Response to Reviewers

true

Dear Editor and reviewers

Thank you all for thoughtful and helpful comments. In response to your queries we have added a new Figure 1 that more clearly articulates (we hope) the need for scale uncertainty. Secondly, we added in new analyses using semi-synthetic datasets to add in a standard of truth. We did this using some of the datasets of, but in a way that was independent of Li et al (1) and observed essentially the same result. In doing this analysis it became clear that the addition of scale uncertainty was much better at controlling FP than fold-change cutoffs and we came to understand why this is so and have included this in the revision.

We have also tried to clean up language and clarify sources of data, etc as requested.

We have included line numbers in many of the responses and have included line numbers in our resubmission for cross-reference.

Finally, we have included a file with the differences (diff.pdf) marked up in the package so that the changes can be identified.

We believe that this revised document is now much stronger because of the reviews and hope it meets approval

The repository for this version is at: https://github.com/ggloor/scale-sim-bio/ and is now public

Best regards

Greg Gloor, on behalf of the authors

Reviewer: 1

Comments to the Author

Gloor, Nixon, and Silverman provide a clear and concisely written manuscript that describes a significant improvement of the ALDEx2 package to incorporate scale uncertainty in analyses of differential expression and abundance. These scale models relieve the reliance on commonly used but difficult to rationalize cutoffs of q-values and log-fold change thresholds, and facilitate adoption through the ALDEx2 R package. This work represents a valuable addition to an already impactful and versatile tool for the community. I have several minor comments that I am confident can be easily addressed by the authors.

1. The authors focus here on the application of scale-informed ALDEx2 to transcriptomic analysis. Can the authors further discuss whether applications to other types of measurements is equally appropriate? If so, there may be an opportunity to generalize the title to use "counts" instead of "RNA-sequencing", as that broadens applicability to other data types (e.g. proteomics, single cell epigenome data, etc).

    - We have changed the title to reflect that it is count data, and have included information on other data types in the introduction. However, the focus of this report is intended to be squarely on transcriptome and metatranscriptome data in an attempt to more tightly target that analytic audience. Companion articles by by McGovern et al. (2) discussing the use in gene set enrichment and by Nixon et al. (3) targeting the microbiome community are published in PLoS Comp. Bio. and under second revision at Genome Biol. respectively.

2. Relevant to statements like this: "The original naive ALDEx2 (39) model unwittingly made a strict assumption about scale through the CLR normalization (3)." - Can "scale-naïve" be used instead to more precisely indicate the distinction between the previous and current defaults in ALDEx2?

- We thank the reviewer for this suggestion of more precise and descriptive terminology.

3. Why this threshold in Figure 1? "The horizontal dashed lines represent a log2 difference of ±1.4" The text goes on to say the following: (line 41-44, pg 6): "Note that there is considerable variation in recommended cutoff values(14). Here, applying a dual-cutoff using a heuristic of at least a 2ˆ1.4 fold change reduces the number of significant outputs to 193 for DESeq2 and to 186 for ALDEx2. This cutoff was chosen for convenience and is in-line with the recommendations of (14) with the fold change limits shown by the dashed grey lines in Figure 1." - Please add a rationale for selection of this cutoff, and briefly state the guideline used here, especially given the variation in recommendations in ref 14.

   - We agree that the explanation for this cutoff was not well explained, and now include a section in the results explaining that all fold-change cutoffs are arbitrarily chosen and this one was chosen to be comparable in number of transcripts identified to our gamma parameter. There is now an example in the text (Figure 3) and in the supplement (Sup. Figure 2) showing that adding scale uncertainty is superior in controlling FDR when compared to fold-change cutoffs.(lines 19-20, 160-169)

4. Supplementary Fig 1 is an excellent demonstration of how scale uncertainty affects the number of DE genes – the recommendation in that section that this is be used a diagnostic plot should be emphasized in the main text as well, possibly on line 50-51, or in the discussion.

   - Thank you we have enlarged the discussion of this and included a second example using a biologically replicated dataset with modeled TP information.

5. This dataset (ref 44) included many technical replicates. How do these conclusions port to datasets that have biological replicates instead?

   - we have included two worked examples that use biological datasets from human-derived studies along with modeled TP data. (lines 177-237)

6. Figure 2B, and pg 9 lines 11-14: "Figure 2B shows a plot of the difference between the $\gamma = 0$ and $\gamma = 1$ data and here we can see that scale uncertainty is preferentially increasing the dispersion of the mid-expressed transcripts that formerly had negligible dispersion; examine the grey line of best fit (overlaid by the red line) for the trend." - Can the 5 lines be separately shown on 5 panels, rather than overlaid? They are difficult to visually distinguish as presented.

   - We realize that this figure was confusing in concept, and have moved it to the supplement as Supp. Figure 5. This figure is now only discussed in support of figures that we believe better make the point that there is not an across-the-board change to dispersion and hence p-values. This is now better described by Figure 2 panels C and D, and Supp. Figure 3 panels C and D. This information is now only briefly discussed in lines 243-245, and is no longer central to our argument.

7. Is the conclusion from the first part of the results that by adding scale uncertainty ($\gamma = 1$), one can do without setting a change threshold (T=1.4 in this case). This should be emphasized in the discussion.

   - Yes, and the inclusion of the modeled data now makes this point explicitly. The discussion has been simplified to more clearly highlight this point. We are explicit in lines 246-251, 367-374 and 381-388

8. The examples outlined in the results would benefit from a more explicit description of key experimental features – e.g. the number of technical replicates in the ref 44 dataset (mentioned in discussion, but would be useful in results); same for the ref 50 data, so the reader can easily contextualize the results, and draw parallels to their datasets of interest to which ALDEx2 may be applicable.

   - We have included more description of the datasets and made clear what is technical and what is biological replication especially at lines 128-134, 177-185, 253-259

Some typos:

Line 25, pg 10, typo: "functions tp be nearly invariant"

Grammatical error: This assumption was often close enough to the true value to be useful, but was not always the a good estimate and could be outperformed by other normalizations (40).

- Thank you, we have tried to catch all typos and grammatical errors.

Reviewer: 2

Comments to the Author Summary: In this manuscript Gloor et al. highlight the application of ALDEx2 and the recent addition of scale models to help improve differential expression analysis in RNA seq experiments. This manuscript builds upon previous work that introduced scale models to ALDEx2 and highlights that differential analysis in high throughput sequencing often suffers from scale assumptions introduced during commonly applied normalization techniques. In this manuscript they used both a transcriptomic and metatranscriptomic experiment to highlight the improvements that can be achieved by using the scale models incorporated in ALDEx2. The manuscript is well written and highlights an important problem within the field, however, there are some major comments that should be addressed by the authors.

Major comments:

1. Throughout the manuscript a clear definition of what the "truth" is for each experimental set up is needed (i.e., is it transcripts relative to species abundance or bulk absolute transcripts or something else?). This is critically important in the metatranscriptomics section where high levels of differing biomass could result in higher levels of absolute transcripts across the board but not higher levels of transcripts relative to the underlying species abundance. The attempt to center the data on housekeeping genes suggests the latter is the objective but this is not clear throughout the manuscript. In addition to this if the authors are interested in identifying the transcript levels relative to species abundance, they may be interested in comparing their tests against a method that normalizes transcript levels against underlying DNA abundances(https://pubmed.ncbi.nlm.nih.gov/34465175/).

   - We have added two semi-synthetic datasets to show how the FDR is controlled much better by scale uncertainty than by fold-change cutoffs. Ultimately, the choice of what to use as the baseline for normalization of scale between samples is a choice that must be made by the analyst. In some situations the biological question may be best addressed by normalizing transcript levels to species abundance (although to be clear, in the method suggested both the RNA and species levels are relative abundances). However, such a normalization is discussed with reference to species-level metatranscriptomics rather than systems-level metatranscriptomics as shown in our report and in (4). We would need to redo all of the analysis in the cited paper in order to make this comparison. Indeed, the scale approach advocated in this report can use any measure of absolute abundance as an anchor and we have been explicit throughout about this. (lines 177-237, 377-380)

2. The claims in Figure 1 should be further validated through simulated data and or a dataset where the true underlying number of differentially expressed transcripts is known. While the reduction in transcripts that the scale model achieves is likely inline with the biology it is unclear if the model is being overconservative.

   - We thank the reviewer for this suggestion while the original paper by Nixon et al. (5) included extensive simulation and theoretic justification we recognize that this is not yet published and a clear demonstration of the utility in transcriptome datasets would be helpful. We have now included two simulated datasets from a recent benchmarking paper [2Li:2022aa] showing that the scale-reliant approach is substantially better in controlling the FDR and, at least in one example, has minimal effect on power (Fig 2, Supp Fig 3). This has resulted in a major re-write of the results section and has made a substantial improvement for how we understand scale uncertainty and its effect on data analys. We again thank the reviewer for pushing us to do this analysis. (lines 177-237)

3. In the metatranscriptomics section it is claimed that biomass differs by 20 fold between the groups that are being compared, however, the scaled model only includes a difference of 14%. Please address this discrepancy. Furthermore, based on Figure 3B, the gamma model by itself was ineffective in this setting, so the manuscript should clarify when only the gamma parameter versus the gamma and mu

parameters must be informed.

  - We have tried to include more insights in the results (Supp Figure 4) and the discussion around when to use an informed vs a simple scale model. In essence, the centre of mass of the difference between groups should be close to 0 (no difference) to avoid FP and FN (6). (lines 358-365, 398-410)

4. Due to the limited number of datasets shown in this manuscript the robustness of the model for various biological and technical conditions is difficult to assess.

- We have added two additional datasets showing generality. (see above)

5. Data availability: Scripts used to run the analysis (beyond just the model itself) should be provided in a well-documented GitHub (or equivalent).

- Thank you for pointing out this oversight. The entire analysis is on github and I forgot to flip this to public access.

Minor comments: 1. Without external information, the Lambda distribution appears entirely determined by the prior. It is unclear how robust the choice of Lambda is across differing HTS experiments. Fig. 3 seems to provide a case in which lambda alone is not enough.

- We have added more guidance and examples on how to choose appropriate parameters as outlined in the response to reviewer 1. (lines 398-410)

2. "The computational methods developed for microbiome analysis have low correlation with the actual scale but are useful (8)." Well this may be true it has been shown that in general coefficients from the relative scale correlate strongly with those from the absolute scale. Especially in cases where biomass is not significantly changing between groups. This should be addressed.

- Unfortunately, this statement is only true in the context of the machine learning model in microbiome analyses. This statement is intended to point out that the inclusion of differences in the location in scale between groups *coupled with uncertainty in scale* are robust. This is also the intent of the Supplementary Figure 7 showing that 14% and 5% scale offset are essentially equivalent as long a some uncertainty is accepted.

3. The authors claim that "the dispersion in the unscaled analysis in Figure 2A reaches a minimum near the mid-point of the distribution" which "makes the counterintuitive suggestion that the variance in expression of the majority of genes with moderate expression is more predictable than highly-expressed genes or of housekeeping genes." However, the dispersion seems essentially constant beyond the midpoint, with potentially a small difference in the tail. Whether or not that is biologically important and to what degree that impacts traditional differential expression analysis is unclear and not presented in a convincing manner. I wouldn't say this small change in the tail is inconsistent with the notation that sufficiently expressed genes all have equally predictable expression.

- we agree that this figure was not optimal for the points we were trying to make. This figure has been moved to the supplement and de-emphasized. As indicated in the response to reviewer 1, the point that the addition of scale uncertainty is affecting transcripts based on dispersion and not on difference between is now made much better by Figure 2 and Supplementary Figure 3. (lines 19-20, 160-169)

4. In the third paragraph of the discussion the authors claim that their method can control type 1 and 2 errors robustly. However, the authors never directly present any evidence that the "hits" their tools are finding are better at controlling these errors than others when a known ground truth is present (they only show that they have reduced numbers of hits and that they may be more reasonable with the underlying biology). The authors may want to benchmark their work on previous simulated data from ref. 18 if they want to make this claim.

  - We have now included two modelled datasets to bolster this claim with biological data and modeled true positive (TP) transcripts.

5. The clause “wet-lab protocols only provide information on the size of the data downstream of the step in the sample preparation protocol where the intervention was made” should be reviewed as previous works have shown that these methods can give relatively accurate log fold changes under real experimental conditions.

   - We agree that spike-ins can be useful, and indeed the use of full scale models is a general tool to do so. However, to give a trivial example, spiking in at the time the RNA library is made will not control for the number of cells in the environment.

6. The authors should make it clear that their works are referring to the absolute and not compositional scale as some researchers may be interested in the composition of transcripts within their sample rather than attempting to infer absolute changes from that composition (which is what is highlighted within this works).

   - We have included additional information around what scale is and what we are trying to measure. Figure 1 is now included to make this clearer. (line 48 now states this explicitly and we give examples throughout regarding what absolute abundance information can be included)

7. In Fig. 2, what is rAbundance referring to? Is it log2(percent abundance)?

   - this is now supplementary Figure 5 and the rAbundance term has been clarified

8. The Fig. 3 caption is inconsistent with the legend in the top of 3A. The text caption seems to be correct, but this should be fixed.

   - We have added information to the legend in now Figure 4 that the first panel has 0 scale uncertainty.

9. I’m not entirely convinced that low dispersion is to blame for issues presenting in HTS data. The example given was of technical replicates which will naturally have lower dispersion rates than true biological replicates. I would suggest that the authors include some evidence that low dispersion is present across biological replicates if they want to make the claim that it is one of the largest sources of HTS data pathologies.

   - We have now included two biological datasets and modeling data

10. In the metatranscriptomic’s example it is not entirely clear why housekeeping genes would have a non-zero log fold change between the groups in the first place. Well the authors suggest that misidentification of these genes is due to inappropriate scale assumptions, it would help to identify what assumptions made by the normalization result in this potential pathology to begin with.

    - The reviewer is correct in that HK genes may not have non-0 LFC, but that the assumption that is almost always made is that these are appropriate references. Note that the analysis in the metatranscriptome is not at the gene level, but at the aggregated functional level. We have tried to clarify that the difference in occurrence is a major driver of this asymmetry. (lines 358-365)

11. Could the authors comment on whether they believe their scale model implementation will work well under other normalization conditions other than CLR.

    - Nixon et al. show that in theory, any interpretable log-ratio will work with this schema and other papers are in preparation showing this. For the purposes of this manuscript the CLR is the normalization used by ALDEx2

Associate Editor: Erb, Ionas Comments to the Author: Dear authors,

Please focus your revision on the following aspects: State what the true scale of the data should describe, i.e., number of transcripts relative to species abundance or similar. Does this change between data sets? Include a comparison with other approaches, e.g., comparing with an approach that uses DNA abundance as described in the paper in Annual Reviews (see reviewer 2).

- we have added a paragraph in the discussion about this. Fundamentally, while the scale approach is generalizable and could be applied to the types of data given in the review paper (the actual method is in (7)), but that approach is designed to ask a different type of question than is addressed by scale. (Lines 374-380 and above)

A comparison with the use of a moderated statistic (see, e.g., the classical Smyth 2004 paper) which also addresses the low dispersion problem in data with low sample size would add considerable value, even if only as a discussion.

- we have added some discussion of the moderated approach in the discussion. However, fundamentally this approach is trying to solve a problem of missing information in the data. The use of scale uncertainty directly addresses this issue because it adds in uncertainty around the missing information rather than try to model it from the data. This is the fundamental problem of partially specified models that we bring into the introduction, when the information is not available, one is better to acknowledge this rather than to try and pull it out of data that is missing it. (lines 51-53)

Additionally, the inclusion of a simulated or otherwise known ground truth (as pointed out by reviewer 2) would validate the claims in Figure 1.

- This has been done with two different datasets and the result has significantly strengthend the paper and our interpretations

Discuss effect of biological replicates vs. technical replicates (see both reviewers).

- This has been added in multiple places and again the introduction of the simulated dataset has helped with this point.

Please indicate if the preprints referenced in (3) and (20) are submitted / currently under review.

- The companion articles by by McGovern et al. (2) discussing the use in gene set enrichment and by Nixon et al. (3) targeting the microbiome community are published in PLoS Comp. Bio. and under second revision at Genome Biol. respectively. The companion article introducing the theory and the initial modeling by Nixon et al. (5) is still under review at Annals of applied Statistics which is a frustration for the authors as much as the reviewers

1. Li,Y., Ge,X., Peng,F., Li,W. and Li,J.J. (2022) Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol*, **23**, 79.

2. McGovern,K.C., Nixon,M.P. and Silverman,J.D. (2023) Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. *PLoS Comput Biol*, **19**, e1011659.

3. Nixon,M.P., Gloor,G.B. and Silverman,J.D. (2024) Beyond normalization: Incorporating scale uncertainty in microbiome and gene expression analysis. *bioRxiv*, 10.1101/2024.04.01.587602.

4. Dos Santos,S.J., Copeland,C., Macklaim,J.M., Reid,G. and Gloor,G.B. (2024) Vaginal metatranscriptome meta-analysis reveals functional BV subgroups and novel colonisation strategies. *Microbiome*, **12**, 271.

5. Nixon,M.P., McGovern,K.C., Letourneau,J., David,L.A., Lazar,N.A., Mukherjee,S. and Silverman,J.D. (2024) Scale reliant inference.

6. Wu,J.R., Macklaim,J.M., Genge,B.L. and Gloor,G.B. (2021) Finding the centre: Compositional asymmetry in high-throughput sequencing datasets. In Filzmoser,P., Hron,K., Martìn-Fernàndez,J.A., Palarea-Albaladejo,J. (eds), *Advances in compositional data analysis: Festschrift in honour of vera pawlowsky-glahn*. Springer International Publishing, Cham, pp. 329–346.

7. Zhang,Y., Thompson,K.N., Huttenhower,C. and Franzosa,E.A. (2021) Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics*, **37**, i34–i41.