Dear Editor,

First of all, I'd like to thank you as well as the two reviewers for your comments. Their advice have definitely improved the manuscript, especially the presentation and clarity. We have addressed the remaining revisions in the new version of the manuscript. Please find below answers to the reviewers comments.

Reviewer 1.
Overall I believe that this manuscript has improved, and the authors are more fair in their comparison of modern methods. Other choices for benchmarking could have been chosen, but the authors have justified their tests.
Thank you for your time and review of the work.

My main remaining issue is on the availability of the software. While the source code is now available (after some effort, but available nonetheless), the galaxy wrapper is not. I would urge the authors to either open access to the galaxy instance linked, or put their code on the galaxy shed so that other galaxy users can most easily benefit from their work.
We use an institutional web platform (sourcesup) as git repository. We are aware that the use of this directory is not optimal. A wiki is now available at this address : https://sourcesup.renater.fr/wiki/ki-s. This wiki includes a procedure for the installation of KI-S and examples for using KI-S with singularity or Galaxy. Academic users can have access to the CIRM-CFP Galaxy instance on demand (https://iris.angers.inra.fr/galaxypub-cfbp/). CIRM-CFP Galaxy instance is registered since April 2019 in the list of public galaxy platforms of the galaxy community (https://galaxyproject.org/use/cirm-cfbp/).

Reviewer 2
The authors have addressed most of my concerns, but there are two important points I would like to see addressed to improve the manuscript.
Thank you for your time and review of the work

Major comments:

First, thanks for expanding on the CLARK analysis in your response, but I still have a few concerns. Does the % classified reads that you refer to refer to just at the species level or in general? It would make sense if it corresponded to just the species level based on your explanation. In either case, a clearer discussion of this result is needed.
We have added the following information in the methods section :
"Classification of nine metagenomic read sets derived from seed, germinating seeds and seedlings of common bean (*Phaseolus vulgaris* var. Flavert) were estimated with Clark version 1.2.4 [16] at the species level. Clark is a method based on a supervised sequence classification using discriminative *k*-mers [16]. These metagenomic datasets were selected because of the high relative abundance of reads affiliated to *Pseudomonas* [27]. The following Clark parameters –k 31 –t <minFreqTarget> 0 and -o <minFreqtObject> 0 were used for the taxonomic profiling. Indeed reducing *k* increase the number of read assignments but also increase the probability of misclassification [16]. Three distinct Clark databases were employed: (*i*) the original Clark database from NCBI/RefSeq (*ii*) the original Clark database supplemented with the 3,623 *Pseudomonas* genome sequences and their original NCBI taxonomic affiliation (*iii*) the original Clark database supplemented with the 3,623 *Pseudomonas* genome sequences whose taxonomic affiliation was corrected according to the reclassification based on the number of shared *k*-mers. For this third database, genome sequences were clustered at >50% of 15-mers, which corresponded to the species level."

From my understanding, this result is a proof-of-concept that clustering genomes into clusters before running taxonomic assignment can improve classification. It could be argued that this is circular because the clusters are based on shared k-mers, which is also what CLARK bases

classification on: if taxa are defined based on shared k-mers then it would always be expected for a k-mer-based classification approach to classify taxa with more resolution. I think the authors should emphasize this k-mer connection between CLARK and the clustering approach and also clearly state that they can only hypothesize that a higher proportion of reads are being correctly classified with their workflow (and that it hasn't actually been demonstrated). Currently the authors discuss this result as showing a clear benefit to microbial ecology in general, which I think would be very misleading to readers.

We propose to add the following to the discussion section
Using the Pseudomonas genus as a use-case, we showed that increasing the breadth of genomic database without investigating the relatedness of genome sequences did not improved the proportion of classified reads. Worse, an unresolved classification may limit the number of species-specific k-mers identified by CLARK and therefore the number of classified reads. Interestingly, an inverse relationship between the number of genome sequences in NCBI RefSeq database and the number of classified reads at the species level was also recently highlighted with other k-mer-based read classifiers [36]. On the contrary, prior classification of the genomic database improve the number of classified reads at the species level. Hence, investigating the relationships between bacterial genome sequences not only benefits bacterial taxonomy but also indirectly microbial ecology.

Secondly, on page 10 the authors state: "Moreover, KI-S includes a friendly visualization interface that could help systematicians to curate whole genome databases.". I was able to get access to the authors' galaxy server to try out the tool thanks to their quick reply to my email and I found it straight-forward to use. However, it wasn't clear to me whether any reader in general would be able to get an account on this server. Based on advertising the link in the manuscript I'm guessing this is true, but this should be clarified either way. If not, then users will need more documentation on how they can use the KI-S code to setup the visualization workflow themselves. I did not find it straight-forward to download the source code and it looks like the only documentation for the source code (the README.md file in the GitHub repository) is in French, which would need to be translated for an English-reading audience. Specifically, looking into this README it appears that the key circle packing visualization step is performed by the generate_packing.pl Perl script. Details on how to prepare the input and look at the output of this script is needed.

A wiki is now available at this address : https://sourcesup.renater.fr/wiki/ki-s This wiki includes a procedure for the installation of KI-S and examples for using KI-S with singularity or Galaxy. Academic users can have access to the CIRM-CFP Galaxy instance on demand (https://iris.angers.inra.fr/galaxypub-cfbp/). CIRM-CFP Galaxy instance is registered since April 2019 in the list of public galaxy platforms of the galaxy community (https://galaxyproject.org/use/cirm-cfbp/). Moreover the readme file has been translated for an English-reading audience.

Minor comments:

Hierarchical clustering is mentioned later on, but it would help readers evaluate the method to know the specific details on how this clustering was performed with the custom R script when it is described in the methods.
We apologize for this mistake in the title of Figure 3. This is now corrected. Indeed genome sequences are clustered according to their connected components at given threshold but this is not a hierarchical clustering.

Minor typo on Pg 4: "was first evaluate" should be "was first evaluated"
This is now corrected