

Reviewer #1

The authors describe a few different log-ratio transformations designed for use with compositional microbiome data, and implemented in the ALDEx2 package. As far as I know, this marks the first time they are described in the literature. I think these transformations, especially the inter-quartile log-ratio transformation, are important contributions to the field. However, I have a few concerns about the way they are contextualized with the compositional data literature.

Notably, I think there is some confusion in the text about what exactly the CLR does. The authors' proposed use of the CLR (and of the IQLR, etc.) is very specific: they wish to use the log-ratio transformation to normalize relative data such that it becomes possible to make inferences about the absolute data. This, of course, requires that the denominator is invariant. But there are two facts about this use case that the authors have not addressed: (a) It is not necessary to have an invariant denominator to perform a solid analysis of compositional data; and (b) Identifying an invariant reference is conceptually similar to the process of effective library size normalization used by edgeR, DESeq2, etc., and thus carries with it similar limitations: that differential abundance estimates can be wildly wrong when the chosen reference failst to "find the centre".

The search for an invariant denominator requires certain assumptions. These assumptions are not unlike those used in effective library size normalization, i.e., that the majority of features remain unchanged. I think the connection between CLR/IQLR and TMM/etc. should be discussed explicitly. (In my opinion, this actually supports the validity of the ALDEx2 approach, rather than undermines it.) A good discussion of the similarities and differences between CLR and TMM can be found in <https://doi.org/10.1093/bioinformatics/bty175> under Section 4.3 "The log-ratio 'normalization', and in the associated Supplementary Material.

Otherwise, I found it difficult to ascertain how exactly the simulations were performed. Some key details are missing or ambiguous. Please ensure that the simulations are described in enough detail to be reproduced from the text!

My comments,

(1) Section 1, the authors write "use a negative binomial or zero-inflated Gaussian model, and assume the features are independent and identically distributed for statistical tests, and that the majority of features are invariant". I'm not sure what the authors mean by "identically distributed", or whether this is technically true. Could the authors clarify what they mean?

clarified the intention of the statement, identified the majority tools and their distributions

Also, it is probably more correct to list these two distributions as examples, since there are many possible distributions that an analyst might use.

(2) Section 1, the authors write "there is no a-priori reason for feature counts to follow any particular statistical model". Although I agree that this is probably true, I think this comment would benefit from having a supporting reference.

Removed the metagenomic allusion as it was not necessary

(3) Section "A ratio based approach", paragraph 2.

I take issue with this paragraph. The authors write that "[the CLR] requires that the denominator used to calculate the CLR be comparable across all samples." This is not true. This requirement only exists if *you wish to use the CLR to perform effective library size normalization*. The CLR confers other benefits to an analysis, such as the sub-compositional dominance of distances and scale invariance, regardless of whether the geometric means are "comparable across samples".

The authors also write “Thus, one shortfall of any ratio-based approach, is to determine the appropriate denominator”. As discussed above, this is only a shortfall if you intend to use the denominator as an invariant feature. I recommend the authors clarify this nuance.

(4) Section “A ratio based approach”, paragraph 3.

The authors write “The asymmetry will also affect the scale-invariance of the data since a value of 0 is not scaled when multiplied by a constant”. I’m not sure if I agree with this statement. I know what the authors mean, but I don’t know if it’s correct to say that sparsity causes a problem specifically with scale-invariance. What I mean is that the ratio $0/4$ and $(a*0)/(a*4)$ are the same...so in a way there is scale invariance even with zeros. Perhaps there is some better way to discuss the consequences of the “zero problem”? e.g., in terms of how increased sequencing depth could feasibly change the results by resolving below detection limit (BDL) zeros?

(5) Section “Choosing the denominator”, paragraph 1.

The example the authors provide of “gene 1: gene 2” is a bit confusing to read. Perhaps the authors could clarify the example to help the reader more easily follow the argument.

(6) Section “Choosing the denominator”, paragraph 2.

The authors write “Much effort is placed on ‘normalizing’ the data to have a consistent read depth instead of treating the data as compositional”. There is some nuance here For all practical purposes the authors’ proposed method is doing the same thing. By seeking to identify an invariant denominator, they are trying to normalize away the effect of sequencing depth, i.e. to use log-ratio transformation as a kind of effective library size normalization.

I don’t think this approach is wrong; in fact, I quite like it. However, I think it is a little misleading to imply that your IQLR/LVHA/etc. approach is fundamentally different than alternatives. The article would benefit from a more careful discussion around this point.

(7) Section “Choosing the denominator”, last paragraph.

Again, the discussion of ALR assumes that the goal of log-ratio transformation is to choose an invariant feature as the reference. This is one very specific application of the ALR. It is completely correct to choose a *variant* feature as a reference for the ALR, for example because this reference makes biological sense or maximizes inter-class discrimination. Choosing an invariant feature is an act of normalization! Since we never know which features are invariant, doing so requires some kind of assumption.

(8) Section “HTS data are sparse”, paragraph 2.

The authors write “The assumption being made...is that most features are either invariant or...”. This is not an assumption of the CLR. This is an assumption of using CLR *for normalization*. As indicated above, there are reasons why one would use the CLR even if the geometric mean were not invariant, for example to compute Aitchison distances.

(9) Section “Alternative denominators”.

This section could probably benefit from being broken further into sub-sections, one for each new transformation, to make it easier to follow. Also

- * The notation for these equations are messy and not consistent (for example compare Eq 4 with Eq 5). The authors should clean this up.

- * The authors do not describe how variance is computed. Based on my knowledge about ALDEx2, I think it is computed using a CLR of the data? Either way, please describe.

- * The notation $\text{var}(A)q_1 > f_A < \text{var}(A)q_3$ is non-sequiter, and hard to follow. Perhaps the authors can define f_A more formally using some kind of set notation? (for both IQVF and NZLR and LVHA), or otherwise express it as an algorithm/pseudo-code.

- * The authors describe $\text{IQVF} = f_a$ or f_b – is this how it is implemented in ALDEx2? I thought $\text{IQVF} = f_a$ in the software package? I didn't think ALDEx2 used f_b .

- * Equation 7 – spell out median, and express the indices over which the median is computed. For example, median $i=1\dots D$?

(11) Questions about simulations...

In general, I recommend describing the simulated data in the Methods section. Also

- * The authors describe simulating 40 significant genes. Is this out of 1000? So 40/1000 positive? Are any of the 2-5 fold changes negative, or all positive? Please clarify.

- * The authors say they simulate “An additional 51 data sets” but do not describe how these differ. I assume you simulate 1 dataset per 1% sparsity? How was sparsity actually introduced into the data? For example, via a model or?

- * I found it confusing to use the word “asymmetry” and sparsity” somewhat inter-changability. If the authors mean sparsity, please use the word “sparsity”. Otherwise, please define explicitly.

(12) Section “Four alternative methods”, paragraph 1

I advise against conflating isometric log-ratio and balance-based approaches in the same sentence. The time and space complexity of selbal does not necessarily apply to other ilr and balance methods. In fact, I would argue that selbal's time complexity is an exception rather than a rule. I understand why the authors would include the first paragraph (it's an obvious thing for a reviewer to comment about!), but I actually think a lot (most) of it can be struck from the final version since it's orthogonal to the main thesis of the paper.

(13) Discussion

I think there are a few points missing from the discussion.

- * A discussion about the difference between log-ratio transformation and “log-ratio normalization”, and how the use of an invariant is helpful but not necessary

- * A reference to previous benchmarking of the CLR and ICLR that shows empirically how that they perform similar to effective library normalization methods, and how ICLR better *normalizes* the data than an ordinary CLR 10.1186/s12859-018-2261-8

- * A discussion about a similar approach to find an invariant center, called robust log-ratio transform (rCLR). The motivation is similar to the NZLR <https://msystems.asm.org/content/4/1/e00016-19.abstract>

(14) Please also make the polyester simulation and post-hoc analysis scripts publicly available!

Reviewer #2

Wu et al. have produced a thoughtful article supporting a useful method (high praise). I find the concept of compositional asymmetry motivating and useful. I believe the IQLR and LVHA are useful tools. Still I think there are necessary improvements the authors should make before publication.

1. Most count data has a mean-variance relationship. RNA-seq and Microbiome data is no different. As a result these transforms which are based on the variance of features will likely be selecting features based on relative abundance. Do the authors intend this behavior? As I see it it is neither good, nor bad, but an important feature in understanding "what is the IQLR and LVHA doing?" This ties into the next comment.


2. The IQLR interested me. In the past few hours I have been playing around with it trying to break it. It seems to hold up nicely in my simulations. Congratulations! I think this is useful. That said, I have very little idea what the IQLR and LVHA are really doing. I understand how they are defined (somewhat, the authors notation is pretty terrible in this section - see below) but not what they are really doing. Why would basing the center on the variance be a good idea? Why not the skew? Depending on your conceptual model for such data, variance itself is compositional (and as I mention above is likely tied to the relative abundance of features through random counting). If you have two groups and define the transform for each group separately you are actually defining that the data exist in two different compositional spaces... that's very odd. What if you had 3 groups? What about N groups? Overall it seems to me these proposed transforms are incredibly heuristic. I think they are useful but even if they work perfectly (I am sure they don't) I think the question is... why? What are the implications of using feature variance. What type of probabilistic model of compositions/counting is reflected in the definition of these transforms? (For example, the ALR and ILR transforms imply a Logistic-normal central limiting distribution... I am not sure the IQLR and LVHA do. I feel like the authors have just dropped a dead fish in my lap and said "here this will tell you the direction to the pharmacy" and sure enough it did... Why? I have no idea, the IQLR and LVHA brings up more questions than they answer. The authors description of "stochastic variance that does not differ between groups" is unclear at best. I have no idea what they mean and even if it was clearer I don't think that is getting at the deeper questions. I would encourage the authors to think critically about these questions and try to rigorously define what they are doing. Perhaps they can answer some of them and that would be great. But to be fair to the authors, even if they can answer none of these questions I think that's fine; in that case I think it is crucial that they make it clear what parts of this are heuristic (my hunch is all) and more research is needed to figure out why this works, what are the underlying assumptions, and most importantly, when it will fail.


3. The authors should give more insight as to why the IQLR and LVHS are recommended for different types of sequence count data. This ties into point 2 above.

4. The authors notation in the section "Alternative denominators" is terrible. What is $S_{\{1...i_A\}}$, G , rab , etc...? I can intuit what they are doing and intending to convey to the reader but that's only with 2-3 reads over this section and some careful guesswork on my part. The notation here needs to be reworked from scratch. Not to mention that defining the IQLR transformation with a variable called IQVF is mildly headache-inducing. Only equation 7 didn't require 2 reads. (I would however recommend they denote the median as "median" not "MED" for convention sake).

5. The authors should formally define compositional asymmetry. I get the concept they are trying to convey (I think), and I think it's a good concept; still, the lack of a formal definition is troubling. It could be as simple as something like: "Let x_i and x_j be two compositional vectors defining the composition of samples i and j respectively. Further let $g(x_i)$ denote the geometric mean of the vector x_i .

Compositional asymmetry is then defined as $g(x_i) \neq g(x_j)$ for some i, j ." Now I think this is likely far too stringent a definition to be useful but, honestly, without more info from the authors I have no idea what they are exactly talking about.

6. Something weird is happening in Figure 2 for Percent Asymmetric Sparsity > 48. Somewhat of an elephant in the room situation in my opinion. I have a feeling it's a pretty innocuous issue related to their choice of performance metric (Median Between Condition Difference) but still, needs to be explained in the text. 

7. The authors' description of "count-based tools" in Section "Choosing the denominator" paragraph 2 is not quite right. Aldex2 (the authors' own software) even makes the second sentence in this paragraph untrue. Not to mention the abundance of multinomial logistic normal tools that have been developed. I think this is just a simple issue of needing to tighten up language + possibly needing to add more diverse citations; still, I got a good chuckle out considering that the authors' own software was not being considered in sentence 2. 

Minor (address/respond at the authors' discretion):

- For maximum impact the authors may want to consider making the figure legend for Fig. 1 more clear. I found it hard to understand what the "red points in the top-right" meant until I realized these were features set to zero in one condition (... I think, still not sure).

- The text under equation (3) "where" should not be indented.

- "the goal is to identify a basis that best represents each sample" -- what does this mean? 

- I had to zoom in on figure 3 to read. I recommend making the figure bigger.