Impact of Disease on the Viruses of the Human Microbiome

Geoffrey D Hannigan, Melissa B Duhaime, Patrick D Schloss

Contents

Abstract

Here we present a global view of phage-bacteria interactions across the human virome. We present our model for phage-bacteria interactions with validation for accuracy and sampling coverage. These networks are valuable because they do not rely on the sub-optimal reference genome datasets, and provide a more accurate view of the relationships within the community. We find that interactive dynamics are associated with disease states and anatomical body sites, using a global virome meta-analysis dataset. Our comprehensive approach to understanding the virome provide new insights not only inro composition and diversity, but their context in the greater community. We find that disease states and anatomical sites are not only linked to altered community composition and diversity, but also represent significant shifts in interactive dynamics.

Introduction

Our Microbial Planet

Studies of the human microbiome in recent years have revealed an unprecidented association between microbial communities and human health. Early efforts showed that a healthy human harbors complex microbial communities that vary between physiological systems (e.g. gut, skin, and oral microbiome). Studies with cohorts of diseased individuals revealed that microbial community composition, diversity, and stability shifted in disease states, providing early evidence for a role of the microbiome in human health. Follow up work has further solidified the role of the microbiome in many diseases and has been used to better inform important clinical practices, as well as to develop therpeutics and prognostic/diagnostic tools.

As the medical and ecological relevance of this research area has become more apparent, and the resources become more accessible and scalable, the focus has shifted from individual, relatively small studies to large comprehensive studies of broad populations. One of the earliest was the human microbiome project, a US National Institutes of Health (NIH) initiative to provide large amounts of data for the general human microbiome. Other global efforts to understand microbial ecology include the JGI metagenome initiative, the Oceanomics initiative, the earth microbiome project, and the incredibly comprehensive tara oceans project. Global, comprehensive approaches have powered informative studies that provided important new insights into microbial communities. These advances have been witnessed for both bacterial and viral communities.

The benefits of such comprehensive studies has been especially apparent in viral communities (the virome). The study of viral communities remains in its infancy and has been hindered by the lack of marker genes (analogous to 16S rRNA in bacteria), incomplete reference databases, uncertain categorization, and less robust computational toolsets compared to bacterial communities (e.g. Mothur and Qiime). Benefits from global studies have included expansions of our global viral catalog, a better understand of global viral diversity, improved categorization of phages, and evaluations of universal disease signatures.

Importance of Phage - Bacteria Analytical Synergy

Regardless of the scale of these studies, bacterial and viral communities are almost always studied in isolation, even when the two populations are referenced in the same study. Even when the two are sampled together in a single dataset, they are often analyzed in parallel instead of conjunction, concluding with cursory associations between the otherwise isolated communities. While still informative, these approaches are far from ideal and leave us wanting for more robust insight into these communities as a whole.

We study bacteriophages (and viruses in general) by their hosts. Phage functionality and replication cycles depend entirely on establishing an infection. Without a host, phages cannot replicate or perform their other other metabolic functions. This is reflected in phage taxonomic classification in which phages are defined by the host they are isolated from. It is therefore imperative that we move on from studying viromes in isolation and begin to study them in the context of their bacterial hosts.

Not only do phages control the metabolic and functional capabilities of their bacterial hosts, but they also control bacterial community ecology which in turn impacts the virome. In fact, the two communities are largely inseperable. Both *in vitro* and *in silico* ecological studies have shown that phages are necessary for maintanance of bacterial community composition, diversity, and stability. Thus a truely informative microbiome study must incorporate both bacterial and viral communities together. Insights from such an approach will extend beyond isolated community composition and diversity, and will provide a more sophisticated understanding of the lacdscape of the greater human microbiome, as well as information on the role disease plays in community fragility, vulnerability, and identity of influential organisms.

Addressing Previous Shortcomings

Here we present the use of a machine learning-based phage-host prediction algorithm and graph theory to evaluate the global disease signatures of the phage-bacterial communities within the human virome. The strength of this graph-based approach is that it allows us to focus on the **relationships** between the microbes. This builds off of extensive previous work that has used network theory to undestand complex ecosystems, including some bacteria-phage communities.

Most microbiome studies rely on three core metrics for understanding their communities: alpha (within sample) diversity, beta (between sample) diversity, and member relative abundance. Our technique will instead rely on seven core metrics that are used to describe the networks in varying states: connectance (fraction of potential links actually established), nestedness, generality and vulnerability (proportion of infected hosts and infecting phages, respectively), network robustness (number of host extinctions required for extinction of half of phages), degree of centrality, network Shannon diversity, and changes in identities of organisms establishing links.

Biological Importance

Network-based approaches are biologically informative and can be used to provide a new biological understanding of the human microbiome as a whole. Networks allow us to understand the stability of a predator-prey system such as is observed in the human virome, based on the connectedness and distribution of nodes. A **highly connected** network is **more stable** as the removal of one or more nodes is less likely to disrupt the flow between nodes. In other words, the path between nodes is more easily corrected when the nodes are more highly connected. Thus this analytical approach will provide us with greater insights into the roles of broader communities in microbiome stability.

Divrsity is often a valuable metric for understanding a microbial community as it provides a metric based on the condensation of the community at large. The metrics most often used are alpha (within sample) and beta (between sample) diversity of a particular population such as bacteria. We can use an ecological network, such as is built in this study, to calucate a new metric of diversity in the context of the greater, interacting community. The **topological diversity** of a system can be calculated using a network adapted version of the Shannon entropy metric[1]. This metric accounts for the number and evenness of distributed nodes within a community. Burt's measure of "structural holes" also provides a method of calculating diversity that relies on open triads (edge holes). This can provide us with a more biologically informative set of measures beyond the virome diversity calculated using an isolated virome system alone.

Not only does this approach provide a community diversity perspective, but also provides greater context for the roles of bacteria and phages in their community. The connectedness of individual bacteria and phages provides insight into their impact on the community and the consequences of removing them. In other words, this allows us to identify keystone microbes or "hubs" within the community. Understanding the distribution of these hubs across communities and in disease states allow us to better understand the biological background beyond "increased bacterial abundance" or "phage presence/absence".

Results

The Global Human Virome Dataset

We leveraged the extensive public sequence archives to assemble a **global human virome** dataset; a robust human virus community metagenomic dataset that spans diverse body site environments. Dataset sampling includes the gut, oral cavity, skin, and urinary tract systems, all of which are associated with healthy and disease states, and were all collected by multiple, independent groups. By working only with virome datasets that were purified for virus like particles (VLPs), we are able to establish confiendence that we are detecting the *active* virome component. The resulting dataset contains data from ten prominant virus metagenomic studies[2–11].

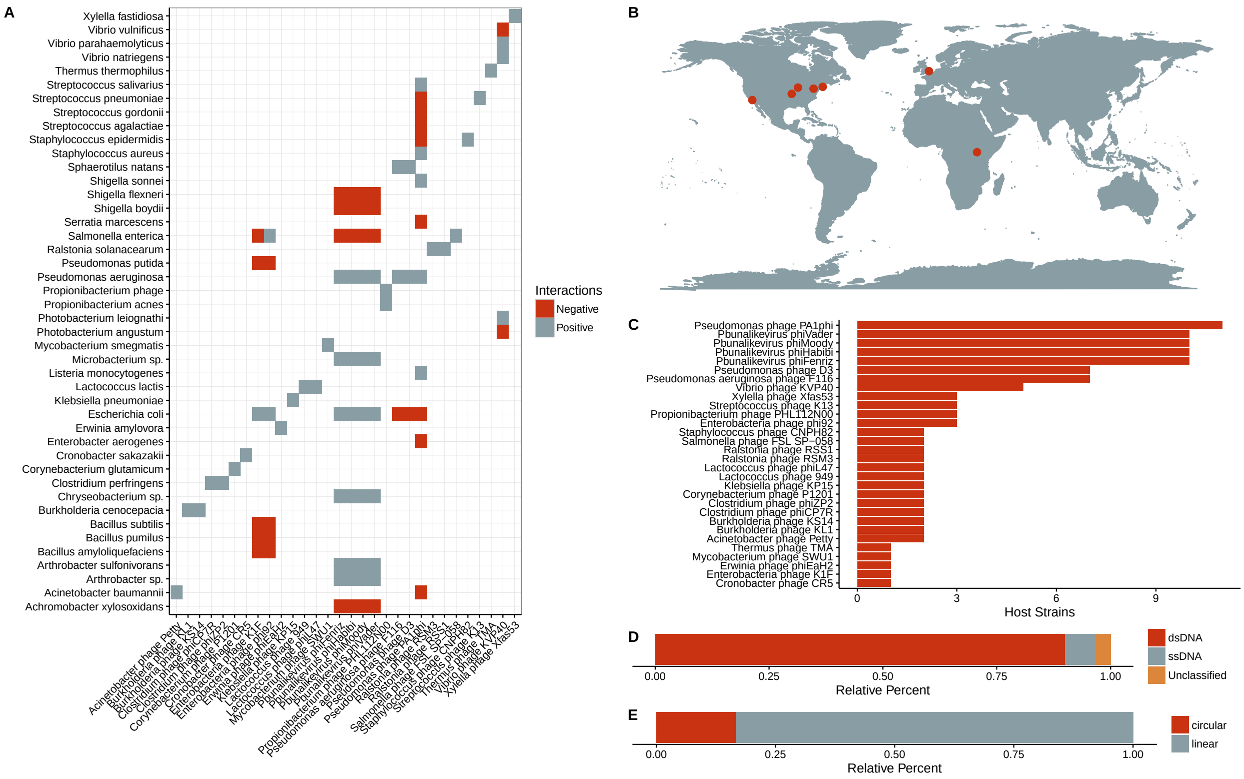

Figure 1: Summary information of validation dataset used in the interaction predictive model. A) Categorical heatmap highlighting the experimentally validated positive and negative interactions. Only bacteria species are shown, which represent multiple reference strains. Phages are labeled on the x-axis and bacteria are labeled on the y-axis. B) World map illustrating the sampling locations used in the study (red dots). C) Quantificaiton of bacterial host strains known to exist for each phage. D) Genome strandedness and E) linearity of the phage reference genomes used for the dataset.

The GHV raw sequences were quality filtered according to our high threshold and assembled into contigs that represent either complete viral genomes or genomic fragments. We assembled approximately 30,000 contigs whose sequencing depth ranged from ten to over ten thousand sequences **(Figure ??)**. Contigs were tens of thousands of base pairs long. A large subset of contigs assembled as complete circles, suggesting complete coverage of a subset of viral genome sequences.

Modeling Phage-Bacteria Interactions

We used Neo4J graph database software to construct a network of predicted interactions between bacteria and bacteriophages. Results from a variety of complementary interaction prediction approaches were layered into a single network. *In vitro*, experimentally validated interactive relationships were taken from the existing literature. Clustered Regularly Inter-spaced Short Palindromic Repeats (CRISPRs) are a sort of bacterial adaptive immune system that serves as a genomic record of phage infections by preserving genomic content from the infectious phage genome. These records were used to predict infectious relationships between bacteria and phages. Infectious relationships were also predicted by identifying expected protein-protein interactions and known interacting protein domains between phages and their bacterial hosts. We finally used nucleotide blast to identify genomic similarity between bacteriophage genomes and sections of bacterial genomes. Such a match is a good predictor of an interaction between the phage and it's bacterial host.

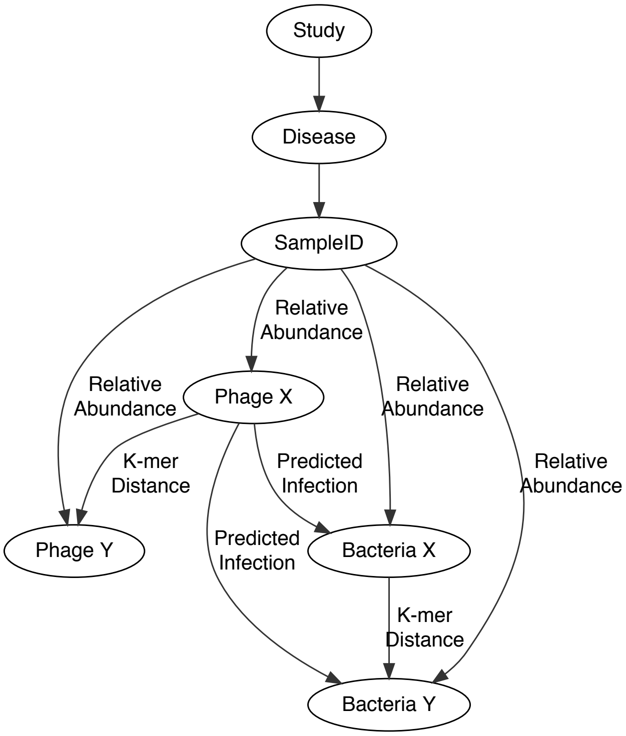

Figure 2: Diagram illustrating the structure of the interactive network.

We began by working in a controled data environment in which the interactions and lack of interactions had been experimentally validated **(Figure 1 A)**. This dataset was extracted from manuscripts published between 1992 - 2015 and includes sampling representation from North America, Africa, and Europe **(Figure 1 B)**[12–17]. Many of the phages are known to target multiple bacterial hosts **(Figure 1 C)**. The majority of the reference phages used contained linear dsDNA genomes **(Figure 1 D-E)**. It is important to note the strength of our approach in that we used data of confirmed non-interactions as well as confirmed interactions. Previous approached have claimed to perform tests of sensitivity and specificity, but assumed a lack of empirical evidence denoted a lack of interactions, which we know to be untrue. Our approach circumvents this problematic assumption.

We used four predictive score categories of the controlled dataset with a tuned random forest model to classify each sample as an interaction or lack of interaction. The model was validated using repeated k-fold cross validation with k = 5 and ten repetitions. The model was optimized using the receiver operating characteristic (ROC) algorithm for the higher area under the curve (AUC) as implemented in R {caret}. The resulting model exhibited an AUC of 0.853, a sensitivity of 0.851, and a specificity of 0.774 **(Figure**

3). These parameters describe only the interactions that were scored. Those that did not have scores were classified as having no interaction prior to predictive modeling. The most important predictor in the model was nucleotide similarity between genes, followed by nucleotide similarity of whole genomes. Protein family (Pfam) interactions were moderately important to the model, while CRISPRs were minimally important. The minimal importance of CRISPRs was primarily due to the low frequency of CRISPR matches to phages compared to the other parameters used.

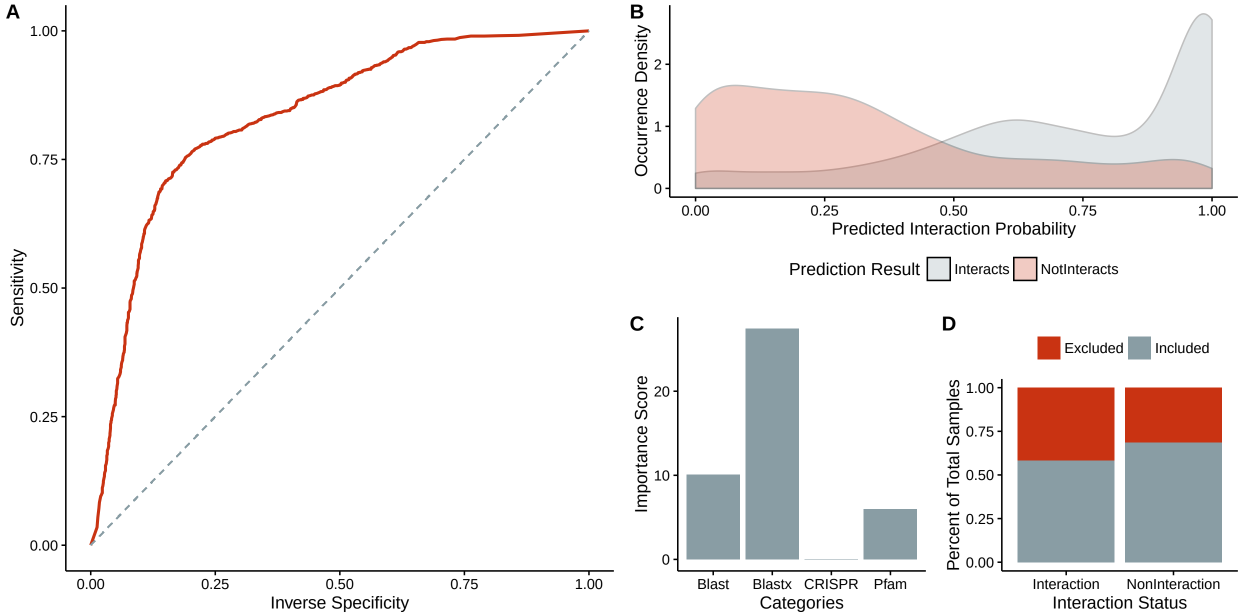

Figure 3: Random forest model for bacteria - phage interactions. A) ROC curve of the ten iterations used to create the prediction model. B) Density plot of the distribution of sample interaction probability. Groups indicate whether the sample represented an interaction. C) Importance scores associated with the criteria used to create the random forest model. D) Proportions of samples excluded from model learning due to a lack of scoring. The true interaction status of the sample is noted on the x-axis and bars are colored by the proportion of sample excluded (red) and included (grey) in model training.

Basic Network Properties

The complete network contained X nodes and Y edges (i.e predicted infectious interactions). The number of nodes and edges per sample and total per study.

Disease and Co-Evolution (Tropism)

Tropic patterns of bacteriophages provide us with an understanding of bacteria/phage co-evolution, as well as a general understanding of the system behavior. In the past, phage tropic patterns have been represented as pairwise adjacency matrices of phages and their predicted hosts.

There are four possible patterns associated with tropism matrices[18]. Each phage may only infect a single or very limited range of bacterial host strains, which results in a nearly diagonal matrix. Groups of many phages may exclusively infect groups of bacteria, resulting in a modular, block matrix. These patterns indicate coevolution that resulted in phage specialization. Coevolutionary pressures that allow for diversification of phage tropism result in a nested matrix structure, in which some specialized viruses remain specialized whereas others exhibit an evolution toward infecting multiple other bacterial strains. The final model is completely random tropism without any distinguishable matrix pattern. These different models are not

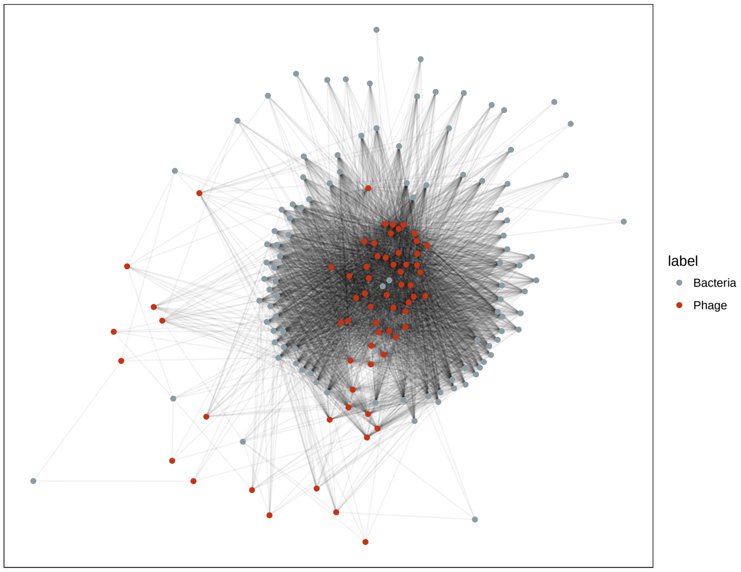

Figure 4: Network diagram of the phage - bacteria relationships.

mutually exclusive and real systems, especially of large size, are likely to exhibit multiple traits. For example, a community may have a modular-nested structure in which modules of interacting bacteria and phages exhibit a nested pattern.

Impact of Disease on Community Fragility

The distribution of members of the human microbiome netwrok determines in part the resliance of that community to significant disruption and eventual disintigration. A network that fits an **exponential model** is distributed such that nodes are equally connected. A network that fits a **power-law model** is conversely distributed to be dominated by few, highly connected nodes that act as hubs between large node clusters. Sequentially removing the most connected nodes of a network that fits a power-law distribution results in a sharp rise in network diamter until it disintigrates into small isolated clusters that are no longer connected. Although vulnerable to the removal of the most connected nodes, power-law distributed models are highly resilient to random node removal, as the probability of randomly removing an important node is often low.

Although the healthy states differed by anatomical location, they all exhibited a power-law model for distribution. As was expected from previous studies, disease resulted in an increase phage diversity and a decreased bacterial diversity. In addition to these previous observed community signatues, the disease states reuslted in a sharply increased network diameter, meaning important hub nodes were removed from the community. A lack of change in diamater would suggest a resilience to randomly removed nodes. Specifically, the nodes lost in disease included X and Y, which had an average connectedness and edge count of Z. These differed by disease. From this we conclude that healthy phage-bacteria communities are susceptible to the loss of their hub members but highly resistant to random loss.

Altered Phage Predation & Disease Recovery

Mention Periodic-Selection vs Constant-Diversity Models here. Also link phage diversity increase in disease (Rowher Nature and Weitz) to predation patterns and what it means for recovery from disease; reference[19].

"Under which circumstances should we therefore expect PS or CD to predominate? The influence of phage predation on bacterial diversity requires that bacterial populations interact with each other; therefore, host-associated niches can act as physical barriers that prevent direct cell competition and phage dispersal."

Would it be possible to run community recovery simulations using this information? Also should be very doable to run simulations of removing nodes (bacteria/phages) from the network.

Phage & Bacteria Clustering

The infectious network of phages and bacteria presents an opportunity to categorize phages and bacteria based on their shared hosts and predatorys, respectively. This provides allows us to functionally condense members of the communities by their shared interaction properties instead of relying on genome alignments.

Discussion

I think there are a couple of important points that I would like to discuss about this work. *First* is that while this marks an improvement in our interaction modeling capabilities, there is certainly a lot of room for improvement. The model will improve as we add more data, and as we validate more metrics. But for now this model is sufficient for beginning to undestand the system. *Second* is that there is a lot that can be done with this approach. It is powerful and can offer a lot of insight into different aspects of the communities. While we focused on answering a couple of questions, we look forward to using this model to really dive into the data in a powerful, unique way. *Third* is that this is not a methods paper, so while it is presented so that it can be reproduced, it is not a teaching tool or tutorial. This does however present us with an opportunity to build such resources that will be distributed as other works.

Materials & Methods

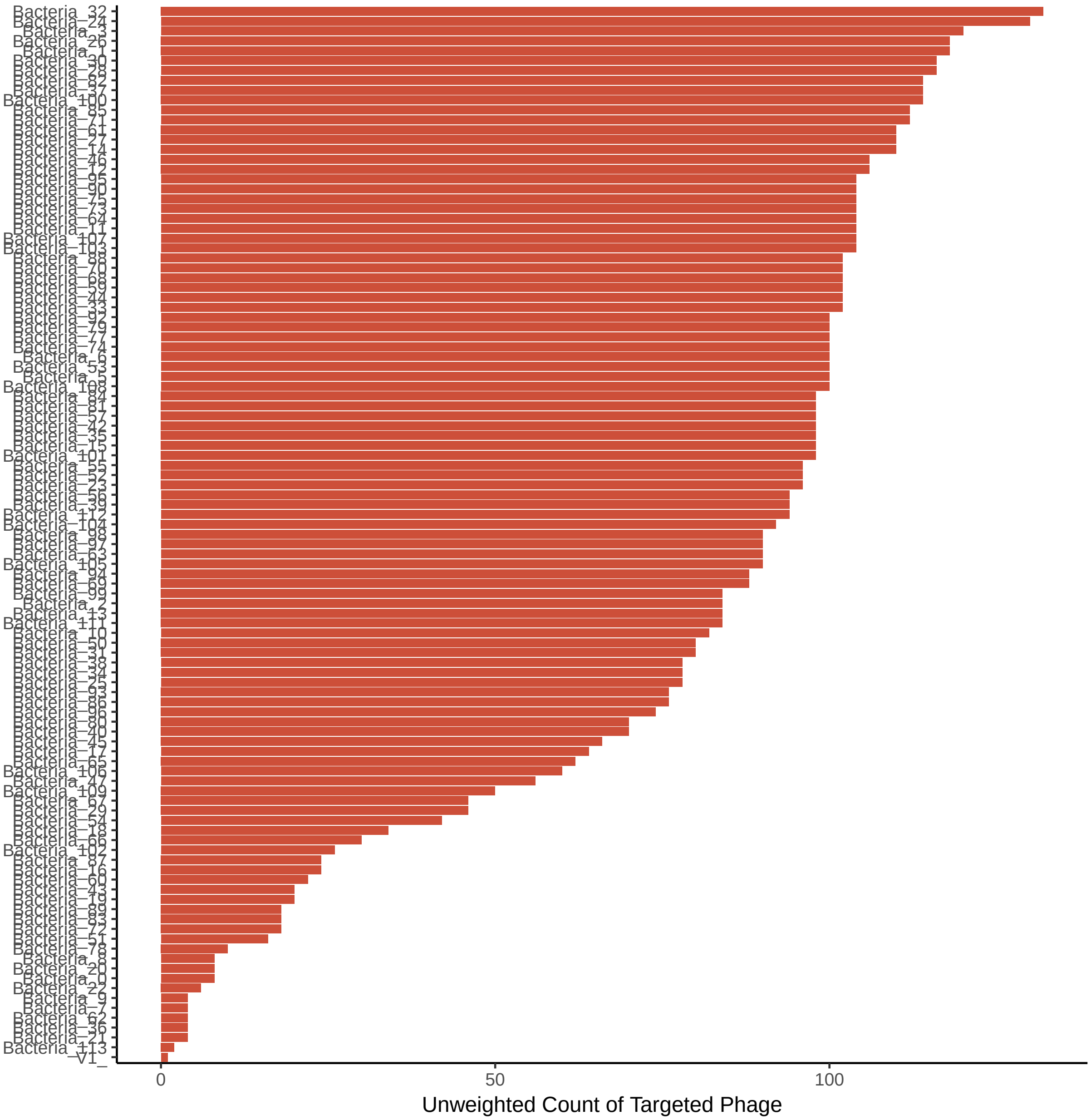

Figure 5: Relative abundance presence of phages by their predicted host target.

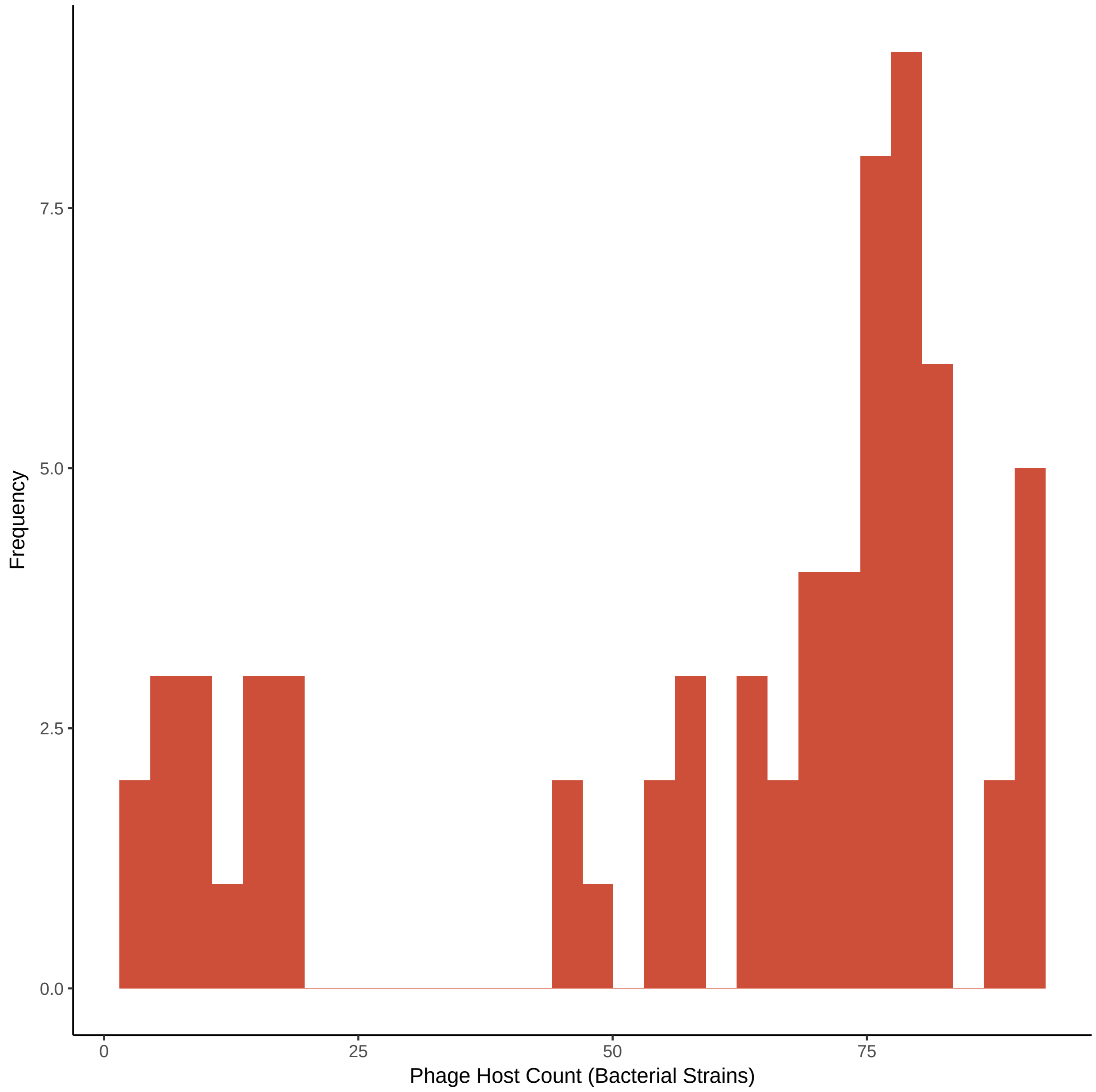

Figure 6: Histogram of the number of bacterial strain hosts identified for each phage contig.

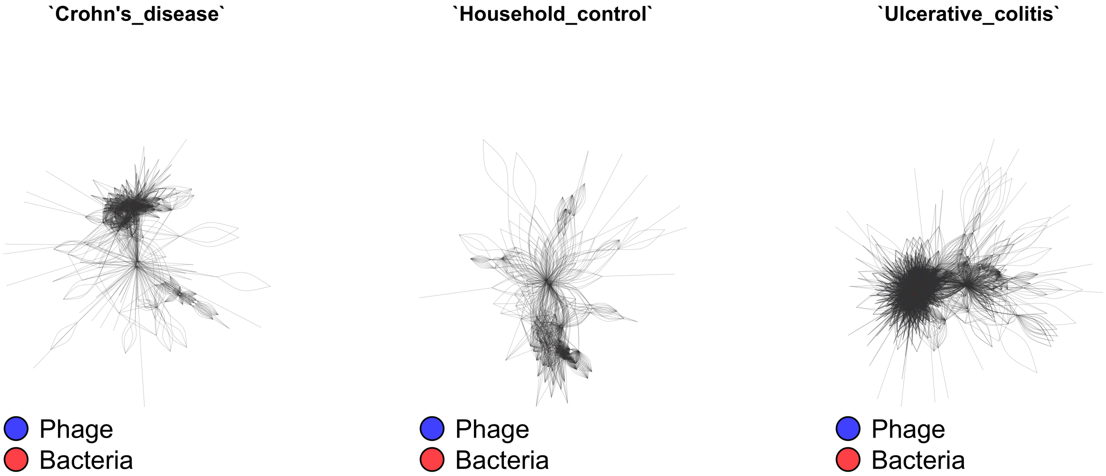

Figure 7: Network visualization for each disease category.

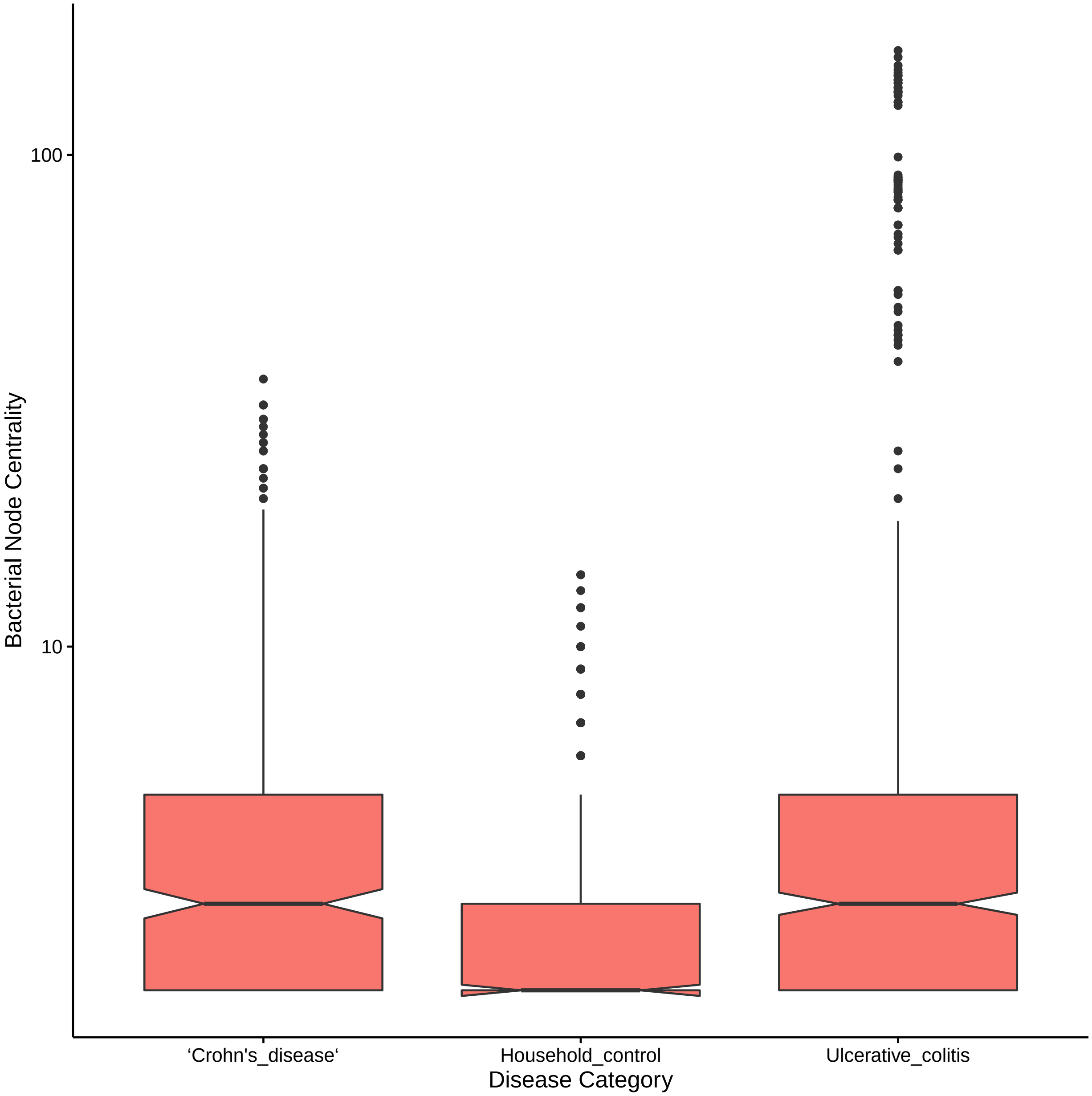

Figure 8: Bacterial centrality across disease states.

References

1. Eagle, N., Macy, M. & Claxton, R. Network Diversity and Economic Development. *Science* **328,** 1029–1031 (2010).

2. Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160,** 447–460 (2015).

3. Monaco, C. L. *et al.* Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host and Microbe* **19,** 311–322 (2016).

4. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* **21,** 1616–1625 (2011).

5. Hannigan, G. D. *et al.* The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *mBio* **6,** e01578–15 (2015).

6. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499,** 219–222 (2013).

7. Ly, M. *et al.* Altered Oral Viral Ecology in Association with Periodontal Disease. *mBio* **5,** e01133–14–e01133–14 (2014).

8. Abeles, S. R., Ly, M., Santiago-Rodriguez, T. M. & Pride, D. T. Effects of Long Term Antibiotic Therapy on Human Oral and Fecal Viromes. *PLOS ONE* **10,** e0134941 (2015).

9. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–338 (2010).

10. Santiago-Rodriguez, T. M., Ly, M., Bonilla, N. & Pride, D. T. The human urine virome in association with urinary tract infections. *Frontiers in microbiology* **6,** 14 (2015).

11. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature Medicine* (2015).

12. Jensen, E. C. *et al.* Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. *Applied and Environmental Microbiology* **64,** 575–580 (1998).

13. Malki, K., Kula, A., Bruder, K. & Sible, E. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology* (2015).

14. Schwarzer, D. *et al.* A multivalent adsorption apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis. *Journal of virology* **86,** 10384–10398 (2012).

15. Kim, S., Rahman, M., Seol, S. Y., Yoon, S. S. & Kim, J. Pseudomonas aeruginosa bacteriophage PA1Ø requires type IV pili for infection and shows broad bactericidal and biofilm removal activities. *Applied and Environmental Microbiology* **78,** 6380–6385 (2012).

16. Matsuzaki, S., Tanaka, S., Koga, T. & Kawata, T. A Broad-Host-Range Vibriophage, KVP40, Isolated from Sea Water. *Microbiology and Immunology* **36,** 93–97 (1992).

17. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews* **40,** 258–272 (2015).

18. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. *Proceedings of the National Academy of Sciences of the United States of America* **108,** E288–97 (2011).

19. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology* **7,** 828–836 (2009).