

# Taxonomic Enrichment Analysis with Isometric Log-Ratios

Quang P. Nguyen<sup>1,2</sup>, Anne G. Hoen<sup>1,2</sup>, and H. Robert Frost<sup>1</sup>

<sup>1</sup>*Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA*

<sup>2</sup>*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

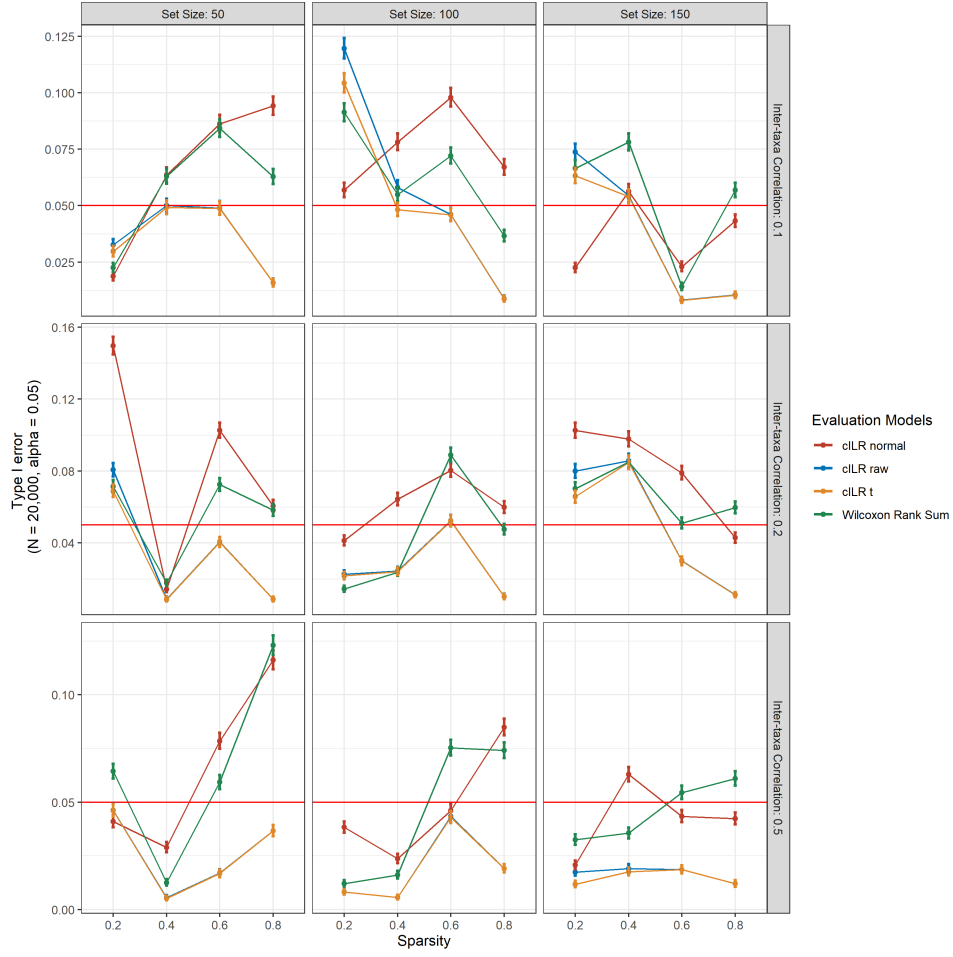
There are various settings where researchers want to test for enrichment of certain groups of microbes in an experiment.

### Type I error control and power

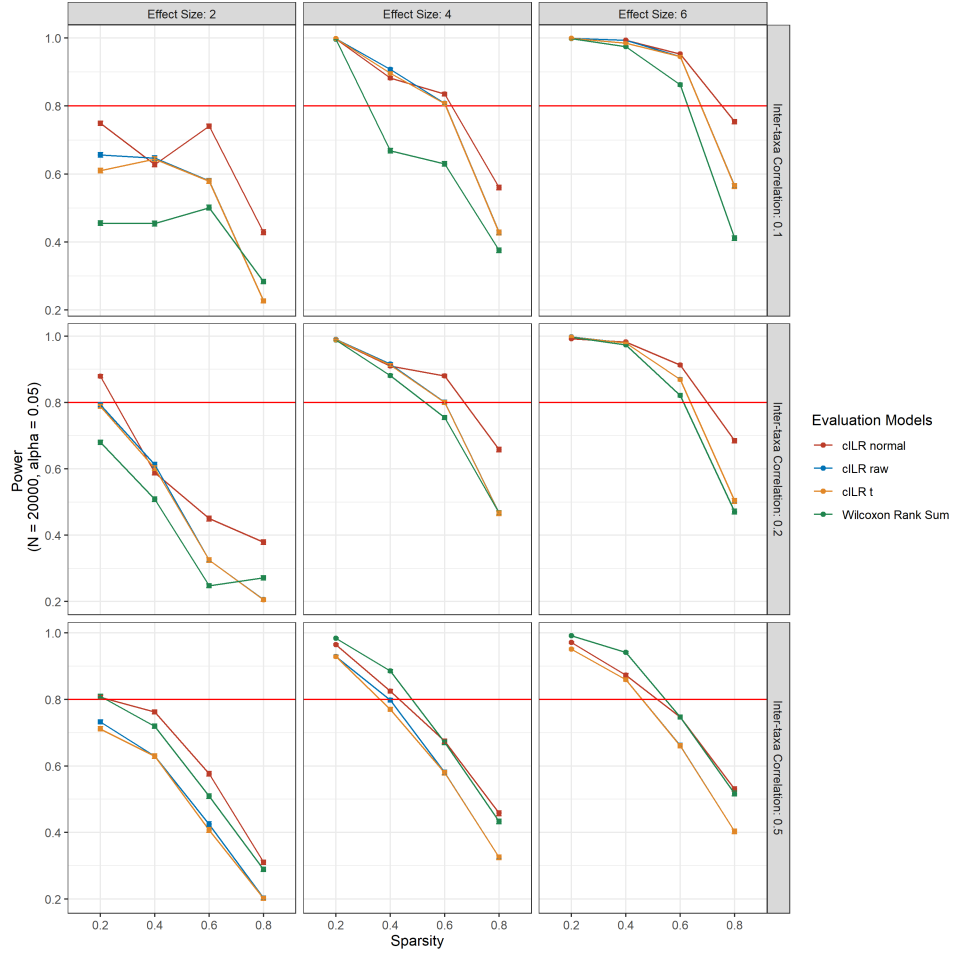
### Differential Abundance Analysis

### 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 cILR and the Wilcoxon rank sum test at  $\alpha$  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  $\alpha$  of 0.05. Each panel represents different effect sizes (horizontal) and inter-taxa correlation (vertical).

## Real data analysis

### Type I error control

We benchmark type I error rate on 16S and WGS data from the Human Microbiome Project (HMP) obtained from the packages *HMP16SData* (ver. 1.9.3) and *curatedMetagenomicData* packages in R. For each data set, we filtered out samples with library size less than 1000, as well as taxa with a proportion of zeroes of 0.9 or more. We then randomly assigned each sample into one of two arbitrary groups.

## References