Response to Reviewer Comments

Machine learning classification using self-reference-based OTU clustering

Courtney R. Armour, Kelly L. Sovacool, William L. Close, Begüm D. Topçuoğlu, Jenna Wiens, Patrick D. Schloss

Reviewer #1:

Armour describe using the closed-reference OTU method OptiRef to predict screen-relevant-neoplasia (SRN) from 16S sequencing data of the fecal microbiome that was originally processed into OTUs by the de novo method OptiClust.

Is this a case study, or something more general? The results here are based on a single dataset, a single set of methods, and a single set of evaluation metrics. Thus, the generalizability of these results is unclear. This concern is exacerbated by the various places in the manuscript that infer generalizable conclusions.

The purpose of this paper is to provide an example of a scenario when the previously publish OptiFit algorithm could be useful. We are demonstrating here that it is possible to use your own data as a reference for consistent OTU classifications without a needing a reference database. We have clarified this throughout the text.

The authors need to situate their methods within the universe of methods commonly used in this area. Closed-reference OTUs and ASVs are widely-used alternative approaches to 16S sequencing data analysis that are already used for ML classification. How does the hybrid approach proposed here of de novo OTUs plus closed-reference OTU assignment compare?

We thank the reviewer for this suggestion and agree that this strengthens the paper. We added the following methods: reference based clustring using OptiFit and the GreenGenes reference, de novo clustering using VSEARCH, and reference-based clustering using VSEARCH and the GreenGenes reference. See lines 47 - 64 and figure 1 for workflow descriptions and lines 87 - 98 and Figure 2 for the results.

Is OptiFit only usable with OptiClust as the precursor?

OptiFit is not dependent on use of OptiClust as the precurson, you could use any algorithm to cluster the reference and then use OptiFit to fit additional data.

What is the relevance of the MCC evaluation? What if an alternative method had lower MCC but better pred

The MCC score is a measure of OTU cluster quality based on the similarity of sequences and whether they are appropriately clustered together or not. Since we added additional methods, we can see that some

methods do have lower MCC but equal model performance. This likely indicates that the model depends on well clustered OTUs. We've added some discussion on this to the paper (lines: 104-107).