cILR: Competitive isometric log-ratio for taxonomic enrichment analysis

PCOMPBIOL-D-21-01648

Response to Reviewers

October 26th, 2021

We greatly appreciate both reviewers for thorough and insightful reviews of our manuscript. We think that you will find its quality much improved as a result of the changes we've made in response. Here, we summarize the major changes, with point-by-point responses to each reviewer comment following:

1) We adjusted the title of our manuscript to better reflect the proposal and its relationship with existing methods. The title of the manuscript will now be "CBEA: Competitive balances for taxonomic enrichment analysis"
2) Throughout, we have provided more clarity and precision in the language surrounding statistical concepts and results.
    a. When discussing our results, we provide additional context of our experimental conditions to ensure proper interpretation and avoid overstatements.
    b. We improved precision in the language surrounding the idea of zero-inflation and compositionality of microbiome data, highlighting the motivation behind our approach and the assumptions we're making.
    c. We added more clarification on the motivation behind our procedure for adjusting inter-set correlation in our null hypothesis inference section. We emphasized the trade-off between type I error and power for enrichment analysis as a discovery tool and provided and clarified that the user can make the choice whether to adjust for correlation when using our method.
3) We reorganized the manuscript to highlight our main contribution which is a new approach for set-based testing for relative abundance microbiome data. This organization is also motivated by the work of Geistlinger et al. [1] on benchmarking standards for gene set testing approaches as suggested by Reviewer 2. We believe this reorganization of the manuscript will make it easier to follow by eliminating redundant material and consolidating the presentation of results.
    a. We agree with Reviewer 1 that enrichment analysis is equivalent to differential abundance for sets and have grouped the "single-sample enrichment testing" and "differential abundance analysis" into one section titled "enrichment analysis". The "single-sample enrichment testing" section is now referred to as "Inference at the sample level" while the differential abundance section is now titled "Inference at the population level".
        i. We added additional evaluations on the real data set for the "single sample enrichment testing" section which are: random [gene] sets and label permutation. We added the random [gene] set analyses for the real data portion of the "differential abundance analysis" section. These additional analyses were motivated by Geistlinger et al. Consequently, we have moved the simulation results for this section to the supplemental.

**Commented [QN1]:** Add references to specific comments below. Remember to re-include where you're adding the section about differential abundance methods

**Commented [QPN2]:** New analysis adding random gene and label permutations for the real data section of single sample enrichment tests (or inference at the sample level).

b. We have also added a new section titled “downstream analysis” where we would expand on using our single sample scores for disease prediction (the “disease prediction” section). We also added a new section on “ordination analyses” using our single sample scores.

c. We have revised the discussion section to highlight the difficulty of evaluating power/phenotypic relevance of enrichment methods and to clarify how our evaluation data set can address those goals.

**Commented [QPN3]:** New analysis here so that disease prediction doesn’t seem to be alone in downstream analysis. This analysis will be using cILR generated scores to perform PERMANOVA and visualizations to distinguish between cases/controls using the Euclidean distance metric.

We believe these changes in response to reviewer comments have made the manuscript more focused and better aligned with previous benchmarking standards, while additionally providing the needed clarity and precision of language around complex statistical terminology. Our revisions also highlight the main contribution of the manuscript which is fundamentally a single sample set enrichment analysis approach (in the same vein as ssGSEA).

**Reviewer 1:**

Summary

In this manuscript Nguyen et al propose cILR for a set-enrichment-like analysis of microbiome sequencing data. This is a scale invariant alternative to the more standard (and problematic) approach of applying GSEA to DESeq2 output or other such approaches. Overall, it’s a nice idea and is fairly well executed. The statistical analysis is largely appropriate, and it could be a nice contribution. My largest comments relate to the lack of precision in the authors' writing, the extent of the authors claims, and the authors handling of zeros and count variation.

We greatly appreciate the reviewer’s detailed comments and suggestions for improving the manuscript. We have made various changes to the manuscript to clarify the concepts we are advancing and provide more context while interpreting our results to avoid confusion. In the following sections we provide detailed responses to each of the reviewer’s concerns.

Detailed Comments

1. Precision of Writing:

   A. The authors repeatedly state this data is "compositional" or "strictly compositional". First, what is "strictly compositional" is there non-strictly compositional data? Second, the data is clearly not compositional; this is a cliche that has been amplified in the literature and is incorrect. The data is count data, it has zeros and is integer valued, both of those features are in strict contrast to the standard definition of compositional data (i.e., positive multivariate continuous data that sums to a constant value and typically is an open set excluding zero to avoid issues with log-ratio transforms). This distinction is non-trivial as the direct application of log-ratio transforms to this data is poorly motivated in this case and fraught with problems (See later comments on handling of the data).

   We would like to thank the reviewer for the comments with respect to the clarity our claims around the compositional nature of the data. We agree that the terminology used in the manuscript can be confusing, and we provide some commentary below with regards to our thought process on framing the issue around compositional data. This is also a response to the comment from the reviewer below with regards to the statement about the constraints of the sequencing instrument making sequencing data compositional.

   We agree with the reviewer that we cannot get absolute abundances of features measured via sequencing data and that count data alone is not compositional according to Aitchison’s definition

[2]. However, despite efforts to sequence equimolar amounts of DNA as the reviewer noted, the total number of sequences per sample (i.e., library size) still varies significantly across samples [3,4] and requires applying normalization approaches to ensure that abundances can be compared across samples [5]. The most used normalization approach is to transform counts into proportions using the total library size per sample as the denominator. Even though some researchers suggest using existing methods in the gene expression literature [6], some of the assumptions that underlie these approaches might not match that of microbiome data. For example, DESeq2's median of ratios method (in the function *estimateSizeFactors*) assumes that the majority of genes do not differ in expression levels across samples. Other studies have also empirically compared different normalization methods, where transformation to proportion is usually the best choice [7]. As such, we mistakenly use the term "strictly compositional" to refer to the fact that microbiome sequencing data, unlike other sequencing data sets, generally prefers a transformation to proportions prior to analysis.

In the case where researchers transform count data into proportions, then the data becomes composition as a sum constraint has been imposed. Even though there are zeroes in the composition which does not fit Aitchison's definition, the imposed sum to one constraint still induces spurious negative correlation between the variables, where log-ratio based methods are well motivated as solutions [3]. Our approach is conceptually a log-ratio based method for aggregating compositional variables. This concept is not novel as it has been advanced prior by the original authors of the ILR transformation [8], where it was termed as balances between groups of parts. Our contribution is towards specifying the "groups" that has a specific interpretation similar to that of the competitive null hypothesis in the gene set testing literature.

We hope that the discussion above has clarified the statistical motivations of our approach. We have restructured the introduction section of the manuscript to reflect this and have amended certain potentially confusing terms such as "strictly compositional" (lines X-Y).

B. What the authors propose is not an ILR transform. Unless I am mistaken, there is no constraint on the matrix A such that the coordinate system is cartesian with an orthonormal basis. In fact, if k does not equal p-1 then it cannot possibly be isomorphic let alone isometric with respect to the Aitchison metric. Unless I am mistaken, the authors should change the name of their method and modify their discussion to be more accurate. I would relate this method is not an ILR transform but it is very similar to phylofactor which takes a similar approach (in phylofactor set membership is dictated by the topology of the tree).

We agree with the reviewer that the method itself is not an ILR transform (as it did not propose a novel binary partition such as PhILR [9] ) and was therefore mislabeled. Our approach still leverages the concept of balances between groups of compositional parts related to the ILR transformation as advanced by the original authors [8]. As such, our approach will be renamed "Competitive compositional balances for taxonomic enrichment analysis" (CBEA). We hope that this new name more clearly reflects the specific advances our method is proposing. For the remainder of this response, we will still refer to the approach as cILR for clarity purposes, though we have changed the naming throughout the manuscript

C. The authors state that the "data is zero-inflated" this is another cliche that I would encourage the authors to remove. Zero-inflation is a particular family of models for these zeros not a objective characteristic of the data. Simply saying there are many zeros would likely suffice in this article. They could argue that the data generating mechanism is well represented by a zero-inflation process, but this has been called into question (see Silverman et al. Naught all zeros in sequence count data are the same.)

We agree with the reviewer that the term "zero-inflated" should be used in reference to the specific class of models instead of as a catch-all term for a characteristic of the data. Since we are

agnostic to the mechanism behind the process of generating zeroes, we have amended the article to use “zero abundant” or “sparse”.

D. The authors state that the data is compositional because the number of reads obtained is constrained by the sequencing instrument. Would an instrument that didn't have this constraint lead to "non-compositional data"? This seems unlikely. For example, standard equimolar pooling protocols explicitly dilute concentrated DNA from each sample to try to equalize sequencing depth. It’s not just an issue of the sequencer. Even sampling from an environment (e.g., taking 5 grams of stool from a larger stool sample) loses the notion of absolute abundance.
We agree with the reviewer that the terminology is confusing and have added additional clarity to the manuscript. We have provided a more comprehensive response on the issue of the compositional nature of microbiome data in the above section (response to comment A), which we believe also addresses the issues raised in this comment.

E. The GSEA method cited on line 51 is not a random-walk like statistic. I think it may be a Brownian bridge but its constrained to be zero at either end -- not a random walk.

We thank the reviewer for the clarification. The “random-walk like statistic” phrase has been clarified and amended in the introduction section (lines X – Y).

F. Between lines 73 and 85 the authors do not properly motivate the multiplicative rather than additive amalgamation. They mention the downsides of the "naive summation-based method" but this is unclear. From later in the manuscript I gather that this statement reflects the perturbation invariance of multiplicative amalgamation: given that some have argued that measurement bias can be modeled as a constant compositional perturbation. This needs to be made explicit. There is no inherent downside of summation (i.e., additive amalgamation) - its a modeling choice and it is not "naive".

We have reworked the introduction section (lines X – Y) to highlight the differences more clearly between product and sum-based aggregations and provide a robust justification for our approach. We have also removed the characterization of sum-based amalgamations as “naïve” as suggested.

G. The authors mention "adjusting for correlation" multiple times throughout the manuscript yet the motivation is not properly clarified. The best I can guess is that they are saying that they need to modify the null-hypothesis to account for a trivial case where something looks differential expressed or set enriched when really it’s just due to the correlation structure between taxa. That said, I think there are many potential sources of confusion that the authors should clarify. Couldn't set enrichment be reflected in those correlations? Isn't the correlations actually a non-trivial part of what the authors are trying to model? In other words if a set of microbes is highly correlated wouldn't that be a sign that that set is potentially enriched or de-enriched? I don't think I understand this point completely but I think it is likely non-trivial. I would encourage the authors to clarify the role of correlation.

We thank the reviewer for pointing out that the motivation behind “adjusting for correlation” was not clearly communicated in the manuscript. We have amended the “Statistical properties” section to provide more commentary on this concept (lines X – Y).

Additionally, we agree with the reviewer that there are situations where highly correlated sets are biologically relevant. As such, we have provided more commentary in the same section (lines X-Y) with regards to that issue and have left the decision whether to adjust for correlation to the user. This also supports the notion (as also recommended by reviewer 2) that set-based analysis is usually exploratory rather than confirmatory so an inflated type I error may be acceptable to achieve higher power.

H. On lines 167-170 the authors state that since the cILR are not orthogonal a correlation can exist between cILR aggregated variables. This is misleading there can be correlation whether or not the cILR's are orthogonal or not, orthogonality and lack of correlation are separate concepts.

The relevant section (lines X – Y, “Statistical properties” section) have been revised to correct for this misconception.

I. Citation on line 187. I don't see how this paper supports this statement. Egozcue et al. take a almost purely mathematical approach as far as I can tell do not discuss central limit theorems or other things that are implied by the authors statement. If I remember correctly the relevant citations are authored by Aitchison while I cannot remember them exactly.

We agree with the reviewer that the source on line 187 did not discuss the distributional properties of cILR. We have amended the citation with the source from Aitchison and Shen [10] which provides details on the logistic normal distribution for compositional data and the original source from Egozcue [11] which talks about the relationship between ILR coordinates and the ALR coordiates that motivated the logistic normal distribution mentioned above.

J. On line 536 the authors mention "inflated counts", I have no idea what this means.

We have added more clarification beforehand in the “Methods” section to detail the meaning of “inflated counts”. In essence, “inflated counts” refers to when sets (or individual taxon) have fold change increase in absolute counts in a certain condition (e.g. IBD) compared to control. This is equivalent terminology to refer to taxa that are differentially abundant across conditions, but also refers specifically to the mechanism of abundance difference (fold change).

K. Lines 572 to 582. I don't understand how this hypothesis makes sense. How does taxa-specific bias relate to the performance of DESeq2 or corncob? The writing here is poor. Also, I am not sure how this could be, are you not basing the gold-standard truth off of permuted data which you know has no signal? This permutation would maintain the measurement bias ... as a result it would seem the data does not support this hypothesis. I expect I am missing something.

We would like to thank the reviewer for bringing this to our attention. In our results, we observed some surprising differences in the performance of DESeq2 and corncob between simulations and the real data analysis. In the simulation analysis, these methods show low type I error and low power, while conversely in real data analyses (i.e., the permutation analyses) these methods show high type I error and high power (when compared against cILR). In the section from lines 572-582, we explained this phenomenon by hypothesizing that this discrepancy might be due to taxa-specific biases. According to McLaren et al. [12], sum-based aggregations are particularly sensitive to this type of bias. In the permutation analyses using real data this bias would be preserved as the reviewer has stated, which explains that high type I error observed when applying DESeq2 and corncob to simulations where this aspect of the data was not considered.

2. Modeling Choices

A. As far as I can tell the authors do not state how they are handling zeros. This is a non-trivial methodologic detail especially if they are simply taking log-ratio transforms of count data. To what extent is the non-normality of the authors results simply a product of directly transforming count data without accounting for the variance of the counts. For example, count data typically have a mean variance relationship that seems largely ignored by the authors approach. Moreover, there has been a number of advances in compositional modeling of microbiome focusing on Multinomial logistic normal models that are not addressed by the authors. In fact in light of the availability of these methods the authors modeling of these counts seems sub-par.

This section of our analysis lacked clarity. Our strategy for addressing sparsity in microbiome data is to use pseudocounts to ensure the validity of the log-ratio transformations. We stated this assumption more clearly in the "statistical properties" section of the revised manuscript. We also acknowledged in our discussion section on the limitations of the approach and mentioned alternative methods that users can apply prior to running CBEA. However, according to our experimental results, the performance of our approach was not significantly affected by data sparsity levels.

As clarified above, we do not model the count data directly but rather model the relative proportions that result from a total sum normalization of the counts. Our approach is to perform an ILR-like transformation to the proportions corresponding to the set annotation and perform inference through empirically modelling the test statistic under the null. Our simulation studies have demonstrated that the empirical distribution of our test statistic is well approximated by a normal distribution. Furthermore, real data analyses also show that normal approximation generate good performance values for all considered situations. We have updated the "statistical properties" section in the revised manuscript to better outline this modelling approach and discuss how statistical features of the underlying count data (e.g. sparsity and mean-variance relationship) may impact the distribution of the transformed proportions.

As the reviewer noted, multinomial logistic normal models are useful for modelling the count data direction, and it may be feasible to apply the multinomial logistic normal model to perform set-based enrichment analysis. Although we are not away of any existing approaches that utilizes this distribution for set-based testing, this is an interesting idea that we hope to explore in future research.

3. Unsubstantiated Claims
There are a number of unsubstantiated claims where the language needs to be altered to be more precise.

A. Line 493: "These results demonstrate that cILR generated scores are informative features in disease prediction tasks." No. These results demonstrate that cILR COULD be informative features in disease predictions tasks. I am not convinced that these are even useful for the case-studies shown in this manuscript let alone other tasks. Moreover, the comparison methods ssGSEA and GSVA seem like odd choices. Are the authors only using methods that can take set-based features? This does not account for the potential that the chosen sets are not informative. The later seems like an important case to establish the motivation of the current work.
This section is motivated by the fact that cILR generates scores at the sample level, performing as a transformation of a $n \; x \; p$ matrix of $p$ taxa and $n$ samples into a $n \; x \; m$ matrix of $m$ sets and $n$ samples. As such, we compare cILR against similar approaches such as ssGSEA and GSVA, which also calculates enrichment scores per sample. ssGSEA and GSVA provides a model-based approach to generate set-based features using the original matrix and set annotation as inputs. For predictive analyses, we can fit a model (in our case, a simple random forest) to these scores to perform predictive analysis using set-based features.

As the reviewer pointed out, the predictive capacity of set-based features would be low if the chosen sets are non-informative or not interesting. In this manuscript we are agnostic as to how the sets are constructed and whether there is a performance increase using sets compared to using the basic features. What we demonstrated in the manuscript is the relative performance of the different approaches to aggregation in instances where the researcher decides aggregation is of interest. As such, our claim that "cILR generated scores are informative features" refers to the fact that given the same sets of microbes, scores constructed by cILR can be informative towards prediction compared to similar approaches, suggesting that it is valid to use set-based features generated using cILR for prediction purposes. However, we agree with the reviewer that this is a

strong statement and have adjusted it accordingly. We also added the context provided in this response to the results interpretation in the manuscript.

B. Line 535-537. The authors show that their model displays Type 1 error control on a set of of simulated datasets. They make some claims about false-discovery control on real data on lines 397-406 but I really don't follow how they know what is non-random or random on this dataset. It seems like they have a strong hypothesis about aerobic vs. anaerobic but that hypothesis seems too weak to serve as a gold-standard reference. Overall these claims are unsubstantiated. I would emphasize that any claim saying a model can be trusted is suspect and bordering on overtly false -- any model can fail and nearly all models are mis-specified there are times at which a model may be useful but that is about it. No model can be globally trusted.

We thank the reviewer for raising this issue. For the simulation results, we updated the manuscript to specify the precise situations of our experiments and to clarify that these scenarios represent only a subset of the data sets that users may encounter in practice. As such, we have modified our performance claims (e.g., lines X – Y).

In terms of our gingival data set, we agree that the aerobic vs anaerobic hypothesis is not strong enough to serve as ground truth. We have clarified and provided further discussion on the lack of standardized gold-standard data sets for enrichment testing in the discussion section. Furthermore, we have added additional type I error results on the real data set (following standards set by Geistlinger et al. [1] – as recommended by reviewer 2). However, we maintain that the results still provide good insight into model performance since the hypothesis does have a clear and straightforward biological interpretation (i.e. based on easy to determine natural characteristics of the microbes) and has been used in prior manuscripts that attempts to validate differential abundance analyses for microbiome data [13].

**Commented [QN4]:** Statement referring to new analyses

4. Other Comments on Clarity

A. The writing after line 215 lacks detail. I kept waiting for a methods section to answer some of my questions (e.g., how was mu or phi chosen in equation 3) but these don't seem to be listed anywhere. Details in the remaining parts of the manuscript are inconsistently given or vague. e.g., Line 249 "all sample sizes were set to 10,000". Do you mean sequencing depth? Number of reads? Number of technical replicates? What is this referring to?

We have adjusted the "Evaluation" section of the manuscript to improve readability and fill in missing gaps in our evaluation methodology. With regards to the specific examples provided by the reviewer, we provide some clarifications as follows:

1) The $\mu$ and $\phi$ parameters were chosen by fitting a negative binomial distribution (using maximum likelihood approach with the *fitdistrplus* package [14]) on non-zero entries in each taxon in the human microbiome 16S data set. The median values across all estimates were chosen as the final estimates for the simulation procedure.

2) 10,000 samples refers to the number of samples (i.e. the number of biological replicates). Since we're attempting to perform inference per sample (assign a p-value per sample), this is equivalent to the number of hypotheses tested for our enrichment analysis procedure.

B. The writing after line 215 is hard to read. In part this relates to the lack of detail but I think it also stems from the fact that the manuscript starts being written in triplicate for Single Sample Enrichment, Differential Abundance Analysis, and then Prediction. It makes the paper repetitive and hard to follow. Further, figures are repetitive and poorly labeled so it's hard, at a glance, to figure out what figure links to what part of the paper. I have never seen a discussion written in the parts but this just adds to the feeling that this is just a paper written in triplicate without a coherent message beyond the initial idea which ends around line 215. Further it was not clear from reading the introduction that the paper would be organized like this; some warning in the introduction may help a bit. In fact, it was not even clear what the distinction between single-

sample enrichment and differential abundance was from the introduction. In addition, this notation is non-standard. Typically enrichment (e.g., as used in gsEa) refers to essentially differential expression but for sets of genes (i.e., it is typically a comparison between groups). This makes the "single-sample enrichment" terminology confusing.

We now include an overview of the manuscript in the introduction. Additionally, we restructured the manuscript to feature the enrichment analysis more prominently and provide clarification on the specific meaning of each section. We also adjusted figures labels and captions to help distinguish across different sections.

**Reviewer 2**:

Summary

Nguyen et al. present a new method for taxonomic enrichment analysis of microbiome data based on an isometric log-ratio transformation of compositional and the competitive null hypothesis borrowed from the gene set enrichment literature. The main strengths are a well-written and structured manuscript, a solid statistical and analytical foundation of the method, and a thorough evaluation of the method on simulated and real datasets. The main weaknesses are installation issues with the R companion package, a lack of adaptation of existing standards for the benchmarking of gene set enrichment methods, and a number of theoretical considerations with the use of a competitive null hypothesis for enrichment testing.

We would like to thank Reviewer 2 for the detailed and encouraging responses. We have made structural changes in the manuscript (as described in the preamble section of this response) to make sure we acknowledge the standards set out by Geistlinger et al. 2021 [1]. Additionally, we have added a discussion on adjusting for correlation, and have also fixed the R package.

1. Installation:

Using a recent R installation (R.4.1.0) and Bioconductor installation (3.13), I was not able to install the package. The error message and my session info is included below. The method looks useful for the community and I would strongly encourage a Bioconductor submission (or at least a CRAN submission) of the package to ensure that the package passes R CMD build, check, and install in a continuous integration setup. (Detailed error message that was attached by Reviewer 2 was excluded from this response for clarity)

We have provided an updated version of the package and submitted it to CRAN/Bioconductor (issue link). The current in development version on GitHub has passed R CMD CHECK --as-cran on Windows, MacOS, and Linux (Ubuntu 20.08) via GitHub Actions. If there are any installation issues, please let us know.

**Commented [QN5]:** Remember issue link here

2. Adapting standards for the benchmarking of enrichment methods:

Geistlinger et al. (doi: 10.1093/bib/bbz158) has recently introduced an extensible framework for reproducible benchmarking of enrichment methods based on defined criteria for applicability, gene set prioritization and detection of relevant processes. This setup consists of compendia of curated and standardized datasets and a number of criteria that apply as-is also for new enrichment methods in the microbiome data realm (such as runtime, proportion of rejected null hypotheses, behavior on permuted sample labels and random gene sets). Although I would really like to commend the authors for using curated and standardized datasets from curatedMetagenomicData and HMP16SData, the authors then proceed with the practice of self-assessment over various scenarios which is typically difficult to transport and apply for new methods. Being one of the first methods for enrichment analysis in the microbiome realm (but very likely not the last one), the paper has the opportunity to very early on set the baseline for how new enrichment methods in the microbiome space should be evaluated building on lessons learned in the gene set enrichment literature. This could be achieved (a) clearly communicating the existence of such

standards, (b) adapting existing standards where possible, and (c) to point out where adaption of such standards would require further work, as there might well be criteria that do not straightforward translate from gene set enrichment to taxon set enrichment.

We thank the reviewer for directing us to the paper by Geistlinger et al. [1]. We agree that set-based approaches from the microbiome field should learn and adapt from existing standards from the gene-set testing literature. After consulting Geistlinger et al, we noticed that many of the existing sections of the manuscript already correspond to the structure recommended in the Geistliner et al. paper. As such, we have reorganized the manuscript to properly communicate the relationship between our evaluation strategy and the standards set by Geistlinger et al. Changes are as follows:

1) We have combined the "differential abundance" and "single sample enrichment testing" sections into one section titled "enrichment analysis". Under this section, "single sample enrichment testing" is now "inference at the sample level", and "differential abundance" is now "inference at the population level".
2) We created a new section titled "downstream analyses" in order to distinguish the primary goal of our approach (which is enrichment testing) from secondary goals of providing sample-level scores that have utility in further analyses. This section now includes the "disease prediction" section and a new section titled "ordination analysis".
3) We have added evaluations for type I error on real data using label permutation and random gene sets for both "inference at the sample level" and "inference at the population level" and have labelled them accordingly.
4) Under both "inference" sections, we have clarified that our power analyses are equivalent to the "phenotypic relevance" mentioned under the Geistlinger et al. manuscript since they assess whether the correct sets were enriched in a certain situation. However, we acknowledge that the sample label is not perfect and have added further clarification in the discussion regarding the differences between the evaluation strategy we used and that of Geistlinger et al., as well as further discussion on the current limitations of the microbiome literature on this issue.
5) We have added specific language to refer the reader to our runtime assessments in the supplementary materials section.

We hope this reorganization will enable more a direct comparisons between existing standards in the gene set testing literature and our own evaluation strategy.

3. On the use of the competitive null hypothesis for enrichment testing:

The authors demonstrate that cILR controls for type I error even under high sparsity and high inter-taxa correlation. However, it has been pointed out that strict type I error rate control might not be a desirable feature for enrichment methods (Goeman and Buhlman, 2009; Wu and Smyth, 2012; Geistlinger et al. 2021). Gene set enrichment analysis is an exploratory process, not a confirmatory, diagnostic process, where strict type I error control augments the lack in power which is well documented for competitive enrichment testing (Goeman and Buhlman, 2009; Wu and Smyth, 2012; Geistlinger et al. 2021) and as the authors demonstrate in their own evaluations. Furthermore, Geistlinger et al. 2021 (Figure 4 therein) has demonstrated that despite controlling the type I error rate, methods might demonstrate widely different rejection rates on real datasets. It is in this context noteworthy that the authors of Camera (Wu and Smyth, 2012), which deliberately abandons strict type I error control by default to compensate for the apparent lack in power of competitive methods.

We agree with the reviewer on the overall goal of enrichment analysis and the discussions presented in the existing literature on the trade-off between power and type I error control. We agree that detecting highly correlated taxa sets can have biological importance (as discussed in Wu and Smyth [15]) and have clarified that the decision for strict type I error control is up to the user. We have also provided additional discussion in the "Statistical properties" section about this topic, including a further clarification on the motivation behind adjusting for inter-taxa correlation.

**Bibliography**


1. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Briefings in Bioinformatics. 2021;22: 545–556. doi:10.1093/bib/bbz158

2. Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological). 1982;44: 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x

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

4. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018;34: 2870–2878. doi:10.1093/bioinformatics/bty175

5. 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. doi:10.1186/s40168-017-0237-y

6. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLOS Computational Biology. 2014;10: e1003531. doi:10.1371/journal.pcbi.1003531

7. 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: 389–400. doi:10.1111/2041-210X.13115

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

9. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. Fodor A, editor. eLife. 2017;6: e21887. doi:10.7554/eLife.21887

10. ATCHISON J, SHEN SM. Logistic-normal distributions:Some properties and uses. Biometrika. 1980;67: 261–272. doi:10.1093/biomet/67.2.261

11. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology. 2003;35: 279–300. doi:10.1023/A:1023818214614

12. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Turnbaugh P, Garrett WS, Turnbaugh P, Quince C, Gibbons S, editors. eLife. 2019;8: e46923. doi:10.7554/eLife.46923

13. 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: 191. doi:10.1186/s13059-020-02104-1

14. Delignette-Muller ML, Dutang C. fitdistrplus: An R package for fitting distributions. Journal of Statistical Software. 2015;64: 1–34.

15. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40: e133. doi:10.1093/nar/gks461