Response to Reviewers

Reviewer #1 (Comments for the Author):

In this manuscript, Sovacool et al. describe a new method for fitting amplicon sequences to existing datasets. This is an interesting problem that emerges in microbial ecology: how do you compare new sequence data to existing results without reprocessing all of the sequences together and creating new de novo sequences. Early (haha) in microbial ecology, one proposed method of generating OTUs was to align new data to an existing reference database. This was done by either a closed (only fitting to the existing data) or open (fitting as much as possible to the existing data but then generating additional de novo OTUs as needed) appraoch. Due to many drawbacks, this method is not widely used these days. However, one benefit it had was the ability to add new data to an existing set of OTUs. This allowed researchers to make comparisons across projects or even to conduct meta-type-analyses. In this manuscript, the authors describe a new method that uses the strengths of modern clustering methods (e.g., OptiClust) and the ability to perform cross study comparisons through reference mapping. The others do a very thorough job of describing the new method, comparing it to existing methods, exploring the potential user options, and summarizing everything in a way that gives the reader an understanding of exactly why they may want to use the new tool, how to use it, and what benefits it will have.

The manuscript is well written and describes a new computational tool without computer science jargon. I think this manuscript and the method will be very well received because it does fill a gap in the computation analysis of amplicon sequences. A have only a few comments and suggestions as outlined below:

Thank you for your enthusiasm!

OTU quality: The main tool for comparing OTU quality in this manuscript is MCC. This metric is central to de novo OTU clustering in the OptiClust approach, and I think it greatly benefits amplicon-based approaches. The MCC metric is briefly described on Lines 54-55. However, since it is so central to the comparisons in the current study, I wonder if the reader would benefit from a bit more detail and, I think the reader needs to be reminded about MCC when interpreting the results. When reviewing Fig 1: the reader should know exactly why an MCC of 0.91 is better than 0.78 and have a way to conceptualizing the difference. Also, it may be helpful to remind the reader how quality was determined (see L172).

We describe MCC in more detail under the description of the OptiFit algorithm, which we expect readers would read just before looking at Figure 1. We added a few more details to the MCC description in this section (L111-113 and L117-118), In the caption of Figure 1 we also added examples of the confusion matrix values for the first and last MCC scores.

We revised the sentence formerly at L172 to note that we mean “quality” in terms of MCC score (L184).

OptiClust: On lines 78-80 the authors state that "previous studies" found OptiClust generate the highest quality OTUs, but then they have only a single reference which is the original paper describing the method (and from the corresponding authors research group). Full disclosure, I use OptiClust. However, I think the authors need to either say "A previous study" or provide additional references (and perhaps one from outside the research group).

Thank you for catching this. We have revised this sentence to say "We previously found" in order to clarify that this claim is supported by a single study from our group (L86).

Computational Resources: Throughout the manuscript, the authors often use compute time as a measure of the computational resource demand. This is an important metric, but I also wonder about RAM. Specifically, at the end of the manuscript the authors state "we recommend using OptiClust for de novo sequencing over the split strategy with OptiFit since OptiClust is simpler to execute and performs similarly in terms of both run time and OTU quality" (L279-281). Earlier they also stated that OptiClust was faster (L244 and L257). Regardless of time comparisons, I'm also wondering if the OptiFit split strategy could be an added benefit for individuals that lack computation performance (e.g., RAM). I don't remember from the OptiClust paper how much RAM is needed or if it scaled-up with data size, but I thought it was interesting that 10% of a sequencing run could be used to generate de novo sequences with OptiClust and then OptiFit could be used to match to other 90% without sacrificing quality. Could this be a benefit for research groups that may be computationally limited?

Saving computational resources is an interesting thought! However, we found that the RAM usage was similar between OptiClust and OptiFit regardless of the reference fraction size used in the split strategy. Both OptiClust and OptiFit consider the distances between all pairs of sequences, so changing the reference fraction and using the remaining sequences as the query doesn't change the total number of distances that are considered in the MCC. We have added statements to note how the RAM usage of OptiFit & OptiClust compared (L217-219) and how the RAM usage compared across reference fraction sizes for the split dataset strategy (L233-235). We chose not to include the RAM data in the figures since we found RAM to be similar, and the figures already contain a lot of information.

Other Reference Based Methods: I think the authors make a good point when the compare VSEARCH and OptiFit (L261-271). Specifically, I found it interesting to think about how VSEARCH can allow more dissimilar sequences to be clustered into the reference OTU. I think this is something that is worth turning into a stronger statement. For example, since VSEARCH could cluster 3% dissimilar to a centroid without considering the other sequences being clustered, does that mean a maximum of 6% dissimilar across all reference clustered sequences is theoretically possible? In other words, is VSEARCH clustering more in the closed approach because the within-OTU maximum dissimilarity is actually very high? This was just a though, but perhaps another way to strengthen the discussion.

We agree, it is theoretically possible that when using VSEARCH for reference clustering, query sequences may be 3% dissimilar to a centroid sequence but possibly up to 6% dissimilar to each other. Since we did not investigate this directly, we hesitate to make *too* strong of a statement in the discussion, but we did add a statement (L289) to better emphasize this point.

Reviewer #2 (Comments for the Author):

The manuscript introduces optiFit, an algorithm that builds on OptiClust that aims to provide an efficient and robust method to fit amplicon data to existing OTUs. The authors utilize OptiClust output as their starting list of reference OTUs, and aim to cluster their new query sequences to the existing output. The results section is logical and very well written: It describes how the algorithm works, the benchmarks its performance with various reference databases (SILVA, GreenGenes, and RDP), compare the results with QIME2 clustering approaches.

The reviewer is by no means an expert in computational biology, and merely has a microbial ecology based background. As such, issues regarding performance metrics, algorithm strategy, coefficient, etc.. are beyond my expertise. Judging merely based on the need for such algorithm, the overall strategy, and the apparent results and benchmarking using reference datasets and databases, I find the results convincing. The need for such addition to the suite of approaches offered by MOTHUR is also well explained. I encourage the authors to define what "high quality OTUs" and "low quality OTUs" means and how they are assessed early on in the introduction.

Thank you for your feedback!

We added a sentence in the introduction to clarify the meaning of quality (L54-57), and provided additional details of the MCC score under the description of the OptiFit algorithm (L111-113 and L117-118).

Reviewer #3 (Comments for the Author):

In this manuscript, Sovacool et. al. describe OptiFit, a new reference-based clustering method for assigning (amplicon) read sequences to representative OTUs. This method (OptiFit) builds upon the concept of OptiClust, a de novo clustering and OTU picking method previously developed by co-author Westcott. When carrying out clustering of a given read sequence, OptiClust utilizes information about all of the sequences within an OTU, instead of only the representative or centroid sequence for that OTU cluster, an approach used by many other clustering methods. This enables OptiClust to generate higher quality OTUs when compared to other de novo methods. OptiFit uses OptiClust's wholistic method (considering all sequences within a given reference OTU) when performing reference-based clustering/OTU picking. All query sequences start as their own OTUs, and the OptiFit algorithm reassigns those sequences to one of the reference OTUs according to

which assignment would maximize the Matthews Correlation Coefficent (MCC) for the dataset; this approach takes into account each of the query sequences' similarities to all of the sequences within each reference OTU. OptiFit is capable of carrying out both open and closed reference-based clustering.

The authors demonstrate OptiFit to be a powerful reference-based clustering method. The quality of OTU assignments obtained by OptiFit is similar to those generated by OptiClust, although there is a run-time penalty for OptiFit when operating open reference mode, although it does not appear to be prohibitive. As the authors note, the quality of OTU assignments is highly dependent upon the nature of the samples being studied, as well as the reference databases being utilized, particularly when carrying out closed reference-based clustering. The inclusion of an option to perform open reference-based clustering was a beneficial decision made by the authors. The authors also compared the performance of OptiFit (and OptiClust) to VSEARCH. OptiFit outperforms VSEARCH in terms of both speed (execution time) and OTU quality (quantified using median MCC scores), although VSEARCH is generally able to classify a larger percent of reads to an OTU in closed reference mode; as the authors note, this is due to VSEARCH only considering an OTU's centroid sequence, as opposed to all of the sequences within that particular OTU, making OptiFit more selective.

The OptiFit algorithm is a natural extension of the OptiClust approach to sequence classification, albeit for situations where a reference-based approach is either required or beneficial. One of the more appealing aspects of OptiFit is that it does not change OTU assignments for samples/data that have been already analyzed when adding new information to a dataset. This functionality would be particularly useful for investigators who are carrying out meta-analysis type studies or who have very large datasets to which new samples are added regularly and reanalysis would be computationally burdensome.

A general question I have is whether the authors have considered suggesting that users make use of the data-splitting approach that they outline in their own paper. OTU classification is a duel task problem (identifying OTUs in a dataset and assigning each read sequence to an appropriate OTU). With de novo classification approaches the same data is used to both identify and classify the reads. From my perspective, the development of OptiFit makes a perfect partner for OptiClust in a Train/Test (Train/Classify) splitting situation, where the users would split data, train using OptiClust and then classify the remaining reads using OptiFit. Of course, pursuing this would be well beyond the scope of this manuscript, but the authors may consider discussing such a possibility for future work.

We are pleased that you made this connection! We agree that it is well beyond the scope of this manuscript, and we are already working on this for another forthcoming paper. Thus, we chose to merely mention machine learning alongside the other potential applications in order to keep this paper narrowly focused on the OptiFit algorithm.

The manuscript is generally clear, concise, and well written. There are only so many

times that one can write "clustered"/"clustering" (or other variants) in a sentence or paragraph and still have it make sense. I can tell (and appreciate) that the authors took great care in the writing process to properly convey their message.

Thank you!

Specific comments:

Thank you for your thorough reading of the manuscript and the detailed comments you provided below.

Line 4 (Abstract): The sentence discussing OptiClust appears to be tangential to the opening of the abstract. Perhaps it would be better to add this context further down in the abstract (line 16-ish).

We have moved the mention of OptiClust near where you suggested (L17). We also trimmed down the abstract for brevity.

Line 83-85: "While other tools represent reference OTUs with a single sequence, OptiFit 84 uses multiple sequences in existing OTUs as the reference and fits new sequences to 85 those reference OTUs." Should "multiple" be "all" in this sentence?

Yes, thank you! (L92)

Lines 129 - 131: Please include references for Greengenes and SILVA, and indicate what versions of all three (Greengenes, SILVA, and RDP) were used, even though this information is included in the Methods section.

References to all three were included at the end of the sentence in the order they were mentioned. To clear up any confusion, we moved each database's reference to immediately after its mention, and also noted the version numbers here as you suggested (L141-143).

Lines 136-138: "Clustering sequences to …" I think it would serve the reader to reiterate/emphasize that OptiFit was used for clustering, i.e. "Clustering query sequences with OptiFit to …".

We agree! (L147-148)

Lines 136-138: I suggest adding the corresponding Figure 3 labels (in parentheses), similar to what was done for the "self-split" tests described in Lines 196-198. This will help guide the reader, as there is quite a bit of information being displayed in Figure 3 as well as what is being described in this paragraph.

Good suggestion. (L150)

Lines 141-142: "Thus, open reference OptiFit produced OTUs of very similar quality as de novo clustering, …" please add "with OptiClust" to be more specific about what the comparison being made is.

Agreed. (L155)

Figure 1: W-Z are designated as the query sequences in the caption. Z does not appear in the distance threshold table at the top of the figure and is not assigned to an OTU at the end of Cycle 4. From Line 115 ("Alternatively, a sequence will remained unassigned…") the reader can infer that sequence Z is different enough from the others that it was designated a singleton (outside of the distance threshold discussed in lines 111-113). If I'm correct, I encourage the authors to make this point explicitly in Figure 1's caption.

That's correct, we added a sentence to the caption as you suggested.

Figure 3: I suggest changing the label for "de novo" to "OptiClust (de novo)" or similar to emphasize that OptiClust was being used for those benchmark tests.

In Figure 3, we show results from *de novo* clustering with both the OptiClust algorithm in mothur (red triangles) and VSEARCH (blue triangles). Changing the label to "OptiClust (*de novo*)" would not be accurate, as VSEARCH does not use the OptiClust algorithm. Instead, we changed the red color label from "mothur" to "OptiClust (_de novo) or OptiFit" in Figure 3.

Figure 3: The second sentence in the caption ("Each dataset…") is difficult to follow due to its structure. I suggest rewriting to make it clear that de novo clustering was performed with OptiClust, and 4 different reference based trials were performed (self-split (referencing the OptiClust-ed reference subset), referencing GG, SILVA, & RDP).

We edited the caption to make the workflow more clear.