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 is a challenging problem in analyzing microbiome relative abundance data. Studies commonly alleviate this problem by aggregating variables into sets, most commonly higher order taxonomic classifications. However, such approaches are often naive and does not consider the hypothesis aggregation problem when testing for significance at multiple taxonomic levels. Here we introduced a novel competitive taxonomic enrichment method based on the isometric log-ratio transformation (cILR) for single samples. We demonstrated that our method controls type I error and power for hypothesis testing at the single sample level, as well as providing more robust results than other single sample enrichment methods for differential abundance and prediction tasks.

Background

Methods

Results

Hypothesis 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

Differential Abundance Analysis

Type I error control

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.

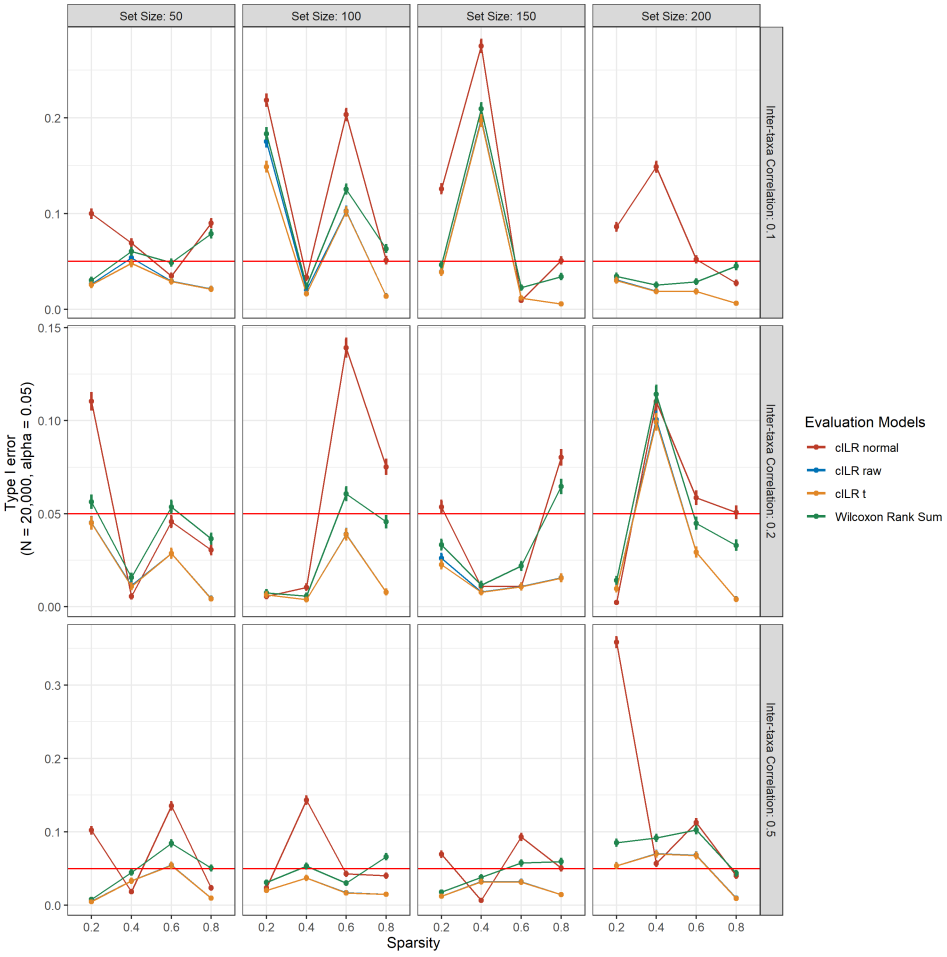

Figure 1. Median type I error rate as a function of data sparsity benchmarked on simulated null microbiome data as enumerated in SI methods. Enrichment of a specified set was tested at the sample level using clLR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different in set size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

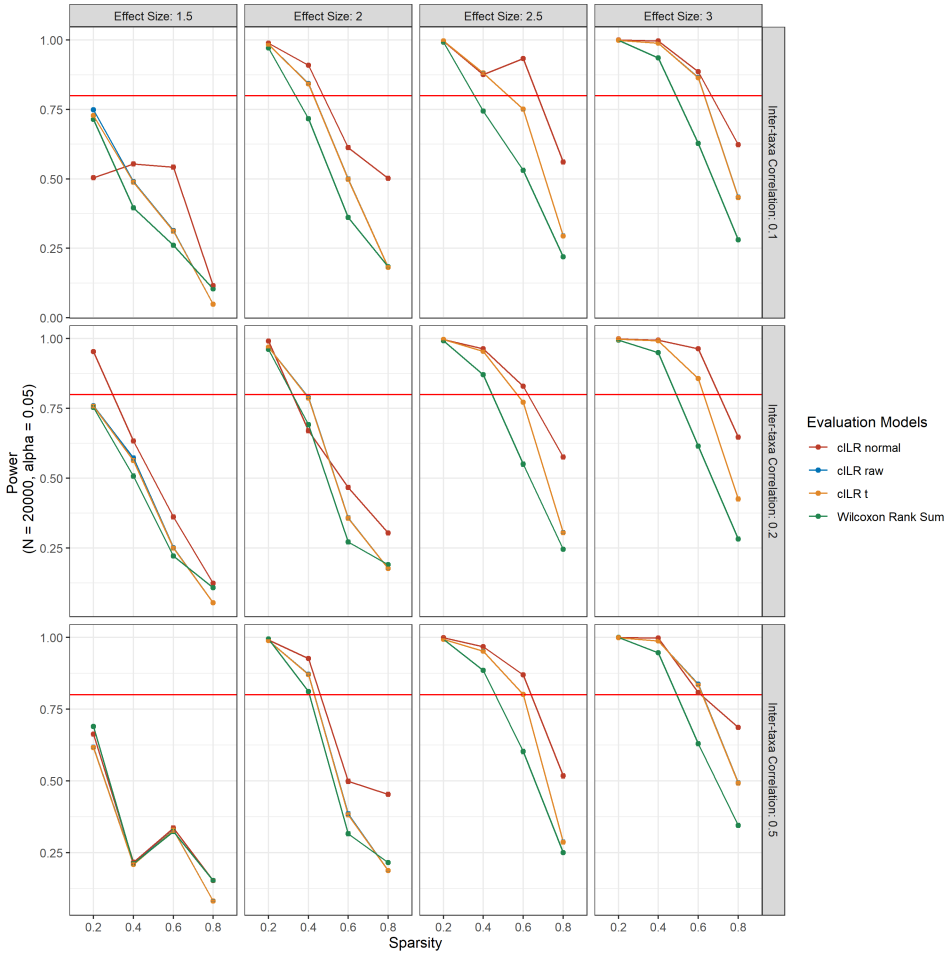

Figure 2. Median power as a function of data sparsity benchmarked on simulated microbiome data as enumerated in SI Methods. Enrichment of a specified set was tested at the sample level using clLR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different effect sizes (horizontal) and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

Real data analysis

Type I error control

We benchmarked type I error rate of the clLR 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

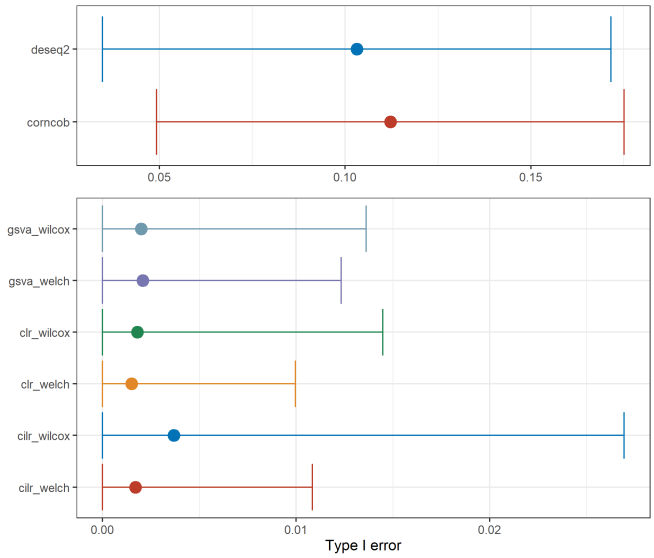

Figure 3. Type I error evaluated on 16S rRNA and WGS stool samples obtained from HMP. Enrichment of genus level taxa sets was tested across different methods where significance was determined at FDR cutoff of 0.05.

References