cILR: Competitive isometric log-ratio for taxonomic enrichment analysis

PCOMPBIOL-D-21-01648

Response to Reviewers

October 26th, 2021

**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 reviewer 1’s detailed comments and suggestions for improving the manuscript. In the following sections we provide responses to each of the reviewer’s concerns.

Detailed Comments

1. Precision of Writing:

* 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 of the claim that microbiome data is compositional. We agree that the terminology used in the manuscript can be confusing, however we maintain the belief that microbiome relative abundance data is compositional.

First of all, we contend that the limitation of the sequencing experiment is evidence to suggest that the data is compositional.

Second of all, we also note that

* 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 [1] ) and was mislabeled. However, our approach still leverages the concept of balances between groups of compositional parts

* 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 a catch all term for this characteristic of the data. Since we are agnostic to the mechanism behind the zero-inflation process, we have amended the article to use “zero abundant”, which hopefully is better at distinguishing the two concepts.

* 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 first comment), which we hope also answers some of the issues raised in this comment.

* 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).

* 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 agree with the reviewer that labelling additive amalgamation as “naïve” is a mischaracterization.

* 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 agree with the reviewer on the simplification of the issue of correlation

* 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 a lack of correlation are separate concepts.

We agree with the reviewer’s comment. We have amended the “Statistical Properties of “ section in Methods against this misconception.

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

1. Modeling Choices

* 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.

1. Unsubstantiated Claims

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

* 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.
* 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.

1. Other Comments on Clarity

* 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?
* 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.

**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.

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. The current in development version on GitHub have passed all R CMD CHECK on Windows, MacOS, and Linux (Ubuntu 18.08) via GitHub Actions. If there are still any installation problems, please let us know!

1. 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.

1. 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.

Bibliography

1. 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