Point-By-Point Reviewer Comment Response

Hannigan *et al.*

We would like to thank the reviewers and editor for taking the time to read our manuscript and offer constructive feedback. We incorporated the suggested edits and feel that this greatly improved the manuscript. Below we included a point-by-point response to the reviewer comments, with our comments highlighted in *italics* and the relevant manuscript change locations highlighted as **bold**. Line numbers were included to guide the reviewers to the relevant sections, and in-text changes are highlighted in red and blue.

Reviewer #1 (Comments for the Author):

Hannigan et al. describe a machine learning-based approach for predicting phage-bacterial host interactions and use it to build ecological networks for human microbiome samples based on published data. They analyze differences in network features as a function of host nutrition (low fat vs high fat), health status (lean vs obese twins) and body regions (diverse skin sites). The authors built their prediction model using data from previous publications as training set, containing 43 diverse bacterial species and 30 diverse phage strains, where the infectiveness relationship had been established experimentally (Fig 1A). The evaluation of the resulting predictive model shows that the amino acid similarity between phage and bacteria was the strongest predictor of infectiveness, followed by nucleotide similarity, shared protein families, and to a negligible degree CRISPRs (Fig 1B). They attribute the lack of predictive power of the latter to the minimal amounts of CRISPRs identified.

The authors applied their model to three publicly available datasets (skin microbiome, twin gut microbiome of lean and obese twins and their mothers, and effect of diet on the gut microbiome) that contain sequences from full metagenomics as well as from purified viral-like particle fractions. Relationships between 298 phages operational genomic units (OGU) and 280 bacteria OGU were predicted from the combined taxa of all datasets and overall network features quantified (Fig1 D-G). Using the phage-bacteria network analysis, they explore the impact of diet and obesity in the gut microbiome structure (Fig 2), compare similarities of intrapersonal and interpersonal networks of skin and gut microbiomes (Fig 3) and test for differences between skin sites with different environmental qualities (Fig 4). All these comparisons are performed by either using measurements of network dissimilarity and network topology (eigen centrality, closeness centrality and degree centrality).

Based on their results, the authors conclude that changes in diet, health state and body site topology can influence the phage-bacterial network stability. As an example they show that high-fat diet and obesity as well as external exposure of skin microbiota to reduce network connectivity. They also find network similarities to agree with the well-accepted notion of differences to be smaller within than between individuals.

The approach taken by Hannigan et al. to construct phage-bacteria networks to analyze host-associated communities is original and will be relevant to many researchers within and also outside of the field of human microbiome research. One missing point the description on how the authors assessed the amount of bacterial DNA in the viral-like particle (VLP) fraction. If bacterial DNA were wrongly labeled as being of viral origin, it would lead to the formation of false phage OGU and result in relationships based on bacterial co-occurrences rather than phage infectivity. Also, the reader is not informed how abundances in the VLP fraction where integrated with the abundances of microbial fractions. In other words, if the DNA of the whole sample was not extracted together, how can the relative abundance of each fraction be compared to each other?

The presented results on the impact of diet and obesity, however, are severely underpowered, as the authors also point out (159, 187, 236). In particular, only one or two data points per group were used to compare the influence of diet and obesity (Figure 2). Although intra- vs. inter-individual and intra- vs. inter-family differences cannot be confirmed by statistical tests, they are claimed to exist. This is unfortunate, since the analytical approach using appropriately powered statistics reveals significant differences between skin sites, highlighting the potential power and impact of the methodology.

Taken together, the machine leaning-based approach would only require some additional tests to exclude or estimate the degree of bacterial contamination in the VLP fraction and some explanations on how abundances in VLP and microbial fractions were combined and normalized. With regard to the application of the methodology, two of the three datasets are severely underpowered to support major claims on the impact of diet and health state. I would thus recommend to put the focus of the manuscript on the analysis of the well-powered skin data set and reduce the diet and obesity-related part as potentially promising case-examples while clearly describing the limitations of the analysis (which is done) and toning down or removing some of the respective claims.

Major Points

1. Conceptual distinction between infectiveness and community stability: For the development of the network approach, the authors use infectiveness data as evidence for interactions between phage and bacteria. For the analysis part, such interactions are used to describe community stability (i.e., more interactions per node, higher stability). If phages are considered as agents grazing on their microbial hosts, is it justified to use degree of infectiveness analogous to community stability? Some discussion or background information would be helpful.

Thank you for pointing this out. This is a good point warranting further discussion. We clarified in the text why we feel this is justified, as the degree of stability refers to the reduced likelihood that a randomly removed node will cause a major divide or lack of connection between agents within the network. The infectiveness refers to the link we drew between the bacteria and phages (how we connected phages and bacteria in our networks). We also provided citations to work that further explains this in metabolic and phage-bacteria systems, if the readers are interested in a more in-depth discussion of the topic. **LINES 176 - 183**

2. Testing for bacterial DNA in the phage fraction: The authors do not describe how bacterial sequences in the VLP fraction were eliminated. If bacterial sequences were present, they would be labeled as phages and lead to false predictions.

This is a good point and we agree that this needed to be better outlined in the text. It is important that we make clear that we are not violating our assumption that that phage OGU pool represents phages and the bacterial OGU pool represented bacteria. We updated the manuscript to address this concern in multiple ways. First, because this was a "data mining" endeavor, we do not have access to the samples for performing further molecular QC. The original study authors did however perform and report a variety of QC measures to confirm the purity of their virome sample sets. To this end, we summarized those QC results into Table S1, which we added to the manuscript and referenced in the results. Furthermore, we performed our own local alignment (blast) QC protocol to assess the degree to which our OGUs were contaminated. We found that many phages had genomic elements similar to reference bacteria genomes, which agrees with the previous findings that the majority of the phages in these systems are temperate. We also identified two complete phages using the stringent Virsorter algorithm. Conversely, there were minimal phage reference genome elements identified in the bacterial OGUs, which also agrees with the previous findings that these sequences are mostly bacterial (cited in Table S1). We also did not identify any phages in the bacterial dataset using Virsorter. The lack of prophage identification, both here and in the previous study analyses, speaks to the need to sequence the bacteria more deeply and create more robust assemblies (their larger size prevents assembly *completion at the level that can be done with viruses).* **TABLE S1; LINES 93 - 119**

3. Combining data from VLPs and microbial fractions: Information on how data from these two individual samples were matched and normalized is not provided. How are potentially different amounts of VLPs in different samples accounted for?

We are not sure we understand this comment? Our goal was to use the edge weights to reflect samples that were highly relative abundant in both the bacterial metagenome and the phage metagenome. The edge weight used was therefore a multiplication function of the two relative abundance values, yielding a high value for those that were highly abundant in both, and low values for those that were not. This is described in the methods section of the manuscript. These values were recorded on a per-sample basis (within the subgraphs) so that each potential combination is accounted for, within each sampling pair. Furthermore, it may be worth noting that the bacterial and viral DNA extractions were done separately so as to maximize purity of the samples. This approach allows us to more accurately measure the relative abundance in the bacterial and especially viral samples, as has been noted previously. We made this clearer in the text and added more relevant citations. **LINES 149 - 153**

4. Datasets underpowered to sustain claims on diet and obesity: Although the authors point out that the data are underpowered, they still make strong claims and the analysis of these data as major points of this manuscript.

We appreciate this point referring to figure 2, and while we attempted to avoid too strong of claims, we realize we fell short. Our goal was to present these as interesting observations, somewhat like a case study. We addressed this in the manuscript by further minimizing the effects of our claims and adding relevant caveats. **LINES 29 - 31; 184 - 201; 268 - 277**

For example, a. the authors describe (158-164) how diet affects the network structure, without using sufficient data points to perform any statistical tests. b. the authors claim (166-169) a difference between obese and healthy individuals by using one or two data points per group. With respect to these analyses, the authors write for example, "We found evidence that diet was sufficient to alter gut microbiome network connectivity" (235), which is a clear overstatement and such claims should be modified here and elsewhere in the manuscript.

Minor Points

- Throughout the manuscript, it is hard to follow how which and how many data points were used for each analysis. A supplementary table listing the samples for each of the studies and providing sample sizes in each figure would be very helpful.

We agree that this needed to be clarified in the manuscript. We added the relevant sample size information to the text and figure legends to make clearer what the data actually represent. **LINES 186; 194 - 196; 219 - 221; 223 - 225; FIGURE LEGENDS 2,3,4**

- Figure 1D is not informative and could be removed since the information is provided in the main text.

It is true that the information is in the text, but we feel this is still an important panel to the figure because it illustrates the degree of connectance between the phage and bacterial communities, and provides a visualization to the text included in the results section. Even more importantly, we intended this panel to act as a visual validation that there are no erroneous connections among phage and bacterial nodes.

- Figure 1 EFG: color labels in legend are wrong and not since y-axis labels are present.

Thank you for noting this mistake. We have fixed it. **FIGURE LEGEND 1**

- Figure 2 could use headers above the plots to indicate which one relate to impact of diet (AB) and which one to obesity (CD).

This was especially confusing because of the mislabeled axes. We fixed the axis labels to make the data clearer. **FIGURE 2**

- Figure 2 panels C and D have wrong x-axis labels. Should be "obese" and "healthy". Here it is not clear why only 2 and 1 data points are available. The original data in Reyes et al. seem to consist of 32 fecal samples from 13 individuals from 5 families with at least two time points.

We fixed the error, and made the causes for the sample size clearer in the text as well as the figure legends. Thank you for pointing this out. **LINES 194 - 195; FIGURE LEGEND 2**

- Figure 3A. Not clear why only 8 data points are available if Minot et al. report 15 samples from 5 individuals.

We only used the subset of data points that had temporal data spaced approximately the same distance apart. We made this clearer in the text. **LINES 186 - 187; 219 - 221; FIGURE LEGEND 1,2**

- Figure 3B. Not clear how data were computed. If the comparison of within individuals (i.e. 1 value if same data as in 3A) with mean of all other individuals, then standard deviations should be shown. Suggest to show only the smallest inter-individual distance, which would greatly strengthen the point on individuality.

Yes, this is a comparison between the interpersonal distance, and the distance between the person's samples and all other samples. We considered calculating the distance differences using standard deviation and pairing the samples as every interpersonal comparison instead of the average. The problem we found was that that pairwise comparison inflated the significance values by artificially creating a large n, all of which were not truly independent. We therefore collapsed the data points into a single average instead of pairing the same sample to many other samples. Showing the smallest distance is also an interesting idea, but we are worried that this may also skew the data by failing to account for the variation in the sample distributions.

- Figure 3C. Not clear which data were used, in particular if same data as in Figure 2 were used. E.g., were only the twins included, or also the mothers?

We made this information clearer in the text. **LINES 223 - 225; FIGURE LEGEND 3**

- Figure 3D. As described for 4B, smallest inter-individual distance should be used if showing individuality should be demonstrated, which is very different from showing on average higher similarity within than between individuals.

This is also a good point, thank you. We hit this point in the related bullet point above.

- Figure 4. Please provide sample numbers for each of the boxplots and data in 4E.

We added this information to the figure legend. **FIGURE LEGEND 4**.

Reviewer #2 (Comments for the Author):

The work presented in the manuscript entitled "Biogeography & Environmental Conditions Shape Phage & Bacteria Interaction Networks Across the Human Microbiome" by Hannigan et al. consists on prediction of host-bacteriophage interactions from skin and gut microbiome public datasets and graph-based analysis of the underlying phage-host networks, assessing the effects of associated metadata (obesity, diet, type of skin site). For the prediction of interactions, the authors trained a random forest classifier using known phage-bacteria interactions using the following features: CRISPR-spacer detection, prophage prediction based on nucleotide similarity (blastn, e-value <=1e-10.), shared gene identification based on aminoacid similarity (Diamond) and detection of pairs of genes whose proteins interact (Pfam interaction information within the Intact database). My main concern with this paper is on the prediction of phage-host interactions. The authors report 72287 infectious relationships among 578 nodes, representing 298 phages and 280 bacteria. If all phages would infect all bacteria, the network would have 83440 links (complete network = total possible interactions). The reported network contains 86,6% of the number of possible interactions. This essentially means that most phages are generalists and can infect any bacteria…The authors fail to report the number of phages and bacteria per microbiome: skin or gut. This number would allow us to estimate the total number of possible interactions within an environment (gut or skin), a more relevant number to calculate the density of the network. One explanation for the high density of the network is the use of aminoacid similarity between phage and bacteria as predictor. Shared genes have been used to link phages to their host, but at nucleotide rather than at aminoacid level (Edwards et al. FEMS Microbiol Rev. 2016). Independently from the classifier performance, when confronted with this high network density, the shared genes should have been further analysed and the procedure revised.

Major comments:

1. Network inference: operational genomic units were linked sample-wise, microbiome type wise or all microbiome together? Aminoacid similarity is not specific enough and is likely the cause of the high density of the network, which suggests that almost all phages infect all bacteria (86% density), meaning the resulting network is (likely) no statistically different from a random network with equal number of links. In the absence of further evidence and controls, I am not convinced with this result.

We would like to thank the reviewer for pointing out this area of clarification. The operational genomic units were linked sample-wise, by linking the OGUs found between each sample. The source of these samples were recorded, so that we knew whether they originated from the gut or the skin, etc. We also agree that amino acid similarity of genes along is not enough, and that is why we incorporated other similarity metrics in our random forest model. As we described, even though the amino acid similarity metric was the most "important" to the model, it was used within the context of the other metrics in the random forest. It is true that this is a dense network, and we hypothesize that that may be because each OGU does not necessarily represent a specific phage or bacterial strain, but is an operationally defined group of similar genomes, similar to the operation taxonomic unit (OTU) often used in microbiome studies. In other words, we are not using the OGUs as individual phage strains, but instead these OGUs are operational units that allow us to condense our dataset by contig similarity. **SEE LINES 114 - 119; 304 - 322**

2. Number of samples per study and metadata groups are not reported but the authors state about the diet effect: "Tests for statistical differences were not performed due to the small sample size". If sample sizes are small for statistical analysis, no comparison should be made in the first place. The plots of Figure 2 feature 2 vs. 3 comparison for high fat vs. low fat and 2 vs. 1 for obese vs. healthy…are

these the nb. of samples per group? Bacterial or phage diversity are likely confounders for network connectivity comparisons and this is overlooked here.

We appreciate this point and while we attempted to report this as more of a case trial observation (as outlined for the related comment above), we unintentionally used too strong of language, thus overstating the finding beyond our intentions. We also needed to make the analysis being performed clearer in the text. It is also true that the network analysis is a function of the diversity of the two communities, and we do not wish to make it sound like they are separate. We would be hesitant however to call this dependence a confounder since it is more building off of the diversity data and incorporating other additional information from the community. We made this clearer in the text. **LINES 186; 194 - 196; 219 - 221; 223 - 225; FIGURE LEGENDS 2,3,4**

3. Graph-analysis: Analysis of network nestedness and modularity (Weitz et al. Trends in Micr. 2012) may be more relevant than diameter, degree and closeness for phage-host networks.

We chose to omit such an analysis as was done in Weitz et al because, as we attempted to describe in the manuscript, we are not using the OGUs as individual phage strains as was done in the Weitz and other papers. Instead these OGUs are operational units that allow us to condense our dataset by contig similarity. We feel that making statements with as high of phylogenetic resolution as Weitz et al would be unfounded at this point.

Other comments:

1. Training set contained on 43 diverse bacterial species and 30 diverse phage strains. Given than these numbers are (much) lower than the ones published in references 41, 47-51, from which the data on known interactions was extracted, I wonder how the curation/filtering was done.

Our limitation was not on how many known interactions that could be found, but instead on how many non-interactions could be found. In other words, we needed to utilize known infectious pairs, as well as phages that have been validated as not infecting certain bacteria. Because we were limited to only a small number of validated non-interactions, we decided to also limit the number of positive interactions included, so as to prevent a large bias in the ratio between the two interaction types in the training dataset. We apologize for this not being clearer in the text and have clarified that text. **LINES 124 - 127**

2. The sensitivity of the classifier is reported as 0.846. However, “Approximately one third of the training set relationships yielded no score and therefore were unable to be assigned an interaction prediction”. This means the sensitivity can not be higher than 0.67…

Here we are reporting the accuracy for those values that were scored. Pairs without scores were omitted from this analysis, and we instead only focused on those that were scored. We made this clearer in the text. **LINES 143 - 144; FIGURE LEGEND 1**

3. Dataset: The authors should report the number of samples per study and associated metadata (e.g low /high fat diet, obese/lean).

This is a very good point and we added this information. **LINES 186; 194 - 196; 219 - 221; 223 - 225; FIGURE LEGENDS 2,3,4**

4. OGU construction, why the OGUs were constructed with CONCOCT without coverage information (CONCOCT-NC)?

We maintained the coverage information but omitted the feature of total sequence counts since we worried this would confound the results. This means that the feature set matrix used by the CONCOCT clustering algorithm only included k-mer frequencies and sequence count (relative abundance) values per sample.

5. Network construction: The master network contained the three studies as sub-networks, which themselves each contained sub-networks for each sample. The authors provide only numbers of phage and bacterial nodes for the master network. For predicting the associations, were all phages and bacteria considered disregarding the sample type (skin vs. gut)?

Yes, this is described correctly by the reviewer.

6. The phages and bacteria in the gut diet and twin sample sets were more sparsely related: each contained fewer than 150 vertices, fewer than 20,000 relationships“. This is a bipartite network, what should be reported is the number of vertices of each type rather than the total nb of vertices (this is relevant to calculate the network density, we cannot know here what is the nb of possible interactions).

This is a good point that makes the manuscript clearer. We added these values to the results section of the manuscript so that the reader can access this information. **LINES 154 - 163**

7. Comparison with Paez-Espino et al. (Nature 2016, doi:10.1038/nature19094) phage-host human associated subnetworks should be done.

This is a cool suggestion. It sounds like this comment is a reference to the host range inferences within the suggested manuscript. Paez-Espino et al used CRISPR and tRNA comparisons to identify bacterial hosts for the bacteriophage genomes and genome fragments. The group’s work suggested that phage host range is broader than has been traditionally appreciated. In our discussion section, we mentioned a comparison of *these and related findings to our own dataset. This information was included in the operational genomic unit section of the discussion.* **LINES 318 - 322**

8. Figure 1. panel B: methods include blastx, in the network inference Diamond was used instead; Panel E-F, colors in figure do not correspond to those in legend.

It is true that we used the Diamond aligner, but we used the blastx algorithm within the Diamond aligner (written as `diamond blastx`). We made this information clearer in the text. We also fixed the color references. **FIGURE 1**

9. Figure 2 legend: it is not clear if/that the dots represent sample-wise networks. "Lines represent the mean degree of centrality for each diet", the only lines in the plots are parallel to the centrality axis, they are just splitting high fat vs. low fat diet fields in the x axis. X-axis from panels C and D are mislabeled.

Thank you for pointing these out as well. We clarified the dot information, and also fixed the incorrect labels. **FIGURE 2**

Reviewer #3 (Comments for the Author):

This research article by Hannigan et.al., focuses on an important and indispensable part of the microbiome- "Virome". Interplay within inter and intra microbial community members is important to get the complete picture. This paper focuses on such interplay and understands that microbiome operates with both the viral and bacterial community members. For this study they worked on previously published metagenomics data, to not only infer based on nucleotide information but also based on protein interactions.

Main Concerns:

The data from studies referenced 13 & 14 even though may have bias due to sequencing, have similarity in approach wherein collection of VLP and bacterial whole genome shotgun sequencing was performed. VLP data in reference 14 was based on Genomiphi amplification of VLP DNA prior to sequencing that is bound to introduce amplification and sequencing errors which would further lead to clustering bias in the current study. The concern is also regarding data from references 38 and 39 that are not only performed by 454 technology. Further on, reference 39 doesn't have VLP data available. The study is based on 16S and whole genome shotgun of gut microbiome.

We referenced this statement in reference 38: "We had previously performed shotgun sequencing of total fecal DNA isolated from the 12 frozen samples obtained at the first time point that were now used to prepare purified VLPs (Table S3)." This reference (12) led to our reference 39, which contained the data in the related SRA submission and was cited accordingly. Between these two published datasets, we were able to pair the whole shotgun sequencing data with the purified virome data.

The question is how were the above biases (including the kitome) taken into consideration while conducting these analyses. The authors may be better off carrying out the analysis only using data from references 13 and 14 for skin and only data from 38 for gut?

This is a good point that there certainly could have been extraction, kit, or other related differences between the 38 and 39 dataset. We feel these findings, which we attempt to present more as case observations that will require more dedicated study, are worth noting in the paper. We altered the text to better clarify our position. We also avoided analyses in which we directly compared datasets, because we are unable to discern between true biological differences and differences between the studies. Such analyses would require further and properly controlled sampling.

Even though the authors have acknowledged some of these biases including MDA, in their discussion, these concerns should be explicitly discussed and measures they have taken to overcome these.

This is a good point. We included a description of these biases associated with these methods. As far as measures to overcome these biases, we acknowledge the biases and only performed analyses within studies where samples were treated the same. The goal is for this work to use these findings as a jumping off point, and performing future studies with more limited biases from techniques such as MDA. **LINES 295 - 303**