clLR: 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. To address this issue, we developed a competitive set enrichment method based on the isometric log-ratio transformation (clLR) 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 (clLR)

The clLR method generates sample-specific enrichment scores for microbial sets using the isometric log-ratio transformation [3]. The clLR method takes two inputs:

- $\mathbf{X}$: 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 sequening. Usually $\mathbf{X}$ is generated from standard sequence processing pipelines such as DADA2 [2] and MetaPhlAn2 [6].
- $\mathbf{A}$: 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 [5], 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 clLR method generates one output:

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

The procedure is as follows:

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

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

   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_o 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 $\mathbf{M}_p$**
4. **Calculate finalized cILR scores as CDF values of the fitted mixture distribution**

Properties of cILR

cILR and the Isometric Log Ratio Transformation

Microbiome data is compositional [4] and as such data transformation strategies from compositional data analysis (CoDA) [1] were often employed in analyses workflows.

Null distribution of cILR

Accounting for inter-taxa correlation

Competitive enrichment tests, particularly those based on feature permutations, are sensitive to inter-feature correlations [7]. This is because the permutation procedure does not preserve the correlation structure, and any estimation based on the permuted null will underestimate the inflation in variance caused by

Simulation Design

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

Our method provides a well-defined null hypothesis for significance testing at the sample level. Researchers who are interested in the enrichment of certain groups of microbes (for example, those of the *Bifidobacterium* genus in infant gut microbiomes)

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.

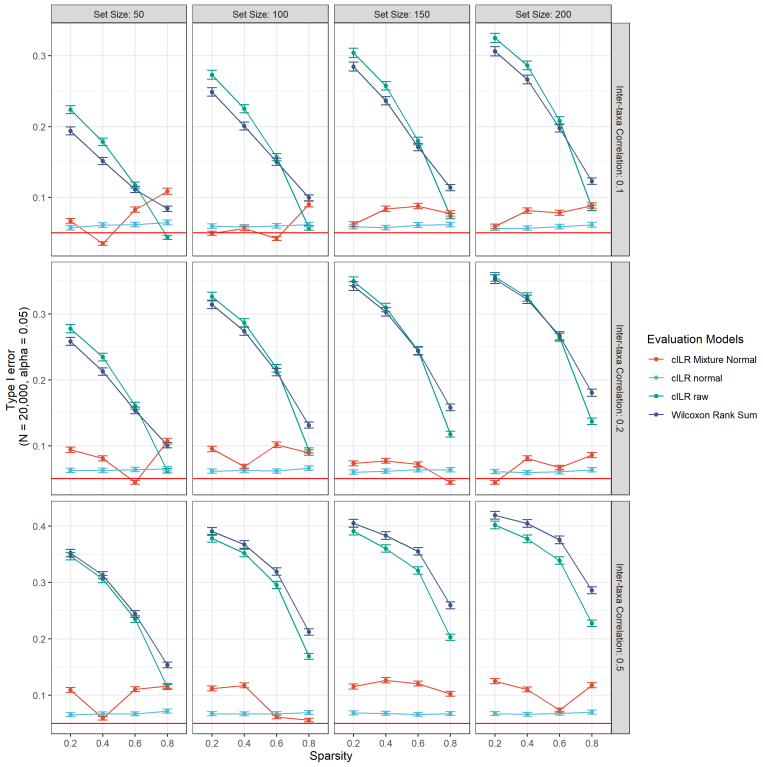

Figure 1. Median type I error rate as a function of data sparsity benchmarked on simulated null microbiome data as enumerated in SI methods. Enrichment of a specified set was tested at the sample level using clLR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different in set size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

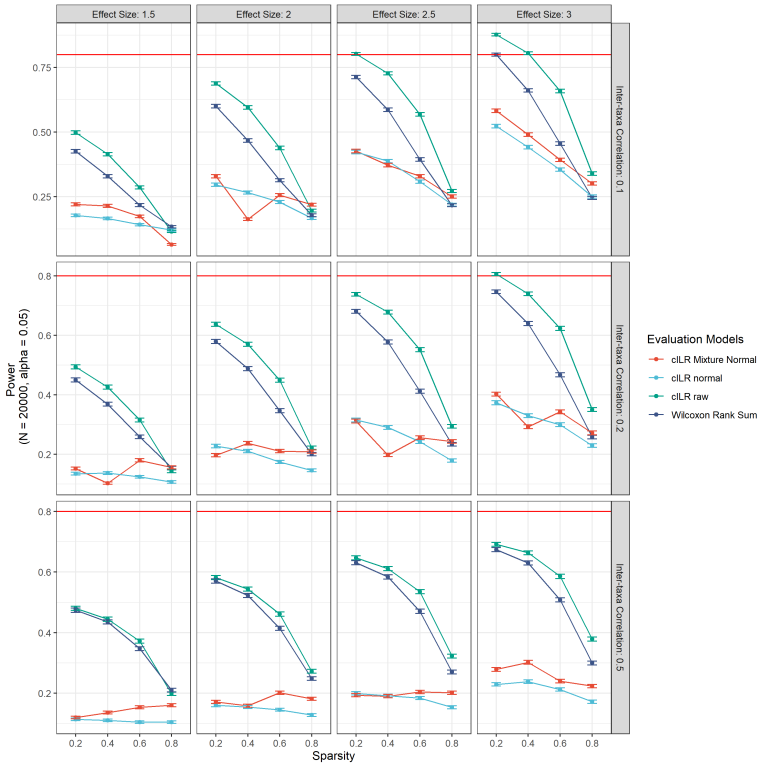

Figure 2. Median power as a function of data sparsity benchmarked on simulated microbiome data as enumerated in SI Methods. Enrichment of a specified set was tested at the sample level using clLR and the Wilcoxon rank sum test at α of 0.05. Each panel represents different effect sizes (horizontal) and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

Differential abundance analysis

Type I error control

We benchmarked type I error rate of the 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

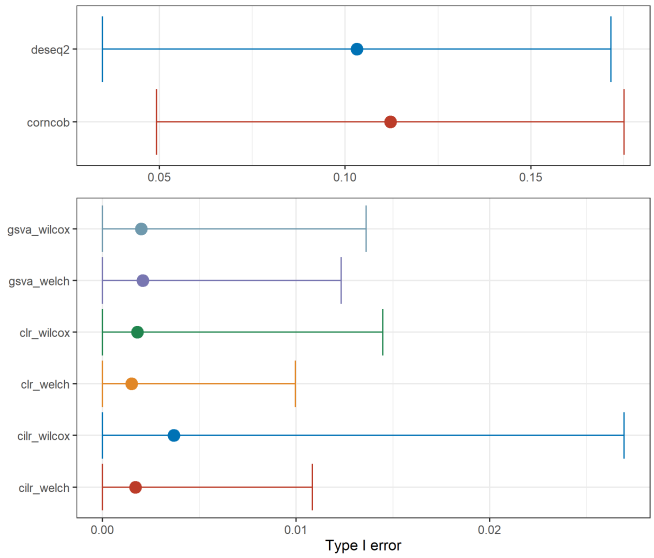

Figure 3. Type I error evaluated on 16S rRNA and WGS stool samples obtained from HMP. Enrichment of genus level taxa sets was tested across different methods where significance was determined at FDR cutoff of 0.05.

References

[1] John Aitchison. A Concise Guide to Compositional Data Analysis. page 134.

[2] 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.

[3] 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.

[4] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8, 2017.

[5] 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.

[6] Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, October 2015.

[7] Di Wu and Gordon K. Smyth. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, September 2012.