

# cILR: 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 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. To address this issue, we developed a competitive set enrichment method based on the isometric log-ratio transformation (cILR) for microbiome relative abundance data. Our method generates sample-specific taxa set enrichment scores 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, including differential abundance testing and prediction.

## Background

## Methods

### Competitive Isometric Log-ratio (cILR)

The cILR method generates sample-specific enrichment scores for microbial sets using the isometric log-ratio transformation [2]. The cILR method takes two inputs:

- **X**:  $n$  by  $p$  matrix of positive counts for  $p$  taxa and  $n$  samples measured through either targeted sequencing (such as 16S rRNA) or whole genome shotgun sequencing. Usually **X** is generated from standard sequence processing pipelines such as DADA2 [1] and MetaPhlAn2 [4].
- **A**:  $p$  by  $m$  indicator matrix annotating the membership of each taxa  $p$  to  $m$  sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [3], or those based on more functionally driven categories such as the tropism of each microbe ( $a_{i,j} = 1$  indicates that microbe  $i$  belongs to set  $j$ ).

The cILR method generates one output:

- **E**:  $n$  by  $m$  matrix indicating the enrichment score of  $m$  pre-defined sets identified in **A** across  $n$  samples.

The procedure is as follows:

1. **Compute the cILR statistic**: Let **M** be a  $n$  by  $m$  matrix of cILR scores. Let  $M_{i,k}$  be cILR scores for set  $k$  of sample  $i$ :

$$M_{i,k} = \sqrt{\frac{\sum_k A_{ik}(p - \sum_k A_{ik})}{p}} \ln \left( \frac{g(\mathbf{X}_{i,j} | \mathbf{A}_{j,k} = 1)}{g(\mathbf{X}_{i,j} | \mathbf{A}_{j,k} \neq 1)} \right)$$

where  $g()$  is the geometric mean. This represents the ratio of the geometric mean of the relative abundance of taxa assigned to set  $k$  and remainder taxa.

2. **Compute the cILR statistic on permuted  $\mathbf{X}$ :** We seek to evaluate the empirical null distribution of the cILR statistic under  $H_0$  that relative abundances in  $\mathbf{X}$  of members of set  $k$  are not enriched compared to those not in set  $k$ . Let  $\mathbf{X}_p$  be the column permuted relative abundance matrix, and  $\mathbf{M}_p$  be the corresponding cILR scores generated from  $\mathbf{X}_p$ .
3. **Fit Gaussian mixture distribution for each column of  $\mathbf{M}_p$**

## Results

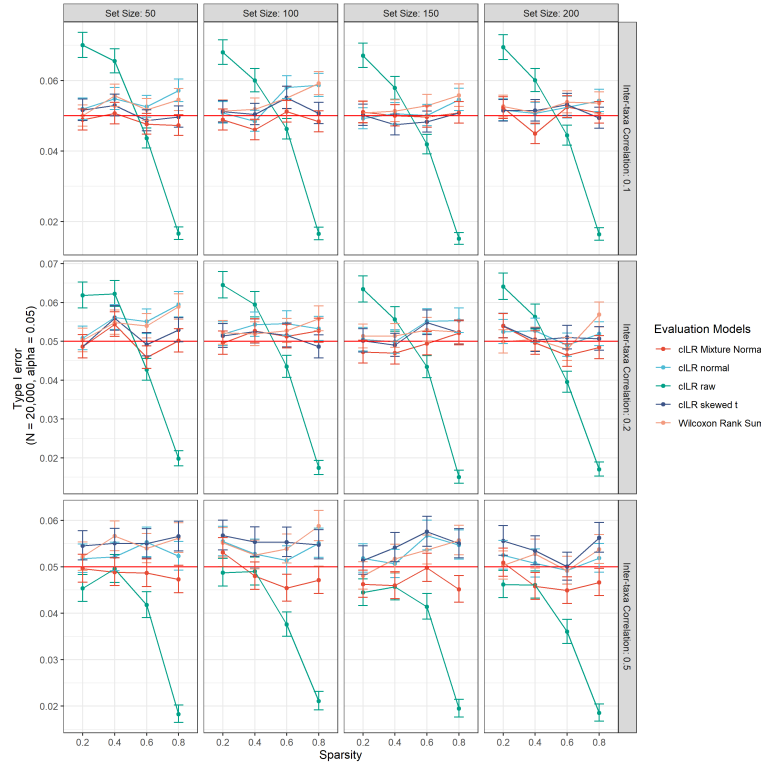
In this section we present the performance of our proposed method for three applicable microbiome analysis tasks: sample level enrichment, differential abundance, and prediction. We obtained these results from both parametric simulations and examples from real data.

### Enrichment testing at the sample level

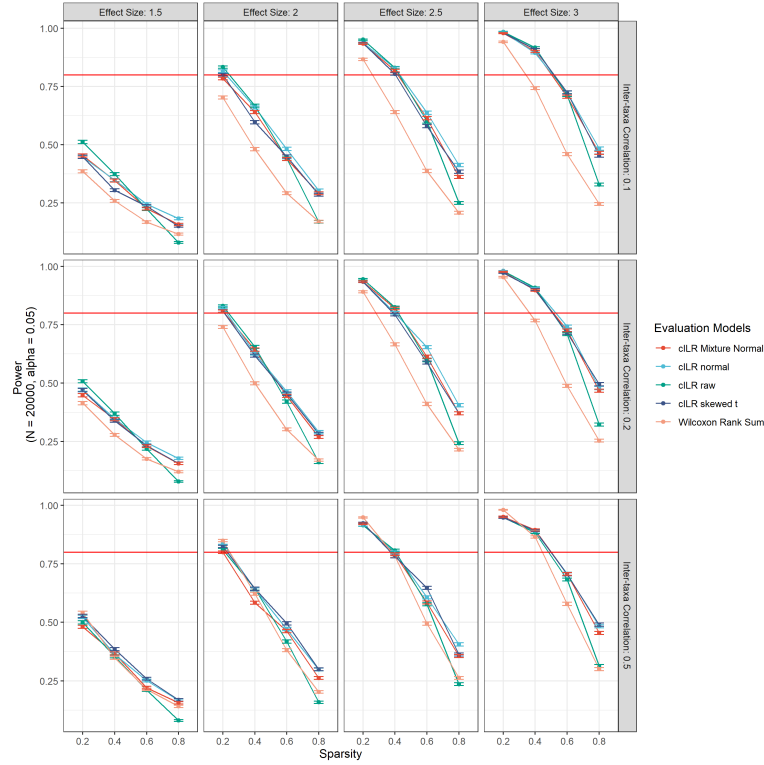
There are various settings where researchers want to test for enrichment of certain groups of microbes in an experiment.

#### 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 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)



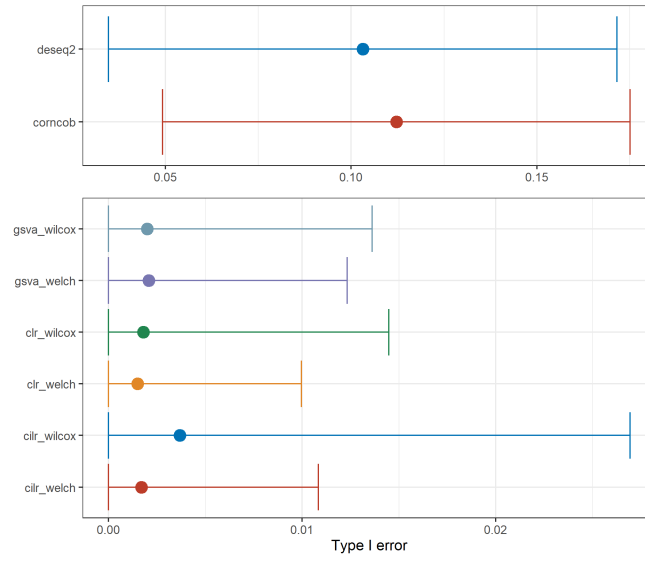
**Figure 2.** Median power as a function of data sparsity benchmarked on simulated microbiome data as enumerated in SI Methods. Enrichment of a specified set was tested at the sample level using cLR and the Wilcoxon rank sum test at  $\alpha$  of 0.05. Each panel represents different effect sizes (horizontal)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

## Differential abundance analysis

### Type I error control

We benchmarked type I error rate of the cLR 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

- [1] Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, July 2016.
- [2] J J Egozcue, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barcelo-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, page 22, 2003.
- [3] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, January 2013.
- [4] Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasoli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, October 2015.