View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Ghent University Academic Bibliography

Bioinformatics, 2020, 1–3
doi: 10.1093/bioinformatics/btaa105
Advance Access Publication Date: 17 February 2020
Applications Note

Gene expression

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

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

[1]Data Analysis and Mathematical Modeling, [2]Biomolecular Medicine, [3]Cancer Research Institute Ghent, [4]Center for Medical Genetics, Ghent University, Ghent, Belgium, [5]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia and [6]Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 19, 2019; revised on January 2, 2020; editorial decision on February 8, 2020; accepted on February 11, 2020

Abstract

Summary: SPsimSeq is a semi-parametric simulation method to generate bulk and single-cell RNA-sequencing data. It is designed to simulate gene expression data with maximal retention of the characteristics of real data. It is reasonably flexible to accommodate a wide range of experimental scenarios, including different sample sizes, biological signals (differential expression) and confounding batch effects.

Availability and implementation: The R package and associated documentation is available from https://github.com/CenterForStatistics-UGent/SPsimSeq.

Contact: alemutakele.assefa@ugent.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The number of computational tools for the analysis of bulk and single-cell RNA-sequencing (RNA-seq) data is growing rapidly (Zappia *et al.*, 2018). Several methods have been introduced for a single task, e.g. testing for differential expression (DE). These tools typically pass through an evaluation process, often focusing on false discovery rate control and sensitivity. Although 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 (Soneson and Robinson, 2018; Weber *et al.*, 2019).

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 levels, e.g. the (zero-inflated) negative binomial distribution (Zappia *et al.*, 2017). Although these parametric simulation methods are flexible and allow simulating various scenarios by generating synthetic data with good fit to the real data (Soneson and Robinson, 2018), such strong distributional assumptions do not hold in general. Due to the intrinsic biological variability and technical noise, single-cell RNA-seq data sometimes show multimodal distributions (Bacher and Kendziorski, 2016). There are also fully non-parametric approaches that employ subsampling from real data (Benidt and Nettleton, 2015). Although non-parametric simulators generate realistic synthetic data, they have limited flexibility and require a large source dataset to subsample from Assefa *et al.* (2018) and Benidt and Nettleton (2015).

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-seq data with reasonable flexibility to simulate a wide range of scenarios. In a first step, the logarithmic counts per millions of reads (log-CPM) values from a given real dataset are used for semi-parametrically estimating gene-wise distributions and the between-genes correlation structure. In particular, the estimation of the probability distributions uses the fast log-linear model-based density estimation approach developed by Efron and Tibshirani (1996) and Lindsey (1974). The method makes use of the Gaussian-copulas (Cario and Nelson, 1997) to retain the between-genes correlation structure, as implemented by Hawinkel *et al.* (2019) for parametric microbiome data simulation. Arbitrarily large datasets, with realistically varying library sizes, can be sampled from these distributions while maintaining the correlation structure between the genes. Our method has an additional step to explicitly account for the high abundance of zero counts, typical for single-cell RNA-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 DE by separately estimating the distributions of the gene expression from the different populations (e.g. treatment groups) in the source data, and subsequently sampling a new dataset from each group.

Our simulation procedure enables benchmarking of statistical and bioinformatics tools with realistic simulated datasets. In Section 3 and Supplementary Material, we demonstrate that the simulated data from our method retains the characteristics of the source data in terms of variability, distribution of mean expression levels,

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btaa105/5739438 by guest on 06 May 2020

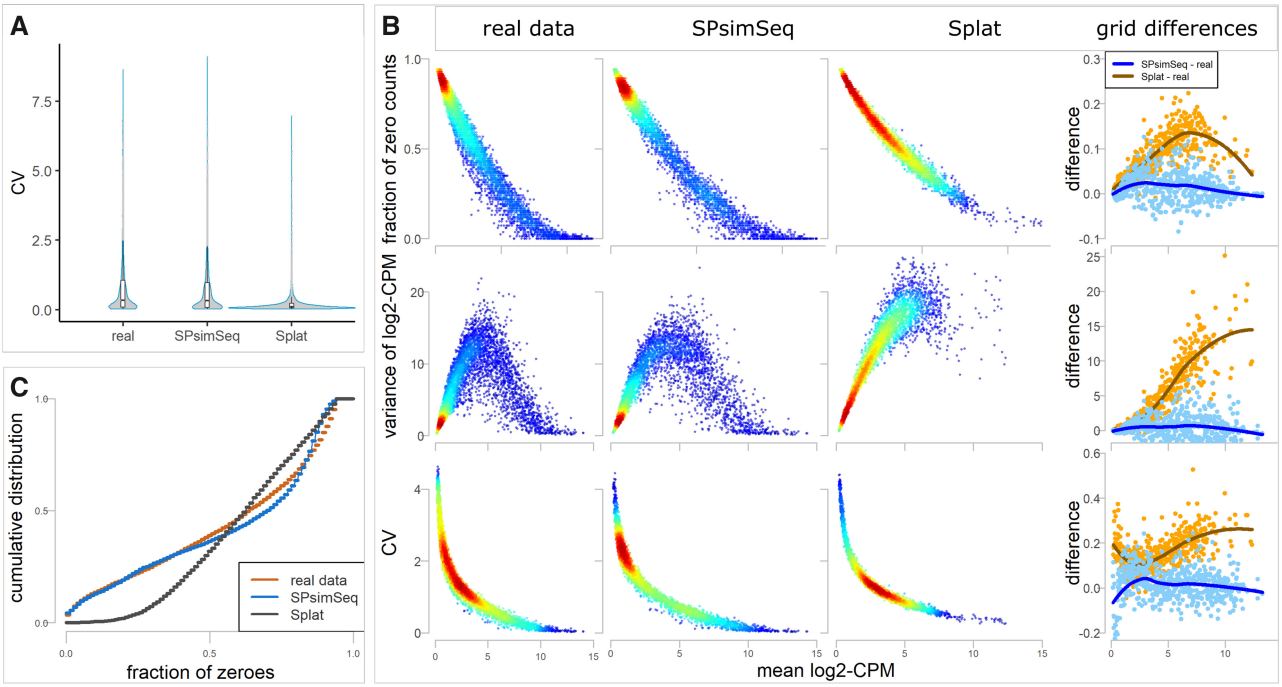

Fig. 1. (**A**) Distribution of the coefficients of variations (CV) from the real and simulated (SPsimSeq and Splat) bulk RNA-seq datasets. (**B**) The relationship between the mean gene expression levels (in log-CPM) and three other gene-level characteristics (fraction of zeroes, variance and CV) from the real and simulated single-cell RNA-seq datasets (read-counts). The right panels show the grid differences between the simulated and real data with respect to the corresponding gene-level characteristic. In particular, the mean log2-CPM is partitioned into 1000 grids and the average characteristic per grid is used to calculate the differences. (**C**) The cumulative distribution of the fraction of zero counts per gene from the real and simulated single-cell RNA-seq datasets (read-counts)

fraction of zero counts and the dependence between genes (Fig. 1 and Supplementary Material). The details of the SPsimSeq procedures, implementations and benchmarking results can be found in the Supplementary Material. Data simulated with our procedure are compared with the original real source data and with data simulated with the parametric Splat procedure (Zappia *et al.*, 2017), which uses a gamma-Poisson hierarchical model (*splatter* R Bioconductor package, v1.6.1; Zappia *et al.*, 2017).

2 Dataset

For the demonstration and benchmarking of the SPsimSeq method, we used three publicly available datasets: (i) neuroblastoma bulk RNA-seq data retrieved from Zhang *et al.* (2015; GEO accession GSE49711), (ii) neuroblastoma NGP cells single-cell RNA-seq data retrieved from Verboom *et al.* (2019; GEO accession: GSE119984) and (iii) peripheral blood mononuclear cell single-cell RNA-seq data retrieved from (www.10xGenomics.com). The details of the datasets can be found in the Supplementary Material.

3 Results

Using the three 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 suggested by Soneson and Robinson (2018) and Zappia *et al.* (2017). To simulate bulk RNA-seq data using Splat, we disabled its feature for adding dropouts (`dropout.type = "none"`), which is specifically designed for single-cell RNA-seq data simulation.

The results generally show that our simulation procedure sufficiently captures the properties of the real data both for bulk and single-cell RNA-seq datasets (Fig. 1 and Supplementary Figs S1–S14). The variability (in terms of variance and coefficients of variation), distribution of mean expression level of genes, the fraction of zero counts (per gene and sample/cells), the relationship between the mean and variability gene expressions, the relationship between the mean expression and fraction of zero counts and the dependence between genes in SPsimSeq simulated data resemble that of the real datasets. When compared with Splat, SPsimSeq generates more realistic data with respect to all the considered metrics. In the Supplementary Material, we present the detailed benchmarking results, including the application of SPsimSeq for simulating single-cell RNA-seq data with read-counts and UMI-counts (unique molecular identifier).

Funding

This work was supported by the UGent Special Research Fund Concerted Research Actions [GOA grant number BOF16-GOA-023].

Author's contributions

A.T.A.: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft and Writing—review and editing. J.V.: Conceptualization, Funding acquisition, Resources, Supervision and Writing—review and editing. O.T.: Conceptualization, Funding acquisition, Methodology, Supervision and Writing—review and editing.

Conflict of Interest: none declared.

References

Assefa,A.T. *et al.* (2018) Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biol.*, **19**, 96.

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

Benidt,S. and Nettleton,D. (2015) SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics*, **31**, 2131–2140.

Cario,M.C. and Nelson,B.L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. *Technical*

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btaa105/5739438 by guest on 06 May 2020

report. Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Efron,B. and Tibshirani,R. (1996) Using specially designed exponential families for density estimation. *Ann. Stat.*, **24**, 2431–2461.

Hawinkel,S. *et al.* (2019) A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.*, **20**, 210–221.

Lindsey,J. (1974) Construction and comparison of statistical models. *J. R. Stat. Soc. B*, **36**, 418–425.

Soneson,C. and Robinson,M.D. (2018) Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics*, **34**, 691–692.

Verboom,K. *et al.* (2019) SMARTer single cell total RNA sequencing. *Nucleic Acids Res.*, **47**, e93.

Weber,L.M. *et al.* (2019) Essential guidelines for computational method benchmarking. *Genome Biol.*, **20**, 125.

Zappia,L. *et al.* (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.

Zappia,L. *et al.* (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.*, **14**, e1006245.

Zhang,W. *et al.* (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.*, **16**, 133.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btaa105/5739438 by guest on 06 May 2020