To the editors of PLOS Computational Biology,

We would again like to thank the reviewers and editors for their encouraging, positive, insightful, and constructive comments. We appreciate the time of the editors and reviewers, and this feedback greatly improved our manuscript. We are resubmitting our manuscript with point-by-point responses to the reviewer comments below. We are also submitting a marked up revised manuscript with additions marked in blue and deletions marked in red. Included in the revisions are four additional figures, one additional table, updated text with further discussion, and an improved quality filtering process for the sequence set, as suggested. All of the analyses were rerun following the additional suggested quality filtering steps.

The reviewer comments centered around four additional points to be addressed. These primary themes were 1) ensuring the purity of the samples, 2) evaluating the broader applicability of the model by testing additional true positive interactions, 3) evaluating the tolerance for in-accuracies in the model predictions, and 4) justifying the benchmarking techniques and outlining the associated limitations. We addressed each of these points in the manuscript text, and included additional data, figures, and dicusssion for each point. The details of how we addressed these points can be found below.

Reviewer #1

The authors have adequately addressed all my concerns. I have only minor remarks left that can be easily addressed by the authors:

1) The list of true positive and true negative host-phage interactions should be shared in a supplemental table that also provides the reference for each.

   This information was originally included as a set of files in the underlying code repository. We included this in the supplment as Table S2.

2) "To avoid further distraction by this figure, we decided to remove it and focus on the suggested matrix approach, which we added to the manuscript. Figure 1" Figure 1 still contains the previous network figure. I personally find the heat map in Figure S8 more informative than Figure 1D, but leave the final decision to the authors.

   We appologize for the confusion on our end. This was an issue we had with the resubmission process. This issue has now been corrected.

I still do not understand why the authors compare graphs via eigenvectors of the adjacency matrix and not directly via the adjacency matrix itself, but I assume that both methods are correlated, so that the choice does not matter.

> *This does not refer to the strict eigenvectors of the matrix but is rather a centrality metric that incorporates the both the influence of the nodes that are highly connected, and the additional influence of the nodes they are connected to. We included this description in our methods section.*

Reviewer #2

In the new version of the manuscript, Hannigan et al. have made a fair effort in addressing all reviewer's concerns. I find an improvement from the previous version. Sometimes these concerns were maybe a bit "over-justified" in the text, in detriment of style, but is fine.

I understand that the main intention of this manuscript is not the computational method, but the methodological approach as a whole and to show how could this approach be used to address interesting biological questions. In this sense, it is a nice contribution the use network theory combined with microbial ecology to generate biological hypothesis about how the interactions of hosts and viruses (bacteria and phages in this case) change in response to different variables and how vulnerable these associations may be to these changes. In terms of this, I find the work valuable and of interest for the scientific community.

However, I am still not quite convinced about the machine learning approach to define bacteria-phage interactions and I didn't find authors' replies on this respect convincing neither. My concern essentially stems from the biological base of their method, i.e., to which extent the connections between bacterial OGU and phage OGU have a biological base or are just computational artifacts.

> *We agree with these concerns and we think that the comments from the reviewer have really helped address them as described in our responses below.*

A first caveat, acknowledged by authors and reviewers, is the fact that the training set is so small and using viruses of highly divergent environments. I understand authors responses to this issue, and I find them valid. However, one could expect that with such a poor training set, almost no links would be found in a highly complex community, with highly divergent bacteria and viruses from those use for training. Yet, many interactions were found. This is simply explained by the features that primarily defined the model.

Figure 1B shows that protein identity between viral and bacterial contigs are the primary reason, followed by

nucleotide identity. This means that the random forest model will mostly look for protein similarity between phage and bacterial contigs, i.e., will predominantly look for prophages integrated in bacterial genomes. Do authors think that this is really a widespread feature between bacteria and their corresponding phages, that can be effectively used in real microbial communities? In a way, the model is only accounting for a limited and narrow sub-fraction of the phage and bacteria spectrum (i.e., only bacteria that are already infected with prophages of a certain viral family/genus, that are ALSO present as free viral particles in the sample).

> *This is a valuable point and deserves further clarrification in the manuscript. These communities have been previously described as consisting primarily of lysogenic phages. Therefore, while our approach could be improved by future work and more comprehensive datasets, we are addressing lysogenic phages, the most abundant component of the virome. Second, it is worth noting that while the alignment approaches were most informative, the protein family interaction approach was also informative. Because the protein family approach relies on protein-protein interactions and not detecting prophage sequences in the bacterial genomes, it is not expected to be limited to lysogenic phages. LINES 180 - 182*

Second, this would only work well if viral preparations were highly pure (which is not the case), otherwise, bacterial contaminant contigs will just get connected with other bacterial contigs, simply because they share similar conserved proteins. The authors try to reply to this searching for 16S rRNA sequences in the viral fraction, and find that only 1.7% of OGU contained 16S sequences. However, they should also show what is the percentage within the bacterial OGU dataset, because since contig data is highly fractionated and incomplete and there is only one or few 16S genes per genome, one could expect that, statistically, also only few bacterial OGU will contain 16S genes. The important here is the ratio between OGU fractions (normalized somehow by the size of each fraction), rather than the absolute number in each of them. The fact that the patterns observed in figures S03 A and B are so similar between the bacteria OGU and phage OGU fractions, also creates doubts about the real nature of the viral fractions. Why about using one of the methods used in the literature to identigy viral contigs (they cite many of these works, e.g., Paez-espino et al. 2016, Roux et al. 2016) and coloring in the Figure S03 the dots that obtained a viral classification? Finding only five phage OGU (1.7%) with blastx hits to phages is a weak evidence. Overall, I think that there is still place for stronger evidences showing that the connections found are really phage-bacteria.

> *This is also a great point and we added more information around this point. We agree that more stringent filtering is warranted here, and the suggestion and citation for using previously described methods is fantastic. We decided to, as previous groups have done (Roux et al, as described above),*

go back and remove bacteriophage OGUs that had similarity to bacterial references but not known bacteriophages. This approach therefore removes OGUs that we do not have high enough confidence in being non-contaminants. This process involved using the tblastx algorithm because phages can be highly diverse at the nucleotide level and the phage reference genomes are limited. We included the results of this filtering in the manuscript, and re-ran the relevant analysis to reflect the changes in the dataset. This was overall a great addition to the paper because it provides us with much greater confidence in the purity of our sample set.

We also thought it would be worth clarifying that we did not intend to communicate that we only found evidence of five matching phages, but rather that those were five OGUs that matched four reference genomes which had previously been validated as having incredibly broad range of infectiousness (cross-phylum infectious ranges). LINES 116 - 147, 493 - 494

Finally, I also agree with referee #3 that once the model is built with the 43 bacteria and 30 phages, the model could be tested on all the remaining phage-bacteria that were not included in the model (i.e., all the positive interactions that were not used because they unbalanced the model) and see how many of them the model is able to link.

We appreciate this suggestion and think that its implementation does add to the manuscirpt. As suggested by the two reviewers, we curated an additional dataset of diverse bacteria and phage interactions. While this approach has been used in previous studies as well, it is worth noting that there is remarkably little information for making additional predictions using interactions at the strain level because reference genomes for the bacterial host strains are not available in most cases. Therefore, as has been done in previous work, we looked to validate our interactions to confirm that bacteriophages are at least being predicted to infect representative bacterial genomes from the correct species. We applied our prediction algorith to hundreds of additional phage reference genomes and thousands of additional bacterial reference genomes and found that each phage was correctly predicted to infect at least one member of the correct bacterial species. This is an exciting finding because it provides an additional level of confidence in the broad applicability of our findings. Also, as we mentioned before, it will be exciting to evaluate this further in future work, with validated cases of phages that do not infect certain bacteria (a negative dataset). LINES 166 - 174, 534 - 546, Figure S5

In brief, I think that, once the step of identifying true phage-bacteria connections in a sample (with the approach proposed here, or even other one) is overcame, the downstream methodology proposed in this

work is a valuable and interesting approach to study the ecological roles and dynamics of phage-bacteria interactions.

We appreciate all of this feedback and we agree that the paper is stronger as a result of the suggested revisions.

Reviewer #3

The author made the point clear that they want to focus on the biological insight brought by their method more than the method itself (minor note on this point: the statement L78: "We implemented an adapted metagenomic interaction inference model that made some improvements upon previous phage-host interaction prediction models." should be removed or changed since they do not compare their model to previous models in terms of results). Nevertheless, to assess the confidence in the potential biological brought by the method, a careful evaluation of it is necessary. In my opinion, the current evaluation still suffers of major weaknesses:

1. Overlooking the impact of wrong predictions in the following of the analysis. The impact of the false predictions is never discussed and the reader have no idea if the false discovery rate of the predictor is likely to change the statistics extracted from each of the networks and how this FDR could affect or even reverse the biological conclusions of the paper. The reported median specificity and sensitivity of the predictor on a balanced dataset lead to an estimated FDR of 0.34. What is the impact of those wrong predictions on the subsequent analysis? Can those false predictions change the final conclusion of the paper? As a suggestion, one can measure the impact on the reported network statistics when randomly adding or removing edges corresponding to the fraction of estimated false negative rate and false positive rate. This would give an estimate of the confidence range on the final statistics of the network and assess if the differences between the samples are significant. Even though this analysis is mandatory for the reported median performances of the predictor, as the authors say there are high variations in the sensitivity and specificity in the nested cross-validation iterations and a complementary analysis of the impact of the wrong predictions on the final results at different sensitivity-specificity trade-offs should also be reported. An estimate of the change in the final results can be computed by tuning the threshold applied on the prediction score (which is responsible for the trade-off sensitivity-specificity).

   This is a fantastic addition to the manuscript. The suggestion is to evaluate the tolerance of the findings to inaccuracies in the network prediction model. The focus of the findings is to compare groups which were trated the same, which means that any false positives or negatives will equally

impact both groups. Therefore it is important to see what level of noise is tolerated by inaccuracies in the model. We accomplished this by implementing the analysis suggested by the reviewer. We iteratively added noise into the network datasets to evaluate how the observed differences in groups was effected. We included the results of this validation in the manuscript text and included relevant figures. As suggested by the reviewer, this confirms the confidence in our findings. LINES 292 - 298, 313 - 316, 393 - 394, 552 - 555, Figure S13, S14

2. Non realistic benchmark. The authors evaluate their predictor on an ideal full-genome case. The authors explain in their answer to reviews that they do not know how to simulate a realistic dataset. One can easily simulate sequencing reads followed by an assembly step. As for the simulation of the bins, a simulation of co-occurrence is not necessary and the effect of the binning errors can be estimated by artificially removing some contigs, and adding some contaminating contig (one can take for example the standard values given in Supplementary Figure S3 of (MetaBat, Kang et al., 2015); it would be for example good to know to what ratio of incompleteness and contamination the predictor is still reliable). Only then the evaluation of the method on this simulated data becomes more realistic and closer to the actual use-case made in the following of the paper.

   This is another valuable comment which certainly warrants further discussion and results. To the reviewers suggestion around simulations, one could definitely artificially create sequence datasets from the reference genomes and assemble them back into contigs. The issue we would like to point out however is that, although it was suggested above that "a simulation of co-occurrence is not necessary", the referenced citation and figure do use co-occurence abundance data in their validation (in fact, MetaBat stands for "Metagenome Binning with Abundance and Tetra-nucleotide frequencies"). The group created artificial contigs (their error free contigs) by randomly shredding reference genomes that were most highly represented in the MetaHIT dataset which they used for co-abundance calculations. In this way the authors could test their binning strategy using artificial contigs with co-abundance data from their MetaHIT dataset. As we mentioned in the last review iteration as well, we are unsure of how to accomplish this without a similar robust empirical dataset with well described interactions. At best we would be building a simulation on a simulation. To address the reviewer comments, and what we interpret to be the heart of the issue, we included a discussion as to how our method relates to this approach, what future work will be pursued, and what the ideal dataset and experimental setup will look like, in the discussion section.

   The second component that we got from this comment was that we should perform evaluation on

the impact of incomplete genomic data in our prediction algorithms. In other words, it is valuable to look at how well the model performs on incomplete references, which more accurately reflect the true applied dataset. We ran this analysis using the diverse validation dataset that we also used in this reviewers point 3, and included the results in the manuscript. Because the issue of contamination would center around negative data (discovering false positives), which are very limited as described above, we focused on the issue of incomplete genomic information. In the end, we included relevant discussion in the manuscript and the resulting figure that we created. Together we feel that this information provides a nice validation for the work. Lines 416 - 429, 460 - 462, Figure S6

3. Overlooking the evaluation on bigger datasets. The author explained why they evaluate the specificity of their tool only on a small dataset (to include validated negative interaction), but did not explain why they do not evaluate the sensitivity of their method on a bigger dataset using the fixed threshold used in the analysis (for evaluating the sensitivity, there is no need of negative interaction information). The question is: is the predictor sensitive enough to detect bacterial-host interaction on a more diverse dataset?

   We appreciate this feedback. This is an important point which we discuss in the related comment above by reviewer 2.

- Minor remarks:

L195: "which similar genomic elements to" lacks a verb.

We fixed this in the text.

Caption S6: "mtry" is not defined.

We added the definition in the caption.