Notes on estimating technical variance in microbiome data

Quang Nguyen

April 6, 2020

1 Existing methods in microbiome field

1.1 Overview

The problem of estimating technical variability in high throughput sequencing studies is intrinsically linked to the process of data normalization [3]. Normalized features should be uncorrelated with technical processes such as library size and PCR amplification differences, and the variability in feature values should reflect true biological heterogeneity. Normalization in microbiome relative abundance data is still a highly debated issue [7]. The current consensus is that microbiome raw counts data should be treated as relative proportions, however studies differ in advocating for using the RNA-seq suite of methods [5] or log-ratio based methods [2]. The issue remains is that estimating technical variability and incorporating such metrics into analyses have not been a primary concern for microbiome researchers.

1.2 Hurdles

It is clear that an quantification of technical variability can be obtained with ample techincal replicates, positive controls and viral spike-ins. However, these approaches are highly inefficient due to sequencing costs, and there is a desire to perform this estimation with available data. However, there exists various hurdles to this tasks owing to the various quirks of microbiome data.

- First, similar to scRNAseq, microbiome data is highly sparse and noisy. The sparsity are due to either biological effects (true zeroes) or technical effects (technical zeroes). Estimating technical variability involves being able to consider true vs technical zeroes.
- Second, unique to microbiome data, there are no features that should be consistent across samples. Unlike RNA-seq data which has housekeeping genes with constitutive expression (however, some papers dispute this consistency due to heterogenous cell populations [6]), there exists no consistent feature for microbiomes. Differences in case/control status might me an a microbe is completely absent (not just differentially abundant) from a community. The lack of consistent features mean that there is no reference of which to attribute variability as technical.

1.3 Most current efforts

To my knowledge, there are currently only two methods that have approached removing and estimating technical variability. The first method is from a paper in 2014 [1], which is also the primary paper for the method ALDEx2 to perform differential abundance analysis (DAA). In this paper, for each sample, Monte Carlo draws of the Dirichlet multinomial distribution was performed (paper stated that 128 draws is enough) with a "non-informative prior" of 0.5 using raw counts.

$$p(n_1, n_2, ...| \sum N) = Dir([n_1, n_2, n_3, ...] + 1/2)$$

where $n_1, n_2, ...$ are raw counts assigned to each taxa (ASV). Each of these draws are then transformed using a standard log ratio transform (centered log-ratio) and analyzed separately. p-values are then combined across each draw (using mean). Essentially, the idea is that for each sample, we get random draws of the probability observing n_i counts assigned to feature i given the total library size. Ensembling these results accounts for the precision of measurements and ensure that technical variability is accounted for (similar to other ensemble methods like random forests). More recently, another method was published to estimate technical variability [4]. However, this study

References

[1] Andrew D. Fernandes, Jennifer NS Reid, Jean M. Macklaim, Thomas A. McMurrough, David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):15, May 2014.

[2] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.*, 8, 2017.

[3] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, December 2019.

[4] Brian W. Ji, Ravi U. Sheth, Purushottam D. Dixit, Yiming Huang, Andrew Kaufman, Harris H. Wang, and Dennis Vitkup. Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nature Methods*, 16(8):731–736, August 2019.

[5] Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, 10(4):e1003531, April 2014.

[6] Thomas P. Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F. Richardson, and Tamsyn M. Crowley. A field guide for the compositional analysis of any-omics data. Preprint, Bioinformatics, December 2018.

[7] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), December 2017.