cILR: 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, USA*
[2]*Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA*

Abstract

Background: 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 the classification of representative sequences. However, most researchers perform aggregation through the pairwise summation of counts, preventing comparison across sets of different sizes.

Methods: We developed a taxa 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 corresponding to the Q_1 competitive null hypothesis in the gene set testing literature. Significance testing was performed by estimating the empirical null distribution accounting for variance inflation due to inter-taxa correlation.

Results: Here we demonstrated the performance of our method using both real data and parametric simulations for multiple microbiome analysis tasks, which are: single sample enrichment testing, differential abundance testing, and disease prediction.

Conclusions: The cILR method provides a flexible way to aggregate taxonomic variables to pre-defined sets, allowing for a comparison of enrichment across sets of different sizes. The statistic corresponds to a well-defined null hypothesis and is designed to address the compositional nature of microbiome data

1 Background

The microbiome is the collection of microorganisms (bacteria, protozoa, archaea, fungi, and viruses) which coexist with their human host. Previous research has shown that changes in the composition of the microbiome have been associated with important health outcomes such as inflammatory bowel disease, type II diabetes, and obesity. To understand the central role of the microbiome in human health, researchers often 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) [14]. Processing raw sequencing data using a variety of bioinformatic pipelines [11, 55] would yield taxonomic abundance tables that would be part of downstream statistical analyses where associations between an outcome/exposure and identified taxa would be found.

However, there exists unique challenges in the analysis of these data tables [36, 35]. 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 [50], allowing for simultaneous model fitting and variable selection. For differential abundance tasks, researchers often utilized multiple hypothesis correction methods [48, 7] or omnibus tests [13] to address hypothesis testing burden.

Second, the number of reads obtained is constrained by the sequencing instrument at an arbitrary limit, and applied inconsistently across samples, resulting in a different number of total read counts per sample. Many normalization methods [57] have been proposed to address these issues, including cross-applying methods from the gene expression literature [39]. 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 [37], which might not be not true in the context of microbiome relative abundance data [46, 47]. As such, microbiome data is strictly compositional [25], 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 [2].

Third, the data is 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 [34], or applying zero-inflated models [13, 30]. 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 [31].

Even though the aforementioned problems are challenging, a very approachable method to address some of them is variable aggregation. Aggregated variables can be less sparse than their constituent elements, and hypothesis testing on a reduced number of variables can alleviate the multiple testing burden. This helps increase power in downstream analyses, as well as interpretability. In practice, microbiome researchers perform variable aggregation to pre-defined taxonomic levels (e.g. phylum, family, order) through element-wise summation of counts. However, this approach contains various downsides: first, it does not allow for comparison of enrichment across sets of different sizes, where larger sets naturally contain more counts; second, aggregating compositional variables using component-wise summations can distort the inter-sample distances before and after aggregation due to the non-linearity of this amalgamation in Aitchison space [21].

Set-based analyses are ubiquitous in the gene expression literature [26] and various methods have been developed for enrichment testing and scoring. Here, we leverage the conception of the Q_1 competitive hypothesis presented in Tian et al. [54], which compares the expression of genes within the gene set against the rest of the genes, in the context of microbiome relative abundance data. The competitive 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 [22], which allows for a log-ratio-based aggregation method that addresses the downsides of the naive summation-based method presented above. The resulting method, titled competitive isometric log-ratio (cILR) is a microbiome-specific approach to set-based enrichment. cILR is unsupervised and can generate sample-specific enrichment scores with a well-defined null hypothesis that allows for significance testing. The resulting scores are also flexible and informative, with applications in downstream analyses.

In this manuscript, we present cILR, a novel method to generate sample-level taxa set enrichment scores for microbiome relative abundance data. We illustrate the benefits of cILR as both a sample-level significance test for enrichment of taxa sets and as a feature engineering approach for downstream disease prediction and differential abundance. We compare the performance of cILR in these respective tasks against standard microbiome data analysis practices, as well as existing sample-level enrichment methods in the gene expression literature such as GSVA [28] and ssGSEA [4]. An R package implementation of this approach can be found on GitHub.

2 Methods

2.1 Competitive Isometric Log-ratio (cILR)

The cILR method generates sample-specific enrichment scores for microbial sets using the isometric log-ratio transformation [22]. Details on the computational implementation of cILR can be found in the supplemental section. 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 16S rRNA) or whole genome shotgun sequencing. Usually $\mathbf{X}$ is generated from standard taxonomic profiling pipelines such as DADA2 [11] for 16S rRNA sequencing, or MetaPhlAn2 [55] for whole genome shotgun sequencing.
- $\mathbf{A}$: 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 [45], 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:

- **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 $\mathbf{M}_{i,k}$ 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 **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 (Figure 1), an empirical null distribution will be estimated from data.

   - **Compute the cILR statistic on permuted and un-permuted 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 **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 [17]. For the mixture normal distribution, parameters were estimated using an expectation-maximization algorithm implemented in the *mixtools* package [6].

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

2.2 Properties of cILR

2.2.1 cILR and the Isometric Log Ratio Transformation

The cILR statistic is a special instance of the isometric log-ratio transformation (ILR) [22]. 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 $\mathbb{R}^{D-1}$ [22, 56]. 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 [21]. 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 [22].

The usage of the ILR statistic is not uncommon in the microbiome literature. They are usually termed "compositional balances", and were leveraged in many recent approaches in variable transformation [56, 51, 40]. The cILR formulation (1) is a special case of (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 [51, 40, 56], 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.

2.2.2 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 [54] 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 [21]. 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 [20] 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 figure 1, we simulated microbiome 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 [20].

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 figure 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 [58]. 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.

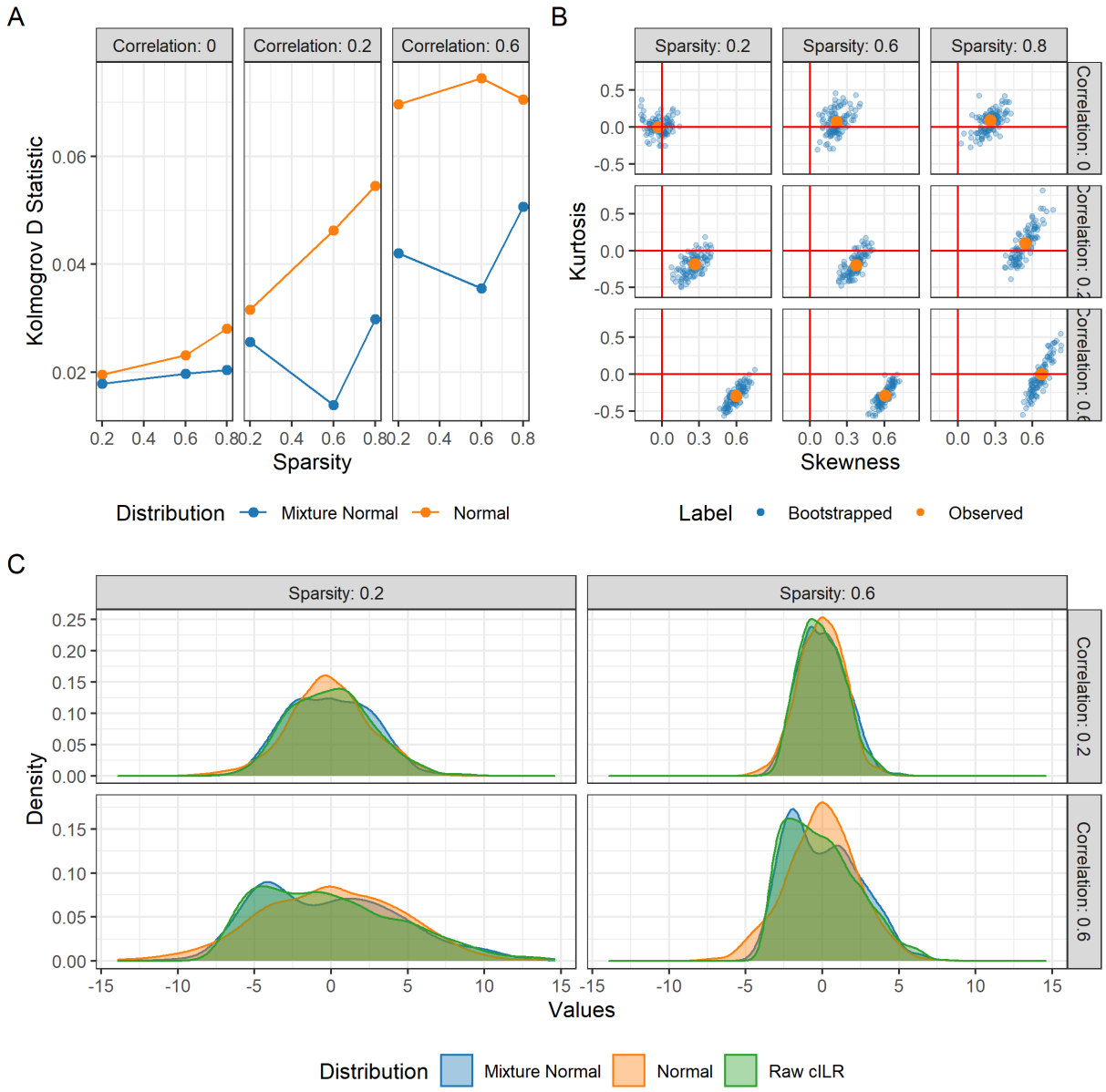

Figure 1. Properties of the null distribution of cILR in different simulation scenarios under the global null. Panel **(A)** presents kurtosis and skewness of cILR scores while panel **(B)** presents the goodness of fit (as Kolmogorov-Smirnov D statistic) for mixture normal and normal distributions. 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.

2.3 Evaluation

2.3.1 Parametric Simulations

To address the performance of cILR under 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 [10]. Suppose X_{ij} are observed counts for a sample i and taxon j,

then we have the following probability model

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

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 [12]. 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 $\mathbf{NB}$ 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 [17]) 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 [1] 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 et al. [23] 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 [18] 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 relative nature of microbiome 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. [29] to ensure the validity of simulation results. All sample sizes were set at 2,000.

Prediction: To assess the predictability, 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 [8]. 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* [33] 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 [59]. 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

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

To ensure a balance of classes, we applied the strategy described in Dong et al. [19] 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.

2.3.2 Real Datasets

In addition to simulation analyses, we also evaluated our method using real data sets across both 16S rRNA gene sequencing and whole-genome sequencing. All data sets are obtained from either the *curatedMetagenomicData* [43] and *HMP16SData* [49] R packages (2020-10-02 snapshot), or downloaded from the Qiita platform [27].

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 [16, 44]. We utilized this data set following the approach outlined in Calagaro et al. [10]. This data set is special because it 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 [53]. Here, we assessed the enrichment of aerobic microbes across all samples, we considered the false discovery 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. [5] and was downloaded directly from the Github repository associated with Calagaro et al. [9].

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 [16, 44]. 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 [42]. 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. [24] which also profiles the gut microbiome of IBD patients and controls but using 16S rRNA gene sequencing. This data set contains Y 16S rRNA gene sequencing samples from a cohort of ABC, where X was classified as having IBD.

2.3.3 Comparison Methods

Single sample enrichment: For type I error and power analyses, we compared the cILR method with a naive Wilcoxon rank-sum test. 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 [28], ssGSEA [4], and the W-statistic from the Wilcoxon rank-sum test.

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. For comparison, we chose representative state-of-the-art methods in differential abundance analysis, namely DESeq2 [37, 39] and corncob [38].

Disease Prediction: We fit random forest on cILR scores, as well as ssGSEA [28] and GSVA [4] similar to single sample enrichment section. Additionally, we also compared performance using enrichment scores 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.

3 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.

3.1 Enrichment testing at the sample level

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

Panel A and B in figure 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 shown in figure 2B. We included only adjusted cILR methods since power is only meaningful when type I error is properly controlled. Results indicated that both sparsity and inter-taxa correlation impacts power, correlation more so than sparsity. The adjusted cILR approaches were able to maintain power above 0.8 even in low effect size settings as long as there was negligible inter-taxa correlation within the set. However, in higher correlation scenarios ($\rho = 0.2$ and 0.5), power was dramatically reduced to 0.25 and below and would require higher effect sizes to maintain power at 0.8, especially at higher sparsity levels.

These results were replicated when assessed on the semi-labeled gingival data set from the Human Microbiome Project as described in Methods. Here, we tested the enrichment of aerobic microbes at each sample using approaches similar to our parametric simulations. As expected in Figure 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 (Figure 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.

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. Figure 3 presents this result. We compared different variants of cILR against

competing methods in the gene set testing space (GSVA [28] and ssGSEA [4]), as well as the U test statistic from the Wilcoxon rank-sum test. Across both simulations (Figure 3A) and real-data applications (Figure 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.

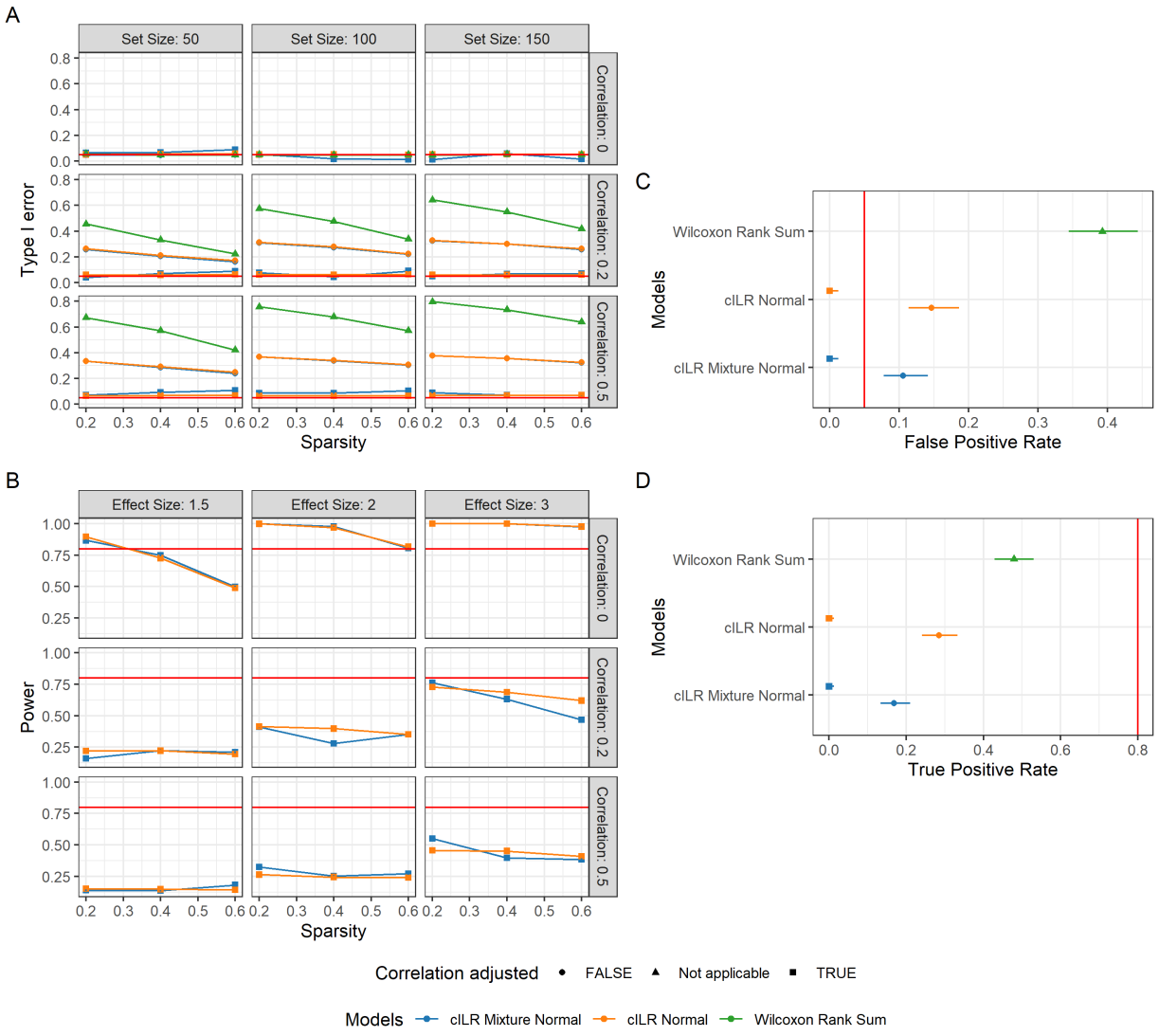

Figure 2. Type I error rate **(A)**, and power **(B)** across different parametric simulation scenarios. Confidence bounds were obtained using Agresti-Couli [1] approach. False-positive rate **(C)** and true positive rate **(D)** evaluation of similar methods on real 16S rRNA data from the oral microbiome of the gingival site. For **(A)** and **(B)**, enrichment of a specified set was tested at the sample level using cILR and the Wilcoxon rank-sum test at α of 0.05. For **(C)** and **(D)**, 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. 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.

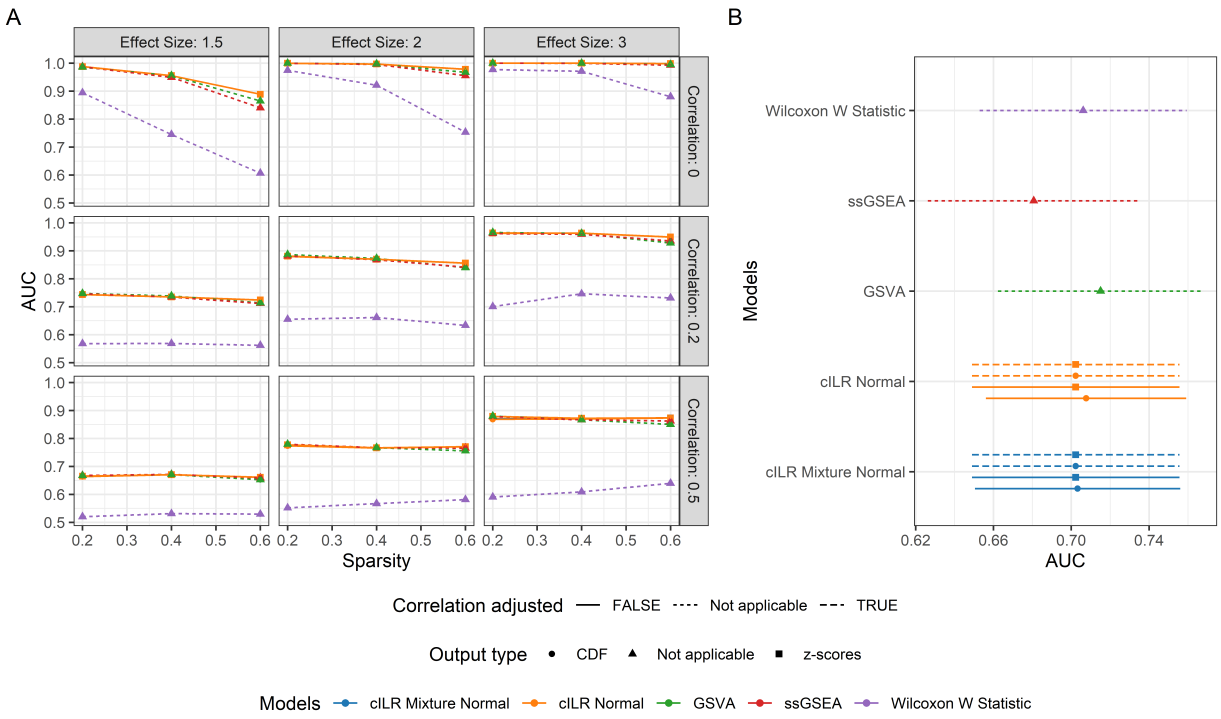

Figure 3. Classification performance via AUC of cILR, ssGSEA, GSVA, and Wilcoxon U statistic on simulated data **(A)** the gingival data set from the Human Microbiome Project **(B)** as detailed in Methods. Performance scores measure whether scores can highly rank samples that are known to have inflated 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 [18]

3.2 Differential abundance analysis

Here we assessed the ability to use cILR with a standard difference of means statistical test (such as Welch's t-test or a Wilcoxon rank-sum test) to perform differential abundance analysis. We compare the performance of cILR related methods to two commonly used approaches in the microbiome literature: DESeq2 [37] and corncob [38].

3.3 Disease Prediction

We can conceive of cILR as a feature engineering step before model fit and prediction. Here we assessed the predictive performance of a standard machine learning model (random forest [8]) using cILR scores compared to scores generated by other single sample enrichment methods in the gene set testing literature (ssGSEA [4] and GSVA [28]). Additionally, we also compared our results with just using the centered log-ratio (CLR) transformation on taxa sets aggregated via count summations.

Figure 4 and 5 showed results from simulation studies as detailed in the Methods section under both classification and regression tasks. As expected, performance across all assessed methods increased with a higher signal-to-noise ratio, however, they remained consistent across different levels of model saturation and inter-taxa correlation. Importantly, cILR methods outperformed both GSVA and ssGSEA, where the difference in AUC increases as a function of sparsity. Conversely, cILR scores did not perform as well as the traditional CLR approach, however, the performance gap decreased at higher sparsity levels. Similar patterns were observed for regression results (Figure 5). Most interestingly, regression performance, in general, decreased at higher saturation levels (proportion of sets associated with the outcome). cILR remained more performant

than both ssGSEA and GSVA, but was still less than the standard CLR, especially at higher signal-to-noise ratios. Similarly, cILR still performed much better in higher sparsity situations, where the results were almost identical to the CLR approach, especially in high saturation scenarios.

In addition to parametric simulations, we also demonstrated predictive performance using real data sets in Figure 6. The learning task across both data sets is to classify patients who are diagnosed with inflammatory bowel disease (includes both Crohn's disease and ulcerative colitis) and healthy controls. The Gevers et al. [24] data set is a 16S rRNA sequencing data set while the Nielsen et al. [42] is a whole-genome shotgun sequencing data set. Similar to simulation experiments, across both data sets cILR methods provide much better performance than both GSVA and ssGSEA. Interestingly, the standard CLR approach only outperformed cILR in the 16S rRNA data set but was marginally worse in the WGS data set. Additionally, we observed that using the normal distribution for cILR generates better predictive performance compared to the other variants.

These results demonstrated that cILR generated scores are informative features in disease prediction tasks. Simulation results indicated that cILR methods perform much better than both ssGSEA and GSVA, but not as good as the standard CLR approach. Interestingly, however, cILR methods were much more competitive with CLR in either WGS data sets or data sets with higher sparsity levels.

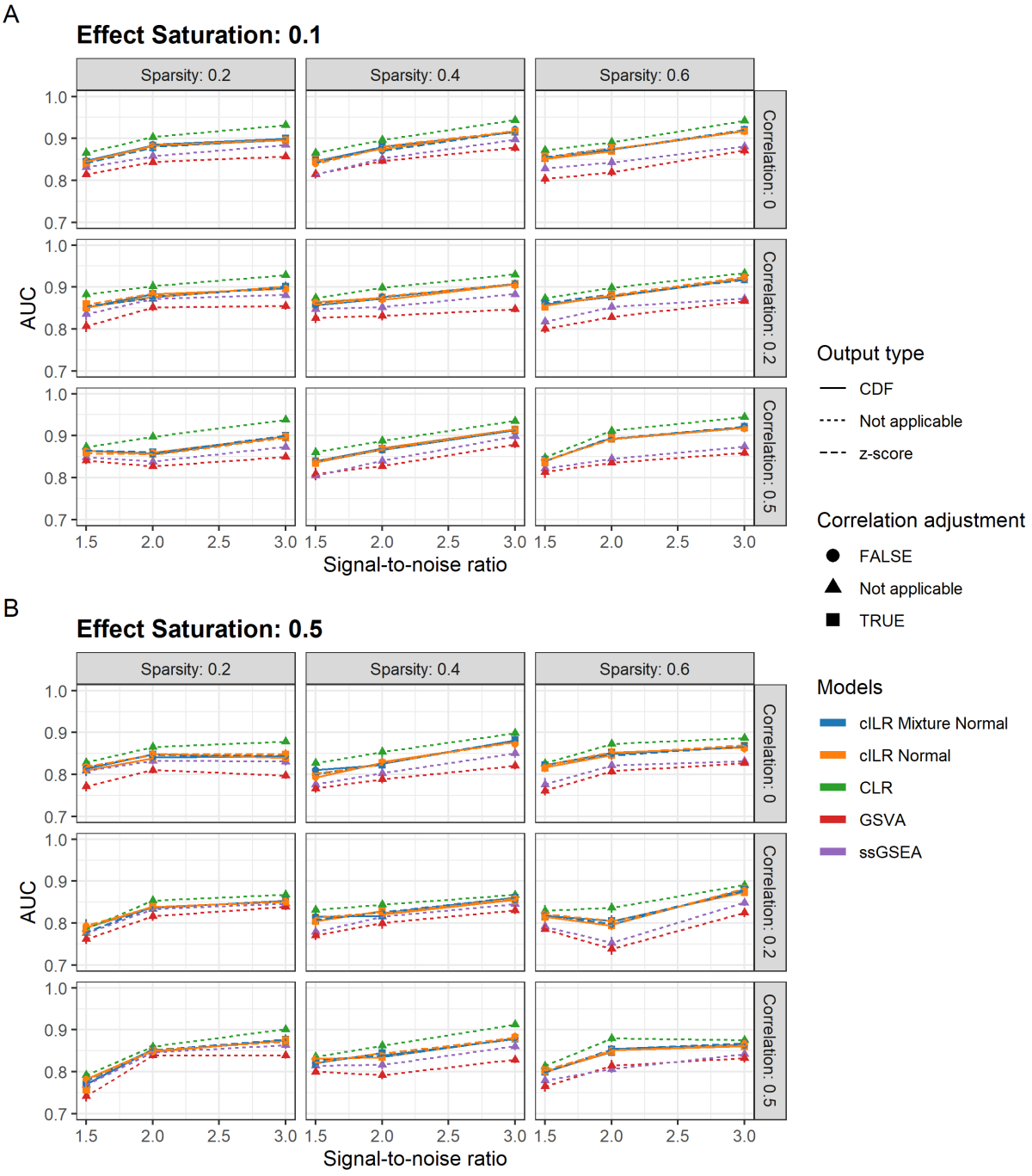

Figure 4. AUC scores of a random forest model for a binary outcome trained on cILR, ssGSEA, GSVA generated scores as well as on standard CLR transformed data evaluated on simulated data across sparsity levels, correlation, and signal-to-noise ratio. Panel **(A)** and **(B)** represent results across different levels of model saturation (proportion of sets associated with the outcome). cILR approaches outperformed GSVA and ssGSEA but not standard CLR.

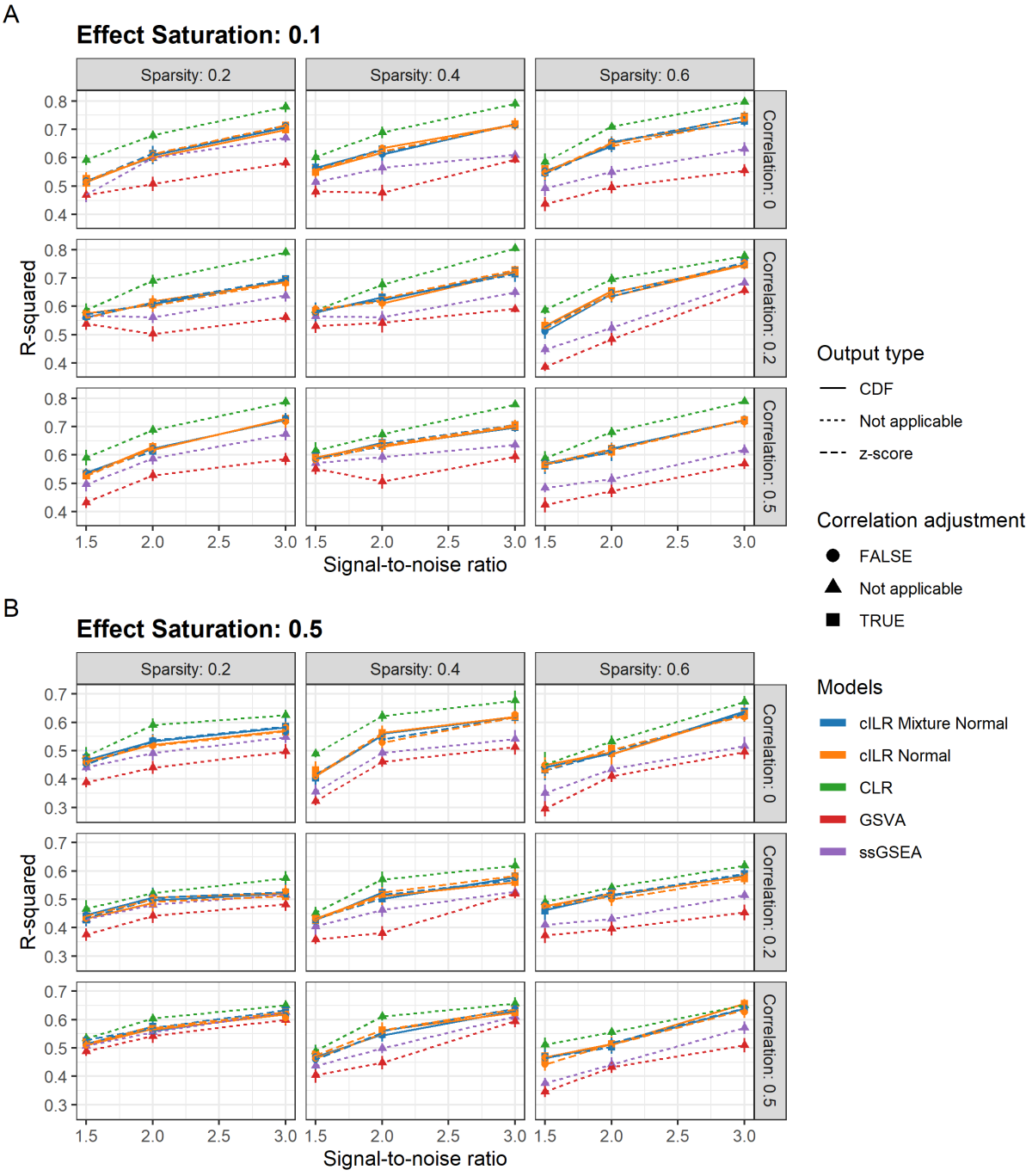

Figure 5. Predictive R-squared of a random forest model for a continuous outcome trained on cILR, ssGSEA, GSVA generated scores as well as on standard CLR transformed data evaluated on simulated data across sparsity levels, correlation, and signal-to-noise ratio. Panel **(A)** and **(B)** represent results across different levels of model saturation (proportion of sets associated with the outcome). cILR approaches outperformed GSVA and ssGSEA but not standard CLR.

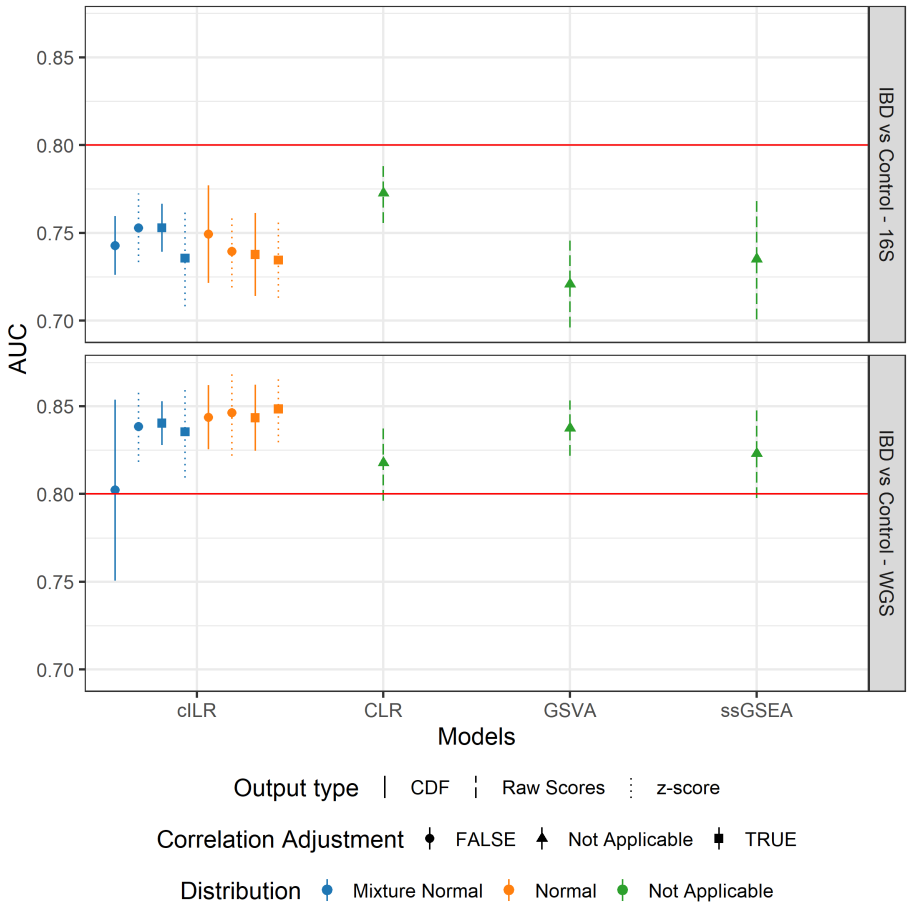

Figure 6. Classification performance of a standard random forest model using cILR scores compared against existing methods in gene set testing literature and the standard centered-log ratio transformation approach. The learning task involves predicting patients with inflammatory bowel disease (including Crohn's disease and ulcerative colitis) versus controls. Data sets used span both 16S rRNA sequencing (Gevers et al. [24]) and whole-genome shotgun sequencing (Nielsen et al. [42]). cILR marginally outperforms GSVA and ssGSEA in both data sets, but only performs better than CLR in the WGS data set.

4 Discussion

4.1 Set-based analysis can enhance microbiome data analysis

Gene-set testing (or pathway analysis) is an important tool in the analysis of high-dimensional genomic data sets [32]. By focusing on a group of variables rather than individuals, this approach reduces the hypothesis testing burden, thereby increasing power as well as replicability. Furthermore, through the usage of functionally informative sets defined apriori based on historical experiments (for example MSigDB [52], and Gene Ontology [3]), gene set analyses also increase interpretability.

In microbiome research, even though no explicit enrichment analysis is performed, researchers often aggregate variables to taxonomic classification levels such as genus, family, or phylum. The process of aggregation involves the element-wise summation of counts of taxa that belong to the same taxonomic group. Even though this allows for a reduction in the number of overall variables (from thousands to only hundreds), there still exist two disadvantages: first, it doesn't preserve inter-sample distances before and after aggregation [21], and second, it doesn't allow for enrichment testing and comparison across sets of different sizes.

As such, there is a need for microbiome researchers to adopt set enrichment methods to perform a more robust set-based analysis that allows for enrichment testing of functionally relevant sets of microbes. However, existing methods in the gene testing literature cannot be directly applied to microbiome data. This is because microbiome relative abundance data is strictly compositional [25] and constrained, whereas expected inputs for gene set testing methods assume data is unconstrained in real space.

Here, we developed cILR, a novel method to perform set-based enrichment analysis tailored for microbiome data. Our method is based on the isometric log-ratio transformation [22], and measures set enrichment relative to those outside the set, which corresponds to a Q_1 null hypothesis in the gene set analysis literature [54]. cILR generates sample-specific enrichment scores, enabling cILR to act as both a significance test for enrichment at the sample level, and as inputs for downstream analyses such as differential abundance testing, predictive modeling, and data visualization. This is different than the taxon set enrichment method implemented in the *MicrobiomeAnalyist* web application [15], which performs enrichment testing across the entire data set through a hypergeometric test following standard differential abundance testing.

4.2 Significance testing with cILR

We can use cILR to test for enrichment of taxa-sets at the sample level. Inter-taxa correlation can result in an inflated variance of resulting test statistics [58], which results in an inflated type I error if not properly controlled for. Our method addresses this issue by combining the mean estimate of scores computed on permuted data and the variance estimate of scores computed on unpermuted data, where correlation is undisrupted. As a consequence, adjusted cILR methods control very well for type I error at the appropriate α level (Figure 2). The conservativeness of the test persisted across different sparsity levels, as well as inter-taxa correlation. Conversely, the unadjusted cILR methods and the naive Wilcoxon rank-sum test has inflated type I error even at low correlation levels ($\rho = 0.2$). However, the trade-off for proper type I error control is lower power. Simulation results showed that power is sensitive to sparsity across all correlation levels.

However, if there is no correlation between members of a set, cILR was still able to maintain 80% power even with low effect size (fold change of 1.5), long as the set of interest is more densely populated, such as a set of core taxa that are prevalent across all samples. Unfortunately, this relationship doesn't hold for higher correlation levels. This means that in sets where associated taxa are highly correlated, a higher fold change is required for cILR to detect enrichment.

We also assessed cILR using different distribution forms, the normal distribution, and the mixture normal distribution. The difference in performance across the two choices was minimal, where both were able to control for type I error, but the normal distribution was slightly better when it comes to power only in certain simulation settings. This runs contrary to our distribution comparison analysis in Figure 1, where the mixture distribution was the superior fit. We hypothesized that this might be due to the difficulty in fitting mixture distributions using the expectation-maximization algorithm (EM), as the convergence rate is slow when there is a lot of overlap between the mixtures and the mixing coefficients of one of the components are low [41]. Furthermore, there is a lot of degeneracy of parameter values in our variance adjustment procedure for the mixture distribution(Supplemental Material). These distribution fitting problems might introduce issues where the mixture distribution might not be the best fit. As such, improvements to mixture distribution fitting might allow for better performance to be extracted.

These results are replicated in our real data analysis, where we attempted to test for enrichment for a group of aerobic taxa among samples in the gingival site from HMP [16]. Even though our labeling is not perfect, it is encouraging to see that the false discovery rate is low (approximately 0.05) when tested among a set of hypotheses where there is a mixture of true null and alternate situations. We believe that a practitioner can identify many interesting samples to follow up on by applying multiple testing adjustment procedures (such as Benjamini-Hochberg [7]) with cILR p-values and using less stringent criteria (e.g. 0.1 false discovery rate).

In short, we demonstrated that cILR can be used to test for significance testing at the sample level. Even though there is not a lot of difference between the distribution choices, ultimately cILR controls for type I error well despite tradeoffs in power. Even then, cILR will still be able to detect enriched sets in high

correlation and sparsity situations insofar as the effect size is large. Researchers can utilize cILR to test for enrichment of a certain set of taxa of interest to perform follow-up experiments or to identify the prevalence of certain sub-population of taxa (for example, oxygen preference as in our real data example). The conservativeness of the test ensures confidence in detected samples.

4.3 Using cILR as inputs for downstream analyses

Since cILR generates enrichment scores for each set at the sample level, they can be used in downstream analyses alongside standard statistical approaches. In this section we assessed the utility of cILR generated scores in two common analysis tasks in microbiome research: differential abundance testing and disease prediction.

For disease prediction, we benchmarked a basic random forest model [8] to predict whether or not a patient has IBD given only their microbiome profile using cILR and relevant methods as inputs. Across both simulations and real data analyses, cILR scores perform better than alternative single sample scoring methods in the gene set testing literature (GSVA [28] and ssGSEA [4]). This is consistent with a previous analysis (Figure 3), where we assessed the informativeness of cILR generated scores by using AUC to measure whether samples with high scores correspond with the true label of set inflatedness. Results indicated that cILR scores is more adaptive to microbiome data sets than GSVA and ssGSEA. However, cILR scores underperforms the standard approach of applying the CLR transformation to count aggregated data, with the exception of the WGS real data analysis. Despite this, there are still benefits to using cILR scores in the context of downstream modelling beyond random forest predictive models. First, CLR transformed variables have a sum to zero constraint, making the covariance matrix is singular [25], impacting methods that rely on matrix decomposition such as principal component regression or canonical correlation analysis. Second, cILR provides significance testing for enrichment for the purposes of sample screening. Third, CLR transformed variables still involve the count summation procedure for variable aggregation, therefore comparision between sets of different sizes is not viable.

Despite such drawbacks, cILR methods still since machine learning predictive models consider modelling all variables jointly (and in the case of random forest, consider interactions as well), good performance also indicated that cILR scores can also capture the joint distribution of considers sets, especially with regards to their relationship with a defined outcome such as IBD status.

Unsurprisingly, different variants of CILR did not differ significantly in performance. Even though we expect Z-scores to perform better due to the added information of directionality of enrichment, using CDF values did not alter performance in both the single sample AUC evaluations (Figure 3) as well as the random forest evaluations (Figure 4).

4.4 Limitations and future directions

5 Conclusion

References

[1] Alan Agresti and Brent A. Coull. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, 1998.

[2] John Aitchison. A Concise Guide to Compositional Data Analysis. page 134, 1999.

[3] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[4] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch, Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov, David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn. Systematic RNA interference reveals that oncogenic *KRAS*-driven cancers require TBK1. *Nature*, 462(7269):108–112, November 2009.

[5] Francesco Beghini, Audrey Renson, Christine P. Zolnik, Ludwig Geistlinger, Mykhaylo Usyk, Thomas U. Moody, Lorna Thorpe, Jennifer B. Dowd, Robert Burk, Nicola Segata, Heidi E. Jones, and Levi Waldron. Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. *Annals of Epidemiology*, 34:18–25.e3, June 2019.

[6] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[8] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[9] Matteo Calgaro. Mcalgaro93/sc2meta: Paper Release. Zenodo, July 2020.

[10] Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology*, 21(1):191, August 2020.

[11] 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.

[12] Marne C Cario. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. page 19, 1997.

[13] Jun Chen, Emily King, Rebecca Deek, Zhi Wei, Yue Yu, Diane Grill, and Karla Ballman. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4):643–651, February 2018.

[14] Ilseung Cho and Martin J. Blaser. The human microbiome: At the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, April 2012.

[15] Jasmine Chong, Peng Liu, Guangyan Zhou, and Jianguo Xia. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature Protocols*, 15(3):799–821, March 2020.

[16] The Human Microbiome Project Consortium, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G.

FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I.-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, YuanQing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, and Owen White. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[17] Marie Laure Delignette-Muller and Christophe Dutang. Fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

[18] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, September 1988.

[19] Mei Dong, Longhai Li, Man Chen, Anthony Kusalik, and Wei Xu. Predictive analysis methods for human microbiome data with application to Parkinson's disease. *PLOS ONE*, 15(8):e0237779, August 2020.

[20] Bradley Efron. Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, 99(465):96–104, March 2004.

[21] J. J. Egozcue and V. Pawlowsky-Glahn. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37(7):795–828, October 2005.

[22] 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.

[23] Hildreth Robert Frost. Variance-adjusted Mahalanobis (VAM): A fast and accurate method for cell-specific gene set scoring. *Nucleic Acids Research*, 48(16):e94–e94, September 2020.

[24] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe*, 15(3):382–392, March 2014.

[25] 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.

[26] Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, 23(8):980–987, April 2007.

[27] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein, and Rob Knight. Qiita: Rapid, web-enabled microbiome meta-analysis. *Nature Methods*, 15(10):796–798, October 2018.

[28] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1):7, January 2013.

[29] Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics*, 20(1):210–221, January 2019.

[30] Abhishek Kaul, Ori Davidov, and Shyamal D. Peddada. Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, 18(3):422–433, July 2017.

[31] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, 8, 2017.

[32] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology*, 8(2):e1002375, February 2012.

[33] Max Kuhn and Hadley Wickham. *Tidymodels: Easily Install and Load the 'tidymodels' Packages*. 2020.

[34] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015.

[35] Hongzhe Li. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.

[36] Hongzhe Li. Statistical and Computational Methods in Microbiome and Metagenomics. In *Handbook of Statistical Genomics*, chapter 35, pages 977–550. John Wiley & Sons, Ltd, 2019.

[37] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.

[38] Bryan D. Martin, Daniela Witten, and Amy D. Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1):94–115, March 2020.

[39] Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, 10(4):e1003531, April 2014.

[40] James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, and Rob Knight. Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2(1), February 2017.

[41] Iftekhar Naim and Daniel Gildea. Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients. page 8.

[42] H. Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, Laurent Gautier, Anders G. Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B. Quintanilha Dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S. Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S. Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W. Ussery, Takuji Yamada, MetaHIT Consortium, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, S. Dusko Ehrlich, and MetaHIT Consortium. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, August 2014.

[43] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B. Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata, and Levi Waldron. Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11):1023–1024, November 2017.

[44] Lita M. Proctor, Heather H. Creasy, Jennifer M. Fettweis, Jason Lloyd-Price, Anup Mahurkar, Wenyu Zhou, Gregory A. Buck, Michael P. Snyder, Jerome F. Strauss, George M. Weinstock, Owen White, Curtis Huttenhower, and The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature*, 569(7758):641–648, May 2019.

[45] 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.

[46] Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(giz107), September 2019.

[47] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: An outlook and review. *Bioinformatics*, 34(16):2870–2878, August 2018.

[48] Kris Sankaran and Susan Holmes. structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data. *Journal of statistical software*, 59(13):1–21, 2014.

[49] Lucas Schiffer, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B Dowd, Nicola Segata, and Levi Waldron. HMP16SData: Efficient access to the human microbiome project through bioconductor. *American Journal of Epidemiology*, 2019.

[50] Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, June 2016.

[51] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, February 2017.

[52] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[53] T. Thurnheer, N. Bostanci, and G. N. Belibasakis. Microbial dynamics during conversion from supragingival to subgingival biofilms in an in vitro model. *Molecular Oral Microbiology*, 31(2):125–135, April 2016.

[54] Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, September 2005.

[55] 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.

[56] Alex D. Washburne, Justin D. Silverman, Jonathan W. Leff, Dominic J. Bennett, John L. Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A. David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969, February 2017.

[57] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), December 2017.

[58] 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.

[59] Jian Xiao, Li Chen, Yue Yu, Xianyang Zhang, and Jun Chen. A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data. *Frontiers in Microbiology*, 9, December 2018.