# Global Interactions & Disease Drivers of the Human Virome

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

Viruses are obligate parasites whose existence relies entires on their ability to infect, function and replicate using their host. In the human microbiome, the most prevalent viruses are bacteriophages; viruses that prey on bacterial hosts. Because bacteria are an essential part of bacteriophage existence, we must utilize a comprehensive analysis that incorporates the virome hosts.

Predator-prey interactions are one of the fundamental pillars to any ecosystem's persistence, diversity, and functionality[1,2]. This is particularily evident in the predator-prey dynamics of bacteria and bacteriophage (bacterial virus) communities. Because phages are incapable of their own metabolic processes, they rely on bacteria as reproductive vessels and functional conduits without which they cannot act or persist. Although bacteria are metabolically capable, their evolution and community stability depend on bacteriophage predation and transduction (phage mediated horizontal gene transfer). Together these communities are capable of stable persistence in a mutually beneficial relationship.

Despite the mutual dependence of phage and bacteria communities, they are often studied in isolation. This is especially true for the human virome and microbiome. The majority of microbiome studies to date have focused exclusively on bacterial community composition and diversity, largely due to technical limitations. Some studies of the human virome have analyzed the bacterial communities in parallel, but often using cursory techniques. Here we present the use network analyses across a global human virome dataset to understand the ecological network signatures associated with disease states.

By understanding the signatures of the interacting communities, we are able to gain new insights into the biology of these systems. The ecological networks can be used to assess community stability and fragility, and components of that network can be used to assess the specific microbial players in that stability. We also know that these networks can be impacted by environmental factors such as resource availability[3].

Until recently, a global human virome analysis has been largely infeasible. Recent advances in sequencing technology and virus purification techinques have allowed for an influx of paired virome and bacterial metagenomic data that have begun to power meta-analysis capabilities.

# 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[4]. 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[5–14].

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 2)**. 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.

We began by working in a controled data environment in which the interactions and lack of interactions had been experimentally validated **(Figure 3 A)**. This dataset was extracted from manuscripts published between 1992 and 2015[15–20]. Many of the phages are known to target multiple bacterial hosts **(Figure 3 B)**. The majority of the reference phages used contained linear dsDNA genomes **(Figure 3 C-D)**. 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 4)**. 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.

Although we have assembled some complete genomes as contigs, the majority of the contigs represent genomic fragments that may originate from the same genome while without sharing nucleotide sequence similarity. K-mer spectrum analyses have been increasingly utilized in recent years as researchers attempt to classify the unknown components of microbial metagenomes. For this study we built our own k-mer spectrum analysis workflow so that we can maintain the most control over the approach as possible. Similarities between genomes/contigs was calculated using the Bray-Curtis dissimilarity metric (perhaps try other metrics here). To account for genomes in different directions, or contigs with inverted regions, our k-mer spectra were calculated as a composite of k-mers in both forward and reverse. Because dissimilarity metrics like Bray-Curtis are sensitive to uneven sampling, the distances are based on equal sampling depths that were normalized by subsampling the contig with the greater number of k-mers down to an equal amount. When considering this approach as a clustering algorith, it is analogous to the de novo OTU clustering approach used in 16S rRNA gene analysis2. The processing was made to run in parallel so as to maximize efficiency.

We began by confirming that k-mer spectra provided informative clustering that accurately reflects known biological properties. To acocmplish this, we collected all of the known bacteriophage reference genomes and classified them by their defined bacterial host. We tested the significance of host classification using an analysis of similarity (ANOSIM) which tests that the composition of kmers within a host class is significantly different from the other sample classes. Using this method, we confirmed that kmer spectra do provide highly significant clustering by bacteriophage host **(p-value=0.001, R=0.6677, Figure 5 A)**. ANOSIM is based on ranked dissimilarity values from the distance matrix. Comparing the distribution of the dissimilarity ranks reveals that some phage taxa are better resolved by kmer spectrum analysis **(Figure 5 B)**. Those phage classes with the least ranked dissimilarities are the best resolved phages, and those at or above the median rank between samples were less well resolved. While Bacillus and Vibrio phages had divergent kmer spectra, Propionibacterium and Streptococcus phages were highly conserved.

As mentioned above, the unique utility of a k-mer spectrum analysis is not in its ability to align, but rather in its ability to infer functional and genomic similarities between biologically related but sequenctially dissimilar genomes. To confirm this benefit over an alignment approach, we assessed the ability of a k-mer spectrum and alignment algoorithm to pair the first and second half of reference genomes. In other words, given the first half of a genome, how accurately can the algorithm identify the matching second half of the genome. We found that alignment performs very poorly at this task and only accurately pairs approximately 20% of the genomes, while the k-mer spectrum algorithm accurately pairs approximately 70% **(Figure 6)**. From this we conclude that k-mer spectrum analyses are able to link genome fragments almost four times more accurately than alignment-based approaches. Together with our biological clustering described above, the data suggest that k-mer spectra do in fact correlate with biological linkages, even when the nucleic acids diverge. This is beneficial when linking contigs (genome fragments) that may have minimal nucleotide similarity despite being biologically linked.

## Evaluating the Effect of Disease on Community Structure

### Impact of Disease on Community Network Stability

### Impact of Disease on Community Network Diversity

The virome has been associated with a variety of disease states across many body sites. Because many of the virome samples within our global virome dataset were associated with diseases, we were able to identify and confirm global virome trends in the human virome. We found that the diversity of disease samples was impacted by the body site. Despite the disease, the body site contributed to the virome diversity signature.

Here I want to get at the fact that at first, given a stable bacterial reference, the interactive dynamics of the networks differ between disease and healthy states.

I can use eccentricity centrality to define the most central microbial nodes of the complex graph.

The diameter of the network is short, suggesting a small-world distribution. Because it follows a scale-free distribution, it is also protected from random attack, but highly susceptible when hub nodes are impacted. I will need to expand on this later.

#### Common & Rare Interactive Signatures

### Identification of Microbial Hubs within the Community

### Periodic-Selection vs Constant-Diversity Models

### Local vs Global Anatomical Signatures

## Discussion

An application that we alluded to here is a graphical approach to microbiome research in general.

## Materials & Methods

## Figures

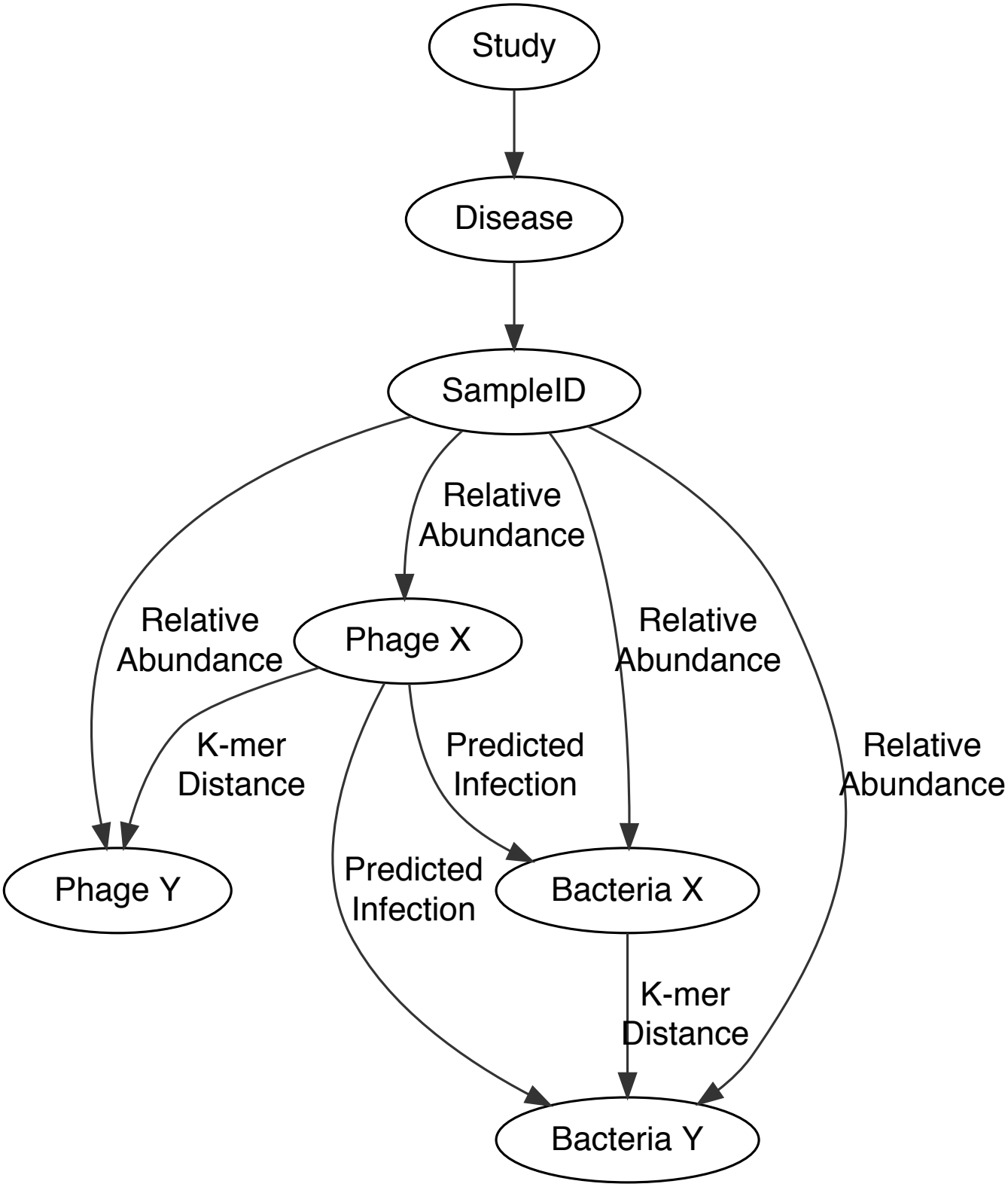


Figure 1: Diagram illustrating the structure of the interactive network.

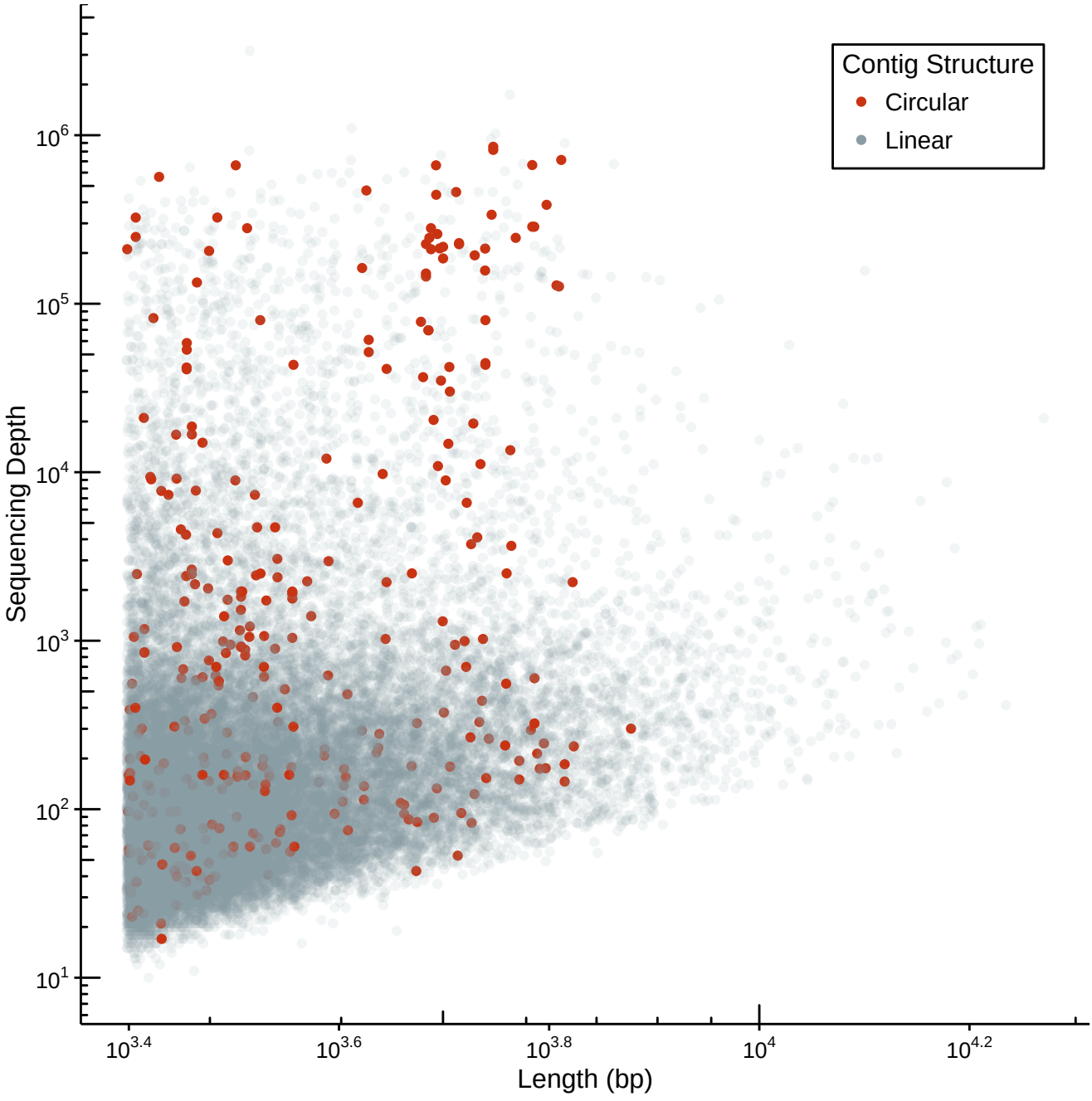


Figure 2: Summary of assembled virus contigs, illustrating contig length (bp) and sequencing depth (number of mapped reads). Probable circular contigs are highlighted in red.

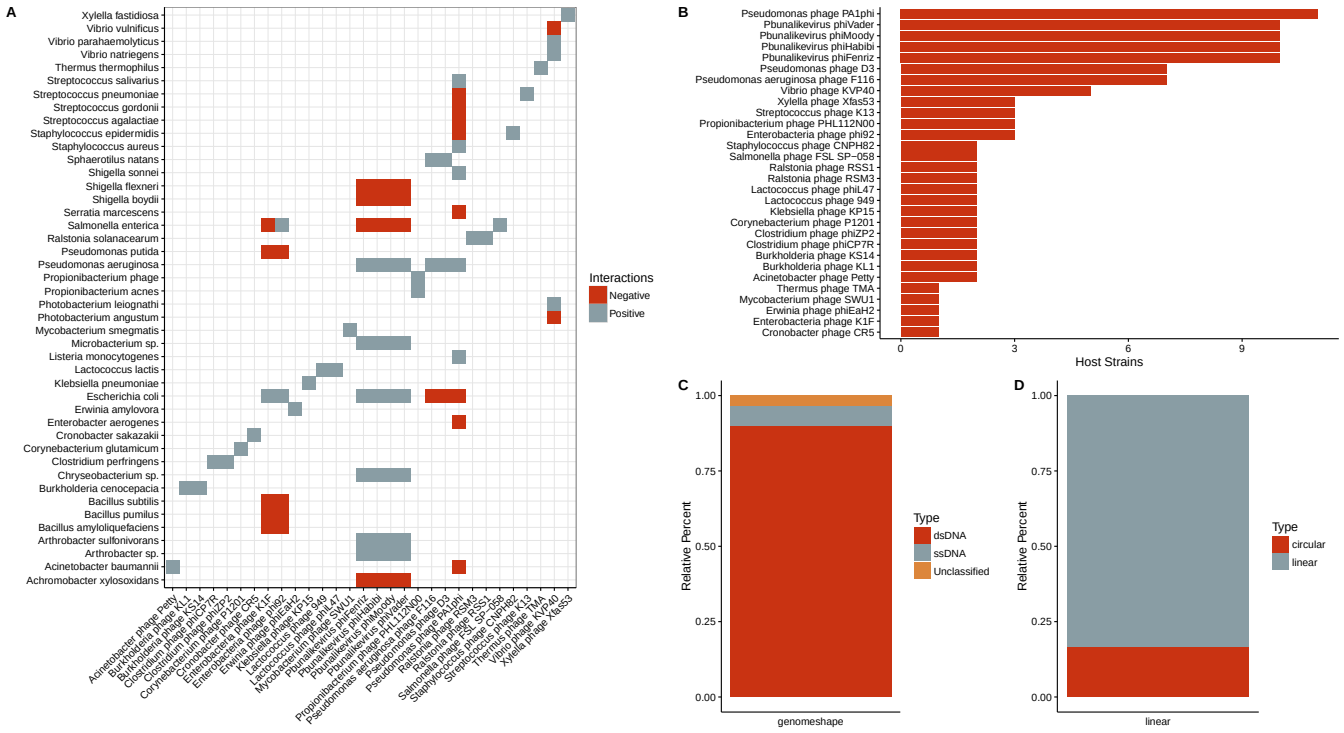


Figure 3: Summary information of validation dataset used in the interaction predictive model. A) Categorical heatmap highlighting the experimentally validated positive and negative interactions. Only bacteria species are shown, which represent multiple reference strains. Phages are labeled on the x-axis and bacteria are labeled on the y-axis. B) Quantificaiton of bacterial host strains known to exist for each phage. C) Genome strandedness and D) linearity of the phage reference genomes used for the dataset.

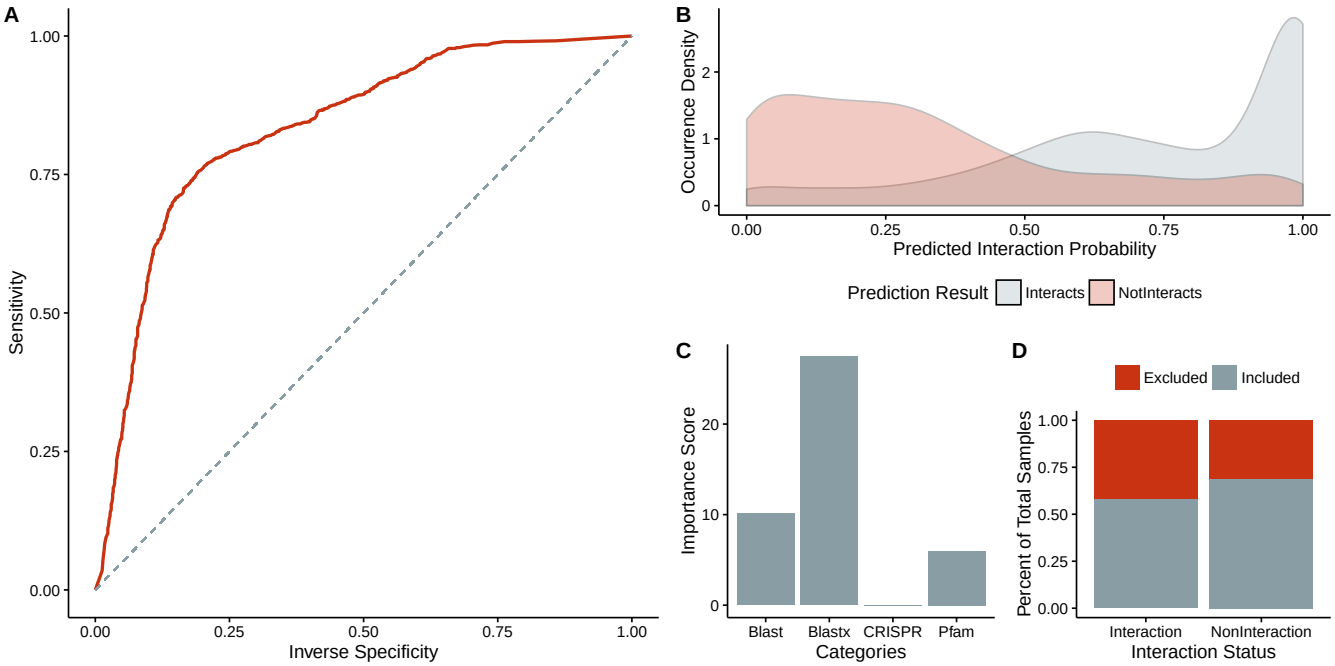


Figure 4: Random forest model for bacteria - phage interactions. A) ROC curve of the ten iterations used to create the prediction model. B) Density plot of the distribution of sample interaction probability. Groups indicate whether the sample represented an interaction. C) Importance scores associated 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.

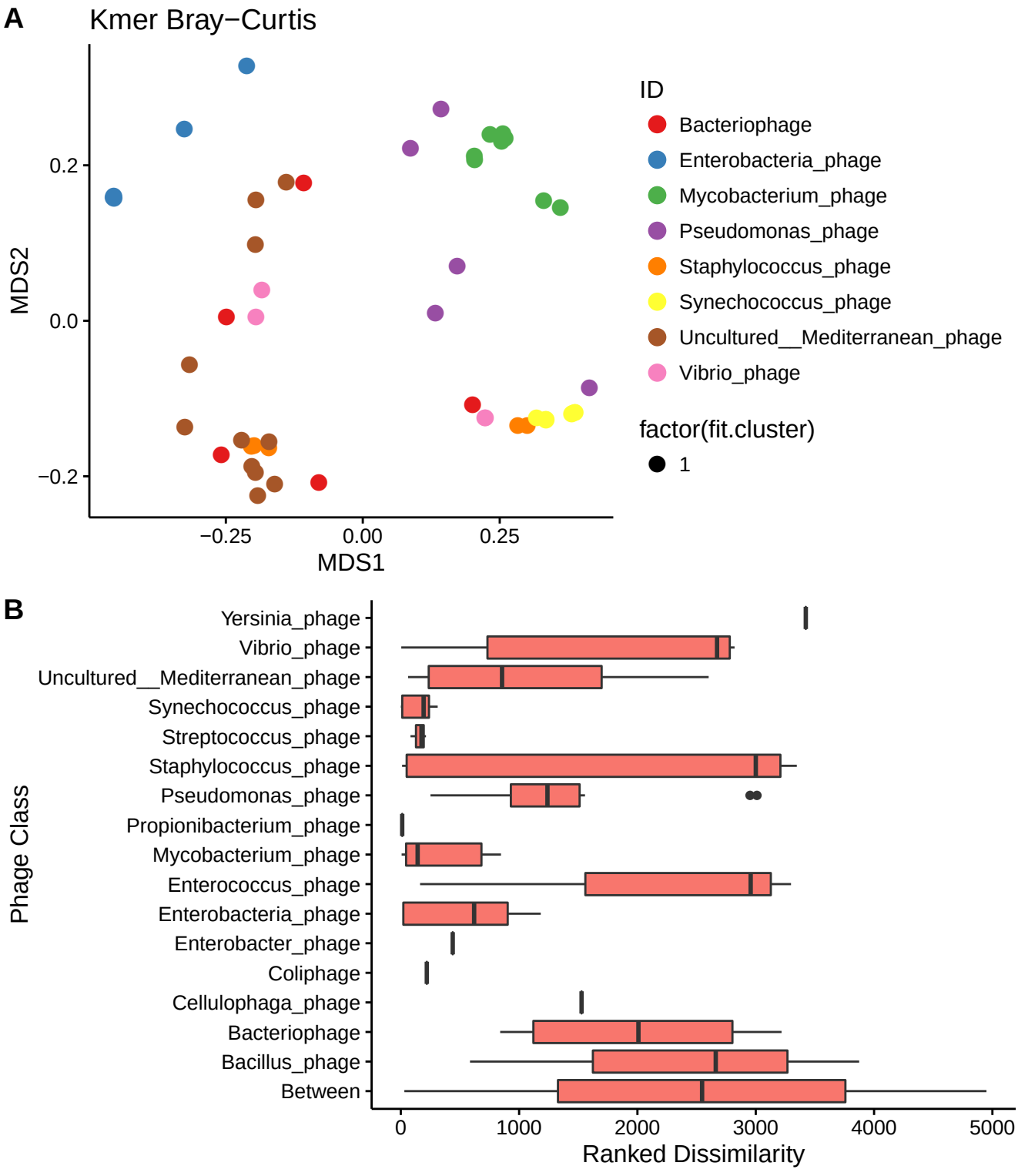


Figure 5: Biological clustering by kmer frequency. A) Comparison of representative phage reference genome kmer spectra. Only genomes with more than three genomes are being shown. Points are colored by their taxonomic host identification. B) Distribution of ANOSIM ranked dissimilarities between samples within each class. Lower value indicates higher degree of kmer conservation.

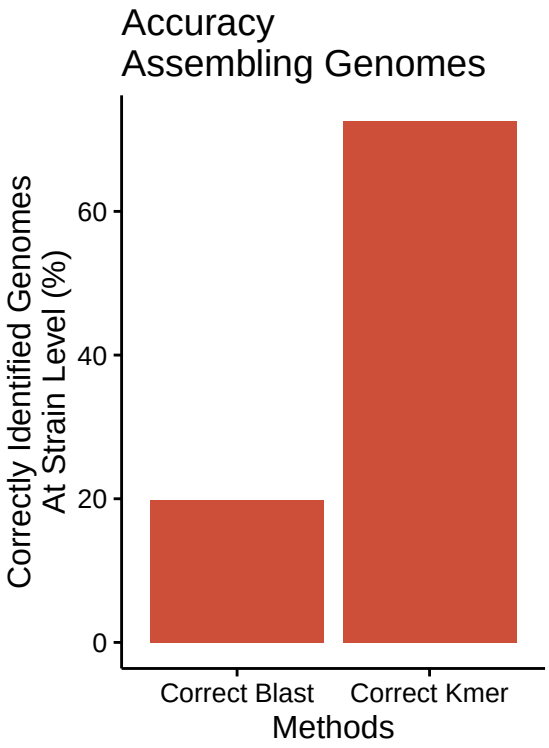


Figure 6: Comparison of kmer spectrum and blast approach to recunstructing reference phage genomes.

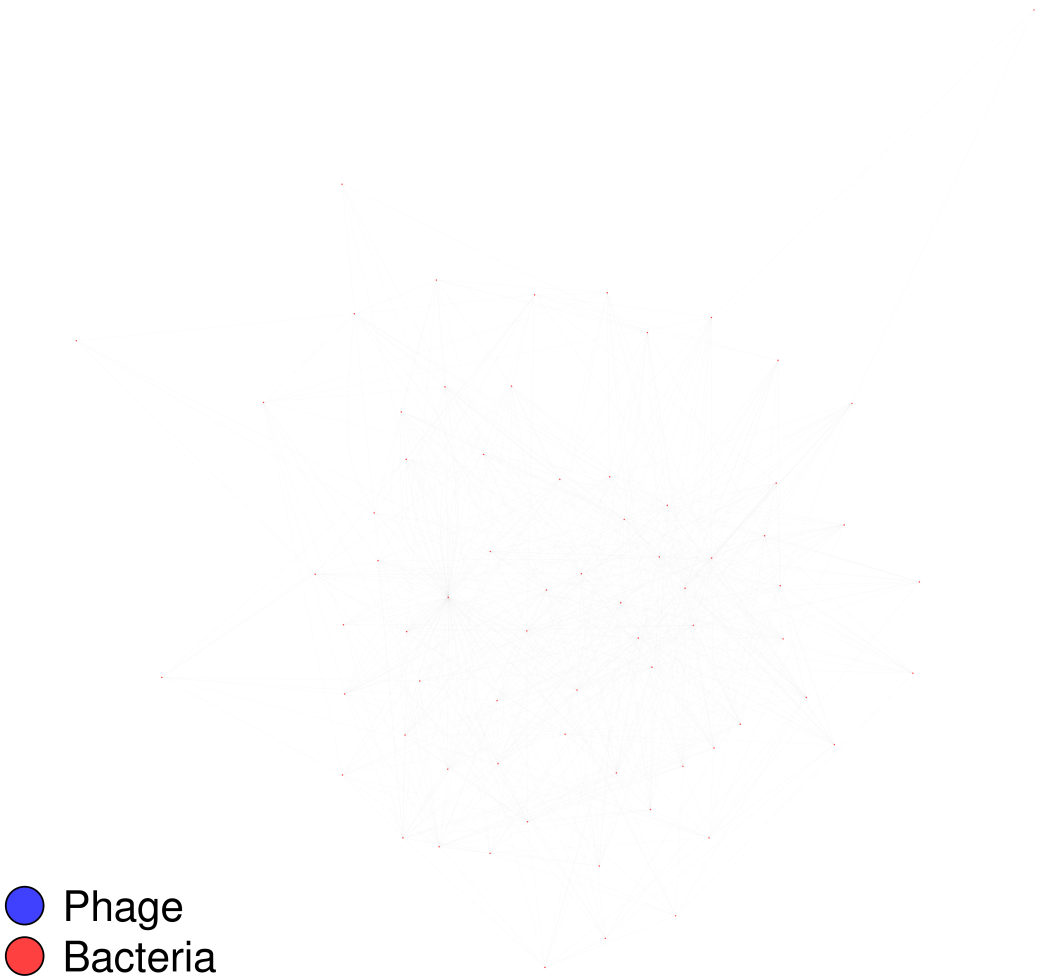


Figure 7: Network diagram of the phage - bacteria network.

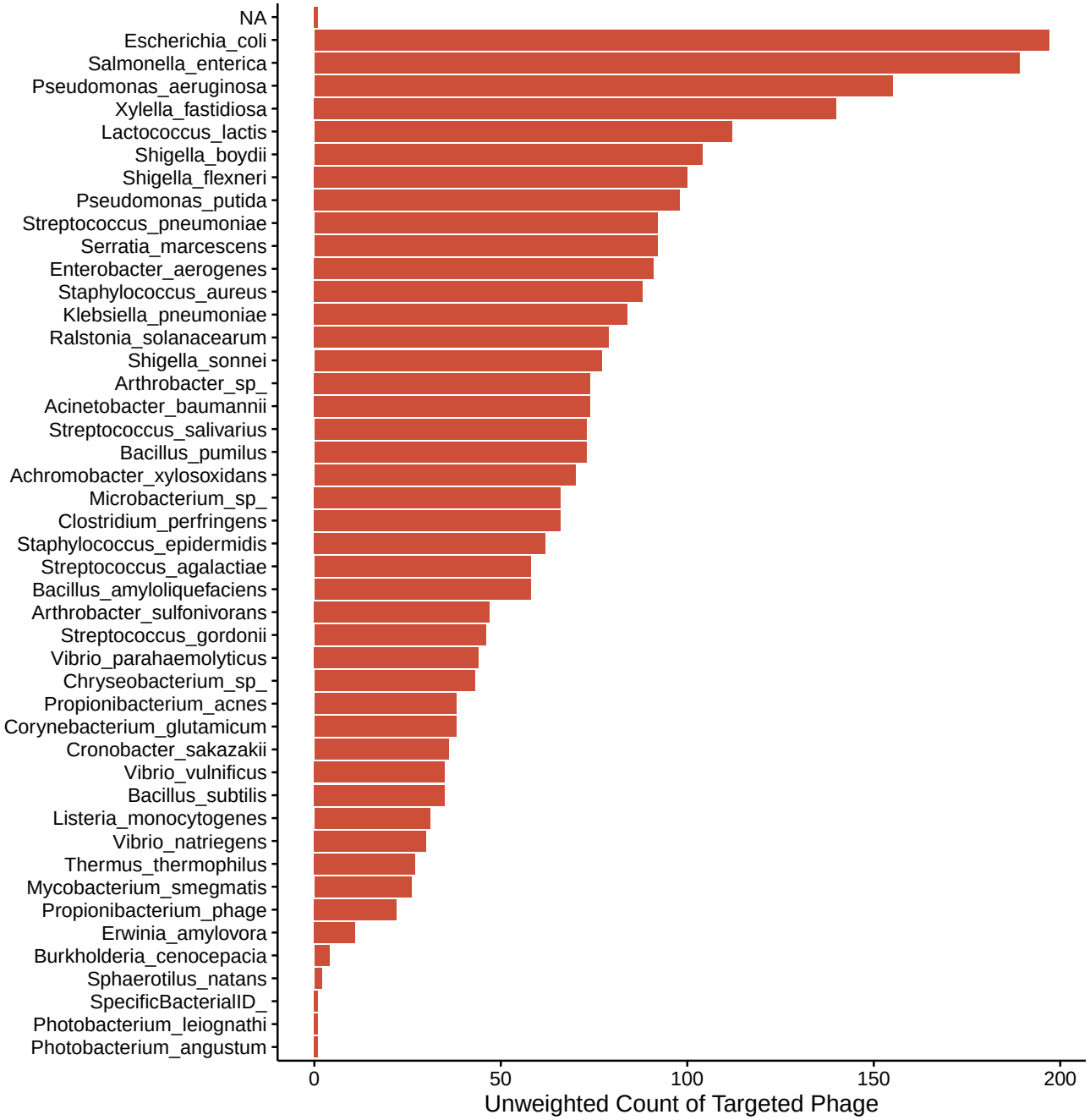


Figure 8: Relative abundance presence of phages by their predicted host target.

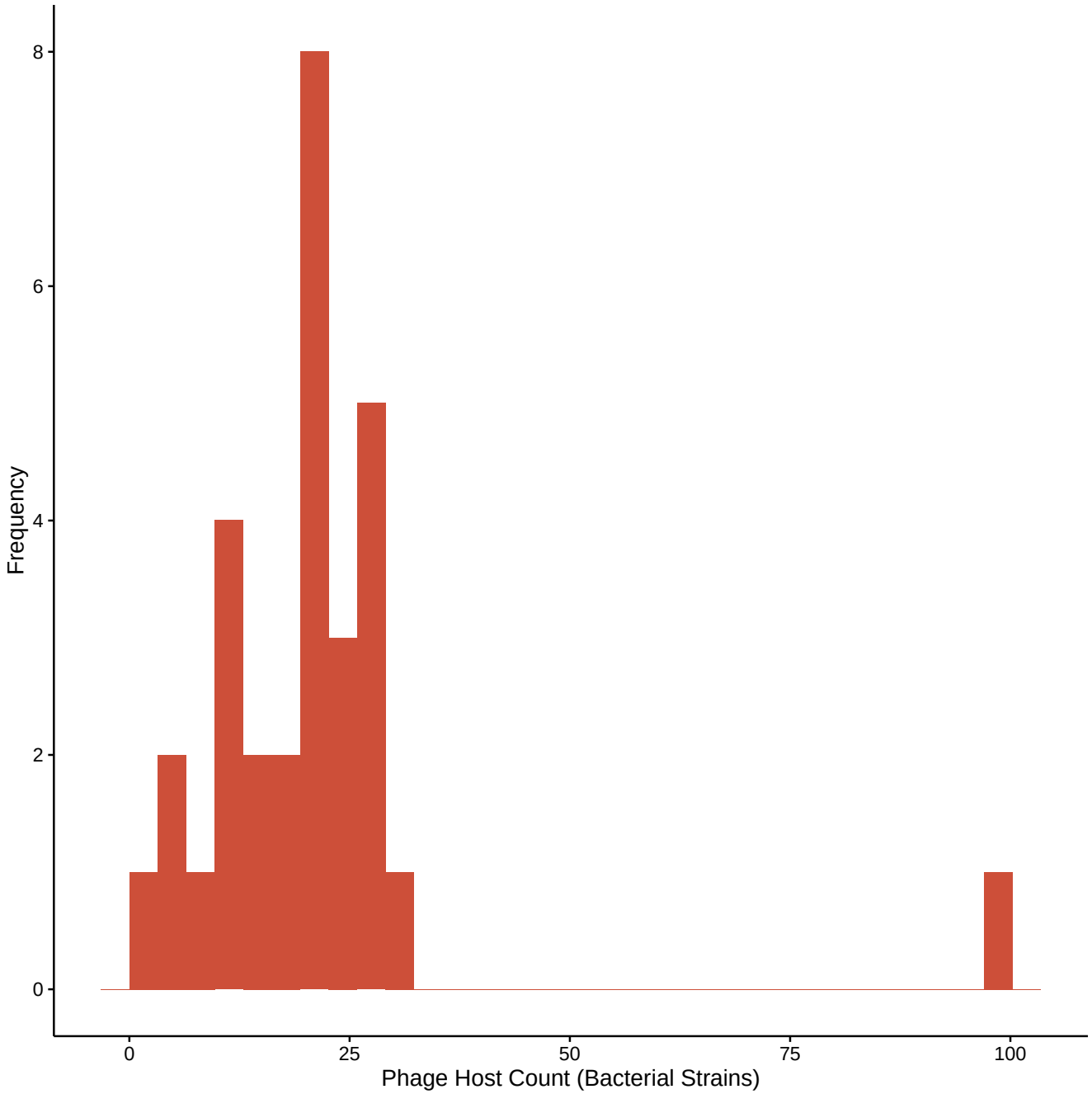


Figure 9: Histogram of the number of bacterial strain hosts identified for each phage contig.

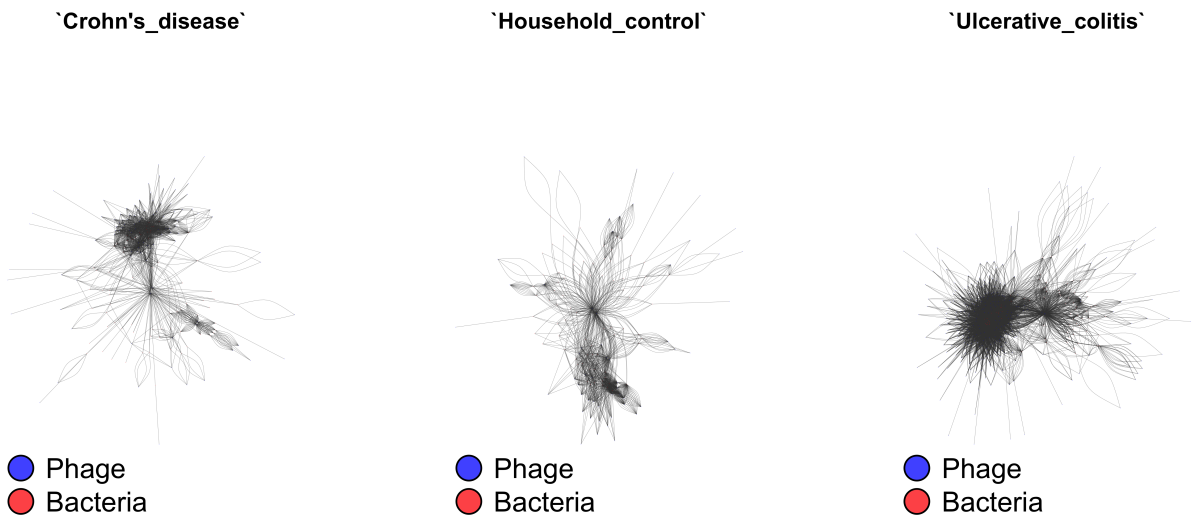


Figure 10: Network visualization for each disease category.

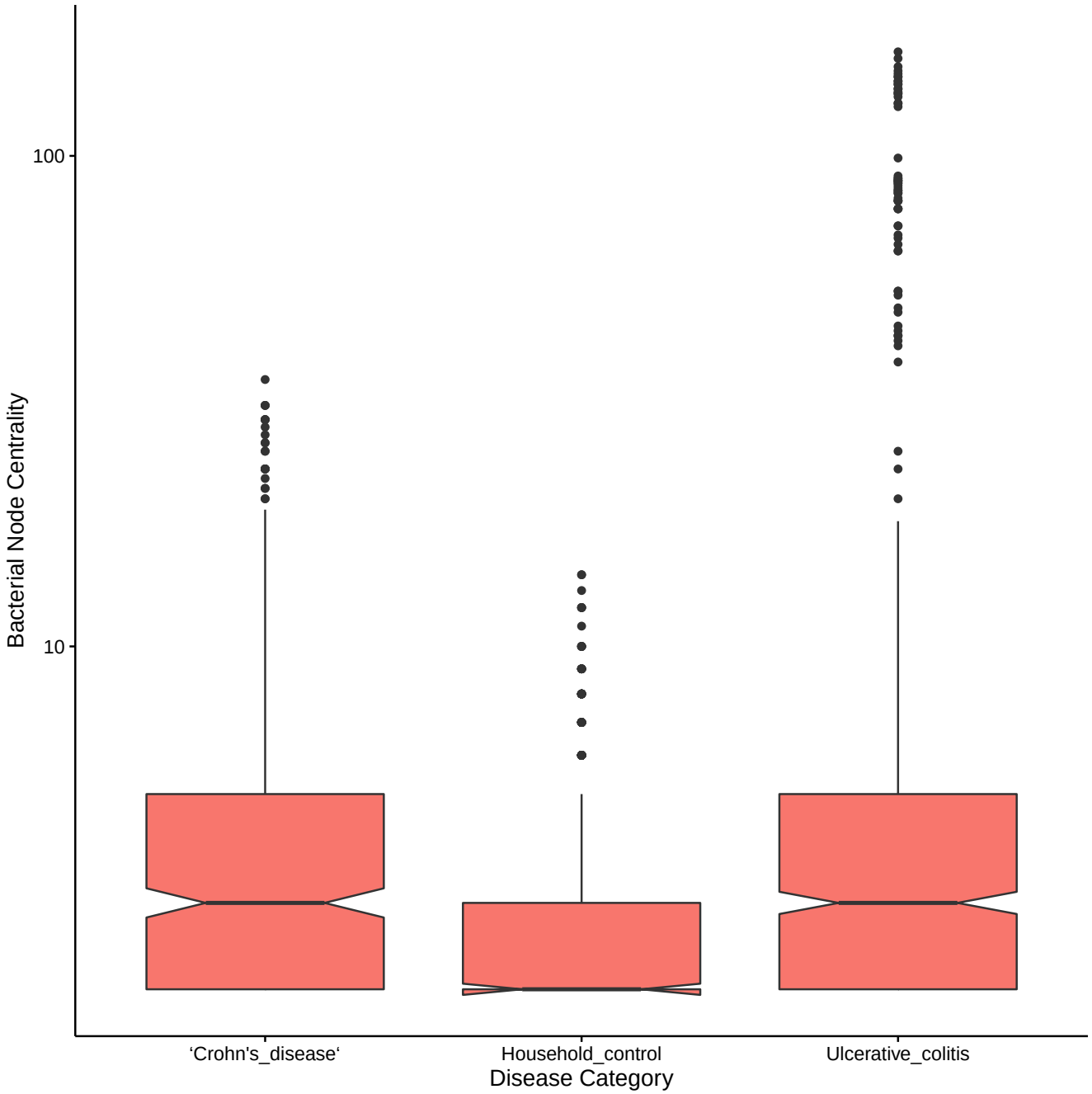


Figure 11: Bacterial centrality across disease states.

# References


1. Poisot, T. & Stouffer, D. How ecological networks evolve. *bioRxiv* (2016).

2. Thompson, R. M. *et al.* Food webs: reconciling the structure and function of biodiversity. *Trends in ecology & evolution* **27,** 689–697 (2012).

3. Poisot, T., Lepennetier, G., Martinez, E., Ramsayer, J. & Hochberg, M. E. Resource availability affects the structure of a natural bacteriabacteriophage community. *Biology letters* **7,** 201–204 (2011).

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

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

6. 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).

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

8. 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).

9. 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).

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

11. 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).

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

13. 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).

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

15. 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).

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

17. 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).

18. 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).

19. 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).

20. 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).