Supplementary Materials

Quang P. Nguyen

October 14, 2020

1 Distribution of cILR

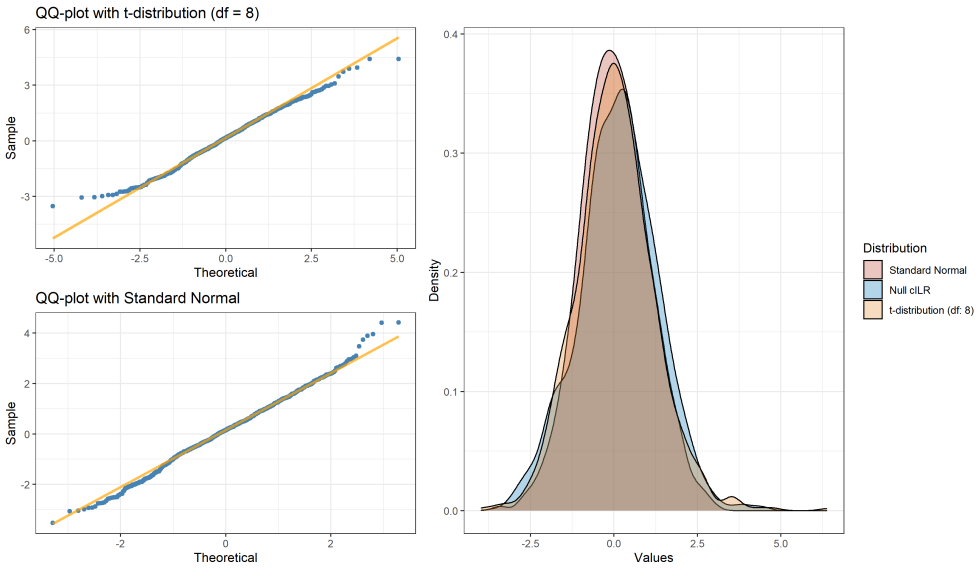

Figure 1: The distribution of cILR statistic under the null. We compared the null distribution of the test statistic and compare it with the standard normal distribution, and the t-distribution with degrees of freedom fitted to cILR scores using the maximum likelihood method

2 Simulation Design

We simulated microbiome relative abundance data using the NorTA method [2]. Using this method, we can generate synthetic microbial counts that incorporates a complex correlation structure and multiple types of marginals. The negative binomial distribution was chosen as it has been shown empirically to replicate the overall distribution of real microbiome count data [3, 1]. We fit the the parameters to our marginal model based on 16S rRNA sequencing of the V3-V5 region from stool samples in the Human Microbiome Project (HMP). This data was accquired via the *HMP16S* package in R[4].

2.1 Hypothesis testing at the sample level

For simulations for type I error control, we simulated 100 data sets of 1000 samples each with 1000 taxa per sample with no taxa being significantly enriched. Models were evaluate on a candidate set with varying sizes (50, 100, 150, 200). We also varied the overall sparsity (0.2, 0.4, 0.6, 0.8) and the degree of correlation (0.1,

0.2, 0.5).

For simulations of power and AUC classification capacity, we simulated 100 data sets of 1000 samples each with 1000 taxa per sample with one set of 50 taxa being significantly enriched in either all samples (power evaluation) or half the samples (AUC evaluation). We varied the overall sparsity (0.2, 0.4, 0.8), the correlation (0.1, 0.2, 0.5) and the effect size (2,4,6).

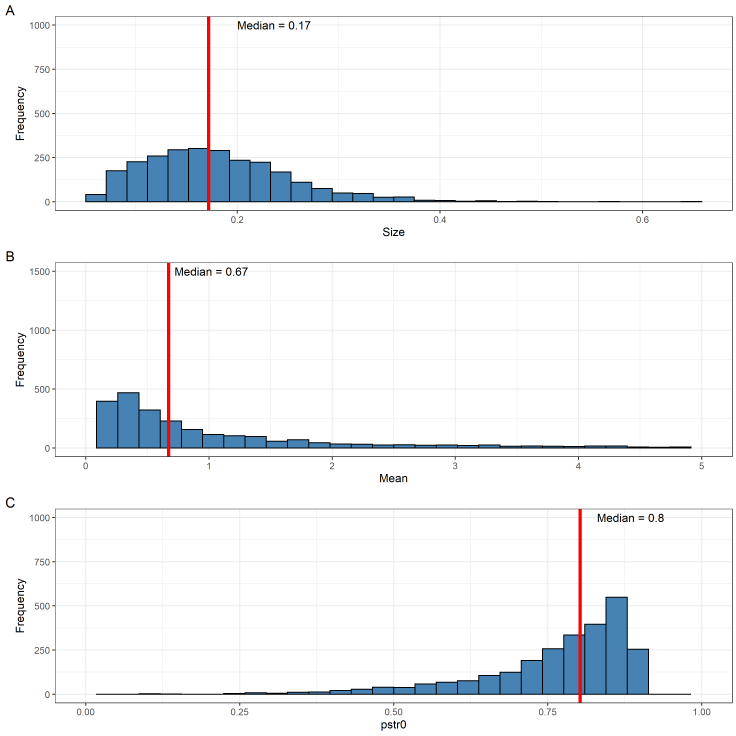

Figure 2: Distribution of each parameter of the zero inflated negative binomial distribution fitted to HMP16S data. The parameters are size (panel A), mean (panel B) and probability of 0 (panel C)

References

[1] Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology*, 21(1):191, August 2020.

[2] Marne C Cario. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. page 19.

[3] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015.

[4] Lucas Schiffer, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B Dowd, Nicola Segata, and Levi Waldron. HMP16SData: Efficient access to the human microbiome project through bioconductor. *American Journal of Epidemiology*, 2019.