Department of Botany and Plant Pathology
Oregon State University, 2082 Cordley Hall, Corvallis, Oregon 97331-2902
T 541-737-3451 | **F** 541-737-3573 | www.science.oregonstate.edu/bpp

May 24, 2016

Dear Dr. Min Zhao,

We thank you and the three reviewers for the comments on the manuscript and the careful assessment of the described online tools. We found the comments and suggestions very helpful in improving the quality of the submitted manuscript.

Please find, as an attachment to this cover letter, a rebuttal (text in blue) to each point from each reviewer. We also addressed the two technical changes by placing the MATERIAL & METHODS section after the INTRODUCTION and before the RESULTS AND DISCUSSION section, and moved the funders' statement from the "Funding statement" to the "Acknowledgements". We uploaded a .docx version with track changes.

We hope that our explanations and changes are sufficient to satisfy the reviewers and look forward to a favorable decision.

Thank you,

Jeff Chang

Reviewer 1

1, The authors need to include the role of phylogeny, how the pathogen can evolve through time, integrated genes or genomic islands from closely related organisms? The bacterial genome evolve very dynamically.

We agree with reviewer 1 that these evolutionary processes innovate genomes. In the pathogens that we highlighted, plasmids are the main horizontally acquired elements that are necessary for pathogenesis. However, the purpose of this work was to develop tools to aid in strain identification. In order to do so, we keyed in on sequences that are conserved and located on the chromosomes. As such, the Gall-typing tools do not allow users to determine whether an isolate is pathogenic or not. Users can make inferences, based on the presence of homologs of virulence genes, which, for the gall causing pathogens, are strongly correlated with virulence. We recommend the user test the criteria described as Koch's postulates to draw conclusions on pathogenicity of an isolate.

2. I want the authors to clarify the affect of the antibiotic resistant genes, genome mobility, CRISPRs, transposable elements on the virulence of this bacteria.

Yet again, we agree that these elements are innovators of bacterial genomes. However, unless we are missing out on a different point, we hope that our rebuttal to point one addressed this request.

3. Genome sizes are more conserved at all tested taxonomic levels than 16S rRNA copy numbers. Only a minority of bacterial genomes harbors identical 16S rRNA gene copies, and sequence diversity increases with increasing copy numbers. How do the authors account for such diversity?

This is yet another good point. We do not account for diversity of 16S sequence. However, as stated in the manuscript, 16S has been extensively used, despite the stated limitations and we indicated its appropriateness for, at best, drawing inferences on genus-level classifications. We also provided tools for higher resolution analyses that overcome the limitations of 16S analysis (MLSA, WGS pipeline). We are providing tools for end users to analyze their data and have stated the appropriateness of the tools.

4. "(MLSA) leverages the phylogenetic signal from four to ten genes" are these the housekeeping genes/ conserved genes ? Why is the set limited to 4-10, what is the rate of conservation of these genes.

The set of housekeeping genes that we used were developed and validated by other groups (please see citations). The 4-10 genes mentioned is not a hard limit, but rather a generalization to the most frequently used range of marker genes. The use of fewer than 4 genes often provides insufficient phylogenetic signal to accurately model species phylogeny, while more than 10 provides diminishing returns and is more time consuming to prepare.

5. Line 67: "phylogenetic analyses can be removed from studies to allow for robust analyses." I am a bit confused about this statement, can the authors please explain this in a better way?

Just to be clear, the statement is that "...sequences that violate assumptions of phylogenetic analyses can be removed…"not that phylogenetic analyses can be removed. One of the key assumptions in phylogenetic analysis is the sequence is vertically inherited. With whole genome sequences, the frequency and linkage of SNPs can be analyzed to infer regions acquired horizontally (HGT or recombination) and removed prior to phylogenetic analysis. To limit confusion, we state on line 69, as a parenthetical clause, (e.g., not vertically inherited).

6. Line 68: Are the authors aware of denovo assembly?, the later being more powerful as it takes into account structural changes in the genome, as bacterial genomes are highly mosaic, there is no dependence on any reference genome. Although sometime it is better to use a combination of de novo assembly and genome-guided assembly.

We agree with the reviewer that a *de novo* assembled genome is more powerful for the purposes stated by the reviewer. However, the tools described in the manuscript are for a different purpose. The tools are used to draw conclusions from conserved and vertically inherited sequences. Aligning reads to a reference is the quickest and best method. If the reviewer is referring to making inferences by aligning reads to search for homologs of virulence genes - then our response is that because we provided web tools, we are limited by the infrastructure and cannot provide online tools for de novo assembly and comparative approaches.

7. Line 77: I would like to stress on the authors need to shed some light on the genomic islands, pathogenicity islands that can be acquired by the bacterial genome? The virulence of the bacteria as stated in this paper is mainly caused by horizontal gene transfer, providing a brief background on that would help the readers.

On line 77, the manuscript describes advantages to the user for generating whole genome sequences. This section was not intended to describe the advantages for the tools that we developed. As stated in responses to previous comments, the tools require vertically inherited, conserved gene sequences. In each of the sections that follow, we indicate that plasmids are critical to pathogen virulence (lines 99, 119, 125, 141).

8. "A Ti plasmid imparts, " change to "A Ti (tumor inducing plasmid) plasmid imparts", how big is this plasmid and which region does this gets integrated to. Please consider including certain important information about this plasmid, At least 25 vir genes on the Ti plasmid are necessary for tumor induction.

Thank you; changed.
We struggled with balancing sufficient details and too much information. As reviewer #1 will see, one of the other reviewers felt there was too much detail. Thus, we felt that we provided

sufficient detail to highlight the virulence of Agrobacterium and have decided against adding more details on the Ti plasmid. Also the reviewer is not privy to some of our data and it is not that straightforward to list sizes and # of core genes. The range in size of Ti plasmids is much greater than previously known and what genes are considered core is also becoming less clear.

9. Line 164: However, one fact that has been overlooked is that multiple copies of this gene are often present in a given bacterium. These intragenomic copies can differ in sequence, leading to identification of multiple ribotypes for a single organism. Is there a way the tool takes into account multiple copies of the gene?

Please see previous response regarding use of 16S to draw conclusions regarding taxonomical classification - we recommend, at best, genus-level.

10. When taking into consideration sequence reads, the authors mention the reads need to be filtered, both quality filtering and removal of PCR and sequencing adapaters, so my question would be how important is the quality of the reads for the tools or does the tool do some sort of sanity check before running any analysis?

This is an excellent point, as quality will affect results. On line 411, we added, “The pipeline down weights reads with a Q score of < 30, requires a minimum depth of 12 and relies on a minimum threshold of 75% for consensus base calling. Users concerned with sequencing quality may, prior to running the WGS pipeline, run programs such as FastQC, Trimmomatic, Sickle, and/or BBDuk to filter reads based on quality threshold (Andrews 2010; Joshi & Fass 2011; Bolger et al. 2014; Bushnell 2016).”

11. Line 290: Can you please provide the depth of coverage, read length or number of reads, such statistics are necessary to evaluate the quality of the sequences needed for such analysis?

We understand this request. The read length is stated in the methods section. The number of reads that passed filter is stated in Table 3. Supplemental figures 2 and 3 show that depth of coverage is not an important metric for quality (assuming that the depth exceed some minimum threshold, which we did not determine and felt was not an important message in this manuscript). As we described in the manuscript, the assembly with the lowest depth of coverage was the best (Fig. S2), which was validated based on a whole genome alignment to a closely related isolate with a finished genome sequence (Fig. S3).

12. Line 305: Denovo assemblers perform really poor with repeat regions, did the authors take into account such issue?

The Vir-search tool includes genes previously identified as important for virulence, none of which have a significant number of repeat sequences within their coding sequences. Therefore, repeats are not expected to be a problem for the intended purpose of the tool. More importantly,

the Vir-search tool does not perform *de novo* assembly. It aligns reads to virulence gene sequences obtained from reference genome sequences.

13. A detailed list of parameters for running Velveth and Spades would be very helpful

We are not clear on this request. Could the reviewer please elaborate? In the material and methods sequence, we thought we had clearly indicated which options were flagged and when default options were used (lines 485-493).

14. Line 482: Please see the bbmap software package, easy to use and really fast (https://sourceforge.net/projects/bbmap/), uses bbduk to do quality trimming and adapter removal.

Thanks for directing this package to us! This looks to be a very useful tool for quality trimming, and we will try it in future analyses.

Reviewer 2

- Even if specific methods for parts of the pipelines are mentioned in the "introduction" section of the paper, the authors do mention or compare their new pipelines with existing ones, used by other groups in similar settings. How does this effort compare to others? Highlighting the differences in terms of performance, accessibility, ease of use or ease of customization may encourage users to adopt this platform.

Thank you. This is a good suggestion. However, we are not claiming that Gall-ID is the “best” platform. Rather, our claim is that this is the only site to provide easy-to-use tools and databases for MLSA/16S based analysis for identifying Gall- and some plant-associated microbes. Also, we are not aware of any platform with similar functionalities that provides the same set of tools that we describe (except for a site that is described in a companion paper that is also under review at PeerJ; Tabima et al., submitted). Lastly, we previously compared our ANI pipeline to the output of JSpecies, and found that our pipeline handles a larger number of genomes with no issues, and is faster due to differences in multithreading.

- It is a good idea to have a "Demo" dataset for people to have a taste of the overall workflow and results. However it would be good to give a general estimation of the time it takes to process the demo file. Since that is a fixed file, it would be helpful to show a little note that says "processing this will take 2 minutes" or whatever the actual time is.

Thank you. We added an estimated time for the demo to the website for clarity.

- After generating the tree for the "Demo", I still see the text "Please wait while the tree is generated" on the page. This is kind of confusing, because the tree is right below. Authors may need to fix this issue.

Thank you. We fixed the issue.

- The authors mention that the databases used in the pipeline are manually curated. There is no mention about updates and maintenance of this information, support for troubleshooting of the pipeline or any other user related query. Authors may need to consider adding this information on the main website.

Thank you. We added the dates for database generated to the website. We are determining genome sequences for ~1000 gall-causing isolates. Updates to the databases will be completed once we acquire meaningful genome sequences or when they become available in NCBI GenBank.

- The generated tree for the "Demo" is very hard to read. Are there parameters that can be changed by the user to make it more readable? If yes, it may be worth to add a link to the document/instructions to do that. If no, it may be worth to add it, as a dynamic exploration of the data from the website could be of value.

We apologize that the reviewer had difficulties reading the tree. There are a variety of tools available that should address this issue. First, from the website, users can manually zoom in on branches of interest. The shiny server dynamically regenerates each figure, so text will be legible at larger or smaller size. Secondly, the subclade with the user-submitted isolate is displayed at a larger size, next to the full tree. Lastly, users can download a newick-format tree file or PDF to receive higher resolution figures.
We are limited by the shiny server, which will hopefully improve with newer versions of the software. We will update shiny when improvements exist.

Minor issues:
- (abstract) "have contributed to increasing the" should be "have contributed to increase the"

Fixed

- In the tab "Auto ANI: Average Nucleotide Identity Analysis" the "Citations" section is empty. The authors may want to update this information.

We apologize for this mistake. Thank you for spotting it. The citation section for the Auto ANI page has been updated with the relevant references.

- Clicking on the different tabs of the website result in a pretty long waiting time before anything appears on the page. I assume that the lag is due to the long loading time of the rshiny frame. However it may be slightly confusing for a user to just see a blank page for 5 seconds. I tested

this with different browsers and the speed seems to be also affected by cache and other browser settings. The authors may think to add a "Loading" kind of image/text in place of the rshiny frame while that is loading.

We appreciate this suggestion and have added a temporary loading image to pages with shiny frames to indicate when the page is still loading.

Reviewer 3

Thank you for the suggestion to include headers. We have followed suggestions.

1. A basic question to the Design is, what is the advantage of these molecular well tool to the proven methods? ("Proven methods for identification have been developed based on discriminative phenotypic and genotypic characteristics, including presence of antigens, differences in metabolism, or fatty acid methyl esters, and assaying based on polymorphic nucleotide sequences .") I noticed the validation were conducted on 14 isolates, how about the performance of these proven methods on them?

The benefits of molecular tools include: ease of use, cost effectiveness, speed of analysis, and ease of sharing data in publically available databases. Importantly, reliance on phenotypic characteristics can be misleading, especially if researchers rely on traits conferred by horizontally acquired genes, such as pathogenicity. This was one of the contributing causes for conflict in the literature on the naming of Agrobacterium, for example. The 14 isolates were provided to us from the director of the Plant Clinic at OSU. The Plant Clinic had previously typed them with the methods that are described in Table 3. Except for one strain, in which we think there was a mix-up, there was good correlation between molecular and more “traditional” approaches.

2. "Users must first select the appropriate taxonomic group, Agrobacterium, Pseudomonas, Pantoea, or Rhodococcus for comparison." What if the user has no idea about this or what if the isolate belongs to non of these 4 groups?

This is a very good question and the reason why Gall-ID includes Phytopath-type tool. With this tool, users can first use a 16S sequence to determine the broad taxonomic group of their isolate. If their isolate is similar to one of the taxonomic groups available for analysis in Gall-ID, MLSA genes can then be used with the specific Gall-ID tools for that group to further characterize their isolate.

Validity of the findings
For timing issue, I just tried several module functions with "DEMO" dataset but not WGS or Vir-Search. For example 16S demo for Agro-type, it takes about tens of seconds to process and finally generated the distance tree with bootstrap and a minimum spanning network.

Not sure the size and time for real data, but I would assume that would be larger/longer than DEMO. It is better to discuss that in the paper.

Analysis of 16S data will take roughly the same amount of time regardless of sequence. MLSA gene sets will take slightly longer to run depending on the number of genes input by the user, however it will not be significantly longer.

Also I can not understand the final generated spanning network, see attached figure. What are these pie chart in the network around each node? Why some pie chart sub-sections have the same color? what does this mean? Please illustrate in details in paper.

Thank you for pointing this out. This feature was included as part of the Microbe-ID framework. However it is not relevant or informative for Gall-ID and was therefore removed. Phylogenetic trees are generated by default.