**I find this commentary of high quality and very helpful in putting the Kopylova et al. paper in the context of complementary evaluation strategies. I have only a couple of suggestions: the first is that in some passages the text should be made a little less harsh against existing evaluation strategies, the second is that the nice Figure 1 is providing an opportunity to discuss the differences in the outcome of the evaluation compared to what was found in the Kopylova et al. paper using different evaluation strategies.

Below you can find few more specific comments with respect to my two suggestions.**

Thank you for your enthusiasm regarding the Commentary article that I submitted to *mSystems*. I have addressed your points in the revised manuscript. In addition, NINJA-OPS was recently updated and I have updated the analysis to reflect the new version of the software. Please do not hesitate to contact me if you have additional comments or concerns.

**- line 23. "This approach is flawed because bacterial taxonomy often reflects those inconsistencies and biases within bacterial taxonomy that OTU-based methods strive to overcome". I agree. However, it is also true that while the biases are definitely impacting the assessment in absolute terms, in practice the "ranking" of methods according to their performance is (or can) still overall meaningful.

- line 29 and following lines. "One could randomly assign sequences to a predetermined number of OTUs. This would be efficient.". I fully agree in principle, but authors that develop a method that randomly assign sequences to OTUs to improve speed would be responsible of a serious scientific misconduct. Moreover, in the available evaluations, the methods are not only assessed for their computational time and memory. So efficient but very imprecise methods can be easily pointed out and the reader would be able to understand if and which methods are "cheating". And I have a similar comment for the "third approach" i.e. comparing "the number of OTUs generated by various methods for a common dataset" for which one can "randomly cluster sequences into a target number of OTUs". I definitely agree that the author's approach based on the Matthew's correlation coefficient is less biased and more accurate, but the number of produced OTUs is a relevant aspect of the clustering process. I thus recommend to edit a bit the tone of these parts.**

I appreciate you encouraging me to be more constructive on these points. I have revised the text to sharpen my language and reduce the harshness of the text.

- It would be very interesting to add a discussion on the results of the evaluation of this commentary (Fig 1) with the results of Kopylova et al. Are the rankings of the methods consistent (despite the different evaluation criteria) between the two studies? Or is the use of another evaluation approach resulting in completely different rankings and recommendations?

I have added two sentences to the end of the fourth paragraph that does a better job of highlighting the differences between the observations of my analysis and that of Kopylova et al.