CBEA: Competitive balances for taxonomic enrichment analysis

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

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

‡These authors jointly supervised this work.
†Corresponding author

* hildreth.r.frost@dartmouth.edu

Abstract

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

Author summary

The study of human associated microbiomes relies on genomic surveys via high-throughput sequencing. However, microbiome taxonomic data is sparse and high dimensional which prevents the application of standard statistical techniques. One approach to address this problem is to perform analyses at the level of taxon sets. Set-based analysis has a long history in the genomics literature, with demonstrable

impact in improving both power and interpretability. Unfortunately, ~~there is not a lot of research in~~ ~~developing~~ new set-based tools for microbiome taxonomic data specifically, given ~~that~~ its unique features compared to other 'omics data types~~microbiome taxonomic data is strictly compositional~~. We developed a new tool to generate taxon set enrichment scores at the sample level ~~by combining the isometric~~ through a novel log-ratio ~~and~~ formulation based on the competitive null hypothesis. Our scores can be used for statistical inference ~~, and as~~ at both the sample and population levels, as well as inputs to other downstream analyses such as ~~differential abundance and~~ prediction models. We demonstrate the performance of our method against competing approaches across both real data analyses and simulation studies.

Introduction

The microbiome is the collection of microorganisms (bacteria, protozoa, archaea, fungi, and viruses) which co-exist with their host. Previous research has shown that changes in the composition of the human gut microbiome are associated with important health outcomes such as inflammatory bowel disease [1], type II diabetes ~~[2]~~ [3], and obesity [4]. To understand the central role of the microbiome in human health, researchers have relied on high-throughput sequencing methods, either by targeting a specific representative gene (i.e. amplicon sequencing) or by profiling all the genomic content of the sample (i.e. whole-genome shotgun sequencing) [5]. Raw sequencing data is then processed through a variety of bioinformatic pipelines [6,7], yielding various data products, one of which are taxonomic tables which can be used to study associations between members of the microbiome and an exposure or outcome of interest.

However, there ~~exists~~ are unique challenges in the analysis of these taxonomic count tables ~~[8–10]. 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 [11], allowing for simultaneous model fitting and variable selection. For differential abundance tasks, researchers often utilize multiple hypothesis correction methods [?,?] or omnibus tests [?] to address hypothesis testing burden.~~

~~Second, the number of reads obtained is constrained by the sequencing instrument at an arbitrary limit, and is inconsistent across samples, resulting in a variable number of total read counts per sample. Many normalization methods [27] have been proposed to address these issues, including cross-applying methods from the gene expression literature [60]. 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 [26], which might not be not true in the context of microbiome relative abundance data [?,23]. As such, microbiome taxonomic data is strictly compositional [12], 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 [13].~~

~~Third, the data are highly zero-inflated, where there is a high number of both structural zeros (truly missing due to biological reasons) and sampling zeroes (due to limits of detection of the sequencing experiment). Researchers often dealt with these issues by imputing zero cells with a pseudocount [?], or applying zero-inflated models [?,?]. 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 [81].~~

[9, 10]. The data is sparse, high-dimensional, and likely compositional [9, 10, 12]. Even though ~~the aforementioned~~ these problems are challenging, a very approachable ~~method to address some of them is through~~ solution is to ~~to~~ use set-based analysis methods, also termed gene set testing in the genomics literature [14, 15]. Aggregated ~~sets are less sparsethan their constituent elements~~variables can be less sparse, and testing on a smaller number of ~~variables reduces the multiple testing burden, thereby increasing~~ features can reduce the multiple-testing burden. As such, gene set testing methods have been shown to increase power and reproducibility ~~. Through~~ of genomic analyses. Furthermore, through the usage of functionally informative sets defined apriori based on historical experiments (for example MSigDB [16], and Gene Ontology [17]), gene set ~~analyses~~ analysis also allows for more biologically informative interpretations.

~~There exists~~ A diverse set of ~~available methods developed to perform such analyses. More traditional set testing~~ methods already developed in this field. Traditional methods utilize the hypergeometric ~~test~~ distribution to test for the overrepresentation of ~~significant p-values for a set of interest~~ a gene set using a candidate list of genes screened based on a marginal model [15]. Unfortunately, these approaches are sensitive to the differential expression test ~~and their generated p-values. The most widely used gene-set analysis method, GSEA [16], 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~~ as well as the chosen threshold when trying to select genes for the candidate list. Aggregate score methods, which are generally ~~more~~ preferred [18], instead assigns a score for each gene set based on gene-specific statistics such as z-scores or fold change. Of these approaches, methods such as GSEA [16] performs a test for each gene set at the population level, ~~incorporating information from~~ summarizing information across all samples. Conversely, methods such as GSVA [19] and VAM ~~[20]~~ [21], generate enrichment scores at the sample level and are more akin to a transformation. ~~This strategy~~ In addition to being able to screen for enriched sets per sample, this strategy also allows for the flexible incorporation of different ~~statistical techniques downstream~~ downstream analyses, such as fitting prediction models, ~~as well as for visualization purposes in ordination plots~~or performing dimension reduction.

In microbiome research, even though no explicit enrichment analysis is performed, ~~standard practice often involves aggregating~~ researchers often aggregate taxa to higher Linnean classification levels such as genus, family, or phylum~~by simple summation of abundances [30]. Even though this allows for a reduction in the number of overall taxa (from thousands to only hundreds), there still exist three disadvantages: first, inter-sample distances are not preserved before and after aggregation [31], second, it doesn't allow for enrichment testing and comparison across sets of different sizes, and third, it increases bias when taxa within the set have different efficiencies in how they are measured through sequencing [30]. As such, there is a need for microbiome researchers to adopt more robust sample-level set enrichment methods. Unfortunately, limited work has been~~ . However, ~~despite this interest~~, there is limited research done to extend existing ~~methods to be more specific to~~ set-based methods to microbiome relative abundance data. Some software suites, such as *MicrobiomeAnalyst*, do offer tools to perform enrichment testing with curated taxon sets [32]. However, the approach used in *MicrobiomeAnalyst* is a form of overrepresentation analysis at the population level and therefore similarly sensitive to the differential abundance approach used ~~.~~ and p-value threshold. One of the primary challenges for adapting gene set analysis to the microbiome context is the compositional nature of the data. Sequencing technologies constrain the total number of reads, and samples are expected to have the same number of reads instead of DNA content [23, 24]. However, different samples still

yield arbitrarily different total read counts [12,25], suggesting the use of normalization methods to allow for proper comparison of feature abundances across samples. However, microbiome data sets do not follow certain assumptions that enable the cross-application of methods from similar fields (such as RNA-seq) [23,24]. For example, DESeq2's *estimateSizeFactors* [26] assumes that the majority of genes acts as housekeeping genes with constant expression levels across samples. As such, practitioners often rely on total sum normalization to transform count data into relative proportions that sum to one [27]. Some studies have provided emprical performance vealuations supporting this normalization schema [28]. Since this approach imposes a sum constraint on the data, post normalization microbiome data sets are compositional [12], which means that the abundance of any taxa can only be interpreted relative to another. Under this scenario, log-ratio based approaches from the compositional data analysis (CoDA) literature [29] are motivated to address this issue.

Unfortunately, the standard practice for aggregating variables using element-wise summations (referred to as amalgamations in the CoDA literature), does not adequately address the compositional issue [30]. First, inter-sample Aitchison distances computed on original parts are not preserved after amalgamation [31]. This means that cluster analyses might show different results depending on the level of amalgamation and differs from the those computed from original variables. Second, amalgamations do not allow for comparison between sets of different sizes within the same experimental condition since larger sets will have larger means and variances. Third, considering that each taxa has specific measurement biases [30], an amalgamation based approach would make the bias of the amalgamated variable dependent on the relative abundance of the its constituents. In other words, if taxon 1 has abundance A_1 and bias B_1, while taxon 2 has abundance A_2 and bias B_2, then the bias of the aggregate variable (for example, a genera) is $(A_1B_1 + A_2B_2)/(A_1 + A_2)$ (see Appendix 1. from McLaren et al. [30]). This means that bias invariant approaches (such as analyses of ratios) would no longer be invariant when applied to amalgamated variables as bias now varies across samples. The alternative would be to multiply the proportions rather than to sum them [31].

Here, we present a ~~novel method that~~ taxon-set testing method for microbiome relative abundance data that addresses the aforementioned issues. Our approach generates enrichment scores at the sample level similar to GSVA [19] and VAM ~~[20]~~ [21]. We leverage the concept of the Q_1 competitive hypothesis presented in Tian et al. [33] ~~, which tests the null that the value of variables within a specific set is equal to the value of measured variables not in the set . The competitive null hypothesis is particularly useful in compositional data analysis, as it naturally assesses enrichment as a ratio between two sets of variables. We incorporated this insight with the isometric log-ratio transformation [39], which allows for a multiplicative aggregation method that addresses the downsides of the naive summation-based method presented above [30,44]. The resulting method, titled competitive isometric log-ratio (cILR), is therefore unsupervised and can generate sample-specific enrichment scores with a~~ to formulate the enrichment of a set as the compositional balance [35] of taxa within the set and remainder taxa using multiplication as the method of aggregating proportions [31]. This well-defined null hypothesis ~~that allows for significance testing . These scores can then act as inputs to differential abundance and predictive modeling tasks downstream~~allows us to perform significance testing with interpretable results through estimating the empirical distribution of our statistic under the null that can also account for variance inflation due to inter-taxa correlation [48].

In the following sections, we ~~provide the formulation of cILR and discuss some~~ present

our approach titled competitive balances for taxonomic erichment analysis (CBEA). First, we presented the step-by-step formulation of CBEA and discuss its statistical properties. ~~We illustrate significance testing at the sample level using cILR and evaluate type I error and power under different simulation scenarios and real data applications. We assess the informativeness of cILR generated scores, and evaluate how it performs as part of downstream analyses, specifically predictive models and differential abundance analysis. We compare the performance of cILR in these respective tasks againststandard microbiome taxonomic data analysis practices,~~ Second, we detailed our evaluation strategy using both real data and parametric simulations, and the methods we're comparing against. Third, we presented results on enrichment testing using CBEA for single samples as well as ~~the existing GSVA [19] and ssGSEA [76], which are single sample methods.~~ at the population level. Fourth, we showed the performance of CBEA in downstream disease prediction. Finally, we discussed our results and the limitations of our method. An R package implementation of ~~this approach~~ CBEA can be found on ~~Github (~~GitHub (qpmnguyen/~~tear~~CBEA).

Materials and Methods

Competitive ~~Isometric Log-ratio~~ balances for taxonomic enrichment analysis (~~cILR~~CBEA)

The ~~cILR~~ CBEA method generates sample-specific enrichment scores for microbial sets using ~~the isometric log-ratio transformation~~ products of proportions [39]. Details on the computational implementation of ~~cILR~~ CBEA can be found in the ~~supplemental materials. The cILR~~ Supplementary Materials. The CBEA method takes two inputs:

- **X**: n by p matrix of positive ~~counts~~ proportions for p taxa and n samples measured through either targeted sequencing (such as of the 16S rRNA gene) or whole genome shotgun sequencing. Usually **X** is generated from standard taxonomic profiling pipelines such as DADA2 [6] for 16S rRNA sequencing, or MetaPhlAn2 [7] for whole genome shotgun sequencing. CBEA does not accept X matrices with zeroes since it invalidates the log-ratio transformation. Users can generate a dense matrix X using a method of choice, however the default mode for CBEA ~~would be~~ add a pseudocount of 10^{-5} if zeroes are detected in the matrix.
- **A**: p by m indicator matrix annotating the membership of each taxon p to m sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [40], 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~~ CBEA 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~~ CBEA statistic**: Let **M** be a n by m matrix of ~~cILR~~

CBEA scores. Let $\mathbf{M}_{i,k}$ be ~~cILR~~ CBEA scores for set k ~~of~~ sample i:

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

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~~ CBEA under the null vary depending on data characteristics (Fig 1), an empirical null distribution will be estimated from data.

   - **Compute the ~~cILR~~ CBEA 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~~ CBEA scores generated from $\mathbf{X}_{perm}$. Similarly, we have $\mathbf{M}_{unperm}$ be ~~cILR~~ CBEA 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 [41]. For the mixture normal distribution, parameters were estimated using an expectation-maximization algorithm implemented in the *mixtools* package [42].

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

CBEA balances groups of parts?

Properties of ~~cILR~~CBEA

~~cILR~~ CBEA and ~~the Isometric Log Ratio Transformation~~balances of groups of parts

The ~~cILR statistic is a special instance of the isometric log-ratio transformation~~ CBEA statistic is based on the multiplication-based aggregation approach used to calculate balances between groups of parts [31]. These balances are computed using the isometric log ratio (ILR) ~~[39]. 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 $\mathbb{R}^{D-1}$ [39,43]. 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~~ transformation [39] formula. For a given balance i splitting variables ~~between~~ across sets R and ~~and~~ S, we have the ~~ILR~~ balance coordinate x_i^* as:

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

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 [31]. 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 [39].~~

~~The usage of the ILR statistic is not uncommon in the microbiome literature. They are usually termed "compositional balances ", and have been leveraged in many recent approaches in variable transformation [43–45]. The cILR formulation in Eq is a special case of Eq 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~~CBEA belongs to a set of methods that seeks to leverage compositional balances for the analysis of microbiome data [35, 43, 45]. Unlike methods such as PhILR [44], CBEA does not provide a sequential binary partition that forms the basis for ILR procedure [39] and is therefore not a subclass of ILR. A similar method to CBEA would be phylofactor [43]. However, ~~unlike the ILR and its variants [43–45], the axes defined by each cILR set are not orthogonal (since the balances are mutually exclusive between setsand 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~~ instead of performing an optimization procedure to identify interesting balances, CBEA constructs balances apriori using pre-defined sets, and formulates the enrichment of a set as the scaled log-ratio between the center of the subcomposition ~~defined by~~ represented by microbes within the set and the ~~set is equal to the~~ center of the subcomposition ~~defined by the complement of the set. This is equivalent to~~ represented by remainder taxa. This formulation aligns with the Q_1 ~~competitive null hypothesis in~~ null hypothesis from the gene set testing literature [33] ~~where the enrichment of a gene set is defined with respect to genes outside the set~~.

~~We can apply prior usage of the ILR statisticin hypothesis testing to cILR by assuming that the null distribution of cILR follows a standard normal distribution [31].~~

Estimating the null distribution

We can assume that the CBEA statistic, similar to other log-ratio based transforms, follows a normal distribution [39, 46]. However, when applying ~~cILR~~ CBEA 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 [47] showed that estimating the null distribution of the test statistic directly (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~~CBEA, we also estimated the null distribution from observed raw ~~cILR~~ CBEA variables.

This assumption is also supported by preliminary simulation studies (detailed below). ~~In panel A of Fig 1,~~ we simulated microbiome taxonomic count data under the global null across different data features and compute raw ~~cILR~~ CBEA 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~~ as suggested by Efron [47].

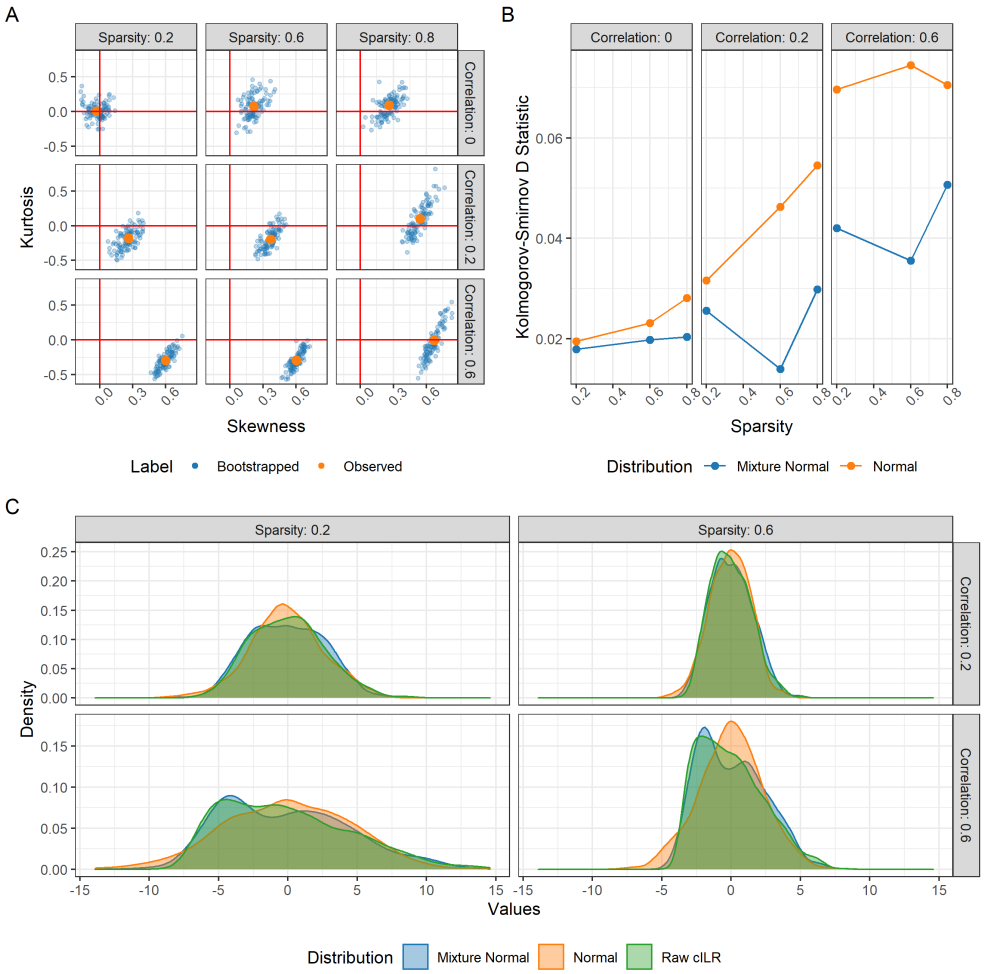


Fig 1. Properties of the null distribution of ~~cILR~~ CBEA under the global null simulations. Panel **(B)** presents kurtosis and skewness of ~~cILR~~ CBEA scores while panel **(A)** presents the goodness of fit (as Kolmogorov-Smirnov D statistic) for mixture normal and normal distributions. Panel **(C)** 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~~two-component normal mixture distribution. ~~Panel B of Fig 1 demonstrates~~ the goodness of fit of the mixture normal and the normal distribution using Kolmogorov-Smirnov (KS) test statistic computed on fitted normal and mixture normal distribution when fitted on

~~cILR~~ CBEA 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~~ CBEA 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 [48]. 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~~

Variance inflation due to inter-taxa correlation

When taxa within a set are highly correlated, the variance ~~estimate from unpermuted data, where the inter-taxa correlation structure remains undisturbed. However, this procedure assumes that the variance of cILR is equal under both the null and alternate hypotheses.~~

~~Evaluation~~

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

~~Parametric Simulations~~

~~To address the performance of cILR for different modeling tasks, we simulated microbiome count data under the assumption that it follows a zero-inflated negative binomial distribution, which is a good fit for real microbiome relative abundance data [78]. Suppose X_{ij} are observed counts for a sample~~ of the sample mean of taxon-wise statistics is inflated. Without loss of generalizability, for a set of taxa with taxa-specific statistics $x_1, ..., x_p$ we have the variance of the mean $\bar{x}$ to be:

$$Var(\bar{x}) = \frac{1}{m^2}\left(\sum_{i=1}(\sigma_i^2) + \sum_{i<j}\rho_{ij}\sigma_i\sigma_j\right) \tag{3}$$

where σ_i is the standard deviation of taxon i ~~and taxon j, then we have the following probability model~~

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

~~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 [?]. Given an n by p matrix of values $\mathbf{U}$ sampled from multivariate normal distribution with correlation matrix ρ~~while ρ_{ij} is the correlation between i and j. The second term of (3) is the correlation dependent variance component, which goes to 0 if there is no correlation. The CBEA statistic follows a similar pattern. Since the geometric mean of a set of variables is equivalent to the exponential of the arithmetic mean of their logarithms, we can ~~generate target~~

~~microbiome count vector $\mathbf{X}_{.j}$ for taxa j following the marginal distribution **NB** characterized by the negative binomial cumulative distribution function $\mathbb{F}_{\mathbf{NB}}$:~~ re-write CBEA score for a set k with size K as follows:

$$\mathbf{X}_{.j} M_{i,k} = {\mathbb{F}_{\mathbf{NB}}}^{-1}(\Phi_{U_i}) \sqrt{\frac{K(p-K)}{K+(p-K)}} \left(\overline{\log X_{i,j|j\in K}} - \overline{\log X_{i,j|j\notin K}} \right) \tag{4}$$

~~In this instance, for each taxon~~ where p is the overall number of taxa, 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 [41]) 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~~ is the index of a taxa and K is the set of indices of taxa in set k. The CBEA statistic then looks similar to a t-statistic for difference in means of log-transformed proportions. As such, the pooled variance of CBEA is dependent on the variance inflation of both mean components $\overline{\log X_{i,j|j\in K}}$ and $\overline{\log X_{i,j|j\notin K}}$. The result of this variance inflation is inflated 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 [64] formulation to calculate 95% confidence intervals. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$) and~~ since highly correlated sets are also detected as significantly enriched.

However, as Wu et al. [48] showed, performing column permutation to estimate the null distribution of a competitive test statistic doesn't allow for adequate capture of this variance inflation factor since the permutation procedure disrupts the natural correlation structure of the original variables. It is important to address this problem since there is strong 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 [20] 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 [68] were obtained for each estimate. Simulation settings for classification performance were identical to power analyses as detailed in the previous paragraph~~microbiome [49]. Our strategy for addressing issue is location (or mean) estimate from the column permuted raw score matrix with the spread (or variance) estimate from the original un-permuted scores. This still allows us to leverage the null generated via column permutation while using the proper

variance estimate taken from scores where the correlation structure has not been disrupted. As such, this procedure assumes that the variance of the test statistic under the alternate hypothesis is the same as that of the null. Details of the computational implementation to this estimation process can be found in the Supplementary Materials.

Differential Abundance Analysis: To assess type I error rate and power for differential abundance testing task, we simulated data based on the schema above, and assessed differential abundance of 50 sets with 100 taxa per set across 20 replicates per simulation condition. Type I error is calculated as the number of differentially abundant sets over the total number of sets for each simulation under the global null. Power is defined similarly, but instead under the global alternate hypothesis. Estimates and confidence intervals for type I error and power are calculated as cross-replicate mean and standard error. A set is differentially abundant when all taxa within a set are differentially abundant with the same effect size. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$), and inter-taxa correlation within the set ($\rho = 0, 0.2, 0.5$). Half of the sets are differentially abundant across case/control status with varying effect sizes (fold change of 1.5, 2, and 3). Due to the compositional nature of microbiome taxonomic data, simple inflation of raw counts would cause an artificial decrease in the abundance of the remaining un-inflated sets . As such, we applied a compensation procedure as described in Hawinkel et al. [50] to ensure the validity of simulation results. All sample sizes were set at 2,000. However, set-based analysis is an exploratory approach that can help generate functionally informative hypotheses, and as such users might not want strict type I error control in favor of higher power. This is especially true for competitive hypotheses, where its stricter formulation compared to the self-contained approach implies that the test naturally has lower power [15, 51]. Furthermore, sets that are highly correlated compared to background can be biologically relevant. Therefore, CBEA provides an option for users to specify whether correlation adjustment is desired.

Prediction: To assess predictive performance, we generated predictors based on the simulation schema presented above and evaluated prediction for both binary and continuous outcomes using a standard random forest model [71]. 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* [73] 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.

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

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 [?]. 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. [?] 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 additionto simulation analyses, we also evaluated our method using real data sets based on both 16S rRNA gene sequencing and whole-genome sequencing. All data sets are obtained from either the~~

Evaluation

We based our evaluation strategy on gene set testing benchmarking standards set by Geistlinger et al. [53] and utilized the same approaches whenever possible. All data sets are obtained from either the *curatedMetagenomicData* [55] and *HMP16SData* [56] R packages (2020-10-02 ~~snapshot), or downloaded from the Qiita platform [57].~~ snapshot), or downloaded from the Qiita platform [57]. All code and data sets used for evaluation of this method is publicly available and can be found on GitHub (qpmnguyen/CBEA_analysis). Additional packages used to support this analysis includes: *tidyverse* [58], *pROC* [59], *phyloseq* [60], *mia* [61], *targets* [62].

~~**Single Sample Enrichment**: To assess the false discovery rate and true discovery rate of cILR in sample-level enrichment testing, we utilized the~~

Statistical significance

We evaluate the inference procedure of CBEA compared to alternate methods using two approaches: randomly sampled taxa sets and sample label permutation. These analyses were performed on the 16S ~~rRNA gene sequencing of the oral microbiome at the gingival subsite from the Human Microbiome Project [1,63]. We utilized this data set following the approach outlined in Calagaro et al. [78].~~ rRNA gene sequencing of the oral microbiome from the Human Microbiome Project [1,63]. This data set contains 369 samples split into two subsites: supragingival and subgingival. We processed this data set by removing all samples with total read counts less than 1000 and OTUs whose presence (at least 1 count) is in 10% of samples or less.

Sample-level inference

Due to CBEA's self-contained null hypothesis, we can perform inference at the sample level for the enrichment of a set. We evaluated this application by generating one random taxa of different sizes $S \in \{20, 50, 100, 150, 200\}$ across 500 iterations. Random sets can act as our estimate for type I error since this matches the CBEA null hypothesis stated in Materials and Methods, where we expect within each sample sets of randomly drawn taxa should not be significantly enriched compared to the remainder background taxa. For this evaluation, we estimated type I error as the fraction of samples ~~that detects~~ our random set as significant at a p-value threshold of 0.05 with confidence bands computed from the standard error across all iterations. Additionally, this analysis also ~~demonstrate~~ whether CBEA is sensitive to different set sizes.

Population-level inference

We can perform enrichment testing at the population level by generating corresponding sample level CBEA scores and performing a two-sample test such as Welch's t-test. In order to evaluate CBEA under this context, we generated CBEA scores of sets representing genus-level annotation in above gingival data set [1,63] and applied a t-test to test for enrichment (similar to GSVA [19]) across a randomly generated variable indicating case/control status (repeated 500 times). Type I error is

estimated as the fraction of sets per iteration found to be significantly enriched with confidence bands computed from the standard error across all iterations. In addition, we also performed a random set analysis assessment, where we generated 100 sets of different set sizes $S \in \{20, 50, 100, 150, 200\}$ and evaluated the fraction of genera that were found to be differentially abundant across the original labels (supragingival versus subgingival subsite). 95% confidence intervals were computed using the Agresti-Couli approach [64].

Phenotype relevance

We want to evaluate whether sets found to be significantly enriched by CBEA are relevant to the research question. To perform this assessment, we relied on the gingival data set mentioned ~~prior~~ [1, 63]. This data set ~~is approximately labeled, where aerobic microbes are~~ was chosen ~~due to~~ its clear biological interpretation ~~that~~ can serve as the ground truth. Specifically, we expect aerobic microbes to be enriched in the supragingival subsite where the biofilm is exposed to the open air, while conversely anaerobic microbes thrive in the subgingival site [65]. ~~Here, we assessed the enrichment of aerobic microbes across all samples, we considered the false positive rate as the number of samples from the subgingival site with significant enrichment, and the true positive rate as the number of supragingival samples with significant enrichment. Microbial tropism annotation at the genus level was from~~ Genus-level annotations for microbial metabolism from Beghini et al. [66] ~~and was downloaded directly~~ obtained from the GitHub repository associated with Calagaro et al. [67].

~~**Differential Abundance Analysis**: To assess type I error using cILR scores in differential abundance analysis, we utilized the 16S rRNA gene sequencing of stool samples from the Human Microbiome Project [1, 63]. 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~~ For sample-level inference, we assessed power as the fraction of supragingival samples where aerobic microbes are significantly enriched. For population-level inference, power is the ~~number of aerobic~~fraction of sets representing genus level taxonomic assignments that were significant across subsite labels.

In addition to statistical power, we also assessed phenotype relevance through evaluating whether highly ranked sets based on CBEA scores are more likely to be enriched according to ground truth. This is represented by the area under the receiving operator curve (AUROC/~~anaerobic genera identified as differentially abundant across all aerobic or anaerobic genera.~~ AUC) scores computed on CBEA scores against true labels (similar approach was used to evaluate VAM [21]). DeLong 95% confidence intervals for AUROC [68] were obtained for each estimate.

~~**Disease Prediction**: To assess predictive power, we utilized the whole~~

Disease Prediction

CBEA scores can also be used downstream analyses such as disease prediction tasks. We utilized two data sets for this evaluation:

1. Whole genome sequencing of stool samples ~~of~~ from inflammatory bowel disease (IBD) patients ~~from~~ in the MetaHIT consortium [69]. 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. [70] which also profiles the gut microbiome of IBD patients and controls but using~~ We processed this data by removing all samples with less than 1,000 total read counts as well as any OTU ~~who~~ was present (with non-zero proportions) in 10% of samples or less. Prior to model fitting, we back-transformed relative abundances into count data (to align the format with our 16S rRNA gene sequencing data set) using provided total number of reads aligned to MetaPhlan marker genes (per sample).

2. 16S rRNA gene sequencing of stool samples from IBD patients in the ~~pediatric~~ RISK cohort [70]. This data set contains 16S rRNA gene sequencing samples from a cohort of pediatric patients (ages < 17) from the RISK cohort enrolled in the United States and Canada. Of the 671 samples obtained, 500 samples belong to patients with IBD. We processed this data set by removing all samples with less than 1,000 total read counts as well as any OTU ~~who~~ was present (at least 1 count) in 10% of samples or less.

Unless there is a pediatric and another one that is different

We evaluate disease prediction performance by fitting a random forest model [71] using inputs ~~as~~ CBEA scores to classify samples of patients with IBD and healthy controls. Random forest was chosen as a baseline learner due to its flexibility as an out-of-the-box model that is easy to fit. In this instance we evaluated predictive performance of a default random forest model (without hyperparameter tuning) AUROC after 10-fold cross validation. Additionally, we utilized SMOTE to correct for class imbalances [72]. Implementation was done using the *tidymodels* suite of packages [73].

~~**Comparison Methods**~~

~~**Single sample enrichment**: For type I error and power analyses, we compared the cILR method with a naive Wilcoxon rank sum test. We added a pseudocount of 1 to all values. This is~~

Comparison Methods

We benchmarked the statistical properties of CBEA against existing baseline approaches. For sample-level inference analyses, utilized the Wilcoxon rank-sum test, which non-parametrically tests the difference in mean counts between taxa from 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~~remainder similar to CBEA. For assessments at the population level, we compared ~~cILR methods against GSVA [19], ssGSEA [76], and the W-statistic from the Wilcoxon rank sum test . All three approaches were applied directly on count data (after pseudocount). For GSVA, the Poisson kernel was used.~~

~~**Differential Abundance**: Since cILR are sample-level enrichment scores, we performed differential abundance by using a Wilcoxon Rank Sum test and Welch's t-test across case/control status on cILR generated scores. We added a pseudocount of 1 to all values. For comparison, we chose representative state-of-the-art methods in differential abundance analysis, namely~~ CBEA against performing a standard test for differential abundance with set-level features generated via element-wise summations instead. We chose DESeq2 ~~[26, 60]~~ [26] and corncob [74] ~~. For DESeq2, we performed a likelihood ratio test against an intercept only reduced model with dispersion estimated with local fit. For corncob, we also performed a likelihood ratio test against~~

~~an intercept only reduced model without bootstrapping.~~ because they represent both methods extrapolated from RNA-seq [60] and those developed specifically for microbiome data.

~~**Disease Prediction**: We fit random forest on cILR scores, as well as ssGSEA [19] and GSVA [76] similar to single sample enrichment section. We added a pseudocount of 1 to all values. Additionally,~~ Since disease prediction models and rankings-based phenotype relevance analyses seek to evaluate the informativeness of CBEA scores instead of relying on computing p-values, we compared performance against other single sample based approaches from the gene set testing literature, specifically ssGSEA [76] and GSVA [19]. Additionally, for evaluating prediction, we also compared performance ~~using cILR~~ against a standard analysis plan where inputs are count-aggregated sets with the centered log-ratio ~~transformation~~ (CLR) ~~was applied to count-aggregated sets as inputs to a machine learning model~~transformation.

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~~results for evaluating statistical significance, phenotype relevance, and predictive performance. In addition to real data, we also evaluated models based on parametric simulations, where results can be found in the Supplemental Materials.

~~Enrichment testing at the sample level~~Statistical Significance

~~cILR~~

Inference at the sample level

CBEA provides significance testing ~~for enrichment~~ at the sample level ~~using the null distribution estimation procedure described in . Here, we present empirical results for this application of cILR assessing type I error, power, and classification capacity~~through a self-contained competitive null hypothesis. Generating random sets approximate the global null setting where within each sample, sets generated by randomly sampling taxa should not be significantly more enriched than remainder taxa.

~~Simulation studies~~

~~Panel A and B in~~

Fig 2 ~~demonstrate~~ demonstrates type I error ~~and power respectively across different simulation conditions. We benchmarked the results of the cILR method against a naive~~ of sample-level inference evaluated using the random set approach. The Wilcoxon rank sum test ~~performed at the sample level, comparing the mean count difference between taxa in the set its complement. All methods demonstrate~~ and unadjusted CBEA under mixture normal assumption demonstrated 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~~ the appropriate α

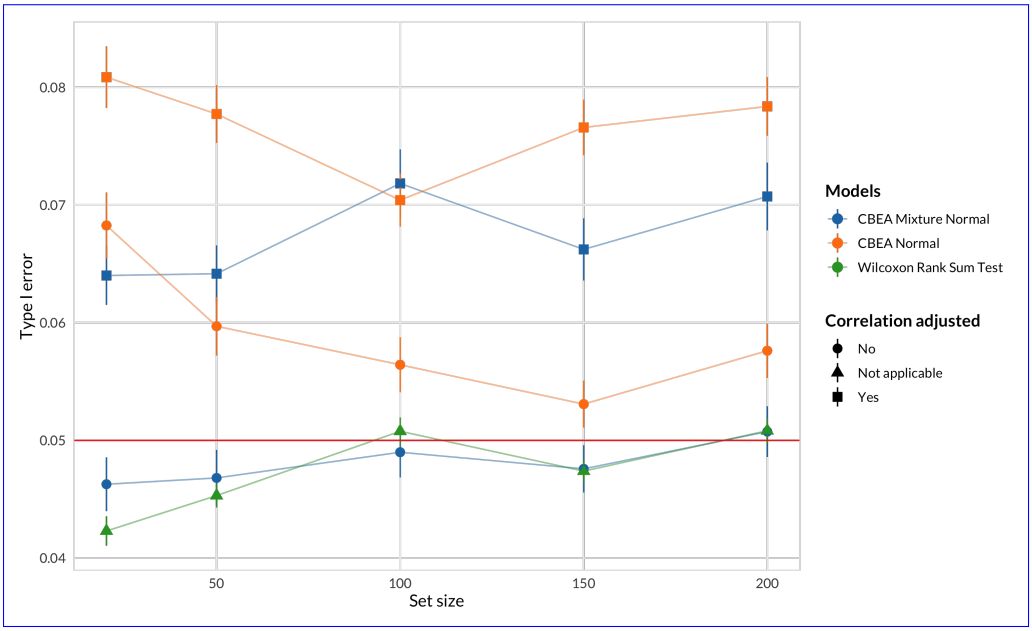


Fig 2. Random taxa set analyses for inference at the sample level of CBEA under different parametric assumptions compared against a Wilcoxon rank-sum test. Type I error (y-axis) was evaluated by generating random sets of different sizes (x-axis) (500 replications per size) and computing the fraction of samples where the set was found to be significantly enriched at $\alpha = 0.05$. Error bars represent the mean type I error $\pm$ sample standard error computed across 500 replications of the experiment. Only the unadjusted CBEA with the mixture normal distribution and the Wilcoxon rank sum test were able to control for type I error at 0.05. All approaches are invariant to set sizes.

level. This fits with our expectations since the mixture normal distribution has much better fit than the normal distribution especially at the tails (Fig 1). However, other variants of CBEA demonstrated inflated type I error, ~~where Wilcoxon test performed the worst. On the other hand, adjusted cILR methods (under both distributions) control for~~ especially correlation adjusted variants compared to their unadjusted counter parts. Encouragingly, all methods demonstrate consistent performance across all set sizes, with a slight increase in type I error at the ~~appropriate α level even at high correlations~~highest levels.

~~Sample-level inference with cILR under parametric simulations ((**A**) and **B**), and real data analysis ((**C**) and (**D**)). In simulation analyses, panel (**A**) shows type I error rate, and panel (**B**) shows power for single sample enrichment test for a specified set and was compared against a Wilcoxon rank sum test at α of 0.05. In real data analysis, panel (**C**) shows the false-positive rate, and panel (**D**) shows the true positive rate. For this analysis, 16S rRNA data from the oral microbiome of the gingival site was used. The set of aerobic microbes was tested for enrichment in all samples and was identified as correctly enriched if a significant p-value was obtained in supragingival samples. Confidence bounds were obtained using Agresti-Couli [64] approach. Adjusted cILR demonstrated control of type I error at the appropriate α level while remaining methods (not included in subsequent power analyses) showed an inflated type I error rate. However, this resulted in lower power for adjusted cILR methods.~~

~~However, the trade-off for good~~ Interestingly, simulation results (S1 Fig) showed an opposite pattern. Adjusted approaches were good at controlling for type I error~~control is demonstrably lower power, as shown in Fig 2B. In situations where there is no~~, especially under the low inter-taxa correlation ~~, cILR still outperforms the wilcoxon rank sum test, however adjusted versions of cILR did not perform as well as un-adjusted ones. However, in higher correlation scenarios, the difference in power is much more dramatic. At the highest effect size (fold change of 3)and correlation ($\rho = 0.5$), adjusted cILR was only performing at 50% power, while unadjusted cILR and wilcoxon~~ values within the set (similar to generating random sets where the natural correlation structure is disrupted). In these simulations, unadjusted approaches and the Wilcoxon rank sum test ~~were able to reach 80%. These results indicate that both sparsity and inter-taxa correlation impacts power, with correlation having a much more dramatic impact especially for adjusted versions of cILR. Most importantly, cILR demonstrate higher power in all scenarios where~~ had significant type I error ~~is properly controlled.~~ inflation with increasing correlation. All approaches seems to be invariant to the level of data sparsity.

~~To further assess the utility of cILR in classifying samples with enriched sets, we generated AUC scores for different cILR scores using true labels of whether a sample has an inflated set. This analysis, therefore, assessed the relative ranking of samples using cILR scores whereby high scores should correspond to samples that are known to be inflated. Fig 3 presents this result. We compared different variants of cILR against competing methods in the~~

Inference at the population level

Similar to other single sample approaches to gene set testing ~~space (GSVA [19] and ssGSEA [76]), as well as the W test statistic from the Wilcoxon rank sum test. Across both simulations (Fig 3A) and real-data applications (Fig 3B), cILR scores perform marginally better especially in low effect size situations but did not stand out in most other scenarios. In simulation studies, classification performance was good (around AUC of 0.8) even at high correlation settings, only requiring medium effect sizes (fold change of 2). Notably, the W-statistic provided the least information for classifying samples with inflated taxa.~~

~~Real data evaluations~~

~~These observations were replicated when assessed on the semi-labeled gingival data set from the Human Microbiome Project as described in~~ such as GSVA [19], we can perform inference at the population level by utilizing a two-sample difference in means test. Here, we ~~tested the enrichment of aerobic microbes for each sample using approaches similar to our parametric simulations. As expected in Fig 2C, the proportion of falsely rejected hypotheses was high in the naive Wilcoxon test and unadjusted cILR methods. Conversely, adjusted cILR controls for false positives adequately at the correct α level of 0.05. Power analysis (Fig 2D) showed similar patterns, where unadjusted cILR methods and the Wilcoxon test have a higher proportion of null hypotheses correctly rejected, however, these results are not useful to a practitioner as the number of falsely rejected hypotheses are also equally high.~~

~~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 . 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 [68].~~

~~Differential abundance analysis~~

~~cILR generates sample-specific scores representing the degree of enrichment of a pre-defined set. As such, we want to assess the ability to use these scores for differential abundance analysis in combination with a standard difference of means statistical test (~~evaluate using CBEA scores generated under different settings with Welch's t-test ~~and Wilcoxon rank sum test). We compared the performance of this approach with cILR and two commonly used methods for differential abundance testing in the microbiome literature: DESeq2 [26] and corncob [74]~~ in a supervised manner to assess whether a set is enriched across case/control status.

~~Simulation studies~~

Fig ~~4 present results for simulation studies for both type I error (panel A) and power (panel B) evaluations. All methods~~ 3 shows results for this scenario using both random sample label and random set evaluations. The random sample label approach (Fig 3A) provides a controlled setting where we can estimate type I error rate controlled at $\alpha = 0.05$. Across all replications, CBEA methods were able to control for type I error ~~well across both sparsity and correlation levels, where the estimated rate was consistently around the~~ at the nominal threshold of 0.05~~pre-defined threshold . Results were similar across all evaluated methods, although in some instances, for example in medium correlation setting ($\rho = 0.2$), the unadjusted cILR resulted in higher~~ , with CBEA raw scores being the most performant. Neither output types, correlation adjustment, nor distributional assumption improved performance values. Surprisingly, DESeq2 and corncob both exhibit significantly inflated type I error~~, regardless of difference in means test and distribution of choice~~.

~~The difference between the methods is more noticeable when evaluating power. All cILR associated variants showed much higher power even when the effect size is limited (fold change is 1.5), and there is a noticeable gap in performance between cILR and both~~ We also assessed the impact of set-size on the inference procedure by testing for enrichment using the original sample labels but with randomly sampled sets of different sizes (Fig 3B). Overall we observed very similar values across CBEA as well as corncob and DESeq2~~and corncob. Surprisingly, this effect is consistent across correlation levels and sparsity, even though we expectedly see performance in power drop as a function of sparsity especially in low effect size settings~~, suggesting that no individual method is systematically identifying too many significant sets. Additionally, similar to analogous analyses at the sample level, no approach ~~were~~ significantly sensitive to changes in set sizes.

Phenotype Relevance

~~**Real data evaluationsIn addition to simulation studies, we also evaluated performance of the methods on real 16S rRNA gene sequencing data set from HMP (Fig 5). For type I error evaluations, we use stool samples and randomly assign them with case/control status and calculated type I error as the proportion of genera identified as significantly different. For true positive rate evaluations, we use**~~ **Inference at the sample level**

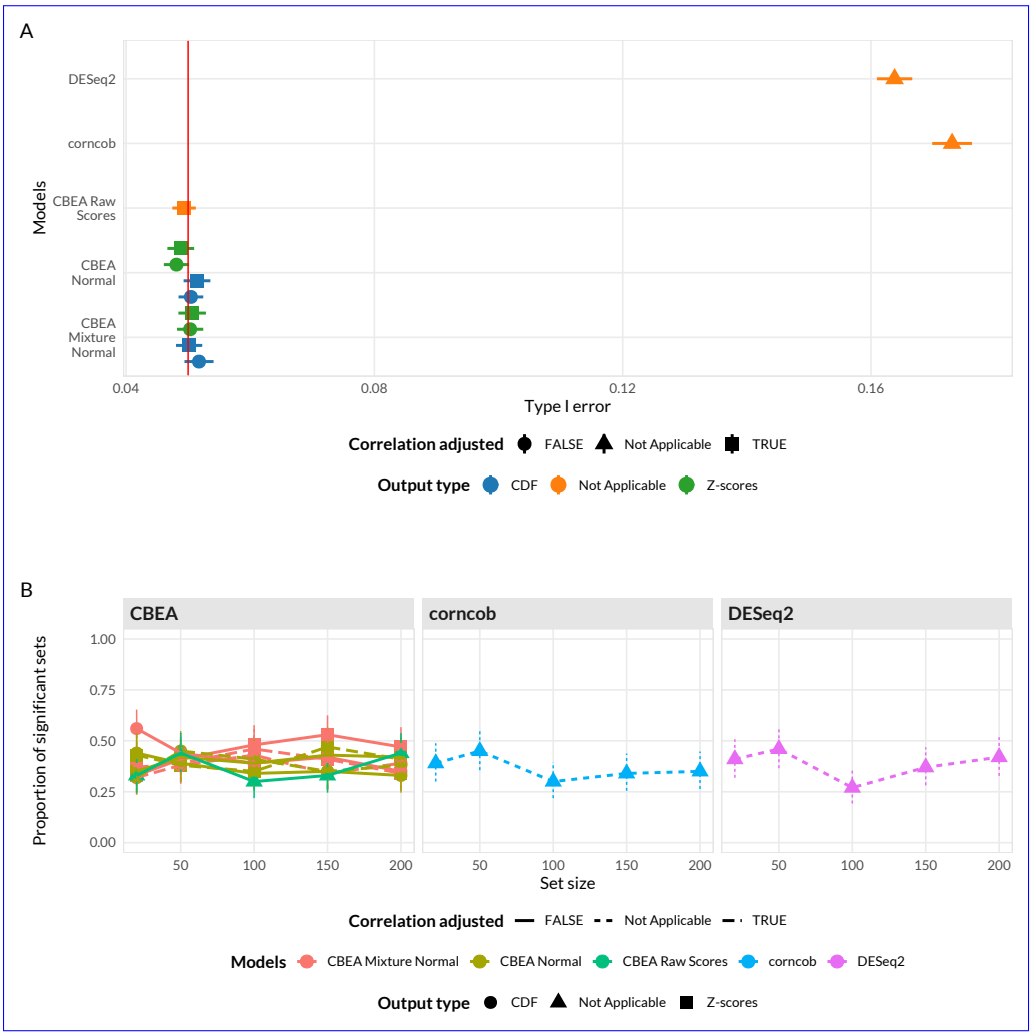


Fig 3. ~~Type I error rate **(A)**~~ Random sample label (**A**) and ~~power **(B)**~~ random set (**B**) analyses for ~~differential abundance test across different parametric simulation scenarios~~population level inference. ~~For cILR methods, differential abundance analysis~~ (**A**) Type I error (x-axis) was ~~performed~~ estimated as the overall fraction of sets found to be enriched $\alpha = 0.05$ using ~~a difference in means test~~ randomly generated sample labels (~~either Wilcoxon rank-sum test or Welch's t-test~~500 permutations)~~across case/control status using single~~ . Error bars represent the mean type I error ± sample ~~scores generated by cILR~~ standard error. (~~across~~ **B**) Proportion of significant sets (y-axis) using 100 randomly generated sets of different ~~output types and distributional assumptions~~set sizes (x-axis). ~~cILR associated methods demonstrated similar~~ Confidence intervals computed using Agresti-Couli method for binomial proportions. For sample label permutation (**A**), all CBEA approaches were able to control for type I error ~~to conventional differential abundance analysis methods~~ but ~~with more power~~ not for corncob and DESeq2. For random set analyses (**B**), all approaches demonstrate similar rate of accepting significant sets and were invariant to ~~detect differences even at small effect sizes~~overall set size.

In Fig 4, we evaluate whether sets found to be significant by CBEA are relevant to the phenotype of interest. We leveraged the gingival data set as ~~detailed in the previous~~

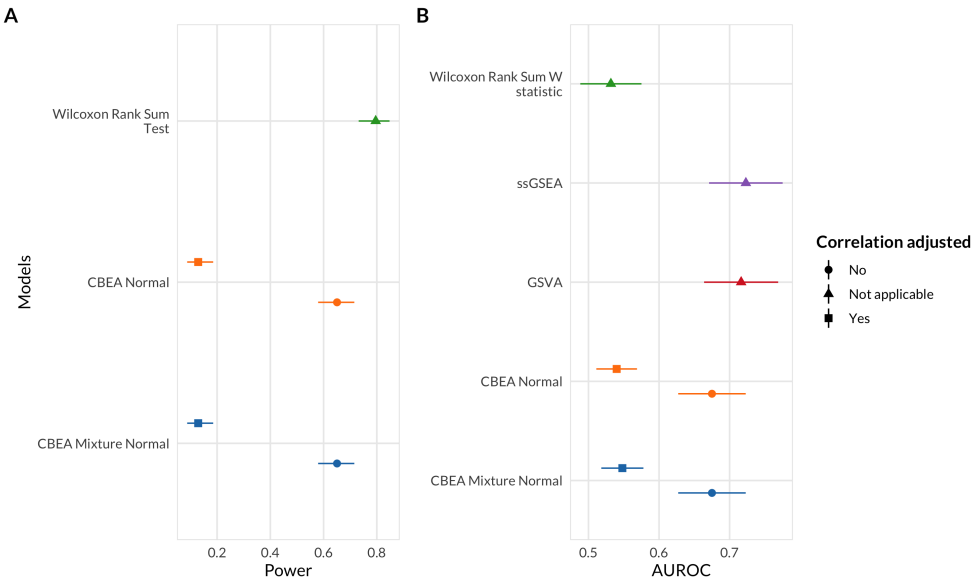


Fig 4. Statistical power (**A**) and score rankings (**B**) to assess phenotype relevance. (**A**) Power (x-axis) was estimated as the overall fraction of aerobic microbes found to be enriched in supragingival samples at $\alpha = 0.05$. 95% confidence intervals were computed using the Agresti-Couli approach for binomial proportions. (**B**) Score rankings were evaluated by comparing computed scores against true values using AUROC (x-axis). DeLong 95 % confidence intervals for AUROC were computed.

~~section , and calculated the true positive rate as the proportion of genera labeled as either anaerobic or aerobic that were found to be significant.~~

~~We observed both corncob and DESeq2 had significantly inflated type I error rate while all variations of cILR were controlling for type I error at the defined α threshold of 0.05~~stated in Evaluation section where we know beforehand that aerobic microbes are more likely to be enriched in supragingival subsite samples and vice versa.

We estimated statistical power using this data set as the fraction of supragingival samples where the set representing aerobic microbes were significantly enriched. We observed that adjusted CBEA approaches demonstrate much lower power compared to the Wilcoxon rank-sum test and unadjusted variants. This is surprising given the ~~consistency of preserving~~ fact that in statistical significance analyses, the adjusted CBEA approach provides inflated type I error~~for both corncob and DESeq2 in all simulation evaluations~~, especially if the normal distribution assumption was chosen, which indicates a mismatch in estimating the null distribution since a high type I error did not result in increased power.

~~In true positive experiments with data from the gingival site, estimated rates were more similar across the different methods. As expected,~~ We also evaluated phenotype relevance by assessing whether enriched sets according to ground truth are preferentially ranked higher using assigned continuous scores (instead of performing a hypothesis test). This aspect is captured through computing AUROC values comparing computed enrichment scores and true labels. Consistent with prior approaches, adjusting for correlation did not improve performance, where the AUROC values are equivalent to using the Wilcoxon ~~rank sum test resulted in lower true positive rate compared to remaining methods , but the difference was not noticable.~~

this sentence is a

bit clunky

~~This is also surprising given that~~ Rank sum statistic at around 0.5. Unadjusted methods were much better at ranking true enriched sets, however the mean AUROC values are lower than alternate single sample enrichment methods (GSVA [19] and ssGSEA [76]) even though this difference is not significant due to overlapping confidence intervals.

The above results were replicated in simulation studies ~~, both corncob and DESeq2 showed markedly lower power across all~~ where we observed that adjusted approaches ~~was~~ very conservative and demonstrated significantly lower power (S3 Fig.) with increasing correlation even at the highest evaluated effect sizes. When assessing score rankings, the performance of CBEA was closer to ssGSEA and GSVA compared to real data evaluations, however all single sample approaches were much better than using the W statistic from the Wilcoxon Rank Sum test.

Inference at the population level

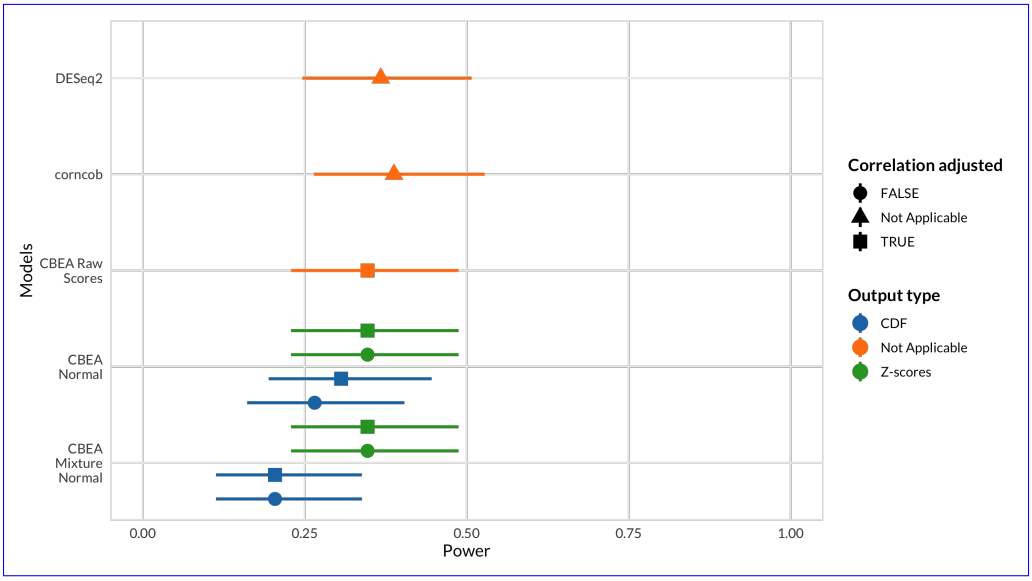


Fig 5. ~~Differential abundance analysis using corncob, DESeq2 and cILR with either Wilcoxon rank sum test or Welch's t-test. Panel (A) shows type I error results as the proportion of significant genera after 500 iterations where case/control status was assigned randomly~~ Statistical power to ~~each sample~~assess phenotype relevance of inference tasks at the population level. ~~Panel (B) shows true positive rate results~~ Power (x-axis) was estimated as the ~~proportion~~ overall fraction of ~~significant~~ sets representing genera ~~who~~ that are ~~either obligate anaerobes~~ aerobic or ~~aerobes~~anaerobic microbes found to be differentially enriched across sample type (supragingival or subgingival). ~~Both evaluations use 16S rRNA gene sequencing data from HMP. Type I error evaluation used stool samples while~~ 95% confidence intervals were computed using the ~~true positive rate evaluation used samples from the gingival site~~Agresti-Couli approach for binomial proportions.~~Results showed that cILR associated methods were able to keep type I error rate at approximately 0.05 while still demonstrating similar power as both corncob and DESeq2~~

We also assessed statistical power for population level inference scenarios using a similar approach. Here, enrichment scores for sets representing all identified genera were computed, and estimated ~~power~~ as the fraction of sets found to be differentially enriched across sample site labels (supragingival or subgingival). We compared these

results against performing a differential abundance test of genus level features generated via sum-based approaches. Results are shown in Fig 5. Some CBEA variants, such as CDF outputs for the mixture normal distributional assumption, did not correctly detect as many significant sets as DESeq2 or corncob despite very close performance values. Using raw CBEA scores was best approach, however it did not exceed values obtained from DESeq2 and corncob.

Disease Prediction

Since ~~cILR~~ CBEA can generate informative scores that can discriminate between samples with inflated counts for a set (Fig 2d), we wanted to assess whether they can also act as useful inputs to predictive models. In this section we assessed the predictive performance of a ~~naive~~ standard baseline random forest model [71] with different single sample enrichment scoring methods as inputs (~~evaluating cILR~~CBEA, ssGSEA, and GSVA). Additionally, we also compared predictive performance of using these scores against the a standard approach of using the centered log ratio transformation (CLR) on taxon sets aggregated via abundance summations.

~~Simulation studies~~

~~Fig 6 shows results for simulation studies as detailed in the section. Panel A presents results for a regression learning task with a continuous outcome while panel B presents results for a classification task with a binary outcome. As expected, performance across all assessed methods increased with a higher signal-to-noise ratio. Both CLR and cILR approaches outperformed both GSVA and ssGSEA across all simulation conditions and learning tasks. This is because both GSVA and ssGSEA are more sensitive to the degree of inter-taxa correlation and sparsity, while cILR and CLR did not experience a similar level of impact. As such, performance gap widens with increasing correlation and sparsity. Interestingly, this difference in performance is not as pronounced under high levels of effect saturation (across both learning tasks), suggesting that when there is a high number of sets contributing to an effect, model choice might not be as important.~~

~~In this analysis, cILR unfortunately did not outperform the CLR approach, which is standard practice within the microbiome literature [12]. This difference in performance is more notable in regression learning tasks compared to classification, and at lower levels of effect saturation. However, the degree of separation between the two approaches is not as dramatic as between GSVA/ssGSEA and cILR/CLR. Moreover, the performance gap decreases with increasing effect signal-to-noise ratio and sparsity. Additionally, we did not observe any performance difference between the different variations of cILR.~~

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

~~Real data evaluations~~

~~In addition to parametric simulations, we also assessed the performance of using cILR scores in predictive models with real data sets. Fig ?? presents results for~~ We fit our model to two data sets with a similar disease classification task of discriminating patients who are diagnosed with IBD (includes both Crohn's disease and ulcerative colitis) using only microbiome taxonomic composition. The two data sets represent different microbiome sequencing aprpaoches: the Gevers et al. [70] data set uses 16S rRNA gene sequencing, while the Nielsen et al [69] data set uses whole genome shotgun sequencing.

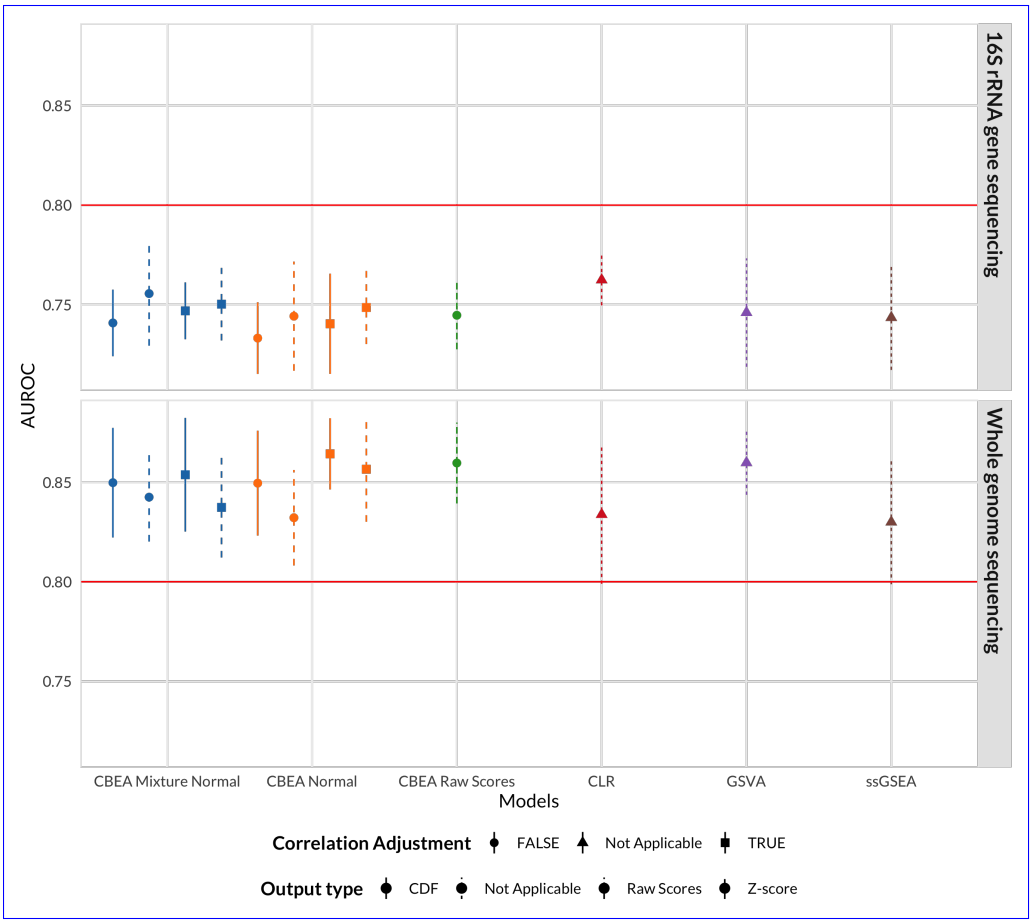


Fig 6. Predictive performance of a naive random forest model trained on ~~cILR~~CBEA, ssGSEA, GSVA generated scores as well as the standard CLR approach on predicting patients with inflammatory bowel disease versus controls using genus level taxonomic profiles. Data sets used span both 16S rRNA gene sequencing (Gevers et al. [70]) and whole-genome shotgun sequencing (Nielsen et al. [69]). ~~cILR~~ CBEA performs better than GSVA and ssGSEA but not as well as CLR, with the exception of the whole genome sequencing data set.

~~Similar to simulation experiments, we also fitted a naive random forest model using CILR, ssGSEA, GSVA, or CLR transformed variables as inputs, and use AUC as the performance criteria. Results also replicated that of the simulations, where across both data sets cILR and CLR methods provide much better performance than both GSVA or ssGSEA~~ Fig 6 illustrates the performance of our model with AUROC as the evaluation criteria. In the 16S rRNA data set, the best performing CBEA variant (CDF values computed from an unadjusted mixture normal distribution) outperforms

both GSVA and ssGSEA but not the standard CLR approach. Interestingly, ~~the cILR approach performed better than CLR~~ in the whole genome ~~data set but did not perform as well in the 16S rRNA gene sequencing data set~~sequencing data set, CBEA outperforms CLR, but ~~were~~ more similar in performance to GSVA. However, ~~these results indicate that cILR~~ due to large confidence intervals, no method ~~were~~ significantly ~~out-performing~~ other approaches ~~evaluated~~. As such, these results indicate~~d~~ that CBEA generated scores are can be informative, ~~providing~~ competitive performance when acting as inputs to disease predictive models. ~~Most importantly, performance values are consistent across both simulated and real data sets.~~

~~These results demonstrate that cILR generated scores are informative features in disease prediction tasks. Simulation results indicate that cILR methods perform much better than either GSVA or ssGSEA, but not as well as the standard CLR approach~~Simulation studies (S5 Fig.) showed similar results, however CBEA ~~was~~ more consistently ~~underperforming~~ compared to CLR across all scenarios. Interestingly, ~~however, cILR methods were much more competitive with CLR in either WGS data sets or data sets with higher~~ the performance gap decreases with increasing sparsity levels.

Discussion

Inference with ~~cILR~~CBEA

CBEA is a microbiome-specific approach to generate sample specific enrichment scores for taxonomic sets defined apriori. The formulation of ~~cILR~~ CBEA as a comparison between taxa within the set and its complement corresponds to the competitive null hypothesis in the gene set testing literature [33]. ~~This allows for conducting inference with cILR even at the sample-level. We assessed the usage of cILR in this type of analysis by evaluating type I error and power across both simulation studies and real data applications. Most importantly~~Since this null hypothesis is self-contained per sample, this allows users perform enrichment testing at the sample level. Additionally, in combination with a difference in means test, CBEA can also test for enrichment at the population level across case/control status similar to GSVA [19].

For single-sample analyses, we demonstrated that ~~our adjusted cILR approach~~ the CBEA approach (unadjusted with mixture normal parametric assumption) was able to ~~address the issue of variance inflation due to correlation [48] by controlling~~ control for type I error at the ~~appropriate α level across different levels of simulated inter-taxa correlation~~ nominal level of 0.05 under the global null (Fig 2) while ~~conversely unadjusted cILR and the naive Wilcoxon rank sum test showed much higher rates of error. This is further encouraged in real data analysis where the false discovery rate was around 0.05 when a collection of true null and true alternate hypotheses were tested. Unfortunately the trade-off of good type I error control is lower power. The conservativeness of the test attenuates with higher sparsity and correlation, where power was not approaching even 50% even at the highest effect sizes. However, when the degree of such data features are reasonable, cILR will still be able to detect a reasonable proportion of samples with inflated counts.~~

~~We also observed that choosing different distributional forms did not alter performance values for cILR. This runs contrary to our comparison analysis in~~ also demonstrating respectable power (Fig 4). This performance is consistent across different set sizes as well as our prior distributional fit analyses (Fig 1~~, where we demonstrated that~~), where the mixture normal ~~distribution had superior fit compared to the simple normal for raw cILR scores computed under the global null . We~~

~~hypothesized that this might be due to the difficulty in fitting mixture distributionto data using the expectation maximization algorithm, as the convergence rate is slow when there is high overlap between the mixtures, resulting in a small mixing coefficient for one of the components [75]. As such, in our implementation of cILR, in order to ensure convergence for the estimating procedure we increased the number of iterations while relaxing the tolerance parameter. Furthermore, there are also possible problems with our adjustment procedure for the mixture distribution that might impact overall fit. In order to combine the scale and location estimate of two mixture distributions, we fixed the overall mean, standard deviation, mixing coefficient and component-wise means and used an optimization procedure to find the component-wise variances. However, this means that we have one equation for the overall variance but two possible parameters to estimate. As such, there is no guaranteed unique solution to component-wise variances. We hypothesized that the instability and degeneracy in component-wise variance estimates might impact the fidelity of estimates at the tails of the distribution, thereby affecting inference.~~

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

~~Downstream analysis~~

~~The sample-level enrichment scores generated by the cILR method can be used in downstream analyses commonly performed in microbiome research: differential abundance testing and disease prediction~~displayed superior fit to the null distribution. Unfortunately, other variants of CBEA demonstrated neither good type I error control nor power. Interestingly, while the adjusted methods showed poor performance in real data evaluations (Fig 2), in simulation studies (S1 Fig, S3 Fig.) these approaches were able to control for type I error well with the trade-off of much lower power.

~~Differential abundance analysis~~

~~For differential abundance testing, we evaluated whether using cILR scores alongside a standard difference in means test (~~For population-level inference task, CBEA also performed very well. Under the permutation global null, representing genera abundance using CBEA scores in combination with Welch's t-test ~~and Wilcoxon rank sum test)~~ is suitable to detect changes in abundance of a set of microbes. ~~We compared cILR against two popular approaches:~~ controls for type I error the correct α threshold while also keeping respectable power. Since population level enrichment test is equivalent to differential abundance using set-based features, we compared CBEA approach against using element-wise summations with corncob [74] and DESeq2 [26] ~~applied on data where taxa were aggregated using the naive sum method~~to test for set-level differential abundance. We chose DESeq2 because it is an older approach from the bulk RNA-seq literature that has strong support for usage in microbiome taxonomic

data [60]. ~~Conversely~~Alternately, corncob is a newer method developed specifically for microbiome taxonomic data sets, where taxonomic counts are modeled directly using a beta-binomial distribution instead of relying on normalization via size factor estimation~~like in DESeq2.~~

~~Surprisingly, we found some conflicting results when evaluating comparisons across parametric simulations and real data analysis. The performance of cILR was consistent across two evaluation criteria, demonstrating good~~ . We observed that using this approach resulted in an inflated type I error ~~and respectable power . However,~~ compared to all variants of CBEA (Fig 2), yet did not improve power (Fig 4). Results for CBEA approaches were replicated in simulation analyses, however for corncob and DESeq2 ~~showed opposite effects~~we observed an opposite effect: in simulation experiments, both methods show good type I error control but low power ~~, while in real data analyses, conversely both corncob and DESeq2 showed inflated type I error but comparable power with respect to cILR methods. Despite such discrepancy, results still indicate good performance of cILR scores when used as inputs to downstream differential abundance analysis compared to using aggregated raw counts, even in methods designed to handle that type of data such as corncob and DESeq2.~~ (S2 Fig, S4 Fig.).

~~We hypothesized that the above behavior can arise due to issues with performing inference in the presence~~ We hypothesized that the discrepancy between simulation and real data evaluations can can arise due to differences in our assumptions regarding the data generating process that informed our simulation schema. For the non-zero component of each taxa we sampled from the same negative binomial distribution where designated enriched taxa were generated with inflated means (but the same dispersion). These marginals were simulated to account for block exchangable correlation within the enriched set only. This might affected our results in two ways: First, our simulation scenario ensures that ~~any~~ designated non-enriched taxa is identical to each other. This is not the case for real data, where our null ~~scenarios~~ involves randomly sampled sets that might by chance all have taxa with inflated means compared to remainder taxa. This is represented in S7 Fig., where the distribution of type I error across ~~our~~ 500 replications are right skewed for underperforming CBEA variants, indicating that these approaches are much more sensitive compared to the Wilcoxon rank sum test or unadjusted CBEA with mixture normal distribution. Second, as described in the Introduction section, we did not consider taxon-specific ~~protocol biases~~ [30]. ~~According to McLaren et al.,~~ biases that distort the observed relative abundance of taxa ~~is different than the true relative abundance due to protocol bias, and importantly this bias is specific to each taxon [30]. This is especially true in compred~~ to true values [30]. In the context of sum-based aggregations, ~~where~~ the resulting bias of the aggregated taxon are dependent on the relative abundances of the contributing taxa (Appendix I in [30]). Conceptually, this means that measurement error for a taxon aggregate is different across samples as relative abundance of contributing changes, leading to issues when attempting to perform inference. As such, we expect methods like corncob or DESeq2 when performed on such ~~aggregates~~ sum aggregates in the presence of taxon-specific biases to have inflated type I error compared to our taxon-ratio based approach.

~~The bias model also helps explain differences in performance of DESeq2 and corncob in simulation analyses compared to real data. Our simulation protocol does not explicitly include bias in our formulation, and all taxa were generated from the same underlying distribution with similar variances across all samples (the only difference is in the mean value where a taxa is expected to have inflated counts). As such, we do not expect our simulated taxa to have any~~ Conversely, in simulation studies, where

taxon-specific biases ~~, which is not the case in real data settings. Therefore, we can expect~~ are absent, corncob and DESeq2 ~~and corncob to retain their expected type I error control in simulation analyses compared to real data. It is still surprising to see lower power for both methods in simulation analyses, which might be due to the fact that the evaluation protocol only considers default settings for both methods and did not attempt to optimize performance~~should perform better.

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

Downstream analysis using predictive models

The sample-level enrichment scores generated by the CBEA method can be used in downstream analyses such as disease prediction. We evaluated whether CBEA can be used to generate set-based features for disease prediction models.

~~Predictive models~~

~~For disease prediction , we~~ We fit~~ted~~ a basic random forest model [71] to predict continuous and binary outcomes using ~~cILR~~ CBEA generated scores as inputs. Similar to our inference analysis, we compared ~~cILR~~ CBEA against both ssGSEA and GSVA. Additionally, we also evaluated ~~cILR~~ CBEA with the approach where counts of a set were aggregated using sums and ~~then~~ applied the centered log-ratio ~~transformed~~transformation (CLR). This is because CLR is considered standard practice in using microbiome variables as predictors for a model [12]. Results ~~indicated that cILR produces good performance values~~ showed that CBEA generate scores perform well across both real data ~~analysis settings~~ and simulation scenarios. Since predictive models consider the effect of variables jointly (and in the case of random forest, consider interactions as well), good performance indicates that ~~cILR~~ CBEA scores can capture joint distribution of sets, enabling both uniset and multi-set type analyses. Comparatively, ~~cILR~~ CBEA generated scores outperformed other enrichment score methods (GSVA and ssGSEA), suggesting that it is more tailored for microbiome taxonomic data sets. This is consistent with our sample ranking analysis (Fig 3d), where ~~cILR~~ CBEA scores are on average more informative when used to rank samples based on their propensity to have inflated counts. However, ~~cILR~~ CBEA did not outperform the CLR approach across our simulation studies, and only marginally performed better in the real data analysis with WGS data.

~~However~~Fortunately, in simulation studies, this performance gap between CLR and ~~cILR~~ CBEA decreases with higher sparsity and correlation, especially in low effect saturation scenarios. ~~Additionally, there are also downsides to applying CLR. First, the singular covariance matrix of CLR transformed variables is singular due to a sum to zero constraint [12], preventing the proper usage of approaches that rely on matrix decomposition. Second, the procedure still relies on using summation of counts prior to transformation, which means that we still can 't compare across sets of different sizes, and any bias might still be propagated [30]. As such, despite benefits in performance for a naive random forest model, there is still space for using cILR as primary inputs into predictive models.~~

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

Limitations and future directions

~~Ultimately, results indicate that cILR~~ These above results demonstrate the applicability of CBEA under different data analysis scenarios. If researchers are interested in performing inference, they can decide between an unsupervised sample level approach (i.e. screen samples for enrichment of certain characteristics) or a supervised population level approach (i.e. identifying characteristics that are differentially abundant across case/control status). For the unsupervised approach, utilizing the unadjusted CBEA with the mixture normal distribution provides a good initial starting point. In the case where researchers only wants to screen samples with mean-inflated taxon sets (instead of additionally detecting taxon sets with increased correlation), they can apply the adjusted approach, which can be effective at conserving type I error even ~~at~~ high correlation scenarios. However, the trade off for this adjustment is power ~~that~~ decreases with increasing correlation. For the supervised analysis, all CBEA variants control ~~well~~ for type I error and provides adequate statistical power. However, using raw CBEA scores with difference-in-means test such as Welch's t-test is preferable since it ~~requires the least amount of computing resource~~ (no estimation process) while still outperforming ~~using~~ sum-based approach with a standard differential abundance test.

Beyond inference, CBEA scores are flexible and can be useful for downstream analysis. We demonstrated that for a given number of set-based features, CBEA can produce informative scores that contribute to competitive performance of prediction models even in low signal-to-noise ratios with high inter-taxa correlation and sparsity. ~~Even though there exists situations where it might not provide maximum predictive values, the flexibility of cILR in various types of analyses enable even though in some scenarios it might not provide maximum predictive values~~This is especially true for whole genome sequencing data sets, where CBEA outperfrorms the standard approach of applying a CLR transformation. Researchers might find CBEA useful under situations of high sparsity and inter-taxa correlation, or if the property of a singular covariance matrix (a byproduct of the CLR transformation [12]) is undesired. Even though we only evaluated prediction models ~~in this manuscript~~, researchers can benchmark their own usage of CBEA for other downstream tasks such as sample ordination.

~~Limitations and future directions~~

~~There~~ However, there are various limitations to our evaluation of ~~cILR~~CBEA. First, our simulation analysis might not capture the appropriate data-generating distributions underlying microbiome taxonomic data. There is strong evidence to suggest that our zero-inflated negative binomial distribution is representative [78], however other distributions such as the Dirichlet multinomial distribution ~~[77]~~ [79] have been used in the evaluation of prior studies. Second, we were not able to evaluate the phenotypical relevance of enrichment results as in Geistlinger et al. [53] due to limited consistent annotations for microbiome signatures in health and disease, especially those that are experimentally verified (and not just from differential abundance studies). We

attempted to perform this evaluation by leveraging the ~~usage of the~~ gingival data set similar to [78] ~~to assess power in differential abundance testing and single sample inference is not perfect. This is because the~~ . However, we acknowledge that this is not a perfect solution, since oxygen usage label of each microbe in the data set is only available at the genus level, and the difference in counts for obligate aerobes and anaerobes across the supragingival and subgingival sites might not be as clear cut. As such, results from power analyses using this data set is only relative between the comparison methods ~~instead of~~ treated as absolute measures of power or phenotype relevance. Finally, we assumed that taxa within a set are all equally associated with the outcome. This limits our ability to evaluate the performance of ~~cILR~~ CBEA when only a small number of taxa within the set is associated with the outcome, or if there are variability in effect sizes or association direction of taxa within a set.

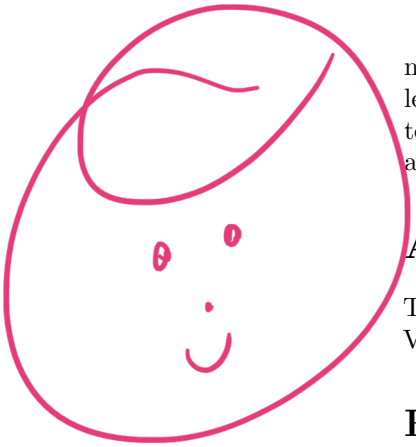
Our evaluation also showed various drawbacks of the ~~cILR method~~ CBEA method itself. First, inference with ~~cILR is limitedin being able to exhaustively detect all samples with significant inflated counts for a set in situations where there is a high degree of sparsity and~~ CBEA at the sample level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

In addition, ~~there are possible extensions cILR can we can consider to provide more flexibility across different data analysis scenarios in data analysis. First, cILR did not address the sparsity of microbiome taxonomic dataand relies on~~ CBEA itself did not consider other aspects of microbiome data. First, across all analyses, we relied on adding a pseudocount to ensure log operations are valid. ~~We can address~~ Users can directly addressing this by incorporating ~~more sophisticated~~ model-based ~~zero-correction methods~~ zero correction methods prior to modelling such as in [80] or [81]. ~~Second, cILR~~ However, in our simulation studies, sparsity seems to not have a significant impact on the overall performance of our approach. Second, CBEA also treated all taxa within the set as equally contributing to the set. Incorporation of taxa-specific weights (similar to PhILR [44]) could reduce the influence of outliers, such as rare or highly invariant taxa. Finally, ~~curating sets based on *apriori*~~ even though for a given set of apriori annotations CBEA can generate useful summary scores, such values are limited in their utility if the ~~annotation~~ themselves are not meaningful. As such, curating and validating sets (~~similar~~ to MSigDB [16]) based on physiological or genomic characteristics of microbes [82] or their association with human disease (in beta BugSigDB `https://bugsigdb.org/Main_Page`) can allow for incorporating functional insights into microbiome-outcome analyses~~while also improving interpretability when compared to using taxonomic categories such as phylum or genus alone~~.

a priori (italics)

Conclusion

Gene set testing, or pathway analysis is an important tool in the analysis of high-dimensional genomics data sets. However, limited work has been done developing set based methods specifically for microbiome relative abundance data. We introduced a



new microbiome-specific method to generate set-based enrichment scores at the sample level. We demonstrated that our method can control for type I error for significance testing at the sample level, while generated scores are also valid inputs in downstream analyses, including disease prediction and differential abundance.

Acknowledgments

The authors thank Becky Lebeaux, Modupe Coker, Erika Dade, Jie Zhou, and Weston Viles for insightful comments and suggestions that greatly improved the paper.

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.
3. 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.
4. 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.
5. 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.
6. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: ~~High~~High-resolution ~~-Resolution~~ Sample Inference from Illumina Amplicon Data. Nature Methods. 2016;13(7):581–583. doi:10.1038/nmeth.3869.
7. 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.
8.
9. Li H. Statistical and Computational Methods in Microbiome and Metagenomics. In: Handbook of Statistical Genomics. John Wiley & Sons, Ltd; 2019. p. 977–550.
10. Li H. Microbiome, Metagenomics, and ~~High-Dimensional~~ High-Dimensional Compositional Data Analysis. Annual Review of Statistics and Its Application. 2015;2(1):73–94. doi:10.1146/annurev-statistics-010814-020351.
11. ~~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.~~

    ~~Sankaran K, Holmes S. structSSI: Simultaneousand Selective Inferencefor Groupedor Hierarchically Structured Data. Journal of statistical software. 2014;59(13):1–21. doi:10.18637/jss.v059.i13.~~

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practicaland Powerful Approachto Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

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.

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.

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.

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.

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.

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.

12. 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.

13. Aitchison J. A Concise Guideto Compositional Data Analysis. 1999; p. 134.

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

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.

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

14. 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.

15. 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.

16. 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.

17. 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.

18. Irizarry RA, Wang C, Zhou Y, Speed TP. Gene Set Enrichment Analysis Made Simple. Statistical methods in medical research. 2009;18(6):565–575. doi:10.1177/0962280209351908.

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

20.

21. 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.

22. 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.

23. 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.

24. 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.

25. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing Microbial Composition Measurement Standards with Reference Frames. Nature Communications. 2019;10(1). doi:10.1038/s41467-019-10656-5.

26. 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.

27. 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.

28. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for Normalizing Microbiome Data: An Ecological Perspective. Methods in Ecology and Evolution. 2019;10(3):389–400. doi:10.1111/2041-210X.13115.

29. Aitchison J. Principles of Compositional Data Analysis. Lecture Notes-Monograph Series. 1994; p. 73–81.

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

31. 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.

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

33. 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.

34.

35. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, ~~Mateu-FigurasG~~Paredes R, ~~Barcelo-Vidal~~Noguera-Julian ~~G~~M, Calle ML. ~~Isometric~~ Balances: A ~~Logratio Transformations~~New Perspective for ~~Compositional Data~~ Microbiome Analysis. ~~Mathematical Geology. 2003;p. 22.~~ mSystems. 2018;3(4):e00053–18. doi:10.1128/mSystems.00053-18.

36. ~~Silverman JD, Washburne AD, Mukherjee S, David LA~~

37. Wu D, Smyth GK. ~~A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data~~Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. ~~eLife. 2017;6:e21887~~Nucleic Acids Research. 2012;40(17):e133. doi:~~10.7554~~10.1093/~~eLife.21887~~nar/gks461.

38. ~~Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic~~

39. Egozcue JJ, Pawlowsky-Glahn V, ~~RNA~~Mateu-Figueras ~~Interference Reveals That Oncogenic~~ G, Barcel*KRAS*ó-Vidal C. Isometric Logratio ~~Driven Cancers Require~~ Transformations for ~~TBK1~~Compositional Data Analysis. ~~Nature. 2009;462(7269):108–112~~Mathematical Geology. 2003;35(3):279–300. doi:~~10.1038~~10.1023/~~nature08460~~A:1023818214614.

40. 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.

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

42. 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.

43. 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~~p. 26.

44. Silverman JD, Washburne AD, Mukherjee S, David LA. A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data. eLife. 2017;6:e21887. doi:~~10.7717~~10.7554/~~peerj.2969~~eLife.21887.

45. 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):e00162–16. doi:10.1128/mSystems.00162-16.

46. Aitchison J, Shen SM. Logistic-Normal Distributions:Some Properties and Uses. Biometrika. 1980;67(2):261–272. doi:10.1093/biomet/67.2.261.

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

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

49. Kurtz ZD, M10.1093/nar/gks461üller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA.

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

Cario MC. Modeling and Generating Random Vectorswith Arbitrary Marginal Distributionsand Correlation Matrix. 1997; p. 19.

Agresti A, Coull BA. Approximate Is Betterthan "Exact" for Interval Estimationof Binomial ProportionsMicrobial Ecological Networks. The American Statistician. 1998;52(2):119–126PLOS Computational Biology. 2015;11(5):e1004226. doi:10.230710.1371/2685469.

DeLong ER, DeLong DM, Clarke-Pearsonjournal.pcbi.1004226DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–845.

50. Hawinkel S, Mattiello F, Bijnens L, Thas O

51. Ackermann M, Strimmer K. A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery RateGeneral Modular Framework for Gene Set Enrichment Analysis. Briefings in Bioinformatics. 2019;20BMC bioinformatics. 2009;10(1):210–221. doi:10.1093/bib/bbx104. 1–20.

52. Breiman L

53. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Random ForestsToward a Gold Standard for Benchmarking Gene Set Enrichment Analysis. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324. Briefings in bioinformatics. 2021;22(1):545–556.

54. Kuhn M, Wickham H. Tidymodels: EasilyInstall and Load the 'tidymodels' Packages; 2020.

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

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.

55. 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.

56. 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.

57. 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.

58. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi:10.21105/joss.01686.

59. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. BMC Bioinformatics. 2011;12:77.

60. 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.

61. Ernst FGM, Shetty SA, Borman T, Lahti L. Mia: Microbiome Analysis; 2021.

62. Landau WM. The Targets R Package: A Dynamic Make-like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. Journal of Open Source Software. 2021;6(57):2959.

63. 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.

64. 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.

65. 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.

66. 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.

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

68. 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.

69. 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.

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

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

72. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321–357. doi:10.1613/jair.953.

73. Kuhn M, Wickham H. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.; 2020.

74. 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.

75. ~~Naim I, Gildea D~~

76. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. ~~Convergence of the~~ Systematic ~~EM Algorithm~~RNA ~~for~~ Interference Reveals That Oncogenic ~~Gaussian Mixtures~~*KRAS*~~with~~ -Driven Cancers Require ~~Unbalanced Mixing Coefficients~~TBK1. ~~Proceedings of the 29th International Coference on International Conference on Machine Learning. 2012;p. 8.~~ Nature. 2009;462(7269):108–112. doi:~~10.5555~~10.1038/~~3042573.3042756~~nature08460.

77.

78. 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.

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

80. 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.

81. 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.

82. Weissman JL, Dogra S, Javadi K, Bolten S, Flint R, Davati C, et al. Exploring the Functional Composition of the Human Microbiome Using a Hand-Curated Microbial Trait Database. BMC Bioinformatics. 2021;22(1):306. doi:10.1186/s12859-021-04216-2.

Supporting information

S1 Fig. ~~Computational performance of cILR.~~ Simulation results for type I error evaluation for CBEA sample-level inference. ~~Computational time (in seconds) as a function of sample size (left panel)and number of taxa sets evaluated (right panel). Evaluation was performed on simulated data sets. For sample size analysis, only 10 sets were evaluated. For taxa set analysis, sample size was fixed at 1, 000. Across all evaluations, the size of each taxa set was also fixed at 50.~~ Placeholder text.

S2 Fig. ~~Distribution of p-values~~Simulation results for type I error evaluation for CBEA population-level inference. Placeholder.

S3 Fig. Simulation results for phenotype relevance evaluation for CBEA sample-level inference. Placeholder text.

S4 Fig. Simulation results for phenotype relevance evaluation for CBEA population-level inference. Placeholder text.

S5 Fig. Simulation results for predictive pefromance evaluation for CBEA. Placeholder text.

S6 Fig. Computational performance of CBEA. ~~Q-Q plot of 10,000 p-values compared against a uniform distribution bounded between 0 and 1. Evaluation was performed on simulated null data sets of 10,000 samples testing for enrichment of a set of size 50. For correlation of 0.5, p-values represent correlation adjusted cILR while for correlation of 0, p-values represent unadjusted cILR.~~Placeholder text.

S7 Fig. Distribution of type I error values across all replications in real data random set evaluations for CBEA inference at the sample-level. Placeholder text.

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

CBEA, simulation analyses, and run-time performance.