To editors of PLOS Computational Biology,

We would like to thank the editors and reviewers very much for the in-depth constructive and positive comments on our manuscript. We deeply appreciate the time of the editor and reviewers, and we think that the manuscript has benefited immensely from the feedback. 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. We included four additional figures, additional data analyses which are included in the text, and further discussion and clarification in the text.

There were many good points brought up by the reviewers. One of the primary themes we did not intend to communicate was a focus on the random forest method itself instead of the biological insights that we gained. We developed our algorithm to predict phage-bacteria interactions to meet the specific needs of our biological hypotheses. While it did improve on other existing methods and performed well, we omitted some components that would otherwise be expected in a methods-centric manuscript. For example, we did not package it for widespread use in other studies (although the source code is available and reproducible). Additionally, while it is important to consider how our method performed in the context of existing methods, we do not wish to claim superiority of our method. It served our purposes and performs as well as other methods.

The other major points of revision focused on more clearly benchmarking and communicating the performance of our model and the resulting network/sub-networks. To this end, we added additional data, figures, discussion and descriptions. The manuscript is certainly more comprehensive and robust as a result of the suggested revisions that we performed.

Reviewer #1: Summary

The authors developed a pipeline that predicts interactions between phages and their hosts from metagenomic data by exploiting similarities on the gene and protein level as well as CRISPRs with the random forest machine learning technique. They then compared the resulting phage-host networks across persons, time points and body sites. In general, this study is a valuable contribution, however some aspects of the network analysis are questionable, as detailed below.

Major

I am suspicious of node removal studies. Even though the accuracy of the authors' inference procedure is much higher than those based on co-occurrence, the resulting networks still contain a number of false positives that can bias the node removal results. Even if the networks were perfect, where is the experimental evidence that an ecosystem will be less stable when separated into several components? Also, less stable in what sense? There are multiple definitions of ecological stability. Which one do the authors have in mind?

> **This is an important point and we agree with the reviewer that this warrants further clarification. Stability was not the best term here, but rather we think *robustness* better communicates our intent and interpretation. Previous work has shown that high connectance of an ecological network is associated with resistance to species extinction events and greater support of interactions through the network, such as horizontal gene transfer. For example, if a phage only has one host in a system, the removal of that host is expected to result in the extinction of the phage. If that phage had multiple hosts in a system, the removal of one host is less likely to result in extinction of that phage. We referred to this as "stability" in our manuscript, but we think that robustness is clearer. We went back and formally defined our use of "robustness" in the manuscript and updated the text to more clearly articulate this point. We had included citations around this relationship between connectance and robustness, and we added text to make this clearer in the manuscript as well. LINES 23, 33-34, 44, 65-66, 92-93, 220-223, 263, 317-318, 328-329, 337-344, 403-405**

In the same context: P. 14, l. 262: "These findings suggested that the community properties inferred from microbiome interaction network structures, such as stability (meaning a more highly connected network is more stable because a randomly removed bacteria or phage node is less likely to divide or disintegrate [30,58] the overall network." The authors define a network with many components after species removal as unstable and a network with a single component as stable. To avoid a tautology here, an external criterion of stability is needed. Assuming an appropriate external criterion is available (e.g. number of species that die upon removal of a species), is there any experimental evidence that species whose removal results in more components destabilize an ecosystem more than species whose removal results in fewer components?

> **Similar to above, we are talking more about robustness of the network to the random removal of nodes (phages and bacteria) instead of true ecological stability, which is a point we needed to make clearer. Previous work has shown that less highly connected**

networks are more likely to experience extinction events resulting from random microbe removal, as we discussed in the point above. Additionally, interactions such as horizontal gene transfer are more limited when there are less "connections" throughout the network. Previous work has supported these model assumptions and we cited some of those relevant pieces of work in the manuscript. LINES 23, 33-34, 44, 65-66, 92-93, 220-223, 263, 317-318, 328-329, 337-344, 403-405

I appreciate that the authors mention that no statement of significance can be made with the small sample numbers in the high-fat/low-fat and obese/healthy groups, but then why comparing network properties between these groups at all?

This analysis was meant to be descriptive and to primarily illustrate to the field what types of questions can be answered once we have more comprehensive datasets. We cannot be certain of the results because of the small sample size, but this illustrates the potential biology we can uncover with more comprehensive microbiome sampling and integrative analyses. LINES 230-231, 247-250

I am curious to know whether the predicted phage-host networks are nested and/or modular (see e.g. the Weitz lab papers cited by the authors). Also, what is the number of phages with 1, 2, ... N hosts and the number of hosts with 1, 2 ... N phages? Do these properties differ from what was observed in the oceans (Flores et al. 2013 cited by the authors)?

This is a good point and while we did show these statistics for the reference set used to train the model, we did not show it from our samples in the original submission. We added a figure that shows a heat-map of the bacteria and phage interactions in our master network. As we alluded to in our manuscript, we are hesitant to make strong comparisons to those other datasets because they were using experimental systems of defined strain-level interactions, which is a resolution we do not feel confident in for our study. Speculatively the infection patterns do seem to suggest a modular nested pattern which has been observed in those previous studies. We included a figure to this point, as well as some more text in the manuscript. Figure S9, LINES 187-189, 391-394

In this context: why not showing in Figure 1D a bipartite matrix in the style of Flores et al.? This would be far more informative than the current visualization, from which we learn nothing much, not even the number of bacteria and phages involved in interactions.

We agree with the reviewer and would add that this figure has been the point of discussion and controversy both within our group and with others. The main purpose of this bipartite network visualization was to visually confirm that there were no erroneous connections between bacteria or between phages. 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

When comparing networks across body sites, the authors want to consider not only network structure but also abundances. So the problem is to compare weighted networks. In the current formula, edges are only indirectly accounted for via centralities. I do not see how to justify this neglect of available information and think that a function that takes the entire edge set into account would be a better choice. How about for instance converting node weights into edge weights (e.g. by summing or averaging) and then summing the entry-wise differences between the corresponding (weighted) adjacency matrices into a single dissimilarity value? The adjacency matrices could include empty rows and columns for absent phages/bacteria. In this context, it would be interesting to know how the beta-diversity of the networks compares to the beta-diversity of the samples.

We apologize for the confusion, but it sounds like the suggested approach is actually what we used. We calculated weighted centrality measures by converting the node weights into edge weights and used that to calculate our dissimilarity matrices. This way were were able to compare the beta-diversity of the networks, as suggested.

It would be interesting to find out whether phage-host interactions change between persons. Universality of microbial networks was the topic of the paper by Bashan et al. (https://www.nature.com/nature/journal/v534/n7606/abs/natur
Here, the authors could check whether hosts and phages present in all selected samples also interact in these samples and what might be the reason if they do not.

This is a very interesting point and we appreciate the reviewer suggesting this addition to our work. The way we built our analysis was to build a universal network like was described by Bashan *et al* and utilize the subnetworks within. We performed our analysis this way because it allowed us to combine the genomic information from all studies to gain deeper coverage of the microbial genomes, as was done previously in the studies we utilized. Unfortunately that approach and the sample depths do not allow us to look at the networks beyond connectivity and abundance within the subgraphs, as suggested. Additionally, it would be important to include time course

data to support these kinds of findings, as has been done in interaction modeling work by Weitz *et al* previously (DOI: 10.1098/rsos.160654). This is certainly an interesting direction of research and would be an excellent topic of future research. LINES 183-184, 253, 339-340

Minor

How was (intermittent) occlusion in skin sites determined?

This was defined in the original publication by dermatologists and other previous work. We clarified this in the text. LINES 281-282, 286-288,

Do the most interconnected skin sites harbor a more diverse microbiome than other skin sites?

This is an important point. As can be seen by the formulas of the metrics we used, the degree of network connectedness is still a function of bacterial and phage diversity. We don't interpret these metrics as mutually exclusive, but rather are ultimately using a metric that integrates the diversity of both the bacterial and viral communities. We included a clarification of this point in the manuscript. LINES 282-289

P. 10, l. 200-201: "These results suggested that the obesity-associated networks are less connected, having microbes further from all other microbes within the community." This statement is problematic. In addition to the problem of small sample numbers, interactions between the hosts themselves are not considered here. For instance, the overall impact of phages on gene exchange in the microbial community may be small compared to other avenues of horizontal gene transfer.

These are good points and we reduced the interpretations that we outlined in our text. LINES 206, 242-250

P. 14, l. 262: "Gut network structure was not conserved among family members" The analysis of the authors does not differentiate between the change of community composition and the change of network structure.

Similar to the point above, the network structure is a function of community composition of the two communities, as is evident in the formulas described in our methods. We clarified this point in the manuscript. LINES 282-286

Reviewer #2

Summary

The manuscript by Hannigan et al. describes a new methodological approach to study the interactions of phage and bacterial communities, which they propose to infer using machine learning approaches trained on known positive and negative interactions of bacteria and phages using different genomic features (nucleotide similarity, amino acid similarity, CRISPR cassettes and known protein-protein interactions). When applied to experimental metagenomic sets of viral and bacterial communities (with contigs clustered in bins or OGU), they were able to use their trained machine learning algorithm to find links between viral and bacterial bins, which are supposed to represent an infective interaction. With all links found between all OGU, they could build interactive networks and finally use different structural or topological features from the resulting networks to conduct microbial ecology analyses.

I overall found the article interesting and I think that it represent a valuable contribution to the long lasting and largely unsolved problem of how to study the interplay between viral and microbial communities in metagenomic studies. Thus, I think that the proposed methodology will be well welcomed by the scientific community and is certainly within the scope of Plos Computational Biology. Although I found also interesting the biological results obtained from the analyzed datasets, I found some of them a bit weak. However, I think that the authors did a very honest and accurate work at the time of highlighting all the limitations throughout the manuscript and most of the concerns I rose while reading, were eventually replied by the authors themselves while moving forward in the reading. I highly appreciated that and I think that presented this way is fine and scientifically sounding.

The article was very well written, explained and scientifically solid.

I have few comments to make, but I don't necessarily consider all of them critical to be potentially ready for publication.

1) The authors do not test whether the interactions between viral and bacterial OGU predicted by their method are likely true or not (i.e., neither they tried to estimate the presence of false positives in the predicted phage-bacteria interactions nor at least pick up some examples and take a closer look to understand why OGUs were connected and whether this makes biological sense). Although I don't consider it essential, I think that the methodology would gain credibility if they could just show few examples of which kind of genomic sequences were linked and why. I would suggest to pick examples of

different nature, e.g., highly or poorly connected nodes, nodes with large or short edges, etc.

We appreciate this point by the reviewer. We apologize for the confusion but we tested whether interactions are likely to be true or not using nested k-fold cross validation approach to determine the false positives and false negatives associated with our model, which are the inverse values of the sensitivity and specificity that we reported for our model. Using this we were able to quantify the degree to which our predicted positive and negative interactions were correct. We liked the idea of including an example so we performed a tblastx search for phage elements within our OGUs that matched the broadly infectious phages isolated from Lake Michigan by Putonti *et al.* We found some broadly infectious OGUs with genomic elements matching these broadly infectious phages, and included that example in the manuscript. LINES 164-175, 203-205

Since the training set was done with a very limited number of genomes (43 bacteria, 30 phages), from highly diverse (ecological) sources (something that it is also acknowledged by authors in lines 288-290), it is somehow hard to imagine how it would perform when tested on complex and highly diverse microbial communities, where most of the bacterial, and specially the viral genomes are highly divergent from those in databases.

This is a very good point and brings up some of the points that we tried to address in our analyses. As we mentioned in our manuscript, this is an important caveat to our approach and why our insights into these dynamics will improve with more robust training set data. The point on the databases is also particularly important. These are indeed highly diverse microbial communities that are highly divergent from those in databases. This was a driving reason for why we chose to base our prediction algorithm on comparisons between phage and bacterial sequences from the samples themselves, instead of relying on alignments to existing reference microbes or the presence of identified marker genes. All of these are important caveats and we added these points to the manuscript text. LINES 164-168

Although authors give an explanation for the high abundance (77%) of bacterial hits in the viral datasets (i.e., reference bacterial genomes contain integrated prophages and the VLP largely correspond to temperate phages sharing elements with those prophages), I believe that in any viral purification protocol there is an important proportion of non-viral contaminant DNA that gets finally sequenced. Although the nature of bacterial hits in viral metagenomes is an ongoing debate or, likely, an inevitable technical issue, I think that

the controls I mentioned above may be a complementary information to validate that the links found are truly bacterial-phage interactions and not unwanted bacteria-bacteria links (or even virus-virus). I think there are probably many different ways to make these controls, but one possibility (that may not be enough by itself anyway) could be to run softwares that identify core bacterial genes (e.g., MetaPhlan2) on the viral contigs and show that the proportion of annotated contigs is significantly low.

This is also a very important point by the reviewer and we agree that it warrants further dedicated discussion in our manuscript. As pointed out, the inclusion of some contaminating sequences in these datasets is an inevitable technical issue that we must consider in our interpretations. This was shown nicely by Roux *et al* in a previous study, which we cited in our text. The group found that viral purification methods only reduce the level of bacterial DNA in the samples and do not completely remove them. Like what was done in Roux *et al*, and as suggested by the reviewer, we went on to identify a low abundance of 16S genes in the dataset which we would expect to be found only in bacterial sequences. This confirms that while the QC mentioned suggests a high level purification, there remains a low level of bacterial background noise. We added this data and discussion to our manuscript. LINES 120-140

To better address this and other points regarding the purity and accuracy of our dataset modeling, we also wanted to include a concept which has been utilized in other classic types of microbiome analyses, which is our tolerance for error. In our study, we are careful to not make any conclusions about the networks themselves, but instead rely on comparisons to other groups which were treated in the same way. This is similar to the classic approach taken in the field of the skin microbiome, whose V4 region 16S analysis could not detect some of the most abundant and important bacteria (*Propionibacterium*). The findings were still informative because the conclusions were not on the *true* composition of those communities, but rather comparisons of overall community structure between the study groups. This is an approach we also aimed to take with our work, so as to most appropriately interpret our data without the assumption that our model or data are perfect. We included a discussion of this in the manuscript as well. LINES 353-373

2) Which would be the results of the different analyses if networks were made only with the OGUs from

whole shotgun datasets or the OGUs from viral datasets? (e.g., would the degree of centrality in the networks from low-fat diets be still lower than the high-fat?) If the observed patterns and interactions are intrinsic of viral and bacterial interplay, results should be markedly different when data from one of both communities is absent.

It sounds like the idea here is to include direct phage and bacterial interactions in our analyses. We expect it would be difficult to identify direct interactions between bacteriophages because they do not have their own metabolic capacity without their bacterial hosts. We could consider modeling bacteria-bacteria interactions, but this is a non-trivial task that we feel is beyond the scope of our current study. Regardless, because this was not a consideration in our network analysis, we are not sure how bacterial interplay could be impacting the observations we made. This would add another interesting layer to future analyses along these lines.

3) May the Figure 1D benefit if the nodes from each of the three studies are colored differently?

This is a good suggestion but, as we mentioned before, we are worried that Figure 1D is a point of distraction in the manuscript. We replaced this with a heatmap matrix visualization. Additionally, because the nodes are all part of the master network which includes data from all studies, the nodes could be associated with more than one study. FIGURE 1

4) Line 131-133: Although I think that taking into account the negative interactions is interesting, I also think that this is a challenging point. While a positive interaction is a strong proof that such viral-host interactions may also occur in natural conditions, I think that negative interactions are more delicate since the infection may not be observed just because the right experimental conditions were not found. I am not saying that negative interactions are not informative, just suggesting that maybe they should have less weight than positive ones.

We are really glad to hear this because we have thought a lot about this in our group as well. Many previous efforts to predict interactions between bacteria and phages have focused on the positive interactions, and many assume that a lack of evidence is sufficient to conclude a negative interaction, which is how accuracy has been determined. We agree that the lack of infection does not definitively show a lack of infectious capability and that we should be conscious of this. This was very informative for our calculations of model sensitivity and specificity. This is an

important caveat that we added to the paper. Additionally, while it is important to consider what the true occurrence is in nature, we are focusing on drawing conclusions as comparisons between states such as different skin sites, as discussed above. LINES 154-159

5) Line 216/218, Figure 3A: The authors propose that, similarly to what happen with microbiome composition (i.e., microbiomes are more similar in the same individual over time, than between individuals), the interaction networks or network structures are also more similar within the same individual. How do we know that this is an intrinsic property of these phage-bacteria interactions and are not a direct consequence of their significantly different microbiome/viral compositions?

This is another very good point which we also discussed above. We can see from our network analysis formulas that the interaction networks are both functions of the community diversity and composition, as well as their degrees of connectedness. Because the bacterial and viral communities were shown to be different under the conditions we studied (outlined well in previous publications), it is unsurprising that the networks, which are integrations of these datasets, resulted in different community dynamics. The value of the conclusions is two-fold. The first is that we confirm that the differences hold true when we integrate the bacterial and viral datasets. The second is that we show which of those are more or less connected than the other, not only showing that they are different. We attempted to make this point clearer in the text. LINES 282-289, 296-297

6) Line 226, 227: Can we really say that the gut microbiome network structure was "strongly" conserved within individuals? Regardless of the very low p-value, the figure 3D does not give me the impression of "strongly", although certainly different. In fact differences seem to be lower than the observed in Figure 3B (which I agree with authors that were probably not significant due to the limited number of samples). I think that the "strongly" in Figure 3D based on the p-value, is just a consequence of the high number of samples, not of the magnitude of the difference.

This is an important point that we agree should be addressed in the manuscript. We updated our wording around this point. LINES 275

Typos and minor details:

- Figure 2 legend, line 5: change "the" for "them"

We fixed this in the text.

- Line 429: add “on” graph structure or “graph structure information”

  We fixed this in the text.

- Line 300: I think there is a typo here: “in biases”

  We fixed this in the text.

Reviewer #3

Summary

The authors present a machine learning tool (random forest) that predicts the host of a phage sequence using a set of features derived from sequence information only (nucleotide similarity, amino-acid similarity, CRISPR and similarity to known phage-host interacting Pfam entries). The authors use their predictor to evaluate the structure of phage-host interaction networks at different body sites across several human cohorts. They draw conclusions relating some metrics of the phage-host networks to different attributes of the samples (cohabitant humans, skin sites, diet, obesity).

General appreciation

The goal of the study is clear and well-stated: even though some studies relate phage or bacterial community independently to environmental conditions, the relation between those communities and their variation in their structure is understudied and could bring highly valuable new insights to better understand the mechanisms, stability, and stress-response of the microbiomes. However, to address this point, the reliability of the prediction of the bacteriophage-bacteria interactions is essential, as it serves as basis for the subsequent analysis. In my opinion the present study suffers from major flaws regarding both the evaluation of the performances of the machine learning method as well as the reliability of its application on previously published mategenomic data.

Major issues

Comparison to existing methods: Existing methods to predict bacteriophage and bacteria interactions from sequence information would suit the purposes of the authors:

- Villarroel, J., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., & Larsen, M. V. (2016). HostPhinder: a phage host prediction tool. Viruses, 8(5), 116.
- Galiez, C., Siebert, M., Enault, F., Vincent, J., & Söding, J. (2017). WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics, 33(19), 3113-3114.
- Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2016). Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic acids research, 45(1), 39-53.

The authors have to show that the method they developed improves on the existing ones (which would be highly valuable for the scientific community) and/or should compare the networks and the associated metrics they would get using these established prediction methods.

> **We thank the reviewer for this point. There are certainly many valuable phage-bacteria interaction prediction tools in the field, and many avenues for improvement. To the point that "The authors have to show that the method they developed improves on the existing ones", we agree this would be important if the purpose of our manuscript was to offer a new and improved tool, but we do not intend this to be a method paper. We intended this to be evidenced by the lack of packaging of the model into an easily distributed tool, or the focus on the tool beyond a means to an end of understanding underlying biology. This is an important point for us to make clearer in our manuscript, so we added a discussion around this. We added this point to our discussion, in addition to referencing the papers. LINES 24-26, 78-87, 164-168**

Benchmark of the method:

1. The size of the benchmark is too small. The authors suggest that the need for having reliable negative bacteriophage-bacteria interaction information together with the positive-negative balance reduces the training set to 43 bacterial genomes and 30 phage strains. If the authors think that fulfilling these criteria at the price of a small training set is worth, they should show that the predictions are actually good on bigger datasets, at least for the evaluating if the known positive interactions are correctly

recalled (in which case there is no lack of data). As in bigger dataset there will be more negative interactions than positive interactions, the precision of the method (and not the specificity) should be reported.

This is an important point and well taken. We also had many internal discussions around this. The problem was deciding between a smaller but balanced dataset and a large but highly unbalanced dataset. The issue we had with an unbalanced dataset is that a highly accurate model could be developed by always stating that interactions are positive and not negative, which is not what we want. Going through the exercise, we would then move to an under-sampling approach to balance the datasets, which is what we presented in our paper. Imputing data through oversampling methods would be problematic because of the very highly imbalanced nature of the dataset. The addition of a larger dataset sounds great, and if we were aware of such a more robust and balanced dataset we would have loved to include it in our model creation and validation. It is true that we could evaluate our model on all of the known positive interactions, but we would still suffer from a lack of negative interaction data. LINES 154-159, 164-168

2. Moreover, since the diversity of the cultivated phages as well as the cultivated bacteria is very low in comparison to real species found in the environment, the authors should evaluate their prediction on reliable known interaction detected in metagenomic samples (e.g. Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., ... & Kyrpides, N. C. (2016). Uncovering Earth's virome. Nature.)

This is similar to a point that we addressed above. This type of dataset would also be a great addition and the inclusion of such information would make our model much stronger, but we are unfortunately unaware of such a dataset. We also considered the suggested Paez-Espino *et al* manuscript but concluded this would be problematic for such an application because it is a whole shotgun metagenomic sequencing dataset whose interactions were inferred taxonomically. As much as we can tell, there was no experimental validation of the interactions, and no virus purification was attempted (all phage sequences were inferred from total metagenomes), so we were worried that we could not have any addition confidence in those findings beyond what we already have. If we included this approach, we would be benchmarking our predictions

against other predictions.

3. On top of the previous remarks, the current benchmark shows how the tool perform only for full-length genomes. Since the presented application cases are dealing with metagenomic bins – likely to be incomplete and to suffer from contamination – it is necessary to evaluate the performances on simulated bins (namely simulate reads from phages and bacteria, run the same quality control+assembly+binning).

   We appreciate this point. This gets back to the previous point about how it would be ideal if we could build the model on metagenomic data. Unfortunately, we are not aware of a reliable and experimentally validated metagenomic dataset that we could use for such an application. Simulating such datasets from the reference genomes would be great, and we considered it, but we are unsure of how to effectively accomplish the task. As pointed out, it would be trivial to simulate contig sampling from the reference genome, but as we pointed out in the manuscript, the metagenomic binning of operational genomic units was also done with co-occurrence information from contig abundance. We are open to suggestions, but we are unsure about how to effectively simulate co-occurrence information because we are unsure about what empirical data that model would be built on. We could create co-occurrence matrices for the simulation binning by giving high and low counts to shared contigs, but this would not include informative error rates or reflect a real sample. In the end, this approach results in our predictive model being built on a simulated dataset that does not reflect reality any better, and had biases introduced by us for co-occurrences. We added more text around this point to the manuscript. LINES 113-116

4. Overall, and most importantly, the method seems to be biased toward a high sensitivity-low specificity trade-off (respectively the median values are 0.952 and 0.615 when cross-validating) meaning that in case of balanced dataset almost 40% of the predicted interactions are expected to be wrong. The application on the skin dataset leads to almost a fully connected graph, while only 60% of those might be real. The authors should discuss explicitly the impact of the false prediction on the subsequent drawn conclusions: since the score threshold (say "s_se") used for predicting the interaction looks biased toward sensitivity, what threshold "s_sp" could be considered as a safe guard on the other extreme (high specificity, with potentially lower sensitivity)? what is the impact of changing the threshold from t_se to t_sp on the network metrics? To what extent does it change the conclusions on real data?

We added additional data and figures to show that the overall values fluctuated over the nested cross-validation benchmarking and that the overall performance is satisfactory. Additionally we feel that this is the ideal balance because the true positive rate is very good and most of the interactions are positive. We also added more discussion around this point to the manuscript. FIGURE S5, FIGURE S6, LINES 170-175

Application to previously sampled cohorts:

Since the number of samples on real-case applications is rather limited, a null model comparison should be used to assess the significance of the findings. For instance, in the case of the obesity-associated networks, the authors could randomly select the same number of nodes in the network as for each of the N=1+2 cases, and report the corresponding centrality values (mean + standard deviation across different random selections) and include it in Figure 2.

This is an interesting idea. The null model seems like it would be useful for showing a difference to a random model, but we are more interested in comparing the obesity and non-obese states. We only intended the obesity data to be an opportunistic observation of interest, and not a formal finding. While it is interesting and warrants further investigation, we recognize that this is only an N=1+2 dataset and we are ultimately going to need more sampling to truly understand the differences in network structure between healthy and obese individuals.

Minor issues

Lines 82-83: this is a strong overstatement since it is only a hypothetical interpretation of the network metrics as mentioned for example in lines 214-215.

We toned down our language here.

Lines 107-114: It is not clear at first sight that the paragraph is describing the phage fraction.

We agree and clarified this in the text.

Lines 115-120: It is not clear at first sight that the paragraph is describing the bacterial fraction.

We agree and clarified this in the text.

Lines 134-138: the description of the methods lacks some important description. It is form example not clear if the nucleotide similarity feature is always computed as the best bitscore obtained on the training bacterial dataset, or if it uses another external bacterial dataset. The same remarks goes for the amino-acid similarity feature.

We agree that the description here is cursory, as we left the reader to see the methods section for more details. No external datasets were used, and we clarified this point in the text.

Line 140: The authors report the median values for sensitivity and specificity when cross-validating their method, but do not report the standard deviations. These standard deviations are of high importance to get an accurate feeling of the reliability of the method.

We agree and added this data as figures and text in the manuscript. FIGURE S5, FIGURE S6

Line 159: the diameter of the graph seems not to be a robust measure of the graph. Indeed, a single spurious branch popping out from the graph will change the value diameter. It would be more informative to report some quantile value for the number of traversed vertices between two vertices.

This is a valuable point. We added the suggested data as a set of historgrams outlining the distributions of node eccentricities associated with each subnetwork and the master network. LINES 199-200, FIGURE S8

Lines 162-163: the percentage of the recalled interactions could be mentioned.

We unfortunately are unsure of what the recalled interactions are referring to.

Line 186: it is not clear what is the limited sample set.

The word limited was confusing here. Instead we used small to communicate the caveat that this is a data mining exercise using a small dataset. LINES 232

Line 340-347: The description of the data is too limited. We do not find for example clearly how many gut samples per individual were included (since on line 223 it is implicit that there are several samples per individual).

We mentioned the numbers of samples used in each part of the analysis when we described them in the results section. We clarified that direction in the text. LINES

415-416