David Landsman
Editor-in-Chief
Database

Dear Dr. Landsman,

We would like to re-submit for publication the revised version of DATABASE-2018-0092 “Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses”.

Firstly, we would like to thank the reviewers for their comments. From the recommendations, we have made substantial changes to the platform’s Antimicrobial Resistance (AMR) reporting, the web interface, and the manuscript. Chiefly, we’ve expanded on the details in the AMR results. In place of generic terms, the AMR results now report the exact percent identity of the match, the full specification for an AMR gene, and a link to the corresponding item in the Antimicrobial Resistance Ontology. The Start/Stop codons have also been fixed. We’ve clarified, added help text, and restructured the web interface to be more intuitive to users, and provided more details on data storage and long term plans in the manuscript. We have addressed each reviewer’s comments individually below:

Reviewer 1 Comments and Issues:

1. “Exist this service during a project period or is there a future plan to keep everything updated? … If there exist long time plans, I recommend adding the information to the manuscript.”

Spfy is part of a larger multi-year grant at the NML and will be used to provide future analyses modules upon their completion. We’ve added additional information to the Funding section of the manuscript to reflect this; specifically Page 9, Line 7-8. For the graph storage itself, it’s possible we may need to scale-out to a different backing graph store with sharding support, but this shouldn’t have a notable effect on end users.

2. “When I tried to apply all Spfy- analyses simultaneously on a genome, the pipeline encountered an error message. Please have a look on this.”

Thank you for bringing this to our attention; it was a bug that we missed prior to submission. It has now been fixed.

Reviewer 2 Comments and Issues:

1. “Page 2, Line 33-36: Can example(s) of this re-computation be provided? Are these instances of where the data is not stored due to space/computational constraints?”

We’ve added an example of the re-computation to Page 2, Lines 33-34 in the Introduction. Currently, all data associated with the analyses pipelines (subtyping, pangenome, etc) is stored in the database. We have tried to store required data in an efficient manner. For example: Spfy does not store redundant entries: there is only one instance of any given AMR or VF gene represented as a node in the database. The platform will create a new edge between any additional genome isolate and existing AMR/VF nodes. We’ve added text to clarify our data storage policies, Page 5, Lines 8-15 in the Data Storage section.

2. “Page 3, line 10: For the spfy database on the web – what browsers have been tested and is there a preferred web browser to use?”

We have updated Page 3, Lines 33-35 in the Functionality section to recommend Chrome as the preferred browser, and noted the different browsers (Firefox, Safari, Edge) Spfy was tested with. The website is also

compiled using Babel with support for any modern web browser, Internet Explorer 11 and up.

3. "Page 6, line 3: How is this quality control conducted? What is being used to verify that submitted genomes are E. coli? Are other tests being performed for quality control?"

To determine whether a submitted genome is of the species *E. coli,* our quality control pipeline tests for the presence of 10 *E.coli* specific genomic markers, which were identified in a previous study by our group (doi: 10.1186/s12866-016-0680-0); the presence of three or more of these sequences are required to pass quality control. Briefly, these *E. coli* specific markers were identified by blast comparison (90% sequence identity and 90% length) to be present only in *E. coli* genomes from GenBank, and not in any other species. All 10 markers are exclusive to *E. coli* and in our validation work, 3 were sufficient to uniquely identify an *E. coli* genome while tolerating moderate levels of genome sequence incompleteness. We've added this information to Page 6, Lines 21-30 in the Results section.

Reviewer 3 Comments and Issues:

1. "It was very difficult to use the search function, and there are no help files or examples on how to use the GUI version."

We have updated the website to include help documentation, and reoriented the website to guide the user in accessing the core functions. There is now help documentation in the Subtyping task, and step-by-step instructions and for using the Statistical Comparisons (previously "Fishers") task. Additionally, the search function had been redesigned to allow queries by accession number. The function now reports the original file name of the submission, identifies all contiguous sequences in the submission, and the associated serotype, virulence factors, and antimicrobial resistance genes. We will look to update the function with the associated probability of matches along with additional metadata.

2. "When I tried to search for known genomes that should be in the database, I failed to find any of them;"

The Spfy database was originally populated with 10,243 *E. coli* genomes from Enterobase. The platform has now been updated with 665 closed *E. coli* genomes from Genbank, and all subtyping analyses have been run on these strains with the results stored in the graph database. We have added a redesigned search function that can be used to query these files by accession number. A complete list of the closed GenBank genomes that are available is located at https://gist.github.com/kevinkle/8e77c05c2057e6e35a9ccaa2561102db.

3. "the meaning of the various functions (e.g., the "fishers" page) are non-intutitve and require explanation somewhere."

We agree with this point, and have renamed some of the functions, in addition to providing more detailed descriptions. Further, we have reoriented the website to separate the core functionality and to highlight the graph database, partitioning it from the additional functionality for specific use cases. We have noted the renaming of features in the manuscript in Page 3, Line 55 of the Functionality Section.

4. "The start/stop codon positions for reported genes are identical in every case (this also occurs in figure 1 in the paper)."

We thank the reviewer for bringing this to our attention. We have fixed the problem on the website, and have updated the Figures in the manuscript.

5-8. General comments, specifics below:

Regarding Antimicrobial Resistance (AMR) gene and mutation detection, Spfy uses the Resistance Gene Identifier (RGI) from the Comprehensive Antimicrobial Resistance Database (CARD) (doi: 10.1093/nar/gkw1004) for identification of AMR markers. We selected CARD because it is one of the most comprehensive AMR resources and is highly curated. Compared to other systems for acquired antibiotic resistance identification, eg. ResFinder, which use a smaller but more targeted set of genes, RGI / CARD takes a systems approach to AMR identification and classification, including components that have additional functions outside of their role in antimicrobial resistance. The additional data provides one of the most comprehensive reports available for AMR markers, but requires a greater degree of interpretation by the user. Based on the feedback from the reviewer, we have redesigned our interface to make it easier to identify relevant AMR components and assess the strength of the matches. These changes are outlined below:

5. "When looking at AMR genes, ACT-7 is not found in the E. coli isolate I tested–and that would be worthy of notice if it did. This is simply a chromosomal ampC (EC family) that is found in nearly all E. coli."

We agree with the reviewer, and the discrepancy is likely due to a nomenclature difference; in CARD, the ACT beta-lactamases may have a greater sequence similarity with the reference sequences of similar ampC beta-lactamases; in particular, "ACT-7" in a genome isolate (https://card.mcmaster.ca/ontology/38230) may be reported as the "*Escherichia coli* specific ampC" (https://card.mcmaster.ca/ontology/41454). We have checked a number of samples, and while not reported as ACT-7, ampC is certainly reported in the results. If there are particular problematic isolates, we would be happy to investigate the discrepancy further.

6. "The resistance genes are often underspecified; for example the gene names (e.g., APH(3')) in the test genome is a 100% identity match to aph(3'')-Ib (from the NCBI reference gene database; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047)."

We thank the reviewer for bringing this to our attention. We have added a new column in the results ("Long Hitname") which reports the full gene specification. APH(3")-Ib, and similar cases, are now reported in this column, which make the results more useful for users of Spfy.

7. "It is unlikely that the E. coli genome I tested has vanG, an Enterococcus vancomycin resistance gene, yet this was recorded as a "Strict" hit (and it's unclear what "Strict" means)."

We agree with the reviewer that using the CARD output of "Strict"/"Perfect" directly was unintuitive. We have changed the output to now report the exact percent identity of a match. This has been updated in the manuscript in Page 3, Lines 23-25 of the Functionality section. If you could provide us with the accession ID for the genome, we can further investigate why the Resistance Gene Identifier was calling the particular vanG hit.

8. "It is not useful to report S. aureus and M. tuberculosis point mutations that putatively confer resistance."

We agree with the reviewer. *S. aureus* and *M. tuberculosis* AMR mutations are not informative in assessing *E. coli* AMR phenotypes. We include all hits from RGI/CARD, because, in the absence of known *E. coli* markers, related species annotations may be relevant and we did not want to limit potential candidate markers *a priori*, for these particular cases. To aid the user in evaluating the CARD hits, we have added a webpage link to the CARD Antimicrobial Resistance Ontology (ARO) page for the annotation for each hit. The link is somewhat long, but can be easily viewed in the "Export to CSV" option available for all results. The page on ARO includes information on the mechanism as well as the prevalence for each marker across bacterial pathogens. We are also currently investigating how we can incorporate CARD's ARO information, including the pathogen prevalence data, directly into the Spfy AMR report. While incorporating the ARO directly into the graph

database would be helpful, it is also a major modification, and would take significant time to implement. For now we provide link-outs to the CARD website where this information is already available.