**Authors’ response to the reviewers**

**Reviewer: 1**

Comments to the Author

The manuscript by Ruijie Xu et al. presents a comparison among three popular software (i.e., Kraken2, CLARK, and CLARK-s) to generate microbial profiles from metagenomics data. The different tools are applied on three sets of wild rodent tissue samples. Results demonstrate a great discrepancy among the different software, which may be responsible of significant differences in microbial identification and diversity within and between samples.

The topic involved in the paper is suitable for publication in Journal of Applied Microbiology.

Overall, the conducted analysis is interesting for the microbiome community.

However, I feel that major modifications are necessary before a possible publication of the paper:

**1. The methodological comparison is done on three software. Although I understand that it is not possible to take into consideration all available tools, a much larger set of tools for taxonomic profiling exist in the literature. As example, please see this recent publication/comparison: "Critical Assessment of Metagenome Interpretation - the second round of challenges" (**[**https://www.biorxiv.org/content/10.1101/2021.07.12.451567v1**](https://www.biorxiv.org/content/10.1101/2021.07.12.451567v1)**). I think that additional tools should be considered in this comparative analysis.**

We thank the reviewer for this suggestion. As such, we have included nine of the most used taxonomic profiling software in the revised version of the manuscript to demonstrate how the selection of software can impact taxonomic profiles and associated downstream analyses. We included the software Blastn, Kraken2, Bracken, CLARK, CLARK-s, Centrifuge, Metaphlan, Diamond, and Kaiju. All the software and databases included in the manuscript as well as associated runtime information are presented in Table1 of the new version of the manuscript.

**2. The analysis is done on shotgun data which can reach species-level resolution. However, major comparisons are done at higher taxonomic levels as summarized in the main figures of the manuscript. It would be interesting/relevant to provide more species-level insights.**

We have added comparisons of classified profiles at the species-level using different software to the current manuscript. These are presented in TableSII.4 and described in the manuscript in lines 268-287. Furthermore, all microbial community characterization (lines 292-311) and the differential abundance analysis (lines 336-346) were performed using species-level metagenomics profiles in the revised manuscript. The differentially abundant taxa identified using taxonomical profiles classified by different software were also reported at the species-level resolution and are presented in Figure 5a.

*[Lines 268-287]:*

*“For the species level classification, the number of reads classified under the same species was by each software is available in Table SII.1. Out of all software, Metaphlan3 classified the least number of species taxa, with only 18 species (Table SII.4) while Kaiju classified the most, 4128 species (Table SII.4). From the species level classifications, 9 species taxa were identified by all nine software (Leptospira interrogans, Leptospira borgpetersenii, Faecalibacterium prausnitzii, Bordetella pseudohinzii, Bordetella bronchiseptica, Bordetella pertussis, Bacteroides uniformis, Phocaeicola vulgatus, and Bartonella elizabethae) (Table SII.1). Centrifuge vs Kaiju had the largest overlap of number of identified species taxa (2285), followed by Kraken2 vs Centrifuge (1737) and kraken2 vs. Kaiju (1723) (Table SII.4). The species-level classification of these software mentioned above shared a total of 1379 species taxa. In addition, BLASTN shared 1253 species taxa with Centrifuge, 1207 with Kaiju, and 1126 with Kraken2. CLARK and CLARK-s’ classification shared 1219 and 1059 species taxa with Kaiju, respectively. To assess if different software had identified the same species taxa as the most abundant taxa, species taxa with at least 10% of the reads from each sample were selected from each software’ classification. Metaphlan3 identified most of the number of unique species taxa (18), while BLASTN and Kaiju identified the least (7). CLARK vs. CLARK-s and Kraken vs. Bracken shared most of the number of taxa in this category (9 and 8, respectively). Two species taxa were identified by all software as the top ten percent most abundant species taxa, which were L. interrogans and Bartonella elizabethae (Table SII.1).”*

*[Lines 292-311]:*

*“For species richness characterization within a community using the Shannon indices, only the indices obtained from minikraken DB were significantly different from the results obtained with the other DBs (Figure 3b, Table SI.4). Moreover, the characterization using the Simpson indices was mostly similar between the results of the four DBs (Figure 3c, Table SI.4). Only the Simpson indices obtained from the results of the standard and customized DBs comparison were significantly different (Figure 3c, Table SI.4).*

*The number of unique observed taxa (Table SII.5, Figure 3d) across different software were largely divergent from each other. Out of the 36 pairwise comparisons between different software, only 6 comparisons were not significantly different (Table SII.5), which were BLASTN’s observed taxa with Kraken2, CLARK, and CLARK-s, comparison between CLARK and CLARK-s, and comparison between Centrifuge and Kaiju. The Shannon indices showed more similarity between software than the unique observed taxa, however, they still had 23 out of 36 comparisons between software significantly different from each other (Table SII.5, Figure 3e). The Simpson indices were least impacted by the differences in classification results across the software. Only 7 out of 36 comparisons were found to be significantly different (Table SII.5, Figure 3f). Most of these were identified in comparisons between CLARK-s (3/7) and Centrifuge (4/7) with other software. The Simpson index between CLARK-s and Centrifuge’s classifications were also significantly different from each other. “*

*[Lines 336-346]:*

*“DA taxa between samples of different tissues were identified to show the most significantly different microbial taxa between the microbiome of two tissues. For DA taxa identified from lung versus kidney samples at the species level, the number of DA taxon identified by the use of different software ranged from 10 (Diamond) to 596 (Centrifuge) (Table SII.7, Figure 5a).  The abundance was significantly higher in the kidney than in the lung samples for all software’ classifications (Figure 5b). Five significantly abundant species (Bordetella pseudohinzii, Bordetella bronchiseptica, Leptospira interrogans, Leptospira borgpeterseni, and Mycoplasm pulmonis) were classified by all software (Table SII.7). Kaiju and Centrifuge had the highest number of distinct DA taxa (390 and 376 taxa, respectively) (Figure 5a). Although Centrifuge identified the largest number of DA species taxa, Kaiju identified the highest number of unique phylum taxa (42), which means that many of Centrifuge’s DA species have the same phylum taxonomy (Figure 5a).*

*“*

**3. It would be also interesting to report some empirical evaluation in terms of computational load and usability of the compared tools.**

Computational resources (CPU and memory) and time used to build databases and to classify each individual sample were presented in Table 1 of the revised manuscript, described in the Results section in lines 183-187, described in detail in Supplementary Text 1, and discussed in the Discussion section in lines 451-465. All analyses were performed using the University of Georgia’s high-performance computing cluster at the Georgia Advanced Computing Resource Center. A total of 12 threads of CPU were used for all software analyses. Time and memory consumption was recorded at the end of each analysis.

*[Lines 183-187]:*

*“Details of the DBs used for each software in this study, as well as the associated computational resources and building time, are available in Table 1 and described in the Supplementary Text 1.  Alignment-free software, CLARK and CLARK-s required the most computational resources and time for DB building (Table 1). On the other hand, alignment-based software, BLASTN and Diamond, took the longest time for microbial profiling.”*

*[Lines 451-465]:*

*“Metagenomics software can be classified into two different categories, alignment-based and alignment-free. The alignment-based software, which suffers greatly from slow speed and the need of large resources, are generally thought to have high sensitivity. On the other hand, the alignment-free software uses relatively small computational resources and significant improvement in speed of the analysis. In our study, the two alignment-based software, BLASTN and Diamond, were the two most time intensive software. They took two and five hours, respectively, on average to complete the analysis for one sample, while other software took at most three minutes for the same task. The time and resources required to build the DBs for the alignment-free software became the trade-off for the speed of the analysis itself. For example, the building of CLARK’s DB took almost two days with 400 GBs of memory used. Fortunately, most of the software included in our study have pre-built DBs distributed with the release of the software (except for CLARK, CLARK-s, Diamond, and Kaiju). However, if the analysis requires the identification of taxa that are not included in these pre-built DBs, the time and resources added to the metagenomics profiling analysis will increase significantly. ”*

**Reviewer: 2**

Comments to the Author

GENERAL COMMENTS

The present manuscript tries to describe the different outcomes of different metagenomic analysis softwares used for microbial profiling and pathogen detection. In general, the manuscript is not badly written and study could be of interest, however, authors failed in several aspects:

1. **The sections within the manuscript are mixed and often repeated: introduction-like sentences are found in the methods or results, methods are found in the results, results are repeated in the discussion and the conclusions are vague and do not add up anything. Results, which are clearly observed in the figures, are almost not explained in the results section nor discussed or compared. Non-scientific terms are also abundant and should be eliminated.**

We thank the reviewer for the feedback. We revised the manuscript accordingly making sure that each section has only the appropriate relevant information and that there are no repetitions across them. We also eliminated all non-scientific terms from the text. The Results section in the revised version of the manuscript was edited to remove all introduction-related sentences and methods descriptions, and all the new results related to the analysis of the extra software were added here and to the Supplementary Material. Figures and tables are described, explained, and compared in the results section. Furthermore, a new paragraph was added to the end of the discussion section (lines 535-553) summarizing the most important findings of the study.

[lines 535-553]:

*“In conclusion, our study found that alignment-based software does not necessarily have better sensitivity in microbial profiling than alignment-free software. Diamond, one of the alignment-based software included in our analyses, reported the lowest sensitivity in DA analyses compared to other software. However, within the alignment-free software included in this study, two index-based software, Centrifuge and Kaiju, were found to be more sensitive than other software in microbial profiling, DA analysis, and pathogen detection. Metaphlan3, developed with a marker-based alignment-free algorithm, was found to have the lowest sensitivity in all the analyses when compared to all the other software included in this study. For the microbial community analyses, the characterization of within-samples microbial richness was largely impacted by software selection, but the impact was less significant if the characterization index used species abundance to weigh the index. A similar observation was found in the microbial community characterization analyses with different DBs. The within-sample richness characterization was mostly consistent when weighed by the species abundances within a sample. In addition, the presence of host genomes in the DBs does not largely impact the microbial profiling, but the overall compositions of microbial genomes included in the DBs impact the microbial classification the most. Decrease in the composition of microbial genomes in the DBs will largely decrease the sensitivity of the microbial classification. Moreover, we also found that the selection of the DBs can impact the ability of pathogen detection.”*

1. **Authors only talked about three software but there is other metagenomics software. Please see the article of Pérez-Cobas et al, 2020. Kraken and CLARK software are reference-based software and non-alignment-based, but there are other approaches. So authors should also discuss if not only the software but the method behind it could be a bias source.**

We thank the reviewer for this suggestion. As such, we have included nine popular taxonomic profiling software in the revised version of the manuscript to demonstrate how the selection of software can impact the results of the downstream analyses. We included two alignment-based software (Blastn and Diamond), and seven non-alignment-based software (Kraken2, Bracken, CLARK, CLARK-s, Centrifuge, Metaphlan, and Kaiju). All the software included in the manuscript as well as associated databases and runtime information are presented in Table 1 of the new version of the manuscript. No clear differences were identified between microbial profiles classified with alignment-based vs. alignment-free software. We added a paragraph to our discussion summarizing that within the alignment-free software, the index-based alignment-free software, Centrifuge and Kaiju showed the best performance in sensitivity, while in the marker-based alignment-free software, Metaphlan3, showed the worst performance in sensitivity (see [lines 534-553] in the response above).

SPECIFIC COMMENTS

L46: please change utilizes by uses.

This change was performed throughout the manuscript.

L47: 16S rRNA amplicon sequencing and shotgun metagenomics approaches.

The sentence has been rewritten from “...approaches: the 16S rRNA and the shotgun metagenomic sequencing-based approach (Jovel” to

[Line 42]: “...approaches: the 16S rRNA amplicon sequencing and shotgun metagenomics sequencing approaches...”

L52: How this is written seems a contradiction. Please, re-write.

The sentence has been rewritten from “Although lower in cost (Breitwieser, Lu and Salzberg, 2019), 16S rRNA markers are only available in the genomes of most bacteria and archaea (Woese, Kandlert and Wheelis, 1990;Janda and Abbott, 2007).” to

​​[Lines 48-50]: “Although lower in cost (Breitwieser, Lu and Salzberg, 2019), 16S rRNA amplicon sequencing are limited to profiling the genomes of bacteria and archaea, and are subject to amplification biases (Woese, Kandlert and Wheelis, 1990; Janda and Abbott, 2007).”

L64: provide references

This sentence has been removed from the introduction section; therefore no references need to be added at this point.

L72-75: ? a dataset is one thing a software is a tool.

This sentence has been removed from the introduction section.

L75:remove real-word, it does not sound scientific. Moreover, the reader does not understand which message the authors want to convey. It is confusing.

This sentence has been removed from the introduction section and all occurrences of the expression ‘real-world’ have been deleted from the manuscript.

L82: please use past tense

This part of the introduction has been shortened in the new version of the manuscript, but we have made sure to use past tense in the new description of our study:

[Lines 102-109]: “*In this study, we compared the microbial profiles of tissue samples from two species of Rattus (Rattus rattus and Rattus norvegicus) using different metagenomic software and DBs. Specifically, we 1) compared the taxonomical profiles classified by four DBs and nine metagenomics profiling software listed above, 2) determined the effect of their differences in the downstream analyses and in the result interpretation; and 3) identified the presence of zoonotic pathogens, Leptospira, in the rat kidneys that we collected using each software’ profiling results and compared the diagnosis of the pathogen with that of the traditional laboratory methods.*”

L85; in rat kidneys that you collected or from a dataset. Please specify!

The paragraph where this sentence has been shortened, but we have included the reviewer’s suggestion and specified from where we identified the presence of *Leptospira* (see reply in previous comment).

[Lines 102-109]: “...*analyses and in the result interpretation; and 3) identified the presence of zoonotic pathogens, Leptospira, in the rat kidneys that we collected using each software’ profiling results and compared the diagnosis of the pathogen with that of the traditional laboratory methods.*”

L80-93: authors should be concise about their objectives. Methods should be explained later.

This part of the introduction has been shortened and the objectives of the study are now written in a much more concise way:

[Lines 102-109]: “*In this study, we compared the microbial profiles of tissue samples from two species of Rattus (Rattus rattus and Rattus norvegicus) using different metagenomic software and DBs. Specifically, we 1) compared the taxonomical profiles classified by four DBs and nine metagenomics profiling software listed above, 2) determined the effect of their differences in the downstream analyses and in the result interpretation; and 3) identified the presence of zoonotic pathogens, Leptospira, in the rat kidneys that we collected using each software’ profiling results and compared the diagnosis of the pathogen with that of the traditional laboratory methods.*”

L95-100: eliminate. These are results!

This sentence has been removed from the introduction section.

L133-141: software characteristics should be explained in the introduction, not here in methods. This sentence has been removed from the methods section and moved to the introduction.

[Lines 58-81]: “*Currently developed shotgun metagenomics sequencing-based taxonomical profiling software can be separated into two groups: the alignment-based and the alignment-free software. Alignment-based software, including BLASTN (Altschul et al., 1990; Johnson et al., 2008; Camacho et al., 2009), which aligns sequences at the nucleotide level, and Diamond (Buchfink, Xie and Huson, 2015), which aligns at the protein level, were thought to have high sensitivity and have been used as the standard for metagenomics profiling. However, these software require a large amount of time and computational resources to build genome alignments for the high number of sequences usually involved in metagenomics profiling studies (Cannings, 2004; Zielezinski et al., 2017). Furthermore, recent investigations in alignment-based methods have reported that alignment-based software decrease in sensitivity with the use of mosaic genomes (ex. viruses) (Zielezinski et al., 2017). To overcome these limitations, multiple software has been developed using alignment-free algorithms. For example: 1) Kraken2 (Wood, Lu and Langmead, 2019, p. 2) and CLARK (Ounit et al., 2015) were designed with k-mer matching algorithms, where only substrings of sequences were matched (Healy and Chambers, 2014); 2) Metaphlan3(Truong et al., 2015; Beghini et al., 2021) was designed to identify unique genetic markers within each microbial taxon; and 3) Centrifuge (Kim et al., 2016) and Kaiju (Menzel, Ng and Krogh, 2016) were designed to optimize the time and resources of profiling by compressing the reference microbial genomes into the index structures for storing and searching (at the nucleotide and protein levels, respectively) (Burrows and Wheeler, 1994). In addition to the software mentioned above, some methods were developed to improve the results of existing software, such as Bracken (Lu et al., 2017) that improves Kraken2’s output by eliminating false positive assignments using a Bayesian framework, and CLARK-s (Ounit and Lonardi, 2016) that improves the sensitivity of CLARK with the use of spaced k-mers.*”

L182-184: these are methods

This sentence has been removed from the results section.

L189-190: it reads more like a discussion

This sentence has been removed from the results section. In addition, discussion sentences in the results section have been moved to the discussion in the current manuscript.

L192: were found significantly different is not grammatically correct. In a higher proportion you mean?

All occurrences in the manuscript of “significantly different” have been replaced by “with no statistically significant differences”.

L192-195: again, this seems like a discussion, because authors are comparing. Please use a normal tone in the results and discuss them only in the discussion section

This sentence has been removed from the results section.

L206: why did the authors do not go deeper into the classification? They would see more differences at phylum level.

Lower-level DB comparisons have been added to the current version of the manuscript in Supplementary Text 2. More differences were described at the genus and species level.

L228-233: these are methods

This sentence has been removed from the results section.

L245-247: same

This sentence has been removed from the results section.

L274-276: same

This sentence has been removed from the results section.

L289-291: same

This sentence has been removed from the results section.

L295-297:same

This sentence has been removed from the results section.

L299-302: same again

This sentence has been removed from the results section.

L306-307: same

This sentence has been removed from the results section.

L322-324: same

This sentence has been removed from the results section.

L327-330: same

This sentence has been removed from the results section.

L365-377: it is a repetition of the methods

This sentence has been removed from the results section.

L377-388: this is a repetition of you rational, which you already wrote in the introduction. Do no repeat information so often, which has been already given!!

This sentence has been removed from the discussion section.

L390-410: this is not a discussion but belongs to the introduction

This sentence has been removed from the discussion section. A similar statement has been added to the introduction at line 95-98.

[Lines 95-98]:” *The effect of using different DBs to classify microbial profiles and their impact on the downstream microbial characterization and pathogen detection have not been addressed in previous benchmarks.*”

L412-424: this is not a discussion. Authors should not provide background info but discuss their results instead.

This sentence has been removed from the discussion section.

L424-426: why would we assume that? This information is vague and does not add anything. Please elaborate or eliminate.

This sentence has been removed from the discussion section.

L427-432: this are methods again!

This sentence has been removed from the discussion section.

L432-436: these are results but are not discussed properly. Please elaborate.

This sentence has been removed from the discussion section. The detection of Leptospira in the discussion has been re-written in line 523-5534.

[Lines 523-534]: *“For comparisons in sensitivity to identify the presence of the zoonotic pathogen Leptospira in all of our tissue samples, Centrifuge and Kaiju were found the most sensitive software in diagnosing Leptospira, where Centrifuge reported the presence of Leptospria in all the samples. Since Leptospira colonizes the kidney of rats (Adler and de la Peña Moctezuma, 2015), we compared the results from  three traditional methods (PCR/DFA/Culture) applied to kidney samples reported in a previous study (Rajeev et al., 2020). We found that most software included in our analysis had similar sensitivity in Leptospira identification with traditional methods, except for PCR. In addition, Centrifuge reported the presence of Leptospira in samples that were not reported by any other software or a traditional method. This identification could be due to Centrifuge’s better performance in sensitivity, or as a result of false positive reporting. Furthermore, we found that Kraken2 with maxikraken DB also reported Leptospira’s presence in all samples.*”

L439: which conventional methods? Where? Elaborate.

Removed from the discussion. In the current version of the manuscript, the conventional method was listed in the discussion in lines 527-528.

[Lines 527-528]: “*we compared the results from three traditional methods (PCR/DFA/Culture) applied to kidney samples...*”

L442-455: this is a mix of intro, methods and repetition of results

This sentence has been removed from the discussion section.

L454: no longer exist? What do authors mean?

Removed from the revised manuscript. More software was included in the study. The pairwise comparisons between CLARK-s and Kraken2’s beta diversity were replaced by beta diversity comparisons across the microbial profiles of all 9 software in line 501-504.

[Lines 501-504]: “*The characterizations of the relationships between-samples were divergent across software, but the most discriminatory relationships within the rat samples (between the lung and other samples) were captured by most of the software (except for Metaphlan3).*”

L455: sorry but this is coming out of nowhere. What is the confounder? Otherwise, it is pure speculation!

This sentence has been removed from the discussion section.

L458-459: popular analysis? It does not sound scientific nor adds anything to your manuscript.

This sentence has been removed from the discussion section.

L459-463: these are sentences without reference but also the reader does not get what the message authors want to convey here is.

This sentence has been removed from the discussion section.

L464-467: again, repetition of the methods, rational etc.

This sentence has been removed from the discussion section.

L468-473; these are results

This sentence has been removed from the discussion section.

L473-475: reference is missing. And what are those conclusions? What do authors propose then? Elaborate!

Instead of saying that selection of software “can produce misleading biological conclusion”, we changed the sentence to the selection of the software could lead to “different biological conclusions” in lines 395-396.

[Lines 395-396]: “...*potentially could lead to different biological conclusions*...”

L477: there is not such thing as a significant biological conclusion. Authors repeat again the same vague conclusion.

This sentence has been removed from the discussion section.

L480: real-world data? Eliminate

All occurrences of “real-world” have been removed from the manuscript.

L484: real-world again

All occurrences of “real-world” have been removed from the manuscript.

L480-498: conclusions are extremely vague and do not add up anything new.

A clearer conclusion was obtained with the new analyses of the additional software. The first sentences of the conclusion were rewritten:

[lines 535-553]:

*“In conclusion, our study found that alignment-based software does not necessarily have better sensitivity in microbial profiling than alignment-free software. Diamond, one of the alignment-based software included in our analyses, reported the lowest sensitivity in DA analyses compared to other software. However, within the alignment-free software included in this study, two index-based software, Centrifuge and Kaiju, were found to be more sensitive than other software in microbial profiling, DA analysis, and pathogen detection. Metaphlan3, developed with a marker-based alignment-free algorithm, was found to have the lowest sensitivity in all the analyses when compared to all the other software included in this study. For the microbial community analyses, the characterization of within-samples microbial richness was largely impacted by software selection, but the impact was less significant if the characterization index used species abundance to weigh the index. A similar observation was found in the microbial community characterization analyses with different DBs. The within-sample richness characterization was mostly consistent when weighed by the species abundances within a sample. In addition, the presence of host genomes in the DBs does not largely impact the microbial profiling, but the overall compositions of microbial genomes included in the DBs impact the microbial classification the most. Decrease in the composition of microbial genomes in the DBs will largely decrease the sensitivity of the microbial classification. Moreover, we also found that the selection of the DBs can impact the ability of pathogen detection.”*