Dear Editor,

First of all, we apologize for the long delay in answering your questions. We don't have any great excuse I'm afraid, just the constant insertion of "other priorities" into our to do list. We hope, however, that you will still be interested in our research paper and that you will have time to review this new version as well as the responses to the reviewers' comments.

Reviewer 1.
In this short manuscript Briand et al describe a workflow which uses k-mer indexing software to compare bacterial genomes. This method generates a similarity measure which is comparable to ANI. They go on to use these relatedness measures to cluster genomes at various thresholds, produce a visualisation of these clusters, and test the use of these clusters in metagenomic read classification. This workflow is deployed on a galaxy server. Overall the methods in the manuscript appear to be sound, as they are mostly based on previously published work. Though the novelty of the algorithm is limited, the implementation and pipeline, being on galaxy, may well be useful to researchers who are more comfortable with a graphical user interface than the command line.

Dear reviewer, thank you for your time and thoughtful review of the work. We totally agree with your synthesis, this work is not a new approach to measure genome similarity but rather a workflow for performing and visualizing the output of such analysis in a simple way. To avoid any confusion we have modified the text to clearly explain our objectives. In this new version of the manuscript, we have also compared several tools for extracting *k*-mers from genomes sequences. One of this tool, Kmer-db (Deorowicz et al 2019 Bioinformatics) outperformed Simka in term of computation time, while producing similar results. Therefore, we added Kmer-db as an alternative in the workflow.

In the lines below, we have addressed your comments in full, point-by-point.

1- This server requires username and password to use, so I was unable to test any of this software myself. Nor was the implementation available on github (or similar), or the galaxy shed, meaning no-one else can use it. This severely limited my ability to review this aspect of the manuscript.
The source code is available on the sourcesup platform at this address: https://sourcesup.renater.fr/projects/ki-s/

2- The comparison with PYANI is not really appropriate. The authors used Simka, which by my understanding is a k-mer indexing package, so is unsurprisingly orders of magnitude faster than nucleotide alignment with mummer and blast. A more modern comparison would be with either other k-mer indexes, or sketch based approaches such as fastANI. These approaches have been around for a number of years, and are the standard now used in this field.
We have compared the computation time and memory footprint (ou RAM usage) of a number of softwares (see Table below).

Software	**Time**	**RAM**	**Reference**
PYANI	3 months	ND	Pritchard et al., 2016
FastANI	11 days	15 Go	Jain et al., 2018
Simka	4 hours	18 Go	Benoit et al., 2015
Kmer-db	40 minutes	25 Go	Deorowicz et al. 2019
Mash	7 minutes	26 Mo	Ondov et al. 2016

To carry out this comparison *k*-mers were extracted a set of 934 genomes sequences. Based on this comparison Simka has a faster computation time than FastANI. However, Kmer-db and Mash clearly outperformed Simka in term of computation time.
The outputs of these different softwares were next compared to Average Nucleotide Identity based on blast (ANIb, **Fig. 2**). FastANI is the best estimator of average nucleotide value as indicated by

the strong linear relationship of average pairwise similarities with ANIb values. However, one small caveat is that sequence similarity values are ignored by FastANI when below 76% of shared *k*-mers, which artificially improved the linear relationship. According to these linear relationships, Simka and Kmer-db performed reasonably well for ANIb values above 0.9, while MASH is restricted to ANIb values above 0.95. In summary, Kmer-db and Simka were selected in KI-S workflow since these tools are the best compromise between quick computation time and robust genome relatedness indexes.
We proposed to add this comparison in the new version of the manuscript.

Figure 2 : Comparison of average pairwise similarity between genomes sequences. Overall genome relatedness indexes of 934 *Pseudomonas* genomes sequences were calculated with PYANI and four different *k*-mers indexing softwares : FastANI, Simka, Kmer-db and MASH.

2) *There is not enough description of the methods, and code is also needed. Describing briefly how components work, what they do and why parameters values were chosen all need to be added. I was not able to find information on simka without following references, and this is an integral part of the method. The difference in how simka and other potential methods work needs to be explained, and why this is expected to lead to large differences in computation time. Likewise, the section on metagenomic read sets needs further description. (What is Clark and how does it work? Why does adding further classifications in helps classify more reads?)*
As stated earlier, the objective of the current manuscript is to provide a workflow for estimating overall genome relatedness index and visualizing the data. One of the output of such analysis is to define clusters of genome that represent coherent groups of bacterial strains. These groups could be ultimately employed for classifying reads derived from metagenome studies. There are a number of sequence classification programs currently available in the litterature with differences in speed and accuracy (eg DOI: https://doi.org/10.1038/nmeth.4458). Clark is one of this read-classifier (https://doi.org/10.1186/s12864-015-1419-2).

3) *The results also lack context. It was difficult to understand what problems were being solved by the presented method, and how much of the method is new compared to e.g. fastANI.*
We agree, the current approach used in our manuscript is not novel and based on previous existing methods. Moreover this research field is rapidly moving and numerous *k*-mers indexing softwares are currently under development. These approaches will allow comparison of more genome sequences in a computational efficient fashion. The purpose of this paper is rather to demonstrate that *k*-mers count could be employ for estimating overall genome relatedness in an efficient way. In addition we are providing a method for visualizing these data.

4) *How did the original Clark database and the newly assigned genome sequences differ in classification, and why exactly did this change the number of reads that could be assigned? How does this relate to the broader issue of misclassification and missing identifiers in RefSeq, which has been noted previously (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1554-6)? More explicit example use cases could be added. In figure 3, describing how to read the figure and e.g. identify misclassiciations would be useful.*
The first classification presented in Fig.4 (red) of our manuscript was based on the original Clark database. The second classification (green) was perfomed with the Clark database amended with all the *Pseudomonas* genome sequences used in this work. This second classification did not improve and even slightly decrease the number of affiliated reads. This is due to closely related genomes sequences that are affiliated to two distinct species name. The third classification (blue) was performed with the Clark database amended with all the *Pseudomonas* genome sequences affiliated to coherent groups. The number of classified reads increased quite significantly between the second and the third classification despite being composed of the same genome sequences.

4) *More care needs to be taken with some of the species and genus name terminology. Particularly, the words 'strain' and 'clique' kept appearing without definition. How do these terms relate to species and genus level differences?*

Based on Bergey's Manual of Systematic Bacteriology : « A strain is made up of the descendants of a single isolation in pure culture and usually is made up of a succession of cultures ultimately derived from an initial single colony ».
Since the term clique is not widely employed, we replaced this term by “group” throughout the text.

5) *Why was the Pseudomonas dataset used? What was the original species classifications breakdown, and how (quantitatively) did this compare with the reclassification? Is this one example sufficient? Other fast distance estimators have been run on all of RefSeq.*
We decided to work on *Pseudomonas* because this genus contains an important diversity of species (*n* = 207), whose taxonomic affiliation is under constant evolution and numerous genome sequences are available in public databases. A total of 350 groups of genomes sequences were obtained based on the percentage of shared *k*-mer. Replacement of SIMKA by Kmer-db allowed the calculation of pairwise similarity between all genomes sequences availables in 40 minutes at the time of analysis.

Reviewer 2. *Briand et al. describe a new approach for computing inter-genome relatedness based on the percentage of shared kmers. The main motivation for this project is that the computation time for computing many relatedness metrics, like the average nucleotide identity, can be prohibitive for many pairwise genome comparisons. I think this tool could be valuable for the field, especially after addressing a few issues which I think currently make the benefits of the tool difficult to evaluate (see below). In particular, I think the authors' tool could be great for running quality control on taxonomy assignments in genome databases. This quality contr this work is not a new approach to measure genome similarity but rather a workflow for performing and visualizing the output of such analysis in a simple way ol can be run using an interactive approach for visualizing genome relatedness that the authors have implemented, which I think could be used for quickly spotting problematic taxonomic assignments.*

Dear reviewer, thank you for your time and thoughtful review of the work. In the lines below, we have addressed your comments in full, point-by-point.

1- I think it would be important to clarify how the results of KI-S clustering in practice differ from other similar tools. One tool in particular is Mash, which was published in 2016. This tool can be used to rapidly compute distances between genomes after performing dimension reduction based on kmer counts. The motivation for Mash was to speed up calculations of inter-genome (and sequences in general) distances. In the Mash paper the authors describe their approach as comparable to ANI while being much faster and so I think it would be important to directly compare to Mash in terms of both the compute time and results. If the authors don't agree that Mash is a comparable tool then this should be explained.
Thank you for this suggestion. In the new version of the manuscript we have performed a comparison of several *k*-mers indexing softwares including Mash (see answer to reviewer 1). Mash was definitely the fastest software in term of computing time and performed well for closely related genome sequences (below the species level with ANIb values > 0.95). However, Mash is not recommended for estimation of genome sequence relatedness of more distantly related strains (ANIb < 0.95).

2- A related issue is that it currently is not clear whether using the % of shared kmers results in comparable genome clusters to existing approaches like average nucleotide identity. It seems like this would likely be the case, but I think this is important for the authors to clearly describe either way so that users can better evaluate the tool.
We have added the following information in the manuscript :
“The 934 genomic sequences were clustered in 329 and 315 groups at an ANIb value of 0.95 and 50% of 15-mers, respectively. The composition of these groups was identical between the two approaches for 302 groups that contained 808 genomic sequences. The 27 additional groups obtained with ANIb were nestled within the 13 additional groups derived from 50% of shared 15-mers”.

3 - *I also do not agree that they have shown evidence that their approach can be used to improve the taxonomic classification of metagenomics samples. This analysis focused on the percentage of classified reads, which cannot alone be used to evaluate how well a taxonomic classification performed*.
We agree with this comment. We only observed an increase of classified reads. This is now clarified in the text.

Other comments

4- T*he way KI-S is mentioned in the abstract makes it seems like it is a pre-existing tool, but on page 5 it sounds like the authors developed it from scratch – this should be clarified. Also, it is unclear whether all the steps like running Simka and the custom R script are run by KI-S itself. Lastly, it would be good to state what KI-S stands for, which I may have missed.*
This work is not a new approach to measure genome similarity but rather a workflow for performing and visualizing the output of such analysis in a simple way. We have clarified this aspect thorughout the text. KI-S stands for **K**inship relationships **I**dentification with **S**hared k-mers, which is a pretty mediocre play on the word "who it this" in French.

5- P3,L45 – I recommend re-wording to make the first few sentences of the Background a little clearer. In particular, it reads like specifically Bacteria vs Archaea are the taxonomic groups being delineated, rather than prokaryotic species in general.
This sentence has been reworded as follows : “Prokaryotic species delineation historically relies on a polyphasic approach”

6- Figure 1 – Axis labels are needed, which might be easiest to do if fewer panels were shown. In particular, it seems like K15-K20 are extremely similar so maybe a couple could be removed. It is also not clear to me from the figure legend what “the number of values by class in the subset of 934 Pseudomonas genomic comparison” refers to on the y-axis. I think this is the ANI / % shared kmers for every pairwise comparison of Pseudomonas genomes, but I think this could be clarified either way.
We have modified the Figure according to your suggestions.

7- When describing the % overlapping species in each peak in figure 1 – how were the cut-offs for which data points to include in each peak decided (e.g. what cut-offs of % shared kmers were used to call data points in peak 2?)
We have reworded this part to highlight that peaks and valleys reflect a genetic discontinuity. We illustrate with the peaks observed for k=15. We have no quantitative criteria to assess the size of these peaks.

8- P7 – The authors imply that using 15-mers is the best or at least equally good as higher kmer values. This decision is discussed in the discussion, but I think it would be useful to explicitly mention this decision here (esp. when contrasting the 15-mer and 20-mer comparisons for instance) – perhaps at the end of paragraph 1 of the results.
We have added the following sentence “Since increasing *k*-mer lengths beyond 15 did not improve the resolution of the multimodal distribution but leads to a more rapid drop in the percentage of shared *k*-mers between strains” in the first paragraph of the results.

9- P7,L132 – I think the paragraph starting with “Fifty percent of 15-mers is close to ANIb value of 0.95” would benefit by making it clear what the goal of these analyses were, possibly with something like this: “We next investigated what percentage of shared kmers corresponds to an ANIb value of 0.95, which is a common cut-off for delineating species”.
Thank you for the suggestion. We have added this sentence to the text.

10- P7,L145-147 – I am not sure what the sentence starting with “In addition, 15-mers allows the investigation of inter and intra-specific…” refers to and I think this should be clarified. One possible way to make the authors’ point clearer might be to contrast why they think this is true specifically for 15-mers and not the 10 or 20-mer distributions also shown in Fig 2.

Increasing *k*-mer lengths inflates the number of specific *k*-mers per genome sequence and then decreases the number of shared *k*-mers between genomes sequences. At high *k*-mers length, the percentage of shared *k*-mers become closed to 0 for high number of pairwise comparison and therefore prevents the study of their relatedness.

11- *P7,L149 – do these run times correspond to running the jobs on a single core? It would be useful to mention the memory usage as well if that's possible.*
These informations have been added in Table 1

12- *P8,L158-159 – "185 cliques were composed of a single genome sequences, therefore highlighting the high Pseudomonas strain diversity" – an alternative explanation would be that KS-I is incorrectly calling those genomes as individual cliques. If there are species (and strain) names for all genomes then that would be one way to evaluate whether these genomes are expected to be in different cliques or not.*
This information has been added in the text (see answer to comment n°2).

13- On a related note to the above it would be useful to compare the cliques identified based on KS-I compared to ANI-b – based on Fig 2 it looks like they would be extremely similar, but I do not think that is clear from the main text.
This information has been added in the text (see answer to comment n°2).

14- P8,L159 – I think using estimates of Chao1 alpha-diversity to estimate the expected number of Pseudomonas clusters would only make sense if you're considering genome cliques in a single environment (and at a particular time). I do not think the numbers of singleton and doubleton genomes in NCBI can really tell you about how many more Pseudomonas genome clusters are out there in general, if only because many Pseudomonas habitats have not been sampled.
We agree, alpha-diversity is not a robust estimate for assessing the overall *Pseudomonas* diversity. We have removed this part from the text.

15- Fig 3 – I really like the zoomable circle packing representation of the data – this seems like a great way to summarize the relationships between many genomes. It is not clear to me how novel this visualization approach is, but if the authors believe that it is novel then I would emphasize that more in the introduction and discussion.
The zoomable circle packing was intially developed by Mike Bostock, a developer of D3.js. In the present work, we only applied this data-visualisation for representing genomes relatedness. We have tried to insist on this visualization method in the new version of the manuscript

16- *P8 – I am not familiar with the term "clique" – maybe "cluster" would be clearer?*
Since the term clique is not widely employed, we replaced this term by "group" throughout the text.

17- P8 – It's not clear to me why changing the taxonomic label of the Pseudomonas genomes added to the database results in a higher proportion of classified reads. Is this because the CLARK algorithm tends not to collapse taxonomy to higher ranks if reads map to genomes associated with different species? If so, that is surprising to me, but I am not sure why else there would be a different in the proportion of classified reads. It would be useful to briefly explain why the authors think this is occurring. Unless I am missing something I also do not think this would make a difference for most metagenomics taxonomic classifiers like centrifuge, kraken2, and MEGAN.
The number of specific k-mers for each groups will be fewer in number if the groups are more heterogeneous in terms of classification (because genomes sequences will be more distantly related). Therefore the number of classified reads will ultimately decrease.

18- P11,L221-223 – I think these are great examples for how this tool could be used to clean up and perform quality control on taxonomy assignments in genome databases.
Thank you for your enthusiasm

19- P11, L229 – end – As mentioned above I do not fully follow why more reads were classified with CLARK after changing the taxonomic labels of the Pseudomonas genomes, but either way I

do not think this is evidence that the taxonomic classifier is actually working better as indicated in the concluding paragraph currently. I think some sort of validation would be needed to be able to state that first organizing the genomes into cliques actually improves taxonomic classification. This is difficult to do because we almost never know the right answer in microbiome datasets. However, one potential way to do this would be to create a simulate dataset enriched for Pseudomonas (ideally with metagenome-assembled genomes from the seed datasets) and then compare the relative abundances of the taxa inferred using the 3 approaches mentioned in Fig 4 to the expected relative abundances.
We have clarified the fact that we only observed an increase of classified reads.

Example of grammatical errors
Lastly, there are numerous grammatical errors throughout the manuscript – I have made a non-exhaustive list of example errors and possible, which hopefully will be useful for the authors.

Thank you for these corrections. We have changed the text.

- P2,L29-30: "…datasets composed of thousand genome sequences" change to "datasets composed of thousands of genome sequences".
- P2,L31 – "kmers counts" should be "kmer counts"
- P3,L64 – "for one pair of genome sequence" should be "one pair of genome sequences"
- P4,L71 – "classifiers differ in term" should be "classifiers differ in terms"
- P4,L72-73 – "for affiliating read to a" should be "for affiliating a read to a taxonomic rank"
- P4,L81 – add "the" before "relatedness"
- P5,L100 – should be "were selected" instead of "was selected"
- P6,L116 – need to add either "the" or "a" in front of "common bean" depending on which is correct
- P7,L143 – "Fifty percent of 15-mers is close to ANIb value of 0.95" should be "Fifty percent of 15-mers are close to an ANIb value of 0.95".
- P7,L153 – "used to investigate relatedness" should be "used to investigate the relatedness"
- P10,L188 – "prohibited its used for comparing" should be "prohibit its use for comparing"
- P11,L216 - "based ANIb" should be "based on ANIb"
- P11,L219 – "Moreover, KI-S tool, provides…" should be re-written, perhaps as "Moreover, KI-S includes…"