SPARSim: taking account of data sparsity in single cell and 16S rRNA gene sequencing

Di Camillo B
06-28
12:15-12:30

Patuzzi I(1,2,3)† , Baruzzo G(1,3), Ricci A(2), Losasso C(2), Di Camillo B(1)

(1) Department of Information Engineering, University of Padova, Padova, Italy
(2) Department of Risk analysis and public health surveillance, Istituto Zooprofilattico Sperimentale delle Venezie, Legnaro, Italy
(3) These authors contributed equally to the present work

† Email: `barbara.dicamillo@unipd.it`

Motivation

Next generation sequencing (NGS) technologies allow for a massive acquisition of information and this has increasingly made RNA Sequencing (RNA-seq) the method of choice for transcriptome and metagenomics studies.

As previously stated by Aitchison [1] and recently recalled [2], data obtained from high-throughput sequencing are compositional, i.e. they are datasets in which the parts (genes or species) in each sample have as a constraint an arbitrary, non-informative sum (i.e. the sequencing depth).

Currently, bulk RNA-sequencing data are modelled as count data using the Negative Binomial (NB) distribution, a compound distribution obtained combining a Gamma distribution, which models the biological variability, and a Poisson model, that accounts for the sequencing process [3,4]. The Poisson model is an appropriate approximation for the binomial model used to describe the sequencing sampling procedure, as in bulk RNA-seq the number of trials (i.e. the sequencing depth) is considerably smaller than the number of original sequences. Under this condition, the effects of compositionality are negligible and the resulting internal dependencies are ignored.

However, the latest developments of NGS in both the recent applications of single-cell RNA-sequencing (scRNA-seq) and the more consolidated 16S rRNA gene sequencing (16S rRNA-seq) require accounting for pronounced sparsity. Not considering this characteristic leads to incorrect analysis and results. In fact, the low efficiency of capture methods used in scRNA-seq and the amplification bias in 16S rRNA-seq, jointly with the reduced sequencing depth deriving from sequencing multiplexing, make no longer acceptable to neglect the competition of original sequences for being sampled for sequencing. This means the sequencing process acts on original sequences as a sampling procedure without replacement and with a limited number of extractions, a process that can be statistically modelled as a Multivariate Hypergeometric (MH). The sequencing depth is represented by the constraint on the number of extractions and the internal competition between features (genes in scRNA-seq and species in 16S rRNA-seq) is accounted for by considering single trials as dependent from each other.

In this work, we show how simulating scRNA-seq and 16S rRNA-seq data following the MH model allows obtaining count data matrices that realistically resemble the typical high sparsity and overdispersion of real data, that are two of the most challenging characteristics to be modelled when simulating scRNA-seq and 16S rRNA-seq count data [5].

Methods

SPARSim simulator generates datasets with a user-specified number of sample groups (e.g. cell types in scRNA-seq and “body sites” in 16S rRNA-seq) and sample replicates. For each desired sample group, the simulation takes as input (or generate based on a given template dataset) a vector of average abundancies (16S rRNA-seq) or expressions (scRNA-seq), from which the biological samples are generated using a gene/species-specific Gamma distribution to model the biological variability. Then, the sequencing step is reproduced by sampling the wanted number of reads from each biological sample accordingly to a MH distribution, whose internal probabilities are defined by sample-specific proportional expressions. The zero count values rise naturally from the sampling procedure, following the real scenario in which rare species or low expressed genes result more frequently than other genes in extra zero counts because they are the most likely features not to be read (i.e. sampled) by the sequencer.

Results

SPARSim is able to generate count data matrices resembling real 16S rRNA-seq and scRNA-seq data. The user can decide whether to specify the required input parameters or to take advantage of a set of pre-coded scenarios that are integrated into the simulator.

SPARSim performances were evaluated in terms of ability in recreate realistic variability between samples, intensity distribution within samples and sparsity, both in an aggregated way (e.g. total count matrix sparsity) and as feature- and/or sample-specific characteristics (e.g. zeros distribution per feature and per sample). Additionally, we also tested the goodness in reproducing realistic alpha and beta diversity values when performing 16S rRNA-seq simulations and the similarity with real data multimodality when generating scRNA-seq datasets. To our knowledge, in 16S rRNA-seq context no count data simulator is available, while in single-cell framework only few simulators have been proposed to recreate count matrices, but they still present some important limitations. Moreover, this is the first time the MH model is applied in simulating compositional NGS data. Thus, we believe that SPARSim could be a valuable tool for researchers involved in developing and testing robust and reliable data analysis methods in the context of 16S rRNA-seq and scRNA-seq.

References

1. Aitchison, J. (1986). The statistical analysis of compositional data. Chapman and Hall, London, UK.

2. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology, 8, 2224. http://doi.org/10.3389/fmicb.2017.02224

3. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140. http://doi.org/10.1093/bioinformatics/btp616

4. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10), R106. http://doi.org/10.1186/gb-2010-11-10-r106

5. Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. Genome Biology, 18, 174. http://doi.org/10.1186/s13059-017-1305-0