cILR: Taxonomic Enrichment Analysis with Isometric Log-Ratios

Quang P. Nguyen[1,2], Anne G. Hoen[1,2], and H. Robert Frost[1]

[1]*Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA*
[2]*Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA*

Abstract

High-dimensionality and sparsity are challenging problems in statistical analysis of microbiome relative abundance data. One approach is to aggregate taxa to sets, most commonly to Linnean taxonomic categories identified through classification of representative sequences. However, most researchers perform aggregation through simple pairwise summation of counts. This approach is naive, whereby downstream analyses still suffer from the same statistical problems. To address this issue, we developed a competitive enrichment method based on the isometric log-ratio transformation (cILR) for microbiome relative abundance data. Our method generates sample-specific taxa group enrichment scores based on a taxa-by-group matrix with a well-defined null hypothesis allowing for inference at both the sample and population levels. Here we demonstrated the performance of our method for multiple microbiome analysis tasks.

Background

Methods

Results

In this section we present the performance of our proposed method for three applicable microbiome analysis tasks: sample level enrichment testing, differential abundance testing, and prediction. We obtained these results from both parametric simulations and examples from real data.

Enrichment testing at the sample level

There are various settings where researchers want to test for enrichment of certain groups of microbes in an experiment.

Type I error control and power

We benchmarked type I error on real stool microbiome data from HMP for both 16S and WGS type data. 16S data was taken from the package *HMP16SData* snapshot 2020-10-02.

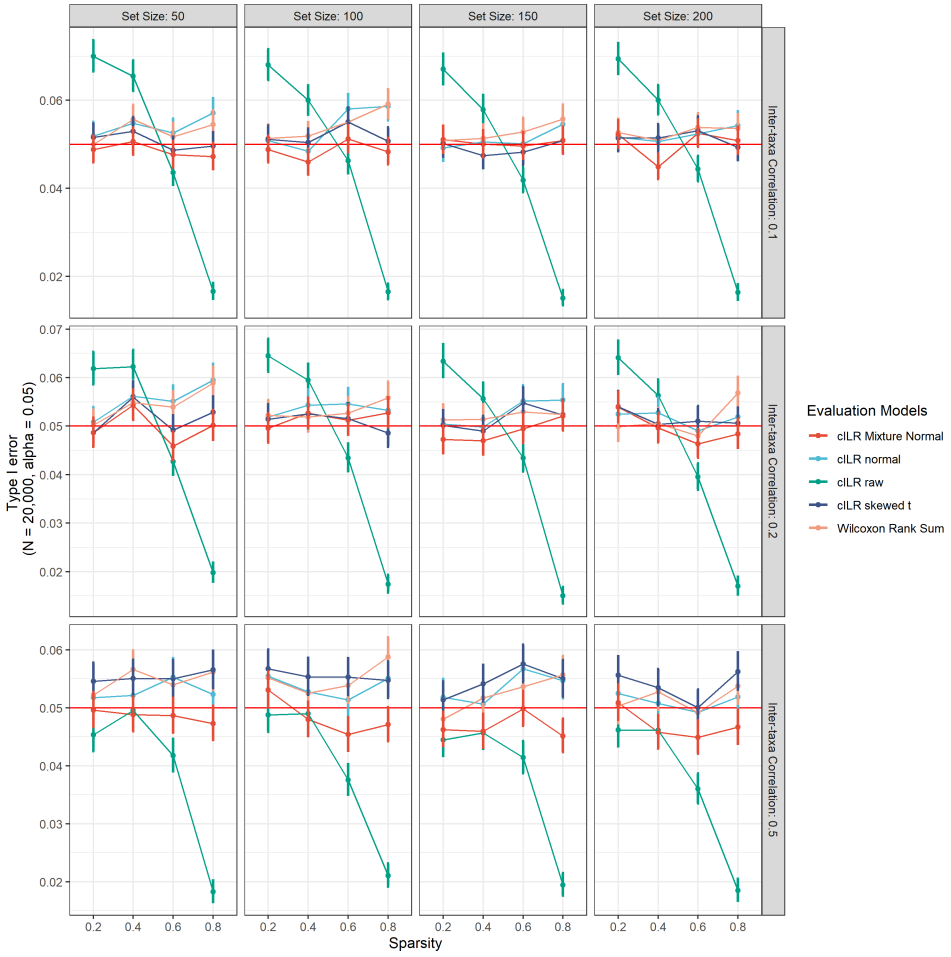

Figure 1. Median type I error rate as a function of data sparsity benchmarked on simulated null microbiome data as enumerated in SI methods. Enrichment of a specified set was tested at the sample level using clLR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different in set size (horizontal) and inter-taxa correlation (vertical)

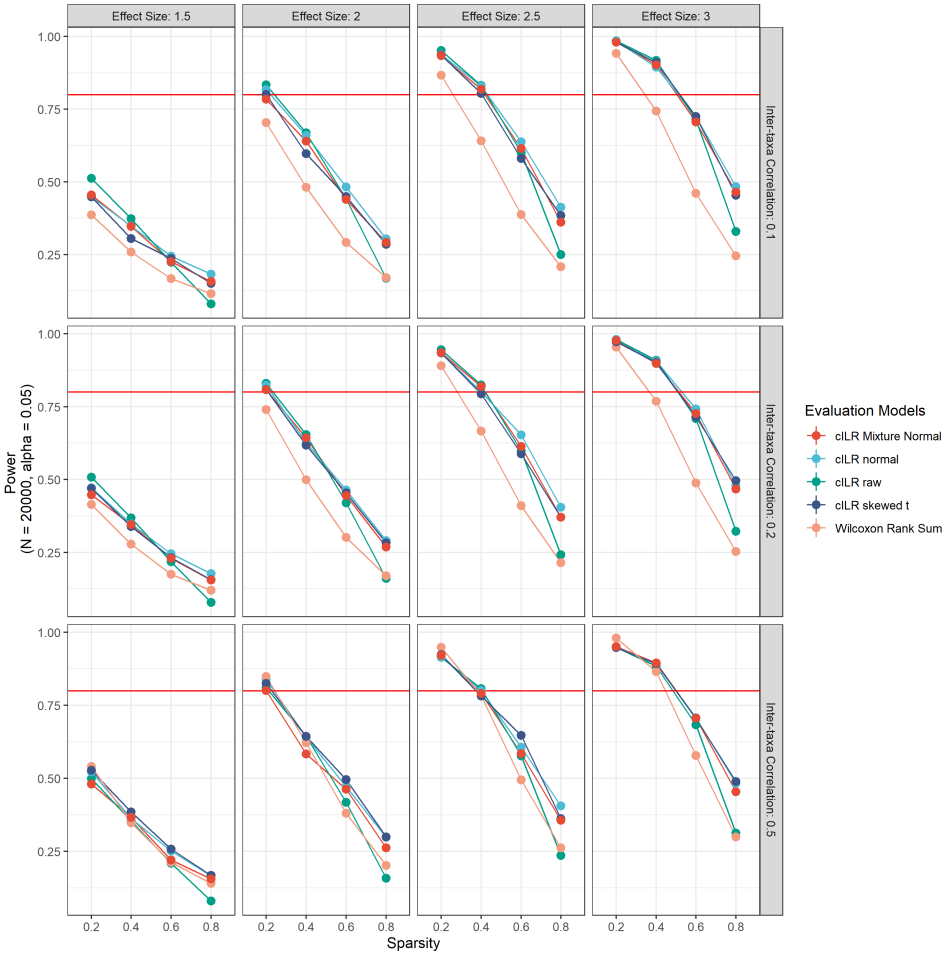

Figure 2. Median power as a function of data sparsity benchmarked on simulated microbiome data as enumerated in SI Methods. Enrichment of a specified set was tested at the sample level using cILR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different effect sizes (horizontal) and inter-taxa correlation (vertical).

Differential abundance analysis

Type I error control

We benchmarked type I error rate of the cILR approach in differential abundance analysis tasks on both real data and numerical experiments. For real data, we utilized 16S rRNA and WGS stool sequencing data from the Human Microbiome project obtained from the packages *HMP16SData* (ver. 1.9.3) and *curatedMetagenomicData* in R. We randomly assigned samples from each data set into two arbitrary groups and evaluated the type I error rate. This procedure was repeated 1000 times. Figure 3 demonstrated these results.
We observed

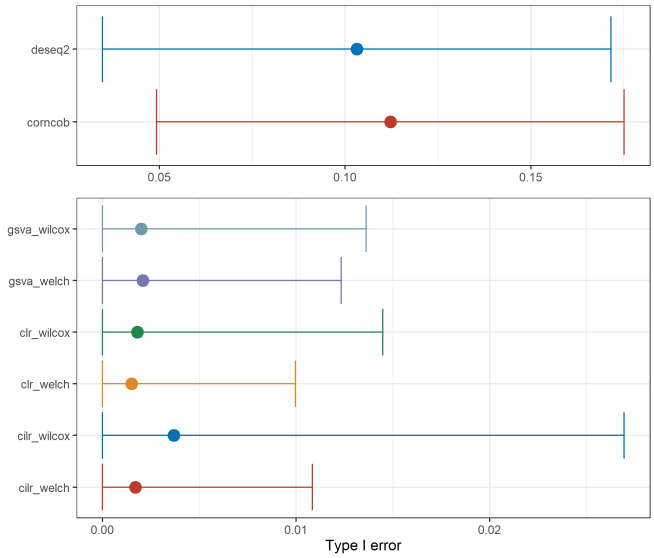

Figure 3. Type I error evaluated on 16S rRNA and WGS stool samples obtained from HMP. Enrichment of genus level taxa sets was tested across different methods where significance was determined at FDR cutoff of 0.05.

References