cILR: Competitive isometric log-ratio for taxonomic enrichment analysis

Quang P. Nguyen[1,2], Anne G. Hoen[1,2‡], H. Robert Frost[1*‡†],

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

‡These authors jointly supervised this work.
†Corresponding author

* hildreth.r.frost@dartmouth.edu

Abstract

Research in human associated microbiomes often involves the analysis of taxonomic count tables generated via high-throughput sequencing. It is difficult to apply statistical tools as the data is high-dimensional, sparse, and strictly compositional. An approachable way to alleviate high-dimensionality and sparsity is to aggregate variables into pre-defined sets. Set-based analysis is ubiquitous in the genomics literature, and has demonstrable impact in improving interpretability and power of downstream analysis. Unfortunately, there is a lack of sophisticated set-based analysis methods specific to microbiome taxonomic data, where current practice often employs abundance summation as a technique for aggregation. This approach prevents comparison across sets of different sizes, does not preserve inter-sample distances, and amplifies protocol bias. Here, we attempt to fill this gap with a new single sample taxon set enrichment method based on the isometric log ratio transformation and the competitive null hypothesis commonly used in the enrichment analysis literature. Our approach, titled competitive isometric log ratio (cILR), generates sample-specific enrichment scores as the scaled log ratio of the subcomposition defined by taxa within a set and the subcomposition defined by its complement. We provide sample-level significance testing by estimating an empirical null distribution of our test statistic with valid p-values. Herein we demonstrate using both real data applications and simulations that cILR controls for type I error even under high sparsity and high inter-taxa correlation scenarios. Additionally, it provides informative scores that can be inputs to downstream differential abundance and prediction tasks.

Author summary

The study of human associated microbiomes relies on genomic surveys via high-throughput sequencing. However, microbiome taxonomic data is sparse and high dimensional which prevents the application of standard statistical techniques. One approach to address this problem is to perform analyses at the level of taxon sets. Set-based analysis has a long history in the genomics literature, with demonstrable impact in improving both power and interpretability. Unfortunately, there is not a lot of research in developing new set-based tools for microbiome taxonomic data specifically,

given that compared to other 'omics data types microbiome taxonomic data is strictly compositional. We developed a new tool to generate taxon set enrichment scores at the sample level by combining the isometric log-ratio and the competitive null hypothesis. Our scores can be used for statistical inference, and as inputs to other downstream analyses such as differential abundance and prediction models. We demonstrate the performance of our method against competing approaches across both real data analyses and simulation studies.

Introduction

The microbiome is the collection of microorganisms (bacteria, protozoa, archaea, fungi, and viruses) which co-exist with their host. Previous research has shown that changes in the composition of the human gut microbiome are associated with important health outcomes such as inflammatory bowel disease [1], type II diabetes [2], and obesity [3]. To understand the central role of the microbiome in human health, researchers have relied on high-throughput sequencing methods, either by targeting a specific representative gene (i.e. amplicon sequencing) or by profiling all the genomic content of the sample (i.e. whole-genome shotgun sequencing) [4]. Raw sequencing data is then processed through a variety of bioinformatic pipelines [5, 6], yielding various data products, one of which are taxonomic tables which can be used to study associations between members of the microbiome and an exposure or outcome of interest.

However, there exists unique challenges in the analysis of these taxonomic count tables [7, 8]. First, like other sequencing-based datasets, microbiome count data is often high dimensional, where the number of detected taxa far exceeds the number of samples usually present. For predictive tasks, microbiome-specific penalized regression approaches have been developed to address this issue [9], allowing for simultaneous model fitting and variable selection. For differential abundance tasks, researchers often utilize multiple hypothesis correction methods [10, 11] or omnibus tests [12] to address hypothesis testing burden.

Second, the number of reads obtained is constrained by the sequencing instrument at an arbitrary limit, and is inconsistent across samples, resulting in a variable number of total read counts per sample. Many normalization methods [13] have been proposed to address these issues, including cross-applying methods from the gene expression literature [14]. However, these methods rely on assumptions specific to the original bulk RNA-seq data sets such as the presence of housekeeping genes with consistent expression levels [15], which might not be not true in the context of microbiome relative abundance data [16, 17]. As such, microbiome taxonomic data is strictly compositional [18], which means that the abundance of any taxa can only be interpreted relative to another. Consequently, log-ratio transformations from the compositional data literature are often utilized [19].

Third, the data are highly zero-inflated, where there is a high number of both structural zeros (truly missing due to biological reasons) and sampling zeroes (due to limits of detection of the sequencing experiment). Researchers often dealt with these issues by imputing zero cells with a pseudocount [20], or applying zero-inflated models [12, 21]. Newer methods developed recently have focused on understanding the different types of zeros in the data, providing more sophisticated heuristics around when pseudocounts can be utilized [22].

Even though the aforementioned problems are challenging, a very approachable method to address some of them is through set-based analysis, also termed gene set testing in the genomics literature [23, 24]. Aggregated sets are less sparse than their constituent

elements, and testing on a smaller number of variables reduces the multiple testing burden, thereby increasing power and reproducibility. Through the usage of functionally informative sets defined apriori based on historical experiments (for example MSigDB [25], and Gene Ontology [26]), gene set analyses also allows for more informative interpretations. There exists a diverse set of available methods developed to perform such analyses. More traditional set testing methods utilize the hypergeometric test to test for the overrepresentation of significant p-values for a set of interest [24]. Unfortunately, these approaches are sensitive to the differential expression test and their generated p-values. The most widely used gene-set analysis method, GSEA [25], instead uses a random-walk-like statistic through a ranking of genes based on a measure of association or effect size. Both of these methods generate enrichment scores and significance testing at the population level, incorporating information from all samples. Conversely, methods such as GSVA [27] and VAM [28], generate enrichment scores at the sample level and are more akin to a transformation. This strategy allows for the flexible incorporation of different statistical techniques downstream, such as prediction models, as well as for visualization purposes in ordination plots.

In microbiome research, even though no explicit enrichment analysis is performed, standard practice often involves aggregating taxa to higher Linnean classification levels such as genus, family, or phylum by simple summation of abundances [29]. Even though this allows for a reduction in the number of overall taxa (from thousands to only hundreds), there still exist three disadvantages: first, inter-sample distances are not preserved before and after aggregation [30], second, it doesn't allow for enrichment testing and comparison across sets of different sizes, and third, it increases bias when taxa within the set have different efficiencies in how they are measured through sequencing [29]. As such, there is a need for microbiome researchers to adopt more robust sample-level set enrichment methods. Unfortunately, limited work has been done to extend existing methods to be more specific to microbiome relative abundance data. Some software suites, such as *MicrobiomeAnalyst*, do offer tools to perform enrichment testing with curated taxon sets [31]. However, the approach used in *MicrobiomeAnalyst* is a form of overrepresentation analysis at the population level and therefore similarly sensitive to the differential abundance approach used.

Here, we present a novel method that generates enrichment scores at the sample level similar to GSVA [27] and VAM [28]. We leverage the concept of the Q_1 competitive hypothesis presented in Tian et al. [32], which tests the null that the value of variables within a specific set is equal to the value of measured variables not in the set. The competitive null hypothesis is particularly useful in compositional data analysis, as it naturally assesses enrichment as a ratio between two sets of variables. We incorporated this insight with the isometric log-ratio transformation [33], which allows for a multiplicative aggregation method that addresses the downsides of the naive summation-based method presented above [29, 34]. The resulting method, titled competitive isometric log-ratio (cILR), is therefore unsupervised and can generate sample-specific enrichment scores with a well-defined null hypothesis that allows for significance testing. These scores can then act as inputs to differential abundance and predictive modeling tasks downstream.

In the following sections, we provide the formulation of cILR and discuss some statistical properties. We illustrate significance testing at the sample level using cILR and evaluate type I error and power under different simulation scenarios and real data applications. We assess the informativeness of cILR generated scores, and evaluate how it performs as part of downstream analyses, specifically predictive models and differential abundance analysis. We compare the performance of cILR in these respective tasks against standard microbiome taxonomic data analysis practices, as well as the existing

GSVA [27] and ssGSEA [35], which are single sample methods. An R package implementation of this approach can be found on Github (qpmnguyen/teaR).

Materials and Methods

Competitive Isometric Log-ratio (cILR)

The cILR method generates sample-specific enrichment scores for microbial sets using the isometric log-ratio transformation [33]. Details on the computational implementation of cILR can be found in the supplemental materials. The cILR 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 of the 16S rRNA gene) or whole genome shotgun sequencing. Usually $\mathbf{X}$ is generated from standard taxonomic profiling pipelines such as DADA2 [5] for 16S rRNA sequencing, or MetaPhlAn2 [6] for whole genome shotgun sequencing.
- $\mathbf{A}$: p by m indicator matrix annotating the membership of each taxon p to m sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [36], or those based on more functionally driven categories such as tropism or ecosystem roles ($A_{i,j} = 1$ indicates that microbe i belongs to set j).

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

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

   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. **Estimate the empirical null distribution** Enrichment scores represent the test statistic for the Q_1 null hypothesis H_o that relative abundances in $\mathbf{X}$ of members of set k are not enriched compared to those not in set k. Since the distribution of cILR under the null vary depending on data characteristics (Fig 1), an empirical null distribution will be estimated from data.

   - **Compute the cILR statistic on permuted and un-permuted $\mathbf{X}$**. Let $\mathbf{X}_{perm}$ be the column permuted relative abundance matrix, and $\mathbf{M}_{perm}$ be the corresponding cILR scores generated from $\mathbf{X}_{perm}$. Similarly, we have $\mathbf{M}_{unperm}$ be cILR scores generated from $\mathbf{X}$.
   - **Estimate correlation-adjusted empirical distribution for each set**. For each set, a fit a parametric distribution to both $\mathbf{M}_{perm}$ and $\mathbf{M}_{unperm}$. The location measure estimated from $\mathbf{M}_{perm}$ and the spread measure estimated from $\mathbf{M}_{unperm}$ will be combined as the correlation-adjusted empirical null distribution $\mathbf{P}_{emp}$ for each set. Two available options are the normal distribution and the mixture normal distribution. For the normal

distribution, parameters were estimated using the method of maximum likelihood implemented in the *fitdistr* package [37]. For the mixture normal distribution, parameters were estimated using an expectation-maximization algorithm implemented in the *mixtools* package [38].

3. **Calculate finalized cILR scores with respect to the empirical null**. Enrichment scores $\mathbf{E}_{i,k}$ are calculated as the cumulative distribution function (CDF) values or z-scores with respect to $\mathbf{P}_{emp}$ distribution. P-values can be calculated by subtracting $\mathbf{E}$ from 1.

Properties of cILR

cILR and the Isometric Log Ratio Transformation

The cILR statistic is a special instance of the isometric log-ratio transformation (ILR) [33]. The standard ILR is a transformation method to address the negative correlation bias inherent in compositional data by providing an isometry between the D-dimensional simplex $\mathbb{S}^D$ and coordinates in the $D-1$ real space $\mathbb{R}^{D-1}$ [33, 39]. This is accomplished by projecting the composition onto a chosen orthonormal basis in $\mathbb{R}$, which can be defined by a sequential binary partition (SBP) of the variables (e.g. a rooted phylogenetic tree). The ILR transformed variables are the coordinates of nodes within an SBP tree of the variables. Without loss of generalizability, in a given SBP with node i splitting variables between sets R and S, we have the ILR coordinate x_i^* as:

$$x_i^* = \sqrt{\frac{rs}{r+s}} \log\left(\frac{g(X_{j|j\in R})}{g(X_{j|j\in S})}\right) \tag{2}$$

where r and s are the cardinalities of sets R and S respectively, $g(z)$ is the geometric mean, and X_j are values of the original predictors with indexes defined by membership in R and S. The ILR confer many important benefits. First, ILR coordinates exist in real space, whereby common statistical methods can be used. Second, ILR aggregated variables preserve inter-sample distances before and after aggregation [30]. Third, ILR variables are not constrained to sum to 0 as that of the centered log-ratio transformation, resulting in a covariance matrix that is not singular [33].

The usage of the ILR statistic is not uncommon in the microbiome literature. They are usually termed "compositional balances", and have been leveraged in many recent approaches in variable transformation [34, 39, 40]. The cILR formulation in Eq (1) is a special case of Eq (2) defined on a node that splits the taxa into two disjoint sets, one representing the set of interest, the other representing the remaining taxa. As such, the cILR transformation inherits the properties of the ILR as a log-ratio method applicable to compositional data sets. However, unlike the ILR and its variants [34, 39, 40], the axes defined by each cILR set are not orthogonal (since the balances are mutually exclusive between sets and do not belong in the same SBP). Hence, a correlation can exist between cILR aggregated variables.

Statistical Properties of cILR

We can perform significance testing on the cILR statistic which corresponds to the null hypothesis that the center of the subcomposition defined by the set is equal to the center of the subcomposition defined by the complement of the set. This is equivalent to the Q_1 competitive null hypothesis in the gene set testing literature [32] where the enrichment of a gene set is defined with respect to genes outside the set.

We can apply prior usage of the ILR statistic in hypothesis testing to cILR by assuming that the null distribution of cILR follows a standard normal distribution [30]. However,

when applying cILR for hypothesis testing at the sample level, it is expected that the researcher would be testing a large number of hypotheses. Under the assumption that the number of truly significant hypotheses is low, Efron [41] showed that estimating the null distribution of the test statistic (termed the empirical null distribution) is much more preferable than using the theoretical null due to unobserved confounding effects inherently part of observational studies. As such, to perform significance testing using cILR, we also estimated the null distribution from observed raw cILR variables.

This assumption is also supported by preliminary simulation studies (detailed below). In panel A of Fig 1, we simulated microbiome taxonomic count data under the global null across different data features and compute raw cILR scores and compute kurtosis and skewness. It can be seen that the characteristics of the null change depending on sparsity and inter-taxa correlation. Sparsity seems to drive the distribution to be more positively skewed while inter-taxa correlation encourages platykurtic (negative kurtosis). The effect is most dramatic under both high inter-taxa correlation and sparsity. This heterogeneity further supports the decision to estimate an empirical null distribution, similar to Efron [41].

Additionally, the degree of kurtosis and skewness also suggests that the normal distribution itself might not be a good approximation of the null. To address this issue, we also evaluated a mixture distribution of two normal components. Panel B of Fig 1 demonstrates the goodness of fit of the mixture normal and the normal distribution using Kolmogorov-Smirnov (KS) test statistic computed on fitted normal and mixture normal distribution when fitted on cILR scores in simulation scenarios under the global null. We can see that the mixture normal distribution is a better fit (lower KS scores) than the normal distribution across both sparsity and correlation settings.

We performed our empirical null estimation by fitting our distribution of choice and computing relevant parameters on raw cILR scores on taxa-permuted data (equivalent to gene permutation in the gene expression literature). As such, the null distribution is characterized by scores computed on sets of equal size with randomly drawn taxa. However, null distribution based on taxa-permutation is sensitive to inter-taxa correlations within the set [42]. Since the permutation procedure does not preserve correlation structures, estimating parameters from empirical scores on permuted data will underestimate the variance inflation due to correlation. We account for this by combining the mean estimate from permuted data with the variance estimate from unpermuted data, where the inter-taxa correlation structure remains undisturbed. However, this procedure assumes that the variance of cILR is equal under both the null and alternate hypotheses.

Evaluation

All code and data sets used for evaluation of this method is publicly available and can be found on GitHub (qpmnguyen/cILR_analysis).

Parametric Simulations

To address the performance of cILR for different modeling tasks, we simulated microbiome count data under the assumption that it follows a zero-inflated negative binomial distribution, which is a good fit for real microbiome relative abundance data [43]. Suppose X_{ij} are observed counts for a sample i and taxon j, then we have the following probability model

$$\mathbf{X}_{ij} = \begin{cases} 0 & \text{with probability } p_j \\ \mathbf{NB}(\mu_j, \phi_j) & \text{with probability } 1 - p_j \end{cases} \tag{3}$$

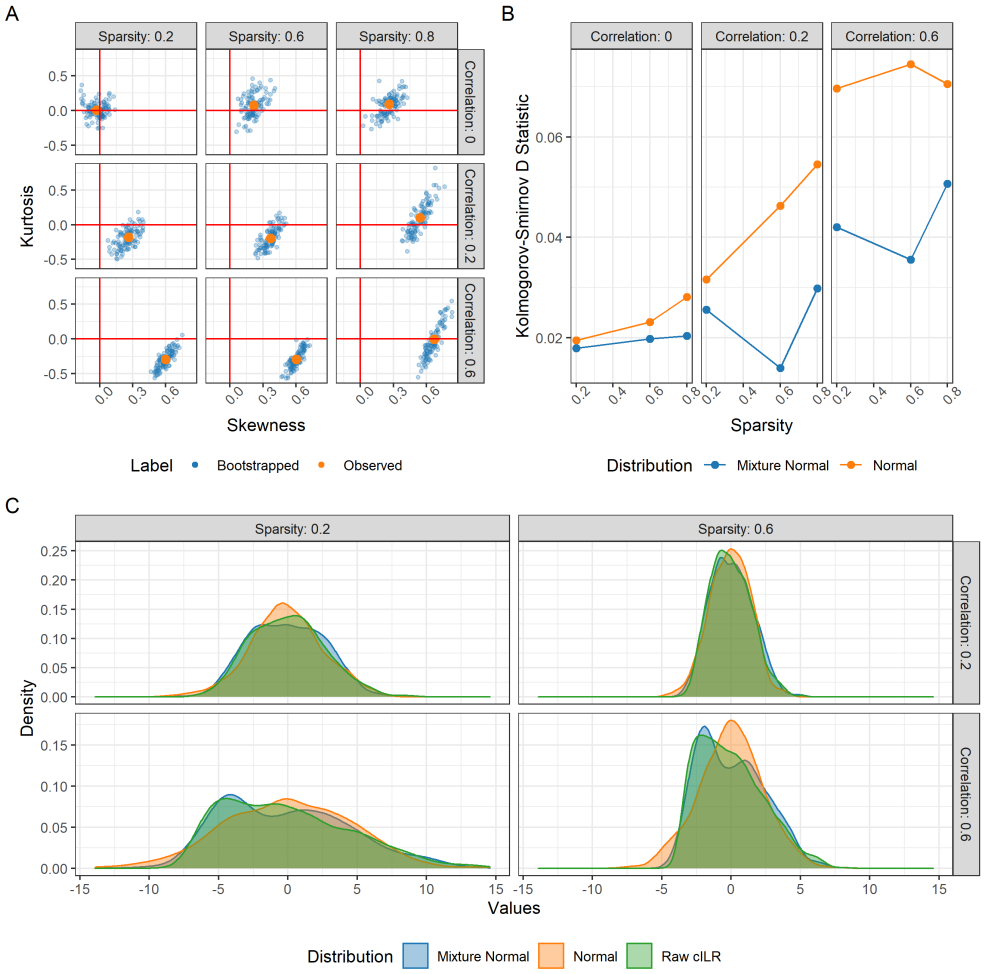

Fig 1. Properties of the null distribution of cILR under the global null simulations. Panel **(B)** presents kurtosis and skewness of cILR scores while panel **(A)** presents the goodness of fit (as Kolmogorov-Smirnov D statistic) for mixture normal and normal distributions. Panel **(C)** is a density plot of the shape of the null distribution. Results indicated the necessity of estimating an empirical null and demonstrating that the mixture distribution was the better fit compared to the basic normal.

where μ_j and ϕ_j are mean and dispersion parameters, respectively. To incorporate a flexible correlation structure into our simulation model, we utilized the NorTA (Normal to Anything) method [44]. Given an n by p matrix of values $\mathbf{U}$ sampled from multivariate normal distribution with correlation matrix ρ, we can generate target microbiome count vector $\mathbf{X_{.j}}$ for taxa j following the marginal distribution **NB** characterized by the negative binomial cumulative distribution function $\mathbb{F}_{\mathbf{NB}}$:

$$\mathbf{X}_{.j} = \mathbb{F}_{\mathbf{NB}}{}^{-1}(\Phi_{U_i}) \tag{4}$$

In this instance, for each taxon j, we set elements in $\mathbf{U}_{.j}$ to be zero with probability p_j and applied $\mathbf{NB}^{-1}(\mu_j, \phi_j)$ on non-zero elements to generate our final count matrix $\mathbf{X}$. To ensure that our simulations match closely to real data, we fitted negative binomial distribution using a maximum likelihood approach (with the *fitdistrplus* package in R [37]) to non-zero counts for each taxon from 16S rRNA profiling of stool samples from

the Human Microbiome Project (HMP). We take the median values of the estimated mean and dispersion parameters as the baseline of our simulations. For simplicity, we assumed that inter-taxa correlation follows an exchangeable structure

Single Sample Enrichment: To assess type I error rate and power for enrichment significance testing at the sample level, we simulated data based on the schema above, and assessed enrichment for one focal set. Type I error was obtained under the global null as the number of samples where the null hypothesis was rejected at $\alpha = 0.05$ over the total number of samples (which represents the total number of hypotheses tested). Power was obtained using the same formulation as type I error rate but under the global alternate. We treated type I error and power as estimates of binomial proportions and utilized the Agresti-Couli [45] formulation to calculate 95% confidence intervals. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$) and inter-taxa correlation within the set ($\rho = 0, 0.2, 0.5$). For type I error analysis, we also varied the size of the set (50, 100, 150). For power analyses, set size was kept constant at 100 but different effect sizes (fold change of 1.5, 2, and 3). All sample sizes were set at 10,000.

For classifiability, we evaluated the scores against the true labels per sample (indicating the sample has a set with inflated counts) using the area under the receiving operator curve (AUROC/AUC). This is a strategy used in Frost [28] which evaluates the informativeness of scores by assessing the relative ranking of samples (i.e. whether samples with inflated counts are highly ranked using estimated scores). DeLong 95% confidence intervals for AUC [46] were obtained for each estimate. Simulation settings for classification performance were identical to power analyses as detailed in the previous paragraph.

Differential Abundance Analysis: To assess type I error rate and power for differential abundance testing task, we simulated data based on the schema above, and assessed differential abundance of 50 sets with 100 taxa per set across 20 replicates per simulation condition. Type I error is calculated as the number of differentially abundant sets over the total number of sets for each simulation under the global null. Power is defined similarly, but instead under the global alternate hypothesis. Estimates and confidence intervals for type I error and power are calculated as cross-replicate mean and standard error. A set is differentially abundant when all taxa within a set are differentially abundant with the same effect size. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$), and inter-taxa correlation within the set ($\rho = 0, 0.2, 0.5$). Half of the sets are differentially abundant across case/control status with varying effect sizes (fold change of 1.5, 2, and 3). Due to the compositional nature of microbiome taxonomic data, simple inflation of raw counts would cause an artificial decrease in the abundance of the remaining un-inflated sets. As such, we applied a compensation procedure as described in Hawinkel et al. [47] to ensure the validity of simulation results. All sample sizes were set at 2,000.

Prediction: To assess predictive performance, we generated predictors based on the simulation schema presented above and evaluated prediction for both binary and continuous outcomes using a standard random forest model [48]. For binary outcomes, we use AUC similar to the classification analyses above. For continuous outcomes, we used root mean squared error (RMSE). All predictive model fitting was performed using *tidymodels* [49] suite of packages. Across both learning tasks, we varied sparsity ($p = 0.2, 0.4, 0.6$), and inter-taxa correlation ($\rho = 0, 0.2, 0.5$). Continuous outcomes Y_{cont} were generated as linear combinations of taxa counts.

$$Y_{cont} = f(\mathbf{X}) + \epsilon \tag{5}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $f(\mathbf{X}) = \beta_0 + \mathbf{X}\beta$. For each simulation, we set β_0 to be $\frac{6}{\sqrt{10}}$

similar to [50]. The degree of model saturation (the number of non zero β values) were varied between 0.1 and 0.5, and signal to noise ratio (SNR $= \frac{\sigma(f(\mathbf{X}))}{\sigma_\epsilon}$) was varied between 1.5, 2, and 3.

For binary outcomes, we generate Y_{binary} as Bernoulli draws with probability p_{binary}, where

$$p_{binary} = \frac{1}{1 + \exp(f(\mathbf{X}) + \epsilon)} \tag{6}$$

To ensure a balance of classes, we applied the strategy described in Dong et al. [51] where the associated β values are evenly split between positive and negative associations. All data sets generated from prediction tasks have 2,000 samples with 5,000 taxa over 50 sets with a size of 100 taxa per set.

Real Datasets

In addition to simulation analyses, we also evaluated our method using real data sets based on both 16S rRNA gene sequencing and whole-genome sequencing. All data sets are obtained from either the *curatedMetagenomicData* [52] and *HMP16SData* [53] R packages (2020-10-02 snapshot), or downloaded from the Qiita platform [54].

Single Sample Enrichment: To assess the false discovery rate and true discovery rate of cILR in sample-level enrichment testing, we utilized the 16S rRNA gene sequencing of the oral microbiome at the gingival subsite from the Human Microbiome Project [1, 55]. We utilized this data set following the approach outlined in Calagaro et al. [43]. This data set is approximately labeled, where aerobic microbes are enriched in the supragingival subsite where the biofilm is exposed to the open air, while conversely anaerobic microbes thrive in the subgingival site [56]. Here, we assessed the enrichment of aerobic microbes across all samples, we considered the false positive rate as the number of samples from the subgingival site with significant enrichment, and the true positive rate as the number of supragingival samples with significant enrichment. Microbial tropism annotation at the genus level was from Beghini et al. [57] and was downloaded directly from the GitHub repository associated with Calagaro et al. [58].

Differential Abundance Analsysis: To assess type I error using cILR scores in differential abundance analysis, we utilized the 16S rRNA gene sequencing of stool samples from the Human Microbiome Project [1, 55]. Here, we randomly assign samples a label of case or control, and repeated this process 500 times, assessing all candidate methods at each iteration. Type I error is then the number of taxa identified as differentially abundant across all tested taxa. For the true positive rate, we used the same gingival data set as described above. However, instead of testing for aerobic microbes as a group, the true positive rate is the number of aerobic/anaerobic genera identified as differentially abundant across all aerobic or anaerobic genera.

Disease Prediction: To assess predictive power, we utilized the whole genome sequencing of stool samples of inflammatory bowel disease (IBD) patients from the MetaHIT consortium [59]. This data set contains 396 samples from a cohort of European adults, where 195 adults were classified as having IBD (which includes patients diagnosed with either ulcerative colitis or Crohn's disease). Additionally, we also utilized a similar data set from Gevers et al. [60] which also profiles the gut microbiome of IBD patients and controls but using 16S rRNA gene sequencing. This data set contains 16S rRNA gene sequencing samples from a cohort of pediatric patients (ages < 17) from the RISK cohort enrolled in the United States and Canada. Of the 671 samples obtained, 500 samples belong to patients with IBD.

Comparison Methods

Single sample enrichment: For type I error and power analyses, we compared the cILR method with a naive Wilcoxon rank sum test. We added a pseudocount of 1 to all values. This is a non-parametric difference in means test, where we compared the abundance of taxa of a pre-defined set and its complement within a single sample. For classification performance, we compared cILR methods against GSVA [27], ssGSEA [35], and the W-statistic from the Wilcoxon rank sum test. All three approaches were applied directly on count data (after pseudocount). For GSVA, the Poisson kernel was used.

Differential Abundance: Since cILR are sample-level enrichment scores, we performed differential abundance by using a Wilcoxon Rank Sum test and Welch's t-test across case/control status on cILR generated scores. We added a pseudocount of 1 to all values. For comparison, we chose representative state-of-the-art methods in differential abundance analysis, namely DESeq2 [14, 15] and corncob [61]. For DESeq2, we performed a likelihood ratio test against an intercept only reduced model with dispersion estimated with local fit. For corncob, we also performed a likelihood ratio test against an intercept only reduced model without bootstrapping.

Disease Prediction: We fit random forest on cILR scores, as well as ssGSEA [27] and GSVA [35] similar to single sample enrichment section. We added a pseudocount of 1 to all values. Additionally, we also compared performance using cILR against a standard analysis plan where the centered log-ratio transformation (CLR) was applied to count-aggregated sets as inputs to a machine learning model.

Results

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

Enrichment testing at the sample level

cILR provides significance testing for enrichment at the sample level using the null distribution estimation procedure described in Materials and Methods. Here, we present empirical results for this application of cILR assessing type I error, power, and classification capacity.

Simulation studies

Panel A and B in Fig 2 demonstrate type I error and power respectively across different simulation conditions. We benchmarked the results of the cILR method against a naive Wilcoxon rank sum test performed at the sample level, comparing the mean count difference between taxa in the set its complement. All methods demonstrate good type I error control at $\alpha = 0.05$ under zero correlation across all simulation conditions. However, under both medium ($\rho = 0.2$) and high ($\rho = 0.5$) correlation settings, both the Wilcoxon test and unadjusted cILR variants show high levels of inflated type I error, where Wilcoxon test performed the worst. On the other hand, adjusted cILR methods (under both distributions) control for type I error at the appropriate α level even at high correlations.

However, the trade-off for good type I error control is demonstrably lower power, as

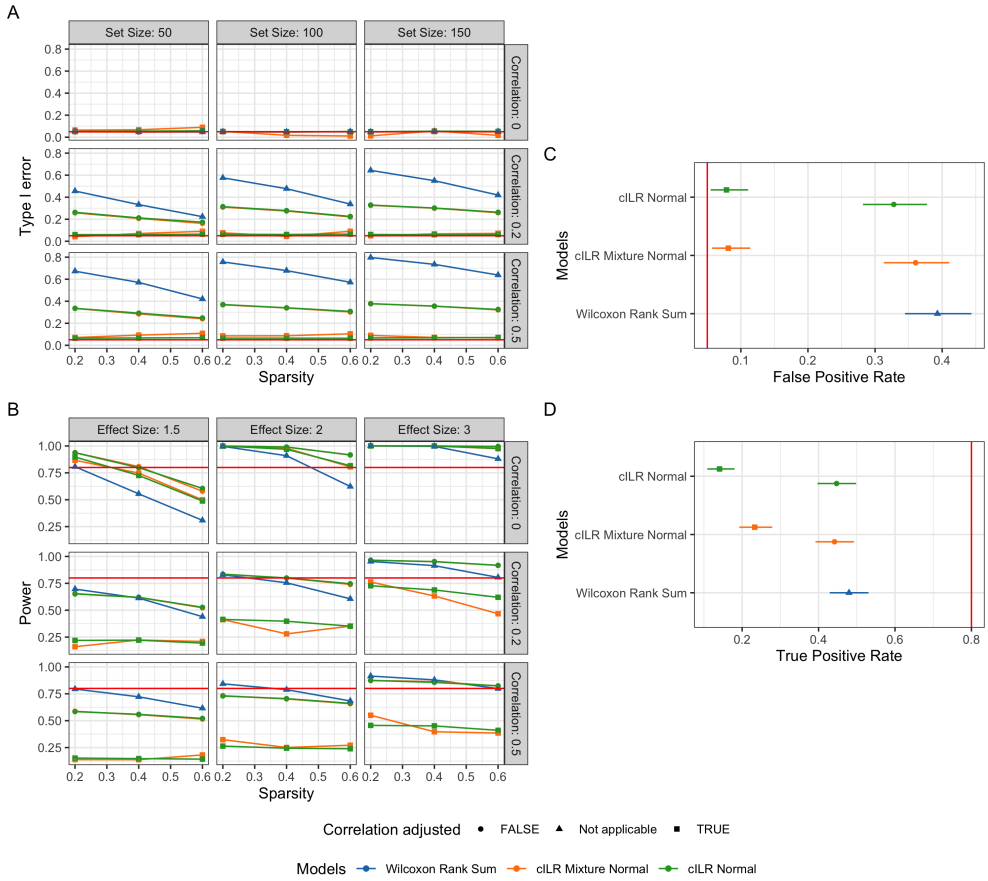

Fig 2. Sample-level inference with cILR under parametric simulations ((**A**) and **B**), and real data analysis ((**C**) and **(D)**). In simulation analyses, panel **(A)** shows type I error rate, and panel **(B)** shows power for single sample enrichment test for a specified set and was compared against a Wilcoxon rank sum test at α of 0.05. In real data analysis, panel **(C)** shows the false-positive rate, and panel **(D)** shows the true positive rate. For this analysis, 16S rRNA data from the oral microbiome of the gingival site was used. The set of aerobic microbes was tested for enrichment in all samples and was identified as correctly enriched if a significant p-value was obtained in supragingival samples. Confidence bounds were obtained using Agresti-Couli [45] approach. Adjusted cILR demonstrated control of type I error at the appropriate α level while remaining methods (not included in subsequent power analyses) showed an inflated type I error rate. However, this resulted in lower power for adjusted cILR methods.

shown in Fig 2B. In situations where there is no inter-taxa correlation, cILR still outperforms the wilcoxon rank sum test, however adjusted versions of cILR did not perform as well as un-adjusted ones. However, in higher correlation scenarios, the difference in power is much more dramatic. At the highest effect size (fold change of 3) and correlation ($\rho = 0.5$), adjusted cILR was only performing at 50% power, while unadjusted cILR and wilcoxon rank sum test were able to reach 80%. These results indicate that both sparsity and inter-taxa correlation impacts power, with correlation having a much more dramatic impact especially for adjusted versions of cILR. Most importantly, cILR demonstrate higher power in all scenarios where type I error is properly controlled.

To further assess the utility of cILR in classifying samples with enriched sets, we generated AUC scores for different cILR scores using true labels of whether a sample has an inflated set. This analysis, therefore, assessed the relative ranking of samples using cILR scores whereby high scores should correspond to samples that are known to be inflated. Fig 3 presents this result. We compared different variants of cILR against competing methods in the gene set testing space (GSVA [27] and ssGSEA [35]), as well as the W test statistic from the Wilcoxon rank sum test. Across both simulations (Fig 3A) and real-data applications (Fig 3B), cILR scores perform marginally better especially in low effect size situations but did not stand out in most other scenarios. In simulation studies, classification performance was good (around AUC of 0.8) even at high correlation settings, only requiring medium effect sizes (fold change of 2). Notably, the W-statistic provided the least information for classifying samples with inflated taxa.

Real data evaluations

These observations were replicated when assessed on the semi-labeled gingival data set from the Human Microbiome Project as described in Materials and Methods. Here, we tested the enrichment of aerobic microbes for each sample using approaches similar to our parametric simulations. As expected in Fig 2C, the proportion of falsely rejected hypotheses was high in the naive Wilcoxon test and unadjusted cILR methods. Conversely, adjusted cILR controls for false positives adequately at the correct α level of 0.05. Power analysis (Fig 2D) showed similar patterns, where unadjusted cILR methods and the Wilcoxon test have a higher proportion of null hypotheses correctly rejected, however, these results are not useful to a practitioner as the number of falsely rejected hypotheses are also equally high.

Differential abundance analysis

cILR generates sample-specific scores representing the degree of enrichment of a pre-defined set. As such, we want to assess the ability to use these scores for differential abundance analysis in combination with a standard difference of means statistical test (Welch's t-test and Wilcoxon rank sum test). We compared the performance of this approach with cILR and two commonly used methods for differential abundance testing in the microbiome literature: DESeq2 [15] and corncob [61].

Simulation studies

Fig. 4 present results for simulation studies for both type I error (panel A) and power (panel B) evaluations. All methods control for type I error well across both sparsity and correlation levels, where the estimated rate was consistently around the 0.05 pre-defined threshold. Results were similar across all evaluated methods, although in some instances, for example in medium correlation setting ($\rho = 0.2$), the unadjusted cILR resulted in higher type I error, regardless of difference in means test and distribution of choice.

The difference between the methods is more noticeable when evaluating power. All cILR associated variants showed much higher power even when the effect size is limited (fold change is 1.5), and there is a noticeable gap in performance between cILR and both DESeq2 and corncob. Surprisingly, this effect is consistent across correlation levels and sparsity, even though we expectedly see performance in power drop as a function of sparsity especially in low effect size settings.

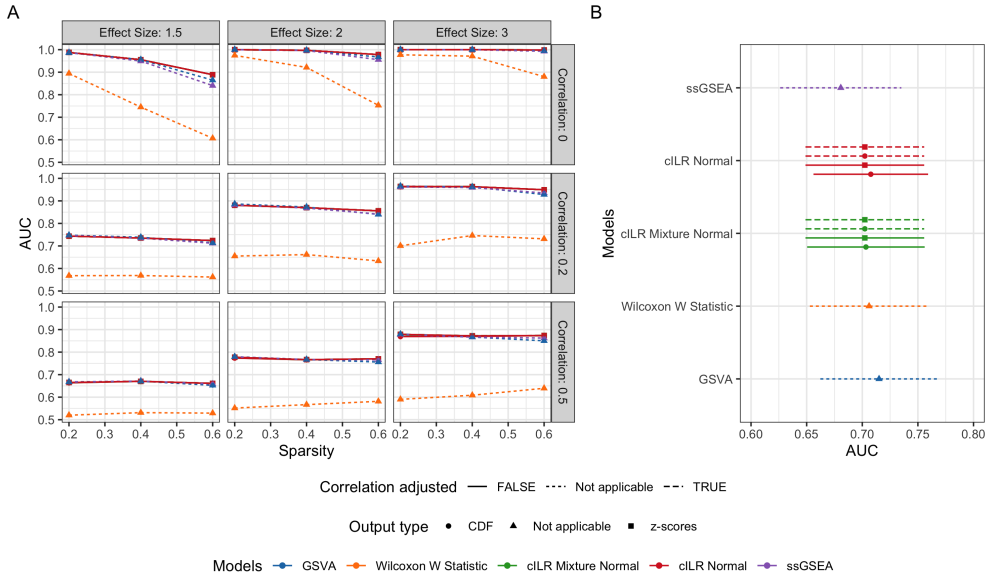

Fig 3. Classification performance via AUC of cILR, ssGSEA, GSVA, and Wilcoxon U statistic on simulated data **(A)** the gingival data set from the Human Microbiome Project **(B)** as detailed in Materials and Methods. Performance scores measure whether scores can highly rank samples that are known to have inflated abundance. In the gingival data set presented in panel **(B)**, samples from the supragingival site are assumed to have an inflated abundance of aerobic microbes. Error bars are the 95% DeLong confidence intervals for AUC [46]

Real data evaluations

In addition to simulation studies, we also evaluated performance of the methods on real 16S rRNA gene sequencing data set from HMP (Fig 5). For type I error evaluations, we use stool samples and randomly assign them with case/control status and calculated type I error as the proportion of genera identified as significantly different. For true positive rate evaluations, we use the gingival data set as detailed in the previous section, and calculated the true positive rate as the proportion of genera labeled as either anaerobic or aerobic that were found to be significant.

We observed both corncob and DESeq2 had significantly inflated type I error rate while all variations of cILR were controlling for type I error at the defined α threshold of 0.05. This is surprising given the consistency of preserving type I error for both corncob and DESeq2 in all simulation evaluations.

In true positive experiments with data from the gingival site, estimated rates were more similar across the different methods. As expected, using the Wilcoxon rank sum test resulted in lower true positive rate compared to remaining methods, but the difference was not noticable. This is also surprising given that in simulation studies, both corncob and DESeq2 showed markedly lower power across all effect sizes.

Disease Prediction

Since cILR can generate informative scores that can discriminate between samples with inflated counts for a set (Fig 2), we want to assess whether they can also act as useful inputs to predictive models. In this section we assessed the predictive performance of a

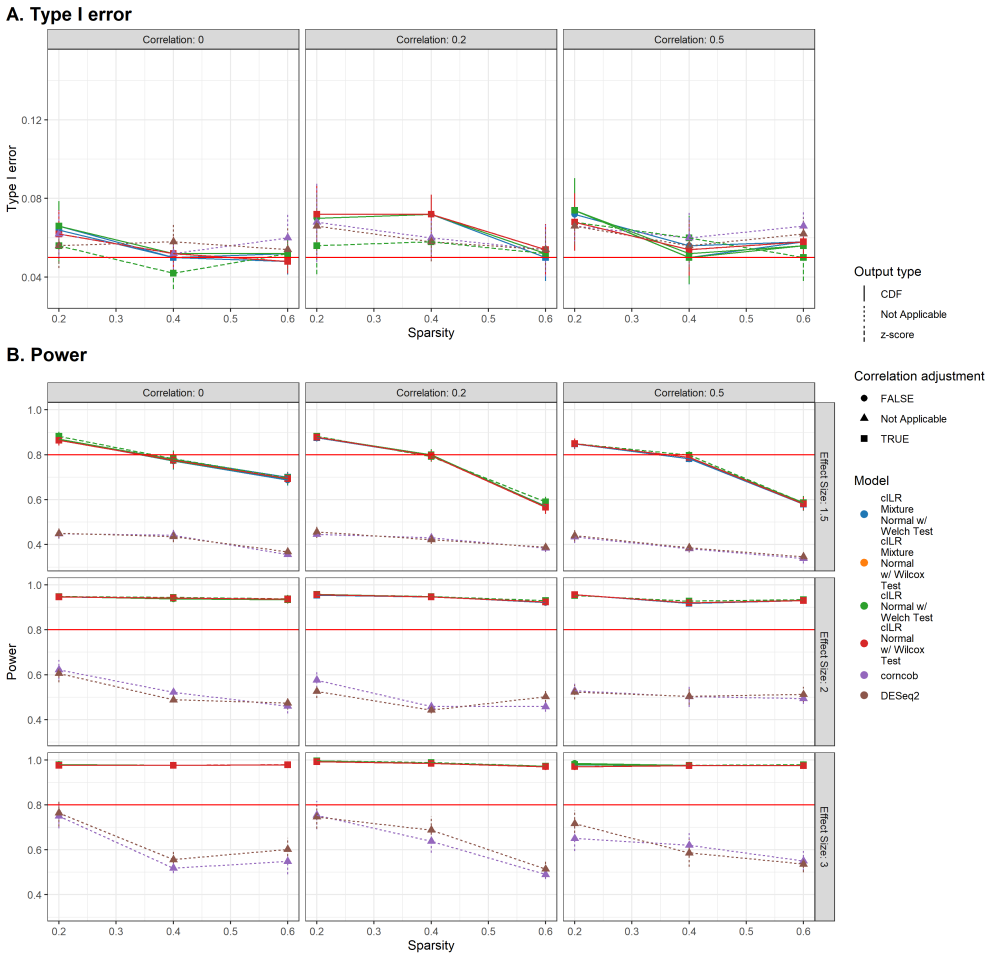

Fig 4. Type I error rate **(A)** and power **(B)** for differential abundance test across different parametric simulation scenarios. For cILR methods, differential abundance analysis was performed using a difference in means test (either Wilcoxon rank-sum test or Welch's t-test) across case/control status using single sample scores generated by cILR (across different output types and distributional assumptions). cILR associated methods demonstrated similar type I error to conventional differential abundance analysis methods but with more power to detect differences even at small effect sizes.

naive random forest model [48] with different single sample enrichment scoring methods as inputs (evaluating cILR, ssGSEA, and GSVA). Additionally, we also compared predictive performance of using these scores against the a standard approach of using the centered log ratio transformation (CLR) on taxon sets aggregated via abundance summations.

Simulation studies

Fig 6 shows results for simulation studies as detailed in the Materials and Methods section. Panel A presents results for a regression learning task with a continuous outcome while panel B presents results for a classification task with a binary outcome. As expected, performance across all assessed methods increased with a higher

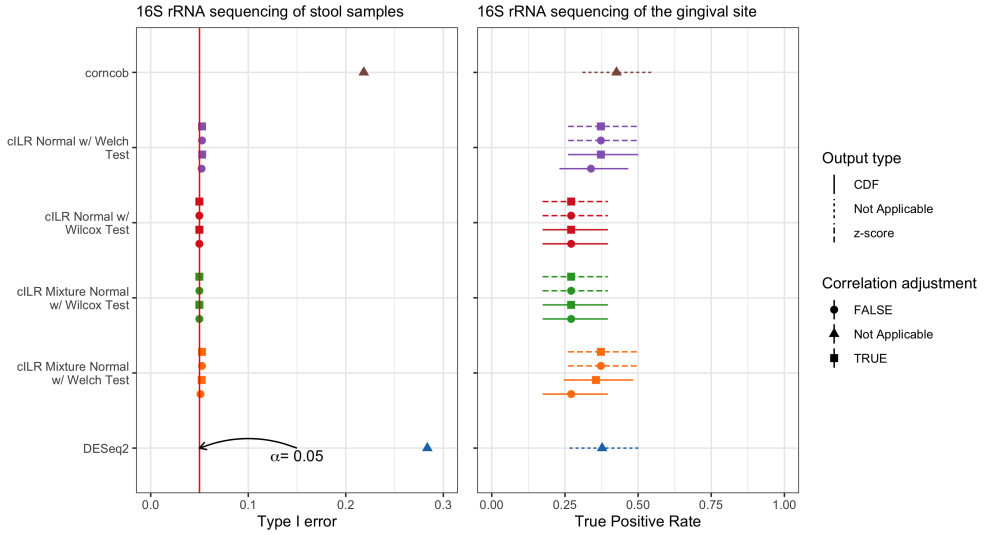

Fig 5. Differential abundance analysis using corncob, DESeq2 and cILR with either Wilcoxon rank sum test or Welch's t-test. Panel **(A)** shows type I error results as the proportion of significant genera after 500 iterations where case/control status was assigned randomly to each sample. Panel **(B)** shows true positive rate results as the proportion of significant genera who are either obligate anaerobes or aerobes. Both evaluations use 16S rRNA gene sequencing data from HMP. Type I error evaluation used stool samples while the true positive rate evaluation used samples from the gingival site. Results showed that cILR associated methods were able to keep type I error rate at approximately 0.05 while still demonstrating similar power as both corncob and DESeq2

signal-to-noise ratio. Both CLR and cILR approaches outperformed both GSVA and ssGSEA across all simulation conditions and learning tasks. This is because both GSVA and ssGSEA are more sensitive to the degree of inter-taxa correlation and sparsity, while cILR and CLR did not experience a similar level of impact. As such, performance gap widens with increasing correlation and sparsity. Interestingly, this difference in performance is not as pronounced under high levels of effect saturation (across both learning tasks), suggesting that when there is a high number of sets contributing to an effect, model choice might not be as important.

In this analysis, cILR unfortunately did not outperform the CLR approach, which is standard practice within the microbiome literature [18]. This difference in performance is more notable in regression learning tasks compared to classification, and at lower levels of effect saturation. However, the degree of separation between the two approaches is not as dramatic as between GSVA/ssGSEA and cILR/CLR. Moreover, the performance gap decreases with increasing effect signal-to-noise ratio and sparsity. Additionally, we did not observe any performance difference between the different variations of cILR.

Real data evaluations

In addition to parametric simulations, we also assessed the performance of using cILR scores in predictive models with real data sets. Fig 7 presents results for two data sets with a similar disease classification task of discriminating patients who are diagnosed with IBD (includes both Crohn's disease and ulcerative colitis) using only microbiome taxonomic composition. The two data sets represent different microbiome sequencing

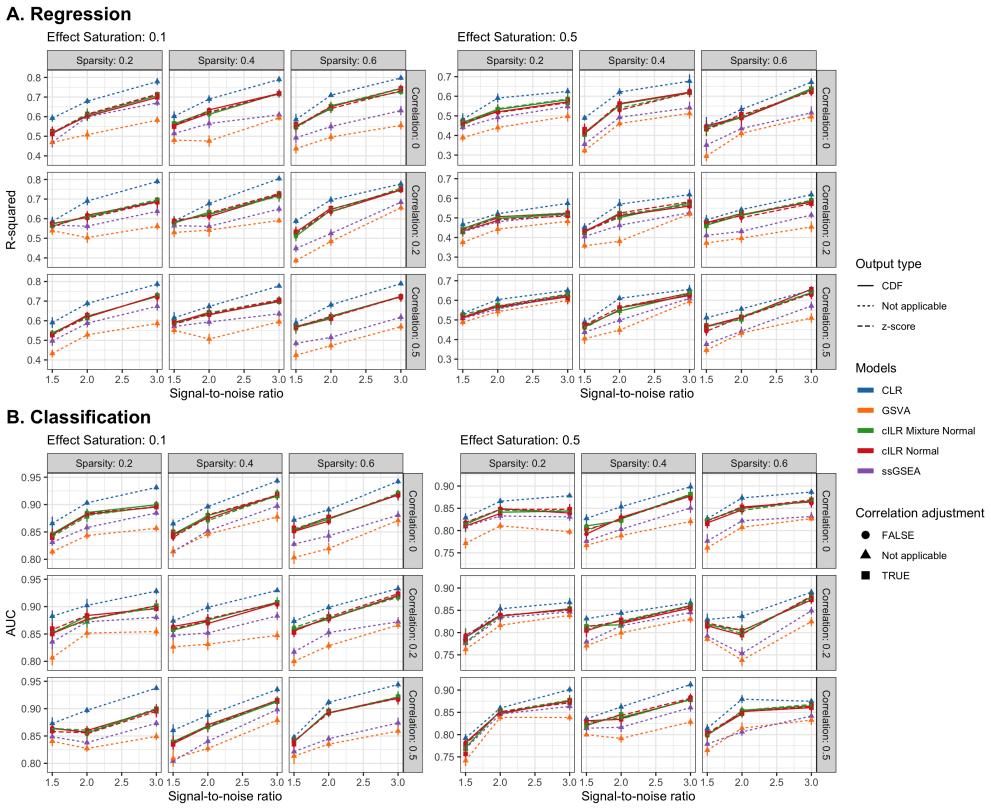

Fig 6. Predictive performance of a naive random forest model trained on cILR, ssGSEA, GSVA generated scores as well as the standard CLR approach on simulated across different levels of data sparsity, inter-taxa correlation, effect saturation, and signal-to-noise ratio. Panel **(A)** presents performance on a regression task using predictive R-squared as the evaluation measure. Panel **(B)** presents performance on a classification task with AUC as the evaluation measure. cILR approaches outperformed GSVA and ssGSEA across all simulation conditions but not the CLR approach.

aprpaoches: the Gevers et al. [60] data set uses 16S rRNA gene sequencing, while the Nielsen et al [59] data set uses whole genome shotgun sequencing.

Similar to simulation experiments, we also fitted a naive random forest model using CILR, ssGSEA, GSVA, or CLR transformed variables as inputs, and use AUC as the performance criteria. Results also replicated that of the simulations, where across both data sets cILR and CLR methods provide much better performance than both GSVA or ssGSEA. Interestingly, the cILR approach performed better than CLR in the whole genome data set but did not perform as well in the 16S rRNA gene sequencing data set. However, these results indicate that cILR generated scores are informative, providing competitive performance when acting as inputs to disease predictive models. Most importantly, performance values are consistent across both simulated and real data sets.

These results demonstrate that cILR generated scores are informative features in disease prediction tasks. Simulation results indicate that cILR methods perform much better than either GSVA or ssGSEA, but not as well as the standard CLR approach. Interestingly, however, cILR methods were much more competitive with CLR in either

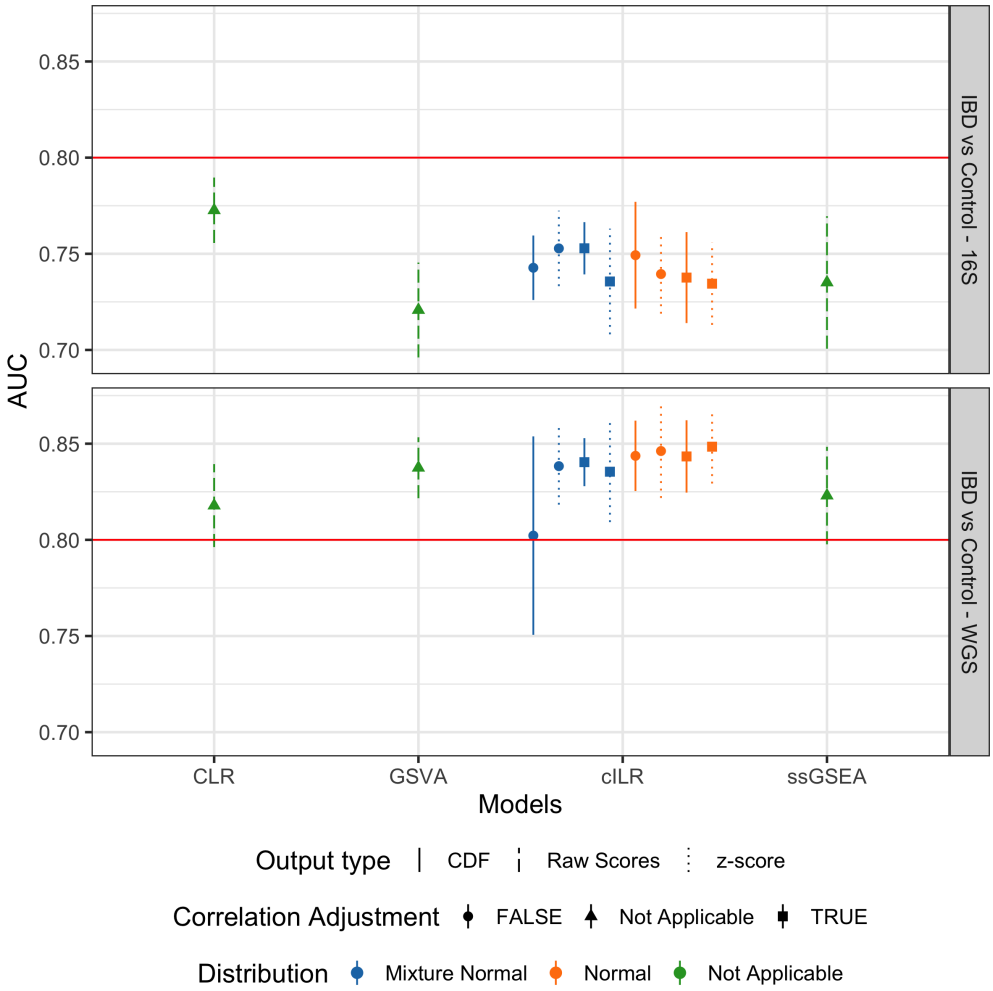

Fig 7. Predictive performance of a naive random forest model trained on cILR, ssGSEA, GSVA generated scores as well as the standard CLR approach on predicting patients with inflammatory bowel disease versus controls using genus level taxonomic profiles. Data sets used span both 16S rRNA gene sequencing (Gevers et al. [60]) and whole-genome shotgun sequencing (Nielsen et al. [59]). cILR performs better than GSVA and ssGSEA but not as well as CLR, with the exception of the whole genome sequencing data set.

WGS data sets or data sets with higher sparsity levels.

Discussion

Inference with cILR

The formulation of cILR as a comparison between taxa within the set and its complement corresponds to the competitive null hypothesis in the gene set testing literature [32]. This allows for conducting inference with cILR even at the sample-level. We assessed the usage of cILR in this type of analysis by evaluating type I error and power across both simulation studies and real data applications. Most importantly, we demonstrated that our adjusted cILR approach was able to address the issue of variance

inflation due to correlation [42] by controlling for type I error at the appropriate α level across different levels of simulated inter-taxa correlation (Fig 2) while conversely unadjusted cILR and the naive Wilcoxon rank sum test showed much higher rates of error. This is further encouraged in real data analysis where the false discovery rate was around 0.05 when a collection of true null and true alternate hypotheses were tested. Unfortunately the trade-off of good type I error control is lower power. The conservativeness of the test attenuates with higher sparsity and correlation, where power was not approaching even 50% even at the highest effect sizes. However, when the degree of such data features are reasonable, cILR will still be able to detect a reasonable proportion of samples with inflated counts.

We also observed that choosing different distributional forms did not alter performance values for cILR. This runs contrary to our comparison analysis in Fig 1, where we demonstrated that the mixture normal distribution had superior fit compared to the simple normal for raw cILR scores computed under the global null. We hypothesized that this might be due to the difficulty in fitting mixture distribution to data using the expectation maximization algorithm, as the convergence rate is slow when there is high overlap between the mixtures, resulting in a small mixing coefficient for one of the components [62]. As such, in our implementation of cILR, in order to ensure convergence for the estimating procedure we increased the number of iterations while relaxing the tolerance parameter. Furthermore, there are also possible problems with our adjustment procedure for the mixture distribution that might impact overall fit. In order to combine the scale and location estimate of two mixture distributions, we fixed the overall mean, standard deviation, mixing coefficient and component-wise means and used an optimization procedure to find the component-wise variances. However, this means that we have one equation for the overall variance but two possible parameters to estimate. As such, there is no guaranteed unique solution to component-wise variances. We hypothesized that the instability and degeneracy in component-wise variance estimates might impact the fidelity of estimates at the tails of the distribution, thereby affecting inference.

Despite these concerns, empirical results still indicate that cILR can confidently identify samples with inflated counts. The conservativeness of the correlation adjustment procedure ensures that significant results can be trusted by practitioners, even if cILR might not be able to exhaustively identify all samples with inflated counts. In situations where either the data is less sparse (e.g. containing a lot of core taxa that are prevalent across all samples), there is less inter-taxa correlation within the set (e.g. taxa that do not participate in common pathways but have shared characteristics like pathogenicity), or if the effect size is large, then cILR will still be able to produce reasonable power. A practitioner can use cILR to screen for samples for subsequent analysis that might involve significant costs, or perform hypothesis generation using a less stringent criteria alongside a multiple testing adjustment procedure (such as Benjamini-Hochberg [11]).

Downstream analysis

The sample-level enrichment scores generated by the cILR method can be used in downstream analyses commonly performed in microbiome research: differential abundance testing and disease prediction.

Differential abundance analysis

For differential abundance testing, we evaluated whether using cILR scores alongside a standard difference in means test (Welch's t-test and Wilcoxon rank sum test) is

suitable to detect changes in abundance of a set of microbes. We compared cILR against two popular approaches: corncob [61] and DESeq2 [15] applied on data where taxa were aggregated using the naive sum method. We chose DESeq2 because it is an older approach from the bulk RNA-seq literature that has strong support for usage in microbiome taxonomic data [14]. Conversely, corncob is a newer method developed specifically for microbiome taxonomic data sets, where taxonomic counts are modeled directly using a beta-binomial distribution instead of relying on normalization via size factor estimation like in DESeq2.

Surprisingly, we found some conflicting results when evaluating comparisons across parametric simulations and real data analysis. The performance of cILR was consistent across two evaluation criteria, demonstrating good type I error and respectable power. However, corncob and DESeq2 showed opposite effects: in simulation experiments, both methods show good type I error control but low power, while in real data analyses, conversely both corncob and DESeq2 showed inflated type I error but comparable power with respect to cILR methods. Despite such discrepancy, results still indicate good performance of cILR scores when used as inputs to downstream differential abundance analysis compared to using aggregated raw counts, even in methods designed to handle that type of data such as corncob and DESeq2.

We hypothesized that the above behavior can arise due to issues with performing inference in the presence taxon-specific protocol biases [29]. According to McLaren et al., the observed relative abundance of taxa is different than the true relative abundance due to protocol bias, and importantly this bias is specific to each taxon [29]. This is especially true in the context of sum-based aggregations, where the resulting bias of the aggregated taxon are dependent on the relative abundances of the contributing taxa (Appendix I in [29]). Conceptually, this means that measurement error for a taxon aggregate is different across samples as relative abundance of contributing changes, leading to issues when attempting to perform inference. As such, we expect methods like corncob or DESeq2 when performed on such aggregates to have inflated type I error compared to our taxon-ratio based approach.

The bias model also helps explain differences in performance of DESeq2 and corncob in simulation analyses compared to real data. Our simulation protocol does not explicitly include bias in our formulation, and all taxa were generated from the same underlying distribution with similar variances across all samples (the only difference is in the mean value where a taxa is expected to have inflated counts). As such, we do not expect our simulated taxa to have any taxon-specific biases, which is not the case in real data settings. Therefore, we can expect DESeq2 and corncob to retain their expected type I error control in simulation analyses compared to real data. It is still surprising to see lower power for both methods in simulation analyses, which might be due to the fact that the evaluation protocol only considers default settings for both methods and did not attempt to optimize performance.

The fact that the performance of cILR remains consistent across both simulation and real data analysis shows that cILR is invariant to taxon-specific biases. Furthermore, our evaluation indicates that even a simple difference in means test when used in combination with cILR scores can preserve type I error while maintaining good power. As such, a practioner can use cILR as a pre-processing step prior to performing a differential abundance test.

Predictive models

For disease prediction, we fitted a basic random forest model [48] to predict continuous and binary outcomes using cILR generated scores as inputs. Similar to our inference

analysis, we compared cILR against both ssGSEA and GSVA. Additionally, we also evaluated cILR with the approach where counts of a set were aggregated using sums and then centered log-ratio transformed. This is because CLR is considered standard practice in using microbiome variables as predictors for a model [18]. Results indicated that cILR produces good performance values across both real data analysis and simulation scenarios. Since predictive models consider the effect of variables jointly (and in the case of random forest, consider interactions as well), good performance indicates that cILR scores can capture joint distribution of sets, enabling both uniset and multi-set type analyses. Comparatively, cILR generated scores outperformed other enrichment score methods (GSVA and ssGSEA), suggesting that it is more tailored for microbiome taxonomic data sets. This is consistent with our sample ranking analysis (Fig 3), where cILR scores are on average more informative when used to rank samples based on their propensity to have inflated counts. However, cILR did not outperform the CLR approach across our simulation studies, and only marginally performed better in the real data analysis with WGS data.

However, in simulation studies, this performance gap between CLR and cILR decreases with higher sparsity and correlation, especially in low effect saturation scenarios. Additionally, there are also downsides to applying CLR. First, the singular covariance matrix of CLR transformed variables is singular due to a sum to zero constraint [18], preventing the proper usage of approaches that rely on matrix decomposition. Second, the procedure still relies on using summation of counts prior to transformation, which means that we still can't compare across sets of different sizes, and any bias might still be propagated [29]. As such, despite benefits in performance for a naive random forest model, there is still space for using cILR as primary inputs into predictive models.

Similar to other experiments in downstream usage of cILR, performance did not change with different underlying distributions, output types, or correlation status. This is surprising since we expect z-scores to perform better as they are able to capture the direction of an association. The fact that this effect persisted even onto our real data analysis suggests that this is not due to a deficiency of our simulation design. As such, practitioners who wish to use cILR in predictive models might be suited to use the settings that is the fastest to compute.

Ultimately, results indicate that cILR can produce informative scores that contribute to competitive performance of prediction models even in low signal-to-noise ratios with high inter-taxa correlation and sparsity. Even though there exists situations where it might not provide maximum predictive values, the flexibility of cILR in various types of analyses enable even though in some scenarios it might not provide maximum predictive values.

Limitations and future directions

There are various limitations to our evaluation of cILR. First, our simulation analysis might not capture the appropriate data-generating distributions underlying microbiome taxonomic data. There is strong evidence to suggest that our zero-inflated negative binomial distribution is representative [43], however other distributions such as the Dirichlet multinomial distribution [63] have been used in the evaluation of prior studies. Second, the usage of the gingival data set similar to [43] to assess power in differential abundance testing and single sample inference is not perfect. This is because the oxygen usage label of each microbe in the data set is only available at the genus level, and the difference in counts for obligate aerobes and anaerobes across the supragingival and subgingival sites might not be as clear cut. As such, results from power analyses using

this data set is only relative between the comparison methods. Finally, we assumed that taxa within a set are all equally associated with the outcome. This limits our ability to evaluate the performance of cILR when only a small number of taxa within the set is associated with the outcome, or if there are variability in effect sizes or association direction of taxa within a set.

Our evaluation showed various drawbacks of the cILR method. First, inference with cILR is limited in being able to exhaustively detect all samples with significant inflated counts for a set in situations where there is a high degree of sparsity and inter-taxa correlation. Second, for downstream analyses, cILR might not always perform better than competing methods, especially when being used to generate inputs to predictive models. We hypothesized that this might be due to the lack of fit for the underlying null distribution in high correlation settings, especially the identifiability problem associated with adjusting the mixture normal distribution. As such, we hope to refine the null distribution estimating procedure by either choosing a better distributional form, or to further constrain the optimization procedure of the mixture normal distribution by fixing the third and fourth moments.

In addition, there are possible extensions cILR can we can consider to provide more flexibility across different data analysis scenarios in data analysis. First, cILR did not address the sparsity of microbiome taxonomic data and relies on a pseudocount to ensure log operations are valid. We can address this by incorporating more sophisticated model-based zero-correction methods such as in [64] or [22]. Second, cILR also treated all taxa within the set as equally contributing to the set. Incorporation of taxa-specific weights could reduce the influence of outliers, such as rare or highly invariant taxa. Finally, curating sets based on *apriori* characteristics of microbes can allow for incorporating functional insights into microbiome-outcome analyses while also improving interpretability when compared to using taxonomic categories such as phylum or genus alone.

Conclusion

Gene set testing, or pathway analysis is an important tool in the analysis of high-dimensional genomics data sets. However, limited work has been done developing set based methods specifically for microbiome relative abundance data. We introduced a new microbiome-specific method to generate set-based enrichment scores at the sample level. We demonstrated that our method can control for type I error for significance testing at the sample level, while generated scores are also valid inputs in downstream analyses, including disease prediction and differential abundance.

Acknowledgments

The authors thank Becky Lebeaux, Modupe Coker, Erika Dade, Jie Zhou, and Weston Viles for insightful comments and suggestions that greatly improved the paper. This research is supported by funding from the National Library of Medicine (NLM) (grants NLM R01LM012723 and NLM K01LM012426). None of the funding bodies had a role in the design of the study or the collection, analysis, and interpretation of data and in writing of the manuscript.

References

1. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641–648. doi:10.1038/s41586-019-1238-8.
2. Sharma S, Tripathi P. Gut Microbiome and Type 2 Diabetes: Where We Are and Where to Go? The Journal of Nutritional Biochemistry. 2019;63:101–108. doi:10.1016/j.jnutbio.2018.10.003.
3. Aoun A, Darwish F, Hamod N. The Influence of the Gut Microbiome on Obesity in Adults and the Role of Probiotics, Prebiotics, and Synbiotics for Weight Loss. Preventive Nutrition and Food Science. 2020;25(2):113–123. doi:10.3746/pnf.2020.25.2.113.
4. Cho I, Blaser MJ. The Human Microbiome: At the Interface of Health and Disease. Nature Reviews Genetics. 2012;13(4):260–270. doi:10.1038/nrg3182.
5. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. Nature Methods. 2016;13(7):581–583. doi:10.1038/nmeth.3869.
6. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling. Nature Methods. 2015;12(10):902–903. doi:10.1038/nmeth.3589.
7. Li H. Statistical and Computational Methods in Microbiome and Metagenomics. In: Handbook of Statistical Genomics. John Wiley & Sons, Ltd; 2019. p. 977–550.
8. Li H. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. Annual Review of Statistics and Its Application. 2015;2(1):73–94. doi:10.1146/annurev-statistics-010814-020351.
9. Shi P, Zhang A, Li H. Regression Analysis for Microbiome Compositional Data. The Annals of Applied Statistics. 2016;10(2):1019–1040. doi:10.1214/16-AOAS928.
10. Sankaran K, Holmes S. structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data. Journal of statistical software. 2014;59(13):1–21. doi:10.18637/jss.v059.i13.
11. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.
12. Chen J, King E, Deek R, Wei Z, Yu Y, Grill D, et al. An Omnibus Test for Differential Distribution Analysis of Microbiome Sequencing Data. Bioinformatics. 2018;34(4):643–651. doi:10.1093/bioinformatics/btx650.
13. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and Microbial Differential Abundance Strategies Depend upon Data Characteristics. Microbiome. 2017;5(1). doi:10.1186/s40168-017-0237-y.
14. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLOS Computational Biology. 2014;10(4):e1003531. doi:10.1371/journal.pcbi.1003531.
15. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biology. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

16. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A Field Guide for the Compositional Analysis of Any-Omics Data. GigaScience. 2019;8(giz107). doi:10.1093/gigascience/giz107.

17. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding Sequencing Data as Compositions: An Outlook and Review. Bioinformatics. 2018;34(16):2870–2878. doi:10.1093/bioinformatics/bty175.

18. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017;8. doi:10.3389/fmicb.2017.02224.

19. Aitchison J. A Concise Guide to Compositional Data Analysis. 1999; p. 134.

20. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology. 2015;11(5):e1004226. doi:10.1371/journal.pcbi.1004226.

21. Kaul A, Davidov O, Peddada SD. Structural Zeros in High-Dimensional Data with Applications to Microbiome Studies. Biostatistics. 2017;18(3):422–433. doi:10.1093/biostatistics/kxw053.

22. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of Microbiome Data in the Presence of Excess Zeros. Frontiers in Microbiology. 2017;8. doi:10.3389/fmicb.2017.02114.

23. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology. 2012;8(2):e1002375. doi:10.1371/journal.pcbi.1002375.

24. Goeman JJ, Bühlmann P. Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues. Bioinformatics. 2007;23(8):980–987. doi:10.1093/bioinformatics/btm051.

25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102.

26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the Unification of Biology. Nature genetics. 2000;25(1):25–29. doi:10.1038/75556.

27. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. BMC Bioinformatics. 2013;14(1):7. doi:10.1186/1471-2105-14-7.

28. Frost HR. Variance-Adjusted Mahalanobis (VAM): A Fast and Accurate Method for Cell-Specific Gene Set Scoring. Nucleic Acids Research. 2020;48(16):e94–e94. doi:10.1093/nar/gkaa582.

29. McLaren MR, Willis AD, Callahan BJ. Consistent and Correctable Bias in Metagenomic Sequencing Experiments. eLife. 2019;8:e46923. doi:10.7554/eLife.46923.

30. Egozcue JJ, Pawlowsky-Glahn V. Groups of Parts and Their Balances in Compositional Data Analysis. Mathematical Geology. 2005;37(7):795–828. doi:10.1007/s11004-005-7381-9.

31. Chong J, Liu P, Zhou G, Xia J. Using MicrobiomeAnalyst for Comprehensive Statistical, Functional, and Meta-Analysis of Microbiome Data. Nature Protocols. 2020;15(3):799–821. doi:10.1038/s41596-019-0264-1.

32. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering Statistically Significant Pathways in Expression Profiling Studies. Proceedings of the National Academy of Sciences. 2005;102(38):13544–13549. doi:10.1073/pnas.0506577102.

33. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology. 2003; p. 22.

34. Silverman JD, Washburne AD, Mukherjee S, David LA. A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data. eLife. 2017;6:e21887. doi:10.7554/eLife.21887.

35. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA Interference Reveals That Oncogenic *KRAS*-Driven Cancers Require TBK1. Nature. 2009;462(7269):108–112. doi:10.1038/nature08460.

36. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. Nucleic Acids Research. 2013;41(D1):D590–D596. doi:10.1093/nar/gks1219.

37. Delignette-Muller ML, Dutang C. Fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software. 2015;64(4):1–34.

38. Benaglia T, Chauveau D, Hunter DR, Young D. Mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software. 2009;32(6):1–29.

39. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, et al. Phylogenetic Factorization of Compositional Data Yields Lineage-Level Associations in Microbiome Datasets. PeerJ. 2017;5:e2969. doi:10.7717/peerj.2969.

40. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, et al. Balance Trees Reveal Microbial Niche Differentiation. mSystems. 2017;2(1). doi:10.1128/mSystems.00162-16.

41. Efron B. Large-Scale Simultaneous Hypothesis Testing. Journal of the American Statistical Association. 2004;99(465):96–104. doi:10.1198/016214504000000089.

42. Wu D, Smyth GK. Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. Nucleic Acids Research. 2012;40(17):e133. doi:10.1093/nar/gks461.

43. Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of Statistical Methods from Single Cell, Bulk RNA-Seq, and Metagenomics Applied to Microbiome Data. Genome Biology. 2020;21(1):191. doi:10.1186/s13059-020-02104-1.

44. Cario MC. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. 1997; p. 19.

45. Agresti A, Coull BA. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician. 1998;52(2):119–126. doi:10.2307/2685469.

46. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–845.

47. Hawinkel S, Mattiello F, Bijnens L, Thas O. A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate. Briefings in Bioinformatics. 2019;20(1):210–221. doi:10.1093/bib/bbx104.

48. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.

49. Kuhn M, Wickham H. Tidymodels: Easily Install and Load the 'tidymodels' Packages; 2020.

50. Xiao J, Chen L, Yu Y, Zhang X, Chen J. A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data. Frontiers in Microbiology. 2018;9. doi:10.3389/fmicb.2018.03112.

51. Dong M, Li L, Chen M, Kusalik A, Xu W. Predictive Analysis Methods for Human Microbiome Data with Application to Parkinson's Disease. PLOS ONE. 2020;15(8):e0237779. doi:10.1371/journal.pone.0237779.

52. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, Curated Metagenomic Data through ExperimentHub. Nature Methods. 2017;14(11):1023–1024. doi:10.1038/nmeth.4468.

53. Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, et al. HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor. American Journal of Epidemiology. 2019;doi:10.1093/aje/kwz006.

54. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: Rapid, Web-Enabled Microbiome Meta-Analysis. Nature Methods. 2018;15(10):796–798. doi:10.1038/s41592-018-0141-9.

55. Consortium THMP, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, Function and Diversity of the Healthy Human Microbiome. Nature. 2012;486(7402):207–214. doi:10.1038/nature11234.

56. Thurnheer T, Bostanci N, Belibasakis GN. Microbial Dynamics during Conversion from Supragingival to Subgingival Biofilms in an in Vitro Model. Molecular Oral Microbiology. 2016;31(2):125–135. doi:10.1111/omi.12108.

57. Beghini F, Renson A, Zolnik CP, Geistlinger L, Usyk M, Moody TU, et al. Tobacco Exposure Associated with Oral Microbiota Oxygen Utilization in the New York City Health and Nutrition Examination Study. Annals of Epidemiology. 2019;34:18–25.e3. doi:10.1016/j.annepidem.2019.03.005.

58. Calgaro M. Mcalgaro93/Sc2meta: Paper Release; 2020. Zenodo.

59. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes. Nature Biotechnology. 2014;32(8):822–828. doi:10.1038/nbt.2939.

60. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe. 2014;15(3):382–392. doi:10.1016/j.chom.2014.02.005.

61. Martin BD, Witten D, Willis AD. Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression. The Annals of Applied Statistics. 2020;14(1):94–115. doi:10.1214/19-AOAS1283.

62. Naim I, Gildea D. Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients. Proceedings of the 29th International Coference on International Conference on Machine Learning. 2012; p. 8. doi:10.5555/3042573.3042756.

63. Wu C, Chen J, Kim J, Pan W. An Adaptive Association Test for Microbiome Data. Genome Medicine. 2016;8(1):56. doi:10.1186/s13073-016-0302-3.

64. Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Model-Based Replacement of Rounded Zeros in Compositional Data: Classical and Robust Approaches. Computational Statistics & Data Analysis. 2012;56(9):2688–2704. doi:10.1016/j.csda.2012.02.012.

Supporting information

S1 Fig. Computational performance of cILR. Computational time (in seconds) as a function of sample size (left panel) and number of taxa sets evaluated (right panel). Evaluation was performed on simulated data sets. For sample size analysis, only 10 sets were evaluated. For taxa set analysis, sample size was fixed at 1,000. Across all evaluations, the size of each taxa set was also fixed at 50.

S2 Fig. Distribution of p-values. Q-Q plot of 10,000 p-values compared against a uniform distribution bounded between 0 and 1. Evaluation was performed on simulated null data sets of 10,000 samples testing for enrichment of a set of size 50. For correlation of 0.5, p-values represent correlation adjusted cILR while for correlation of 0, p-values represent unadjusted cILR.

S1 File. Supplemental derivations. Includes additional details on addressing variance inflation due to correlation in cILR, as well as computational performance and p-value distribution of the method.