Assessment of single cell RNA-seq statistical methods on microbiome data

Matteo Calgaro[1]*, Chiara Romualdi[2], Levi Waldron[3], Davide Risso[4,5], Nicola Vitulo[1]
*(lead presenter) matteo.calgaro_01@univr.it
[1] Department of Biotechnology, University of Verona, Italy
[2] Department of Biology, University of Padova, Padova, Italy.
[3] Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA.
[4] Department of Statistical Sciences, University of Padova, Padova, Italy.
[5] Division of Biostatistics and Epidemiology, Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA.

Keywords: differential abundance, microbiome, single-cell RNA-seq, false discovery, simulation.

The correct identification of differentially abundant microbial taxa between experimental conditions is a methodological and computational challenge. Recent work has shown that commonly used methods do not control the false discovery rate due to the peculiarity of these data (e.g. high sparsity), leading to an abundance of false positive results.

Since single-cell RNA-seq shares some of these peculiarities, we apply methods developed for single cell differential expression to microbiome data. We compare these approaches to methods developed for bulk RNA-seq and microbiome data, in terms of suitability of distributional assumptions, ability to control false discoveries, consistency, replicability, and power. We benchmark using 100 manually curated datasets from 16S and a whole metagenome shotgun sequencing. A simulation framework is developed to assess the impact of experimental design in power analysis.

Our analyses show that DESeq2 and limma-voom have the best performance. We recommend a careful exploratory data analysis prior to application of any inferential model and we present a framework to help scientists make an informed choice of analysis methods in a dataset-specific manner.