Response to reviewers:

First of all, we would like to thank the Editor and reviewers for your comments, because the paper is in much better shape with your suggestions being added to it. Point-by-point responses are listed below. Thank you!

Editorial comments:
Editorial Changes
Changes to be made by the Author(s):

Thank you for your comments and considerations.

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.

Modified accordingly.

2. Please adjust the numbering of the Protocol to follow the JoVE Instructions for Authors. For example, 1 should be followed by 1.1 and then 1.1.1 and 1.1.2 if necessary. Please refrain from using bullets or dashes.

Modified accordingly.

3. JoVE policy states that the video narrative is objective and not biased towards a particular product featured in the video. The goal of this policy is to focus on science rather than to present a technique as an advertisement for a specific item. To this end, we ask that you please reduce the number of instances of "_______" within your text. The term may be introduced but please use it infrequently and when directly relevant. Otherwise, please refer to the term using generic language. For example, R software, RStudio, ProkEvo, etc.

Modified as best as we could.
We did the best we could here because the R scripts and analytical platform was designed specifically to complement the ProkEvo pipeline (which is freely available). We are not promoting the use of the R software or RStudio, but all the scripts and software are completely free for any user anywhere in the world. We tried as best as we could to minimize the number of times we mentioned the names though, but in certain key places we could not avoid it.

4. The Protocol should contain only action items that direct the reader to do something. Please move the discussion about the protocol to the Discussion.

Modified accordingly.

5. Please revise the text to avoid the use of any personal pronouns (e.g., "we", "you", "our" etc.).

Modified accordingly.

6. Please include a single line space between each step, substep, and note in the protocol section. Please highlight up to 3 pages of the Protocol (including headings and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader. Please keep in mind that software steps without a graphical user interface (GUI) cannot be filmed.

Modified accordingly.

7. Please refrain from using bullets in the representative results.

Modified accordingly.

8. Please include all the Figure Legends (including supplementary figures and tables) together at the end of the Representative Results in the manuscript text.

Modified accordingly.

Reviewers' comments:

Reviewer #1:

Manuscript Summary:

The manuscript presents a protocol for the analysis of bacterial pangenomes using the ProkEvo workflow previously published by the same authors. It is demonstrated how from a set of whole genome sequencing files one can generate diagrams with population structure, classify bacteria into sequence types, and obtain a list of the identified genes known to confer antimicrobial resistance. The presentation is extensive and accompanied with R scripting code snippets and anticipated results. The protocol may be potentially useful if the authors make it simple to use and demonstrate that it is better in some way than other existing analytical pipelines in the field.

Thank you for your comments and considerations.

Minor Concerns:

1. The original publication in the PeerJ journal presents ProkEvo as a complex analytical workflow that employs many third-party tools. I am not sure whether the presented protocol in this manuscript covers all dependencies in the section where installation steps are described. Maybe a table of dependencies and corresponding

steps to install them will give a better view on the pre-requisites to this protocol to successfully execute.

This is an important point that we have clarified now. ProkEvo is the bioinformatics platform that processes WGS data into .csv outputs that we are now using in this paper to show the R-based statistical analysis. For this paper the user does not need to run ProkEvo but we point it out because the outputs are all generated through it (previous publication of the ProkEvo paper which is cited in this paper). Our R-based scripts basically complement ProkEvo, but its premise and approach could be used with any output that contains the same kind of information (ST classification, AMR mapping, etc.). ProkEvo does not create any problems with dependencies. In fact, for consumers of WGS data, ProkEvo helps tremendously because in a single installation step the user gets the most up-to-date version of all packages contained in it. Now, as for the R scripts, there are R libraries that need to be installed, and those are now all clarified inside of the paper. Thank you for highlighting this point because it helped us clarifying the purpose of this paper more objectively which is to conduct a population-based analysis assuming the user has those outputs to begin with (which are produced by ProkEvo as stated).

2. The authors objectively discussed the shortcomings on the protocol. However, I think 2 major points are still missing or under-described in the capabilities of the protocol: (1) The identification and report of mutations in the AMR genes. Frequently, mutation in the targeted gene drives the antimicrobial resistance. (2) The use of this protocol to analyze a single sample is unclear. For example, contrasting a given sample to the collection of the already annotated strains would be useful.

Thank you for bringing these points up. Answer to point 1: ProkEvo is not set up for identification of mutations or alleles as it is, although it could, but this type of analysis is out of the scope of this paper. We clarified that limitation in the text.
Answer to point 2: Yes, ProkEvo can be run with a single sample to get genotyping and genome annotation including AMR gene and plasmid mapping, but the purpose of this paper is to carry out a population-based analysis. That is also now clarified in the text.

3. The authors should discuss or demonstrate in results as to why their protocol is any different or advantageous compared to the existing tools, for example
PATRIC at https://patricbrc.org/
iTOL at https://itol.embl.de/
BacWGSTdb at http://bacdb.cn/BacWGSTdb/

We have highlighted the potential for users to use web-interface or GUI type of tools in comparison to ours, but there are tremendous advantages in using ProkEvo and these R scripts, such as: customizing your own analysis, deploying the software in high-performance computers or private cloud environments, adding or removing programs, exploring the data and conducting statistical analysis using domain expertise, changing the visualization as the user wishes, etc. We certainly acknowledge the importance of other tools that are out there, and have also used them, and ProkEvo is not here to be a silver bullet, but instead it is in our view that this approach is very useful for those who would like to deploy a scalable program for WGS analysis and tailor it to their specific needs.

Reviewer #2:

Manuscript Summary:

The authors discussed the application of their ProkEvo computational platform for performing population-based genomic analysis. They used R to discuss the hierarchical analysis based on a sample dataset. In general, the protocol, applications, and results are well-explained in the manuscript.

Thank you for your comments and considerations.

Major Concerns:

I don't have any major concerns.

Minor Concerns:

I have a couple of comments/questions.

1. The authors should include system requirements (processor, memory, etc.) to run their R-based framework.

We clarified that, but the .csv files used as an input are very light, therefore, this analysis could be done virtually in any laptop, and certainly in any high-performance platform.

2. Their computational platform supports analysis for a limited number of genomes. What can be the maximum input size? Is it possible to overcome this limitation?

ProkEvo has one major limitation (with the tools it contains) that we can detect which is the number of genomes used, if the user cares about generating a core-genome alignment through Roary (which is inside of ProkEvo). If the user turns off that option, it scales to many thousands of genomes at once. We don't have a cap right now because we have only run a 10-30,000 at once and had no problem. But we have clarified that in the text.

3. The authors mentioned that their framework can be generalized for python and other languages. How can it be generalized? The manuscript needs to add some explanations to it.

Thank you for bringing that up as well. What we meant with that was that if the user does not want to use R, he or she could simply apply the same principles or concepts with Pandas in Python. But we have clarified that in the text as well.

4. The manuscript should include any limitations regarding time & space complexities? Is it possible to perform parallel computation using their framework?

ProkEvo is already set up for that and the original publication which is cited in this paper demonstrates that very aspect. But that is not part of this paper, because what the work presented in this paper focus on using the .csv files produced by ProkEvo to conduct a statistical analysis of the data, which as it is has not required parallelization, because the output files are really light and easy to work with in a simple laptop setting.

5. If possible, please improve the visualization of the statistics/results.

Thank you for suggesting that, but we have kept our visualizations as they are. We totally respect your comment though.