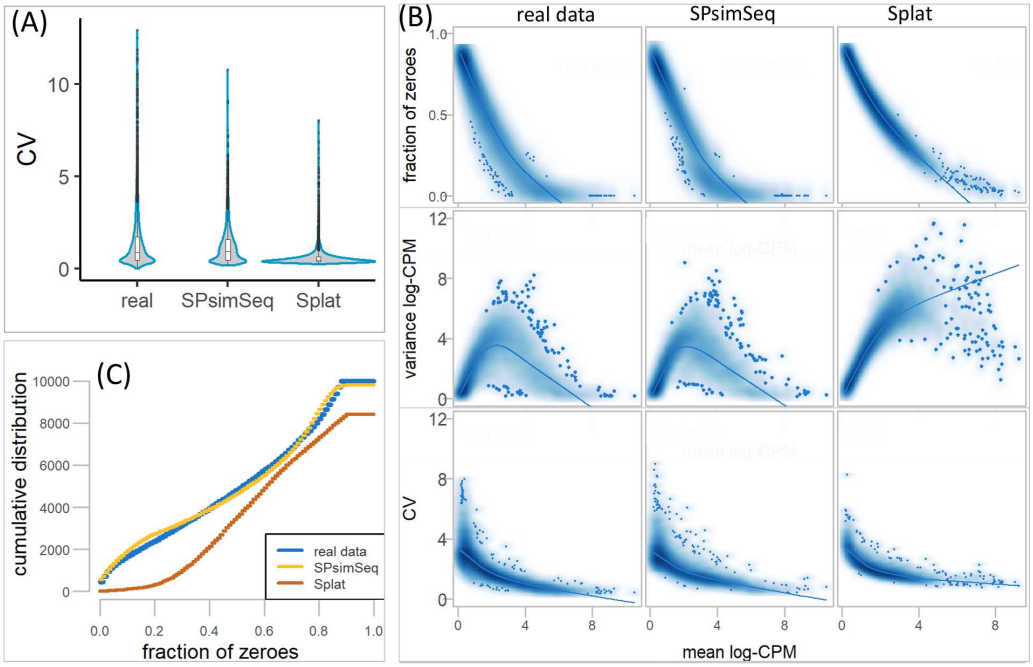

Figure 1: (A) Distribution of the coefficients of variations (CV) from the real and simulated (SPsimSeq and Splat) bulk RNA-seq datasets. (B) The relationship between the gene specific mean expression (in log-CPM) and three characteristics (fraction of zeroes, variance, and CV of each gene) from the real and simulated scRNA-seq datasets (read-counts). The curves show the smoothed relationship using *LOESS* regression. (C) The cumulative distribution of fraction of zero counts per gene.

the supplementary file, we present the detailed benchmarking results including the application of SPsimSeq for simulating scRNA-seq data with read-counts and UMI-counts (unique molecular identifier).

Funding

This work has been supported by the UGent Special Research Fund Concerted Research Actions (GOA grant number BOF16-GOA-023).

References

[1] A. T. Assefa, K. D. Paepe, C. Everaert, P. Mestdagh, O. Thas, and J. Vandesompele. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biology*, 19(1), jul 2018.

[2] R. Bacher and C. Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, 17(1):63, Apr 2016.

[3] S. Benidt and D. Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.

[4] B. Efron, R. Tibshirani, et al. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

[5] C. Soneson and M. D. Robinson. Towards unified quality verification of synthetic count data with countsimqc. *Bioinformatics*, 34(4):691–692, 2017.

[6] L. M. Weber, W. Saelens, R. Cannoodt, C. Soneson, Y. Saeys, and M. D. Robinson. Essential guidelines for computational method benchmarking. *arXiv preprint arXiv:1812.00661*, 2018.

[7] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[8] L. Zappia, B. Phipson, and A. Oshlack. Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLoS computational biology*, 14(6):e1006245, 2018.