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

Research in human associated microbiomes primarily involve high-throughput sequencing of samples, resulting in count tables of taxa abundances. Unfortunately, the analysis microbiome data is complex as it is high-dimensional, sparse, and strictly compositional. Various statistical tools have been developed to address these issues, but a very approachable method is to aggregate taxa into pre-defined sets, which is called gene set testing or enrichment analysis in the genomics literature. However, there has been a considerable lack of interest in developing set-based analysis methods tailored specifically for microbiome relative abundance data. Here, we present a new sample-level taxon set enrichment method based on the isometric log ratio transformation and the competitive null hypothesis in the gene set testing literature. Our approach, titled competitive isometric log ratio (cILR), generates enrichment scores per sample as the scaled log ratio of the subcomposition defined by taxa within a set and the subcomposition defined by its complement. We also provide sample-level significance testing by estimating an empirical null distribution of our test statistic with valid p-values. Herein we demonstrate under both real data applications and simulations that cILR controls for type I error even under high sparsity and high inter-taxa correlation scenarios, while additionally providing informative scores that can be inputs to downstream differential abundance and predictive tasks with good performance. These results demonstrate how our approach can be used to generate compositionally valid taxonomic aggregation and enrichment.

Background

The microbiome is the collection of microorganisms (bacteria, protozoa, archaea, fungi, and viruses) which co-exist 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 [1], type II diabetes [2], and obesity [3]. 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) [4]. Processing raw sequencing data using a variety of bioinformatic pipelines [5, 6] 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 [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 utilized 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 applied inconsistently across samples, resulting in a different 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 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 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 [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 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 [23].

Set-based analyses are ubiquitous in the gene expression literature [24] and various methods have been developed for enrichment testing and scoring. More traditional gene-set testing methods utilize the hypergeometric test to test for the overrepresentation of significant p-values for a set of interest. 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. More recent methods, such as GSVA [26] and VAM [27], generate sample level enrichment scores more akin to a transformation. Generated scores can then be used in data visualization, as well as flexible scores that can be utilized in downstream analyses.

Here, we developed a new sample-level set-based analysis method specific to microbiome relative abundance data. We leverage the conception of the Q_1 competitive hypothesis presented in Tian et al. [28], which compares the expression of genes within the gene set against the rest of the genes. 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 [29], 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 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 this manuscript, we present cILR, provide its formulation and discuss some statistical properties. Then, 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 using both real data and simulation studies. 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 [26] and ssGSEA [30]. 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 [29]. 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 [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 taxa p to m sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [31], 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 (Figure 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 [32]. For the mixture normal distribution, parameters were estimated using an expectation-maximization algorithm implemented in the *mixtools* package [33].
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. Valid 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) [29]. 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 $\mathbb{R}^{D-1}$ [29, 34]. 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 [23]. 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 [29].

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 [34, 35, 36]. 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 [35, 36, 34], 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 [28] 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 [23]. 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 [37] 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 B 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 [37].

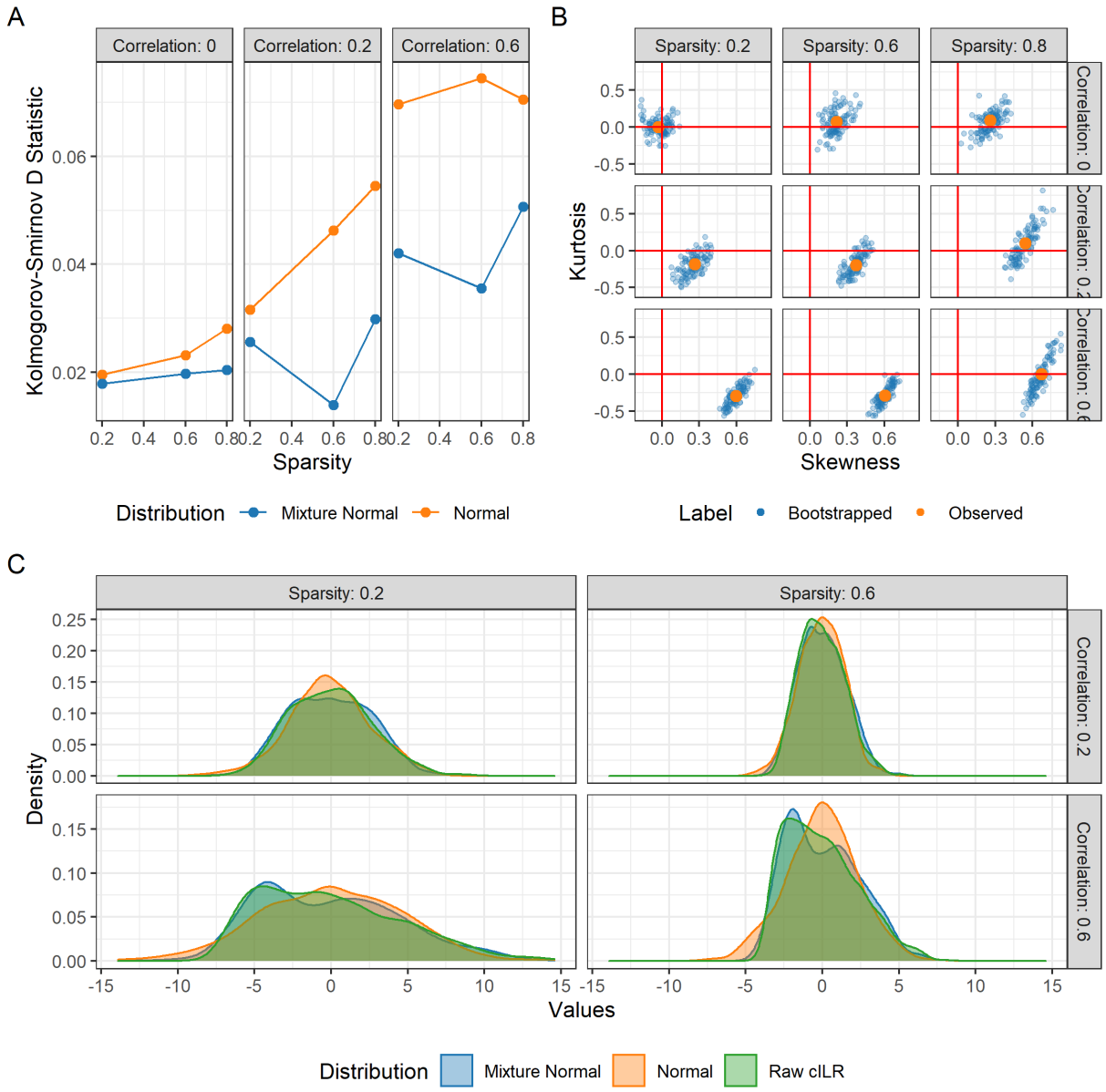

Figure 1. Properties of the null distribution of cILR in different simulation scenarios under the global null. Panel **(B)** presents kurtosis and skewness of cILR scores while panel **(A)** presents the goodness of fit (as Kolmogorov-Smirnov D statistic) for mixture normal and normal distributions.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

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 A 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 [38]. 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

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 [39]. 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 [40]. 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 [32]) 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 [41] 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. [27] 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 [42] 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. [43] 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 [44]. 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* [45] 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 [46]. 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. [47] 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 across both 16S rRNA gene sequencing and whole-genome sequencing. All data sets are obtained from either the *curatedMetagenomicData* [48] and *HMP16SData* [49] R packages (2020-10-02 snapshot), or downloaded from the Qiita platform [50].

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 [51, 1]. We utilized this data set following the approach outlined in Calagaro et al. [39]. 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 [52]. 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. [53] and was downloaded directly from the Github repository associated with Calagaro et al. [54].

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 [51, 1]. 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 [55]. 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. [56] 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 North America. 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 zero entry 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 [26], ssGSEA [30], 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. We added a pseudocount of 1 to all zero entry values. For comparison, we chose representative state-of-the-art methods in differential abundance analysis, namely DESeq2 [15, 14] and corncob [57].

Disease Prediction: We fit random forest on cILR scores, as well as ssGSEA [26] and GSVA [30] similar to single sample enrichment section. We added a pseudocount of 1 to all zero entry values. 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.

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.

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.

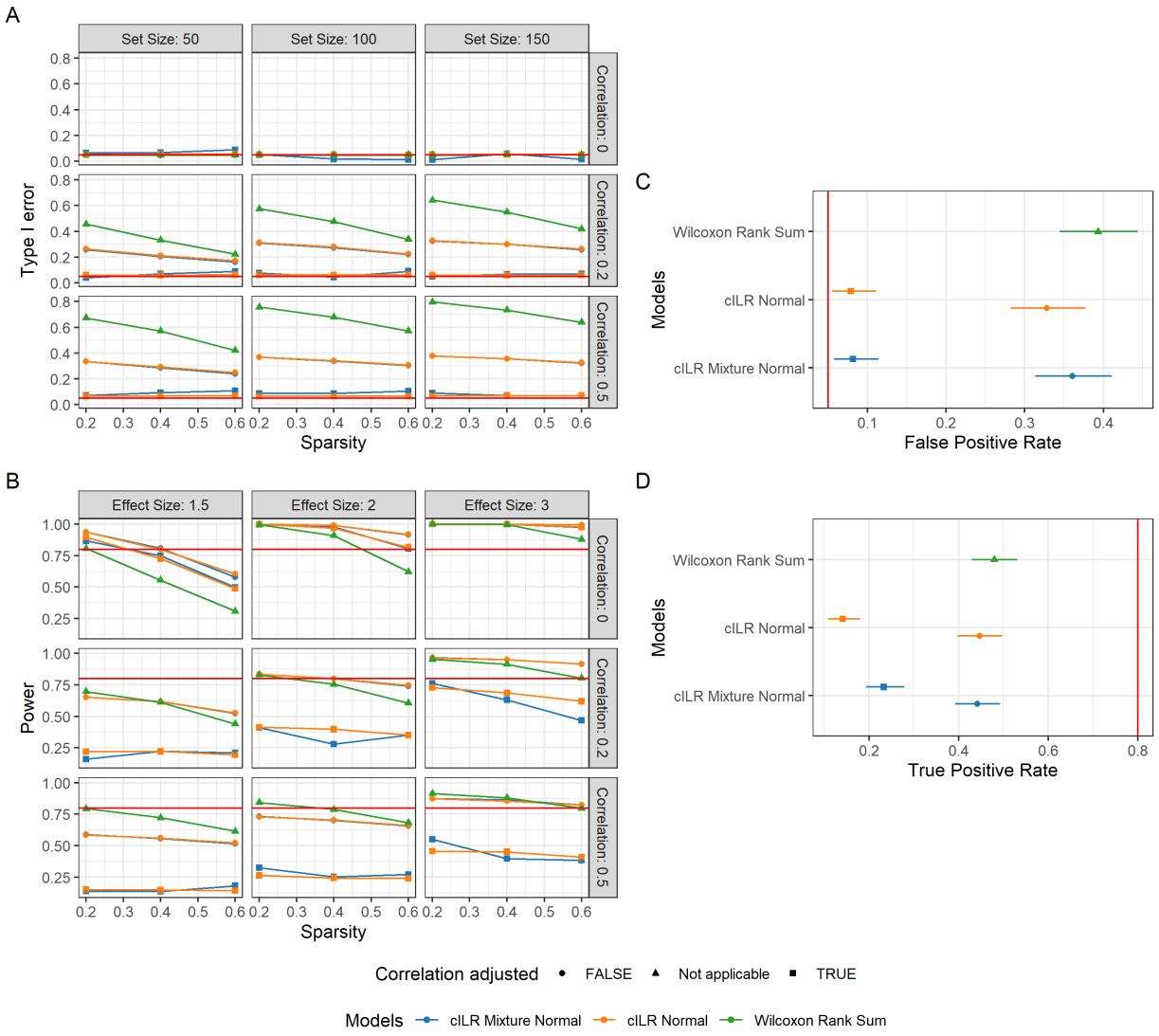

Figure 2. Type I error rate **(A)**, and power **(B)** across different parametric simulation scenarios. Confidence bounds were obtained using Agresti-Couli [41] approach. False-positive rate **(C)** and true positive rate **(D)** evaluation of similar methods on real 16S rRNA data from the oral microbiome 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.

However, the trade-off for good type I error control is demonstrably lower power, as shown in figure 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 mus much more dramatic. At the highest effect size (fold change of 3) and correlation ($\rho = 0.5$), adjusted cILR was only perfomring at 50% power, while unadjusted cILR and wilcoxon rank sum test were able to reach 80%. These results indicated that both sparsity and inter-taxa

correlation impacts power, with correlation having a much more dramatic impact especially for adjusted versions of cILR.

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 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 [26] and ssGSEA [30]), 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.

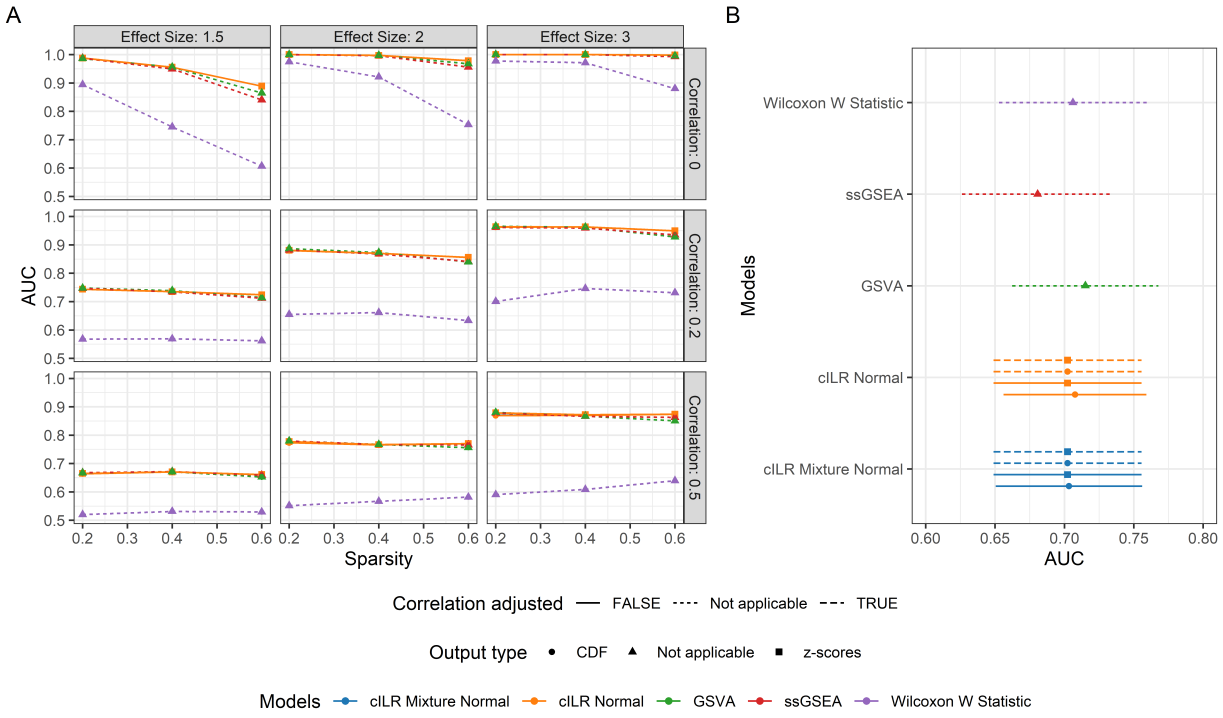

Figure 3. Classification performance via AUC of cILR, ssGSEA, GSVA, and Wilcoxon U statistic on simulated data **(A)** the gingival data set from the Human Microbiome Project **(B)** as detailed in Materials and Methods. Performance scores measure whether scores can highly rank samples that are known 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 [42]

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 [15] and corncob [57].

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 [44]) using cILR scores compared to scores generated by other single sample enrichment methods in the gene set testing literature (ssGSEA [30] and GSVA [26]). 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 Materials and 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.

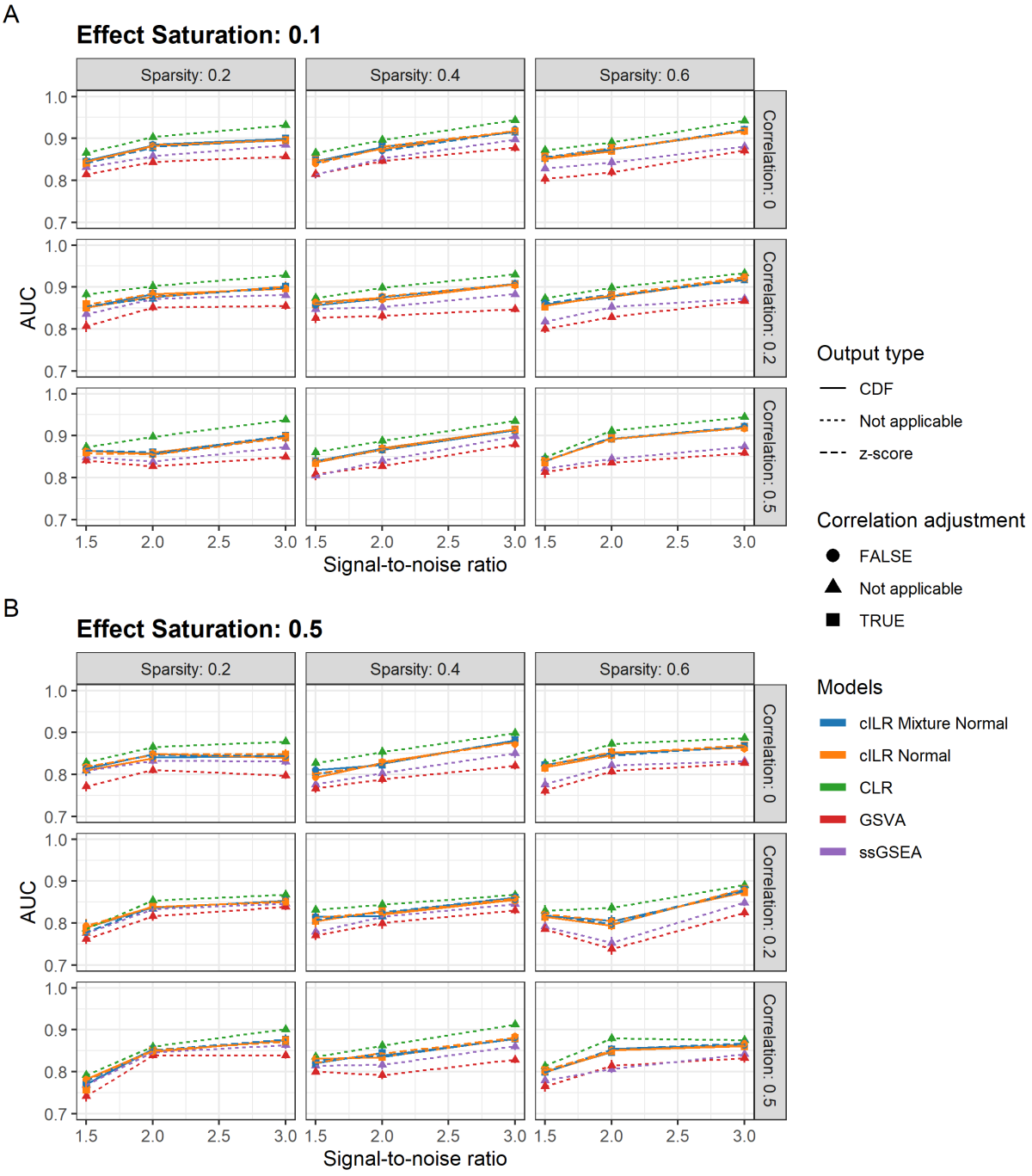

Figure 4. AUC scores of a random forest model for a binary outcome trained on cILR, ssGSEA, GSVA generated scores as well as on standard CLR transformed data evaluated on simulated data across sparsity levels, correlation, and signal-to-noise ratio. Panel **(A)** and **(B)** represent results across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

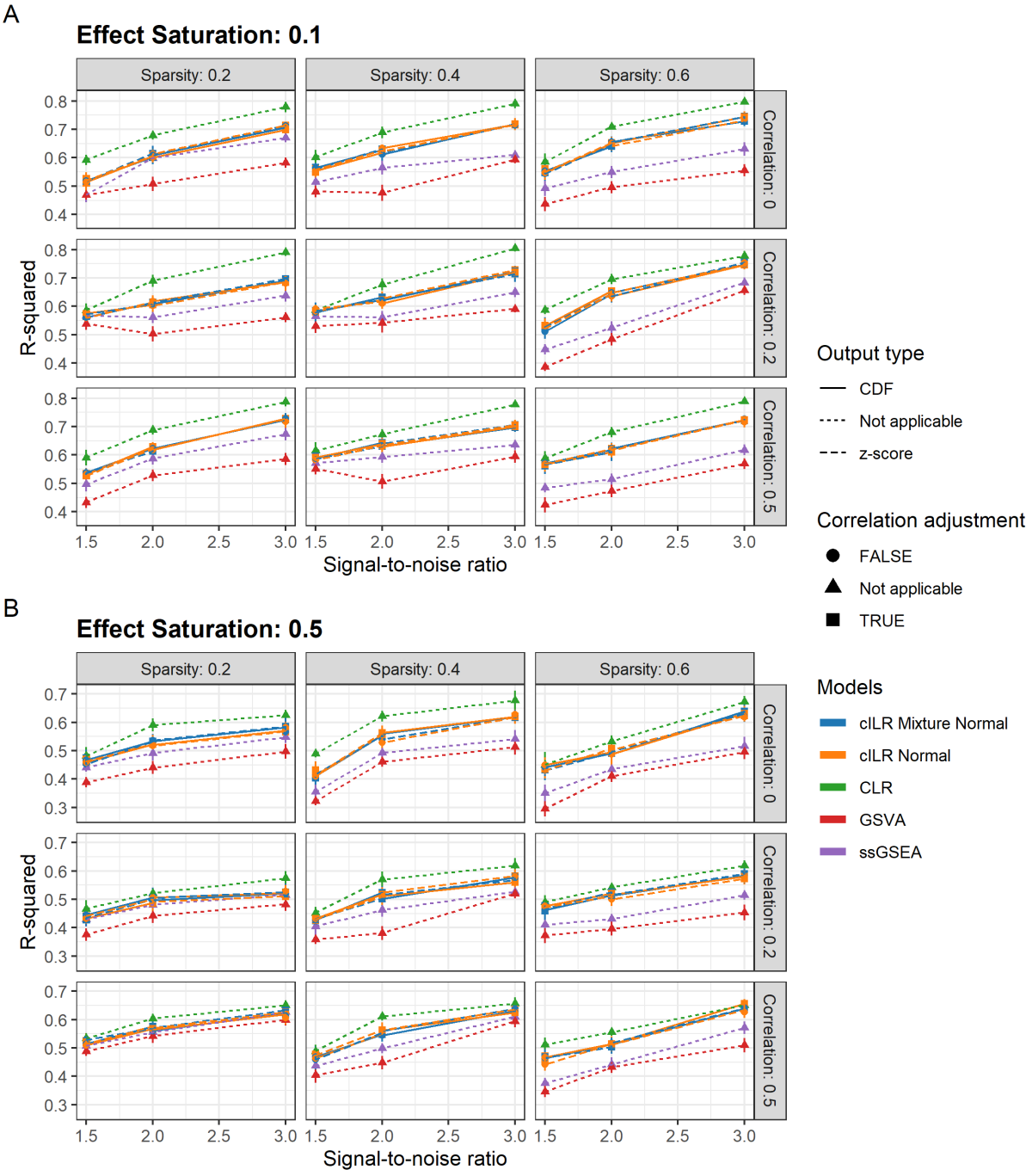

Figure 5. Predictive R-squared of a random forest model for a continuous outcome trained on cILR, ssGSEA, GSVA generated scores as well as on standard CLR transformed data evaluated on simulated data across sparsity levels, correlation, and signal-to-noise ratio. Panel **(A)** and **(B)** represent results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

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. [56] data set is a 16S rRNA sequencing data set while the Nielsen et al. [55] 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.

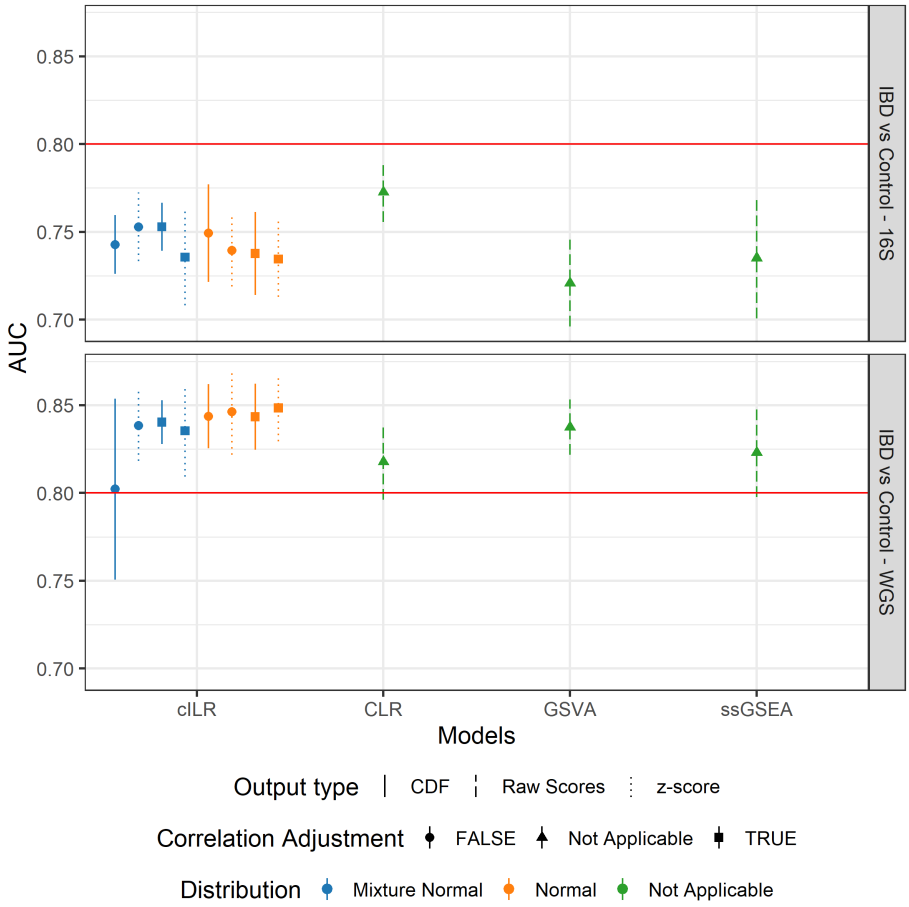

Figure 6. Classification performance of a standard random forest model using cILR scores compared against existing methods in gene set testing literature and the standard centered-log ratio transformation approach. The learning task involves predicting patients with inflammatory bowel disease (including Crohn's disease and ulcerative colitis) versus controls. Data sets used span both 16S rRNA sequencing (Gevers et al. [56]) and whole-genome shotgun sequencing (Nielsen et al. [55]). cILR marginally outperforms GSVA and ssGSEA in both data sets, but only performs better than CLR in the WGS data set.

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.

Discussion

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 [58]. 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 [25], and Gene Ontology [59]), 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 [23], 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. Some software suites, such as *MicrobiomeAnalyst*, have begun to provide an option to perform such analyses with curated taxa sets [60]. However, the approach used in *MicrobiomeAnalyst* is based on a hypergeometric test of significant hypotheses and is sensitive to the method of differential abundance used. More recent methods in set-based testing generate enrichment scores at the sample level [26, 27]. This approach allows for the flexible incorporation of different statistical techniques downstream, but unfortunately is not currently tailored to the compositional nature of microbiome relative abundance data [18].

Here, we developed cILR, a novel microbiome-specific method to perform set-based enrichment analysis at the sample level. Our method is based on the isometric log-ratio transformation [29], and measures set enrichment relative to those outside the set, which corresponds to a Q_1 null hypothesis in the gene set analysis literature [28]. 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.

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 [38], 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 [61]. 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 [51]. 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 [11]) 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.

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 [44] 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 [26] and ssGSEA [30]). This is consistent with 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 an inflated set. Results indicated that cILR scores are more adaptive to microbiome data sets than GSVA and ssGSEA. However, cILR scores underperform the standard approach of applying the CLR transformation to count aggregated data, except for the WGS real data analysis. Despite this, there are still benefits to using cILR scores in the context of downstream modeling beyond random forest predictive models. First, CLR transformed variables have a sum to zero constraints, making the covariance matrix is singular [18], impacting methods that rely on matrix decomposition such as principal component regression or canonical correlation analysis. Second, cILR provides an additional value of significance testing for enrichment for sample screening. Third, CLR transformed variables still involve the count summation procedure for variable aggregation, therefore comparison between sets of different sizes remains unviable.

Despite such drawbacks, cILR methods still produce competitive performance values across both regression and classification tasks even in low signal-to-noise settings. Most interestingly, the performance gap between cILR and CLR was smaller in high sparsity settings ($p = 0.6$), which is a feature of microbiome relative abundance data [7]. Furthermore, these predictive models consider the effect of variables jointly (and in the case of random forest, consider interactions as well), and 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 similar to significance testing experiments. 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).

In short, users can apply cILR generated enrichment scores as inputs to downstream analyses, especially predictive models and differential abundance analysis. cILR scores are informative in discriminating samples between classes even in low effect size situations. Even though cILR did not outperform a standard analysis of count aggregation followed by the CLR transformation, the additional benefits listed above would prove to be attractive in certain use cases, especially in high sparsity situations.

Limitations and future directions

There are various limitations to using cILR to generate sample-level taxonomic aggregation for microbiome data. First, significance testing at the sample level is not powered to assess inflated sets with low effect size. Practitioners can be confident in the validity of a significant result (following proper multiple testing correction) due to cILR's type I error control, but should not expect cILR to be able to identify all samples with inflated counts. Second, even though cILR is flexible and addresses the compositional nature of microbiome data, the performance of cILR might not be consistently better than competing methods. Careful evaluation of methods should be done before application to ensure that the value of cILR scores is appropriately extracted in suitable models. Third, cILR did not directly address the zero-inflatedness of microbiome data and utilizes a pseudocount to ensure log operations are valid. As such, practitioners are encouraged to choose a different pseudocount or an appropriate method of choice.

As such, we hope to address the aforementioned issues with cILR in later projects. First, we hope to refine the distribution fitting procedure for the mixture distribution due to its superior fit and address the identifiability problem that might underlie the lack of power in our experiments. Second, we hope to incorporate some of the sophisticated model-based zero correction methods [62, 22]. Third, we can account for the difference in importance of certain taxa by incorporating taxa-specific weights. Finally, we also hope to curate more interest apriori sets that are functionally informative, which would generate more interesting insight and improve interpretability compared to using taxonomic categories such as Phylum or Genus.

Conclusion

Gene set testing, or pathway analysis is an important tool in the analysis of high-dimensional genomic data sets. However, there has not been a lot of set-based analysis methods developed specifically for microbiome relative abundance data. In this manuscript, 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.

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] 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.

[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] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

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

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Calgaro M. Mcalgaro93/Sc2meta: Paper Release; 2020. Zenodo.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.