Reviewers' comments on MMETSP paper

Reviewer #1

The manuscript submitted by Lisa Johnson and colleagues is a well written and comprehensive work aimed at reanalyzing a fairly enormous dataset. I feel that the work is important for the following two main reasons:

1. Assembly methods have improved substantially since the original datasets were analyzed, and as the authors point out - these new analyses recover new transcripts that might be useful to the original researchers and to the broader community.

2. Applying standardized and reproducible methods - at scale - is challenging, and the authors provide an example for how this could be done. I can imagine others using these ideas (or the actual code) to assemble other datasets in a similar fashion.

In terms of the manuscript itself, it is sound, with just a few areas where improvements will make for a more readable paper. Interspersed with this, I have a few more pedantic suggestions that the author should feel free to ignore if deemed unhelpful.

- [x] Line 91: replace 'higher' with 'more favorable' or even 'better'.

Changed to: Here, we show that our re-assemblies had better evaluation metrics and contained most of the NCGR contigs as well as adding new content.

- [x] L102: The link to the code does not seem to be active. I would have loved to review it.

It appears to be active in our version. I apologize that the link was not active for you in the manuscript version you received. The link is here: https://doi.org/10.5281/zenodo.249982 (not sure how to fix this, since it appears to be active in the pdf and url works woth copy/paste too...)

- [x] L111: You are using 50bp reads. Do you think your conclusions would have been any different had longer reads (100-150bp) been available? More novice readers might wonder if these methods are just as applicable to them with longer reads as they are to you. I'm sure the answer is yes - your new assemblies might have been even better had you had longer reads.

Thank you for bringing this point to our attention. This is an important point to mention in the discussion. Added to the end of the second section of the discussion as an extension of subheading: "Reassembly with new tools can yield new results." (L327):

We predict that assembly metrics could have been further improved with longer read lengths of the original data since MMETSP data had only 50 bp read lengths, although this would have presented Keeling et al. [31] a more expensive data collection endeavor. Chang et al. [25 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3988101/)] reported a consistent increase in the percentage of full-length transcript reconstruction and a decrease in

the false positive rate moving from 50 to 100 bp read lengths with the Trinity assembler. However, regardless of length, the conclusions we draw would likely remain the same that assembling data with new tools can yield new results.

- [x] L180-182: If I didn't know khmer already, I might struggle with the HyperLogLog estimator. Maybe just a sentence or 3 more might be useful to explain what this is and why it is used.

L182 added: "We used the HLL function to digest each assembly and count the number of fixed-length substrings of DNA (k-mers)."

- [x] L245ish: I keep wondering about your BUSCO scores, and the fact tat they are lower on average in your new assemblies compared to the older ones. Why is this? How do you reconcile this with the more general statement you are trying to make about 'more genes being recovered' in the new assembly. I see that BUSCO is just one of the available metrics to assess this, but it's a little strange I think, given that I'm convinced that these assemblies are actually better.

Complete BUSCO scores were lower than over half of DIB vs. NCGR. The degree of the differences were not very dramatic, compared to differences between the number of contigs and the CRBB differences.

Re-checked BUSCO v3 scores, eukaryota and protista and had to change the mean, sd, and k-s test numbers listed in the text, although they were still significantly higher in NCGR vs. DIB.

Edited sentence and emphasized the less dramatic differences in ORF and BUSCO relative to contig number and CRBB differences (not sure if this belongs more in the Discussion?):

""
Therefore, although the number of contigs and amount of CRBB content were dramatically increased in the DIB re-assemblies compared to the NCGR assemblies, the differences in ORF content and BUSCO matches compared to the eukaryotic (Figure 5 C,D) and protistan (Supplemental Figure 3) databases - while they were significantly different - were less dramatic. This suggests that content was not lost by gaining extra contigs. The extra content contained roughly similar proportions of ORFs and BUSCO annotations. Therefore, the re-assemblies may contribute more biologically meaningful information.

""

I'm wondering, are the contigs found within the BUSCO databases we're querying biased in some way? Like the BUSCO contig is a fragment that our fragment is just not matching with? It's interesting that some assemblies improve while others go way down. (line plots)

```
    SampleName  Complete_BUSCO_perc_NCGR  Complete_BUSCO_perc_DIB
18   MMETSP0121                 65.016502                31.683168
134  MMETSP0932                 85.148515                 3.630363
232  MMETSP0045                 73.927393                35.313531
282  MMETSP0169                 68.646865                 6.270627
451  MMETSP0232                 82.508251                 0.660066
475  MMETSP0439                 80.858086                 3.630363
654  MMETSP0329                 80.198020                 5.280528
661  MMETSP0717                 61.716172                17.821782
```

- [ ] Could you (did you) do a CRHB against Swiss-prot? I imagine that for each assembly pair (old assembly vs new assembly), you'd see more hits to unique Swiss-Prot genes in the newer assembly.

Did they nean CRBB?

The dammit pipeline we used already did a CRBB with Pfam, Rfam, Orthodb. (Added version numbers of each database used to the text). I don't think that CRBB directly with Swiss-prot would give additional information, since Pfam is a collection of protein family alignments which were constructed using Swiss-prot. I think this would be redundant, and more noisy. Pfam (version 28) contains domains generated using HMM of Swiss-prot (version 2014_07, which has 16,230 families, 65,484,326 sequences, and 15,576,887,997 residues). Add this to the text?

- [x] L254: "less significantly different" Do you mean "significantly less"?

The difference in BUSCO was significant, but to a lesser degree than other distribution comparisons (p=0.002 rather than p < 0.001). I changed the wording here, see comment above.

- [ ] L306: I'm also confused about the TransRate scores. As best as I can tell the "NT" assemblies were the raw assemblies, while the "CDS" assemblies were further filtered. If my understanding is correct, then the opening statement for this paragraph (DIB assemblies were more inclusive) is incorrect, given that transrate metrics were higher for the NCGR nt assemblies that they were for the DIB assemblies. I'm also worried about the statements about DIB assemblies being better, while transrate scores were on whole, worse. Should reconcile this.

Yes, the "nt" assemblies were raw assemblies, whereas the "cds" assemblies that the NCGR published were filtered for only coding sequences, as far as we can tell. Transrate scores for may not be the best metric for comparison. I think we should argue that the metrics we have available to us to measure differencs between assemblies are lacking. The transrate score is a measurement of how well the original reads support the the final assembly. We have more unique k-mers, but we have worse transrate scores. We need better metrics to decide what assemblies are "good" or "good enough".

Edited to "suggesting that both pipelines yielded equally valid contigs," to "suggesting that both pipelines yielded similarly valid contigs," so that the significant BUSCO and ORF differences between assemblies can be acknowledged.

To the discussion, added:
""
Moreover, even though the number of contigs and the CRBB results between the DIB

and NCGR assemblies were dramatically different, both the fraction of contigs with ORFs and the mean percentage of BUSCO matches were similar between the two assemblies, suggesting that both pipelines yielded similarly valid contigs, even though the NCGR assemblies were less sensitive.
""

- [x] L315: I'm not sure that you are "directly" evaluating the de Bruijn structure.

We did not directly evaluate the de Bruijn graph structure. The sentence in the discussion mentions:

Metrics directly evaluating the underlying de Bruijn graph data structure used to produce the assembled contigs may be better evaluators of assembly quality in the future.

Clarified with:

In future studies, metrics directly evaluating the underlying de Bruijn graph data structure used to produce the assembled contigs may be better evaluators of assembly quality in the future.

- [x] L320: I'm not sure you show "Biologically meaningful". You show that you have recovered new stuff that is likely real (not an artifact of asembly), but not sure you can claim it's meaningful.

True. Would it suffice to add the word "relevant"?

- [x] L364: In your discussion of kmer content (and other metrics) the idea that some of these datasets might in fact be meta-transcriptomes should be discussed. Lots of marine microeukaryotes associate with bacteria, viruses, etc, and unless extreme care was takes with the target species, to grow in sterile conditions, some of patterns of kmer distrib. might be because the datasets contain more than 1 species.

Added a paragraph:

""
The DIB re-assemblies, including the additional biologically relevant information, are likely to be meta-transtriptomes. Even though these samples were cultured purposefully to target a particular strain or species, conditions were likely not sterile. Sequencing data, and unique k-mer content, may include bacteria, viruses, and other constituents that occurred within the sample community. We did not make an attempt to de-contaminate the assemblies.
""

- [x] Table 1. Can you include the BUSCO results here?
- [x] Fig3 needs a y-axis label
- [x] Fig 5c and a few other places. There are a few DIB assemblies that are WAY worse than the original assembly. Why? This could benefit from some explanation.

It's like this person is channeling Titus! Added to Discussion:

""
For some samples, the DIB re-assemblies had lower metrics than the NCGR assemblies. Complete BUSCO scores were lower than over half of DIB vs. NCGR. This could be an effect of the BUSCO metric, given that these samples did not perform

poorly with other metrics, such as % ORF and number of contigs compared to the NCGR. For other samples, MMETSP0252 (Prorocentrum lima) in particular, assemblies required several tries and only four contigs were assembled from 30 million reads of data. The fastqc reports were unremarkable, compared to the other samples. In such a large dataset with a diversity of species with no prior sequencing data to compare make it challenging to speculate why each anomaly occurred. However, further investigation into the reasons for failures and peculiarities in the evaluation metrics may lead to interesting discoveries about how we should be effectively assembling and evaluating nucleic acid sequencing data from a diversity of species.
""

- [ ] In the end, what I took away from this paper is that the new assemblies had different transcripts, and this is great and potentially helpful for researchers. Saying that, on a whole, both BUSCO and TransRate scores trended toward lower, which is maybe surprising, especially because the original assemblies were assembled (best as I can tell) using a general genome assembler (ABySS/MIRA) rather than software specialized for transcriptomes.

Yes. Added sentence mentioning ABySS assembler and cited Keeling et al. 2014 for methods. (since we do not have the exact methods from the NCGR)

Reviewer #2

The manuscript by Johnson et al. describe the re-analysis of the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP). The authors have generate a new computational pipeline for the de novo assembly (using Trinity de novo) of the RNA-Seq reads of several hundred transcriptomes as well as downstream a set of scripts to compare the outcome with the results of the original publication (which used Trans-ABySS for the assembly).

The current manuscript is a great example that shows the value of revisiting old data sets with new computational tools. The authors put strong focus on reproducibility of their analysis. The effort for this should not be underestimated and the work can serve as a blueprint for similar data re-analysis projects.

I see no major issue in this work but still would like to have a few smaller ones addressed:

- [ ] The manuscript is currently rather descriptive and has only a few explanations why there are certain differences in the presented assembly approaches. E.g. what are the reasons for the observation displayed in Figure 4 that there so many more unique k-mers in the DIB than in the NCGR set? Maybe not all results can be explained mechanistically but least at some potential reasons could be discussed.

Added to L316

""

The relative increase in number of unique k-mers from the NCGR assemblies to the DIB re-assemblies could be an effect of having more contigs. Within the data, the Trinity assembler found evidence for building alternative isoforms. Whereas the ABySS assembler and transcriptome pipeline that NCGR used may not have preserved that variation, in an attempt to narrow down the contigs to a consensus transcript sequence.
""

- [x] The authors write: "We used a different pipeline than the original one used to create the NCGR assemblies, in part because new software was available [8] and in part because of new trimming guidelines [27]". Is [8] really the correct reference here? If so this has to be further explained.

This was a typo. Fixed to reference [18].

- [x] I think figures 2, 3 and 5 are not red green blind safe.

Changed to blue and brown.

- [ ] In the script collection uploaded to Zenodo I personally would have removed the "pycache" folder and the containing Python byte code files (*pyc). Or do they have any purpose / contain useful information?

True. Needs to be fixed.

- [ ] The supplementary notebooks could additionally be uploaded as ipyn files.

True. Needs to be uploaded.

- [ ] The authors have a configuration file for user specif paths but this is not strictly used. In "[dibMMETSPconfiguration.py (https://github.com/ljcohen/MMETSP)](https://github.com/ljcohen/MMETSP)" another "basedir" variable is set and in trimqc.py even the full path for Trimmomatic is set ("/mnt/home/ljcohen/bin/Trimmomatic-0.33/trimmomatic-0.33.jar"). This make the reuse of the framework harder.

True. Needs to be fixed.

- [ ] While I understand that it is sometime needed due to dependencies on old libraries I would like to discourage the use of Python 2.7 (aka "legacy Python") in currently research projects and would strongly recommend to use a current Python version (3 and higher) instead.

True. At the time when I wrote the scripts, they were in Python 2.7. The scripts I run now are in Python 3+, but they are not the ones used to generate these assemblies.

Reviewer #3

In the manuscript, Johnson et al have reassembled RNA-seq data from 678 samples generated from MMETS Project using a pipeline, which follows the Eal Pond mRNA seq protocol. The pipeline (DIG) starts by quality trimming the data followed by digital normalization and assembly using the Trinity assembler. The authors have compared their re-assemblies against assemblies generated from the method suggested by the National Center for Genome Resource (NCGR). For comparison, they have used difference evaluation metrics like Conditional Reverse Best BLAST (CRBB), BUSCO scores, annotation using the Dammit pipeline and ORF content in the assembly. They argued that their pipeline is able to provide additional biologically meaningful content as compared to the NCGR pipeline. While the work overall is quite interesting and the large set of assemblies appear useful, I feel that there are some improvements and clarifications necessary:

Major comments:

- [ ] The core reason behind the observation that DIG pipeline being better than the NCGR pipeline is not clear. It might be due to the core algorithm behind the assembler used by the pipelines (DIG using Trinity and NCGR uses AbySS). But this should be explained in more detail why their pipeline performs better. For example, is the performance increase linked to sequencing coverage of the read data sets? Or transcriptome complexity of the sample? Or is it the fact that the NCGR pipeline seems to use a custom build pipeline that uses multi-kmer ABySS but not the de novo transcriptome assembler trans-ABySS, which may be more suited?

Main differences were the assemblers used (we used Trinity) and trimming parameters (MacManes 2014). In postprocessing, we did not filter contigs for ORF content whereas NCGR did. Not sure what else to say here. We downloaded their same data and were directly comparing our assemblies to theirs, from the same sample. Sequencing coverage was the same, transcriptome complexity should have been the same. The details of the custom built pipeline used by NCGR are unknown. We did not compare ABySS to trans-ABySS. We compared the NCGR pipeline to our pipeline.

- [x] The other major difference between the pipelines is the additional step of digital normalization which DIG uses. Normalization generally removes kmer information, which affect the overall assembly. It is not clear why normalization in case of DIG should improve the assemblies. Normally the expectation would not that the digital normalization leads to an improvement. So I assume the authors do it simply to reduce the computational costs of the many assemblies, which is plausible but should be stated.

Yes, normalization removes redundant k-mers for the purposes of reducing computational resources required. It does not remove unique content. Or seek to improve the content. We have stated in line 144: "To decrease the memory requirements for each assembly, digital normalization was applied prior to assembly."

- [x] Also, Trinity by default performs in-silico normalization. So, the additional normalization step is redundant. Is the option for normalization switched off in the assembler. If yes, the authors should comment on why they are using Diginorm instead of using Trinity's built-in normalization, is there any indication that this works better for the assemblies they have done?

We used an older version 2.2.0 which did not have this turned on by default. Added: "This version of Trinity (2.2.0) did not include the “in silico normalization” option as a default parameter.

Titus: Has anyone tested Trinity in silico norm vs. khmer diginorm? Should I do that?

- [ ] It is not clear which version of NCGR assemblies ("nt" or "cds") the authors used for calculating the mean ORF% in Table 1. If they have used the "nt" version, then the number can be misleading. The "cds" version of the NCGR assemblies contains contigs that have been predicted to show coding potential and hence might have a higher mean ORF content (as this is computed as percentages). I suggest the authors compare the mean ORF% content of the two NCGR version against the assemblies generated using DIG for full transparency and then discuss the differences regarding these two NCGR version and their assemblies.

I clarified the ORF% in Table 1, which was originally only "nt" version of the NCGR assembly, and added the "cds" version of the NCGR assembly. The DIB re-assemblies weremore comparable to the "nt" versions since we did not filter contigs based on

ORF content, which the NCGR did in their "cds" version. Our point is that when filtration steps are performed, potentially useful content is lost. Added this to the results and discsussion.

- [ ] I think the line plots used in the paper can be improved, because it is hard to quantify the amount of overlapping lines. For example I think that Figure 2A, 3A,5A,5C are probably more easy to interpret when made as a scatterplot, e.g. Fig2A where the number of contigs is compared between NCGR and DIB assemblies.

In this context, the line plots are appropriate to visually draw for the reader the relationship between the same metric in each NCGR and DIB assemblies. Scatter plots will not show that relationship.

- [ ] I would not say that the distribution in Figure 2c looks like a Normal distribution as the right tail is much heavier than the left one. If you want to make that statement, use a test of normality, however I feel this is not important for the paper.

Changed to:
""
The frequency of the differences between Transrate scores in the NCGR ‘nt’ assemblies and the DIB re-assemblies is to centered around zero (no change) (Figure 2C).
""

Minor comments:

- [x] Typo in reference 25 .. de ovo assembly ..

Fixed.

- [ ] line 336: I was not able to understand what the (see op-ed Alexander et al. 2018) refers to, as there is no such reference in the bibliography and no footnote

How to fix citation on this since it is not published yet, and we're not sure volume, page numbers yet?