Global Trends & Disease Drivers of the Human Virome

Geoffrey D Hannigan, Patrick D Schloss

Abstract

Introduction

The microbiome is a primary driver in global chemical processes, including environments as diverse as the ocean and the human body. Organisms such as bacteria perform metabolic processes but are controlled in large part by the associated virus community (the virome). Viruses control microbiome metabolism through predation, gene expression control, and horizontal gene transfer. In humans, this translates into virus control of disease and metabolism. Despite being crucial members of the human microbiome, we understand remarkably little about the human virome.

In addition to studying specific virome cohorts, it will be important to investigate the global trends associated with these communities. Such large scale virus studies have proven valuable for understanding the global marine virome, and we are leveraging a similar approach here for the human virome (cite sullivan science and hurwitz pnas). This view from ten thousand feet provides us with a new tier of understanding the ecology of the human virome, and how it relates to human metabolism and disease.

Results

The Global Human Virome Dataset

We leveraged the extensive public sequence archives to assemble a robust human virome dataset that spans diverse body site environments. Our sampling includes the gut, oral cavity, skin, and lungs, all of which were collected by multiple, independent groups. The dataset contains over X number of sequences from Y manuscript archives. See the figure for the project metadata. (Here I want to include a figure that has a map of where the samples were taken both geographically and on a human model, as well as a table with patient metadata like male vs female. This could also be represented in graphical form. I'll try it out and see how it looks.) To effectively navigate this dataset, we relied on a graph database data structure. We used this dataset to assemble A contigs, B of which represented complete circular genomes. (I could also add a representation of contig abundance, coverage, and highlight the complete circular genomes). (Be sure to include information about how many diseases were represented in the dataset, and how many samples were associated with each state).

Operational Protein Family Richness Across the Core and Pan Virome

Because the virome sequence space is dominated by unknown and poorly annotated genomes, we focused on the Operational Protein Families (OPFs) within the assembled contigs of the dataset. Operational protein families are functionally similar to operational taxonomic units, in that they are groups of open reading frames with similar nucleotide sequences.

We began by assessing the current progress in sampling the human virome OPF space. A rarefaction analysis revealed that virome OPFs can be sufficiently covered in a single study, however the average sequencing depth to achieve such coverge is relatively high at approximately 10 million sequences **(Figure 1 A)**. We quantified the rarity of the OPFs by showing that the majority of OPFs only had sequence countes between 10-1000 **(Figure 1 B)**. These results provide important information that informs both the technical and biological aspect of the human virome. Technically, the general sequencing depth for a human virome study should be YYY. At this point the majority of OPFs have been sampled and only very rare OPFs are being detected. Biologically, this lends confidence to our current understanding of the virome because we are able to sample the majority of virome operational protein families.

We investigated the core and pan viromes by calculating the shared distribution of phage OPFs across the samples. As expected, the core virome tapers off rapidly as the sequencing depth increases. Individual body sites harbor different core viromes. The majority of the core genes were housekeeping and structural genes, however we did identify some auxilary metabolic genes throughout the human virome. Auxilary metabolic genes are phage-encoded genes that convey a bacterial or eukaryotic function. These genes represent the tools that phages use to manipulate their bacterial host population functionality.

Genomic Insights Into 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 **(Figure 1)**. *In vitro*, experimentally validated interactive relationships were taken from the existing literature. Clustered Regularly Interspaced 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 protien-protien 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 validated our predictive graph model by quantifying the sensitivity and specificity using a manually curated dataset of experimentally validated positive and negative interactions. Experimental results were extracted from manuscripts published between 1992 and 2015 **(Figure 2)**[1–6]. This allowed us to both evaluate the utility of the model, as well as determine the optimal decision thresholds to use for predictions.

The resulting model had an AUC of 0.605, an optimal sensitivity of 0.89, and an associated optimal specificity of 0.310 **(Figure 3)**. These low false positive and low true positive rate mean that our model is unlikely to identify incorrect phage - bacteria interactions, but is also prone to missing existing relationships. Therefore we are confident in the diverse observed interactions while also understanding that we are only observing some of the ongoing interactions.

Distribution of Pan and Core Phages by Bacteria-Phage Ineraction Networks

Phages are known to transfer genetic content between bacteria in the process of transduction. This has great medical importance when considering transduction of antibiotic resistance genes and other virulence factors. In a dense microbial community, transduction is likely to play an important role in bacterial fitness and virulence. To date, we have a minimal understanding of the interactions phages are facilitating between bacteria. Furthermore, the roles of broadly infecting phages have yet to be considered. Our graph approach allows us to begin predicting and understanding these interactions.

We predicted the phage-mediated relationships between bacteria by executing traidic closures as (bacteria)-[phage]->[bacteria]. Triadic closure theory states that a strong relationship of two entities to a shared intermediate suggests a relationship between the two previously unrelated entities. In our case, we are assigning relationships between bacteria based on shared strong relationships to a phage intermediate.

One of the most powerful aspects of this analysis is that it allows us to evaluate the global interactive properties of the interactive networks across the body and thus provide insight into the complex ecological dynamics. We found that the phage-bacteria interactive network follows a scale-free distribution instead of a random exponential distribution. Not only does this indicate a lack of randomness in the population, it also suggests the hub is composed of hubs that are highly interconnected to the remaining nodes.

The diamter 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.

Ecological Signatures of Global Virome Diversity and Composition

In addition to insights into core and pan virome functionality, OPFs provide an effective approach to evaluating phage diversity (cite Sam's modular function manuscript). We found that body sites are indeed associated with different degrees of diversity.

Disease Drivers of the Core Human Virome

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.

Discussion

Materials & Methods

Figures

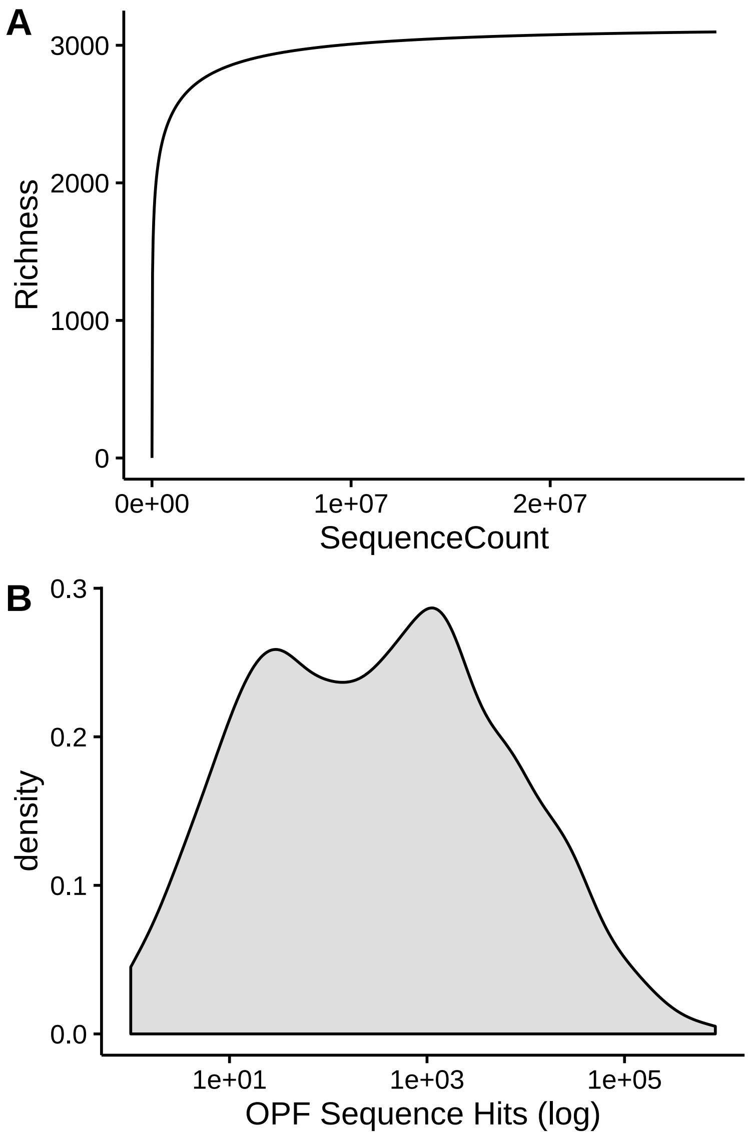

Figure 1: Analysis of sequencing coverage required to sufficiently sample the human virome. A) Rarefaction analysis of the number of OPFs detected (richness) as more sequences are used from the dataset. B) Distribution of the number of sequences that mapped to each OPF.

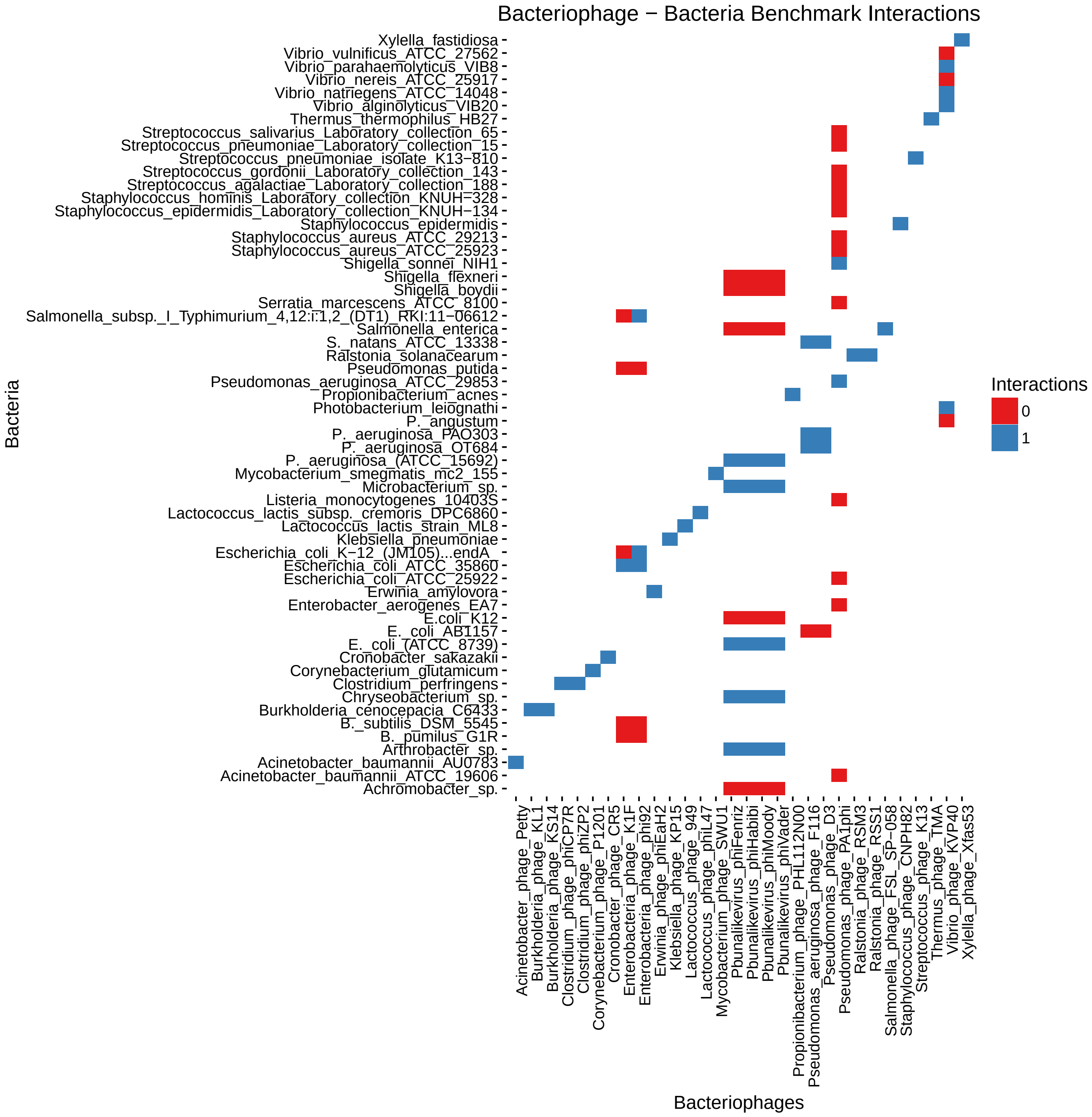

Figure 2: Positive and negative interactions of our reference dataset.

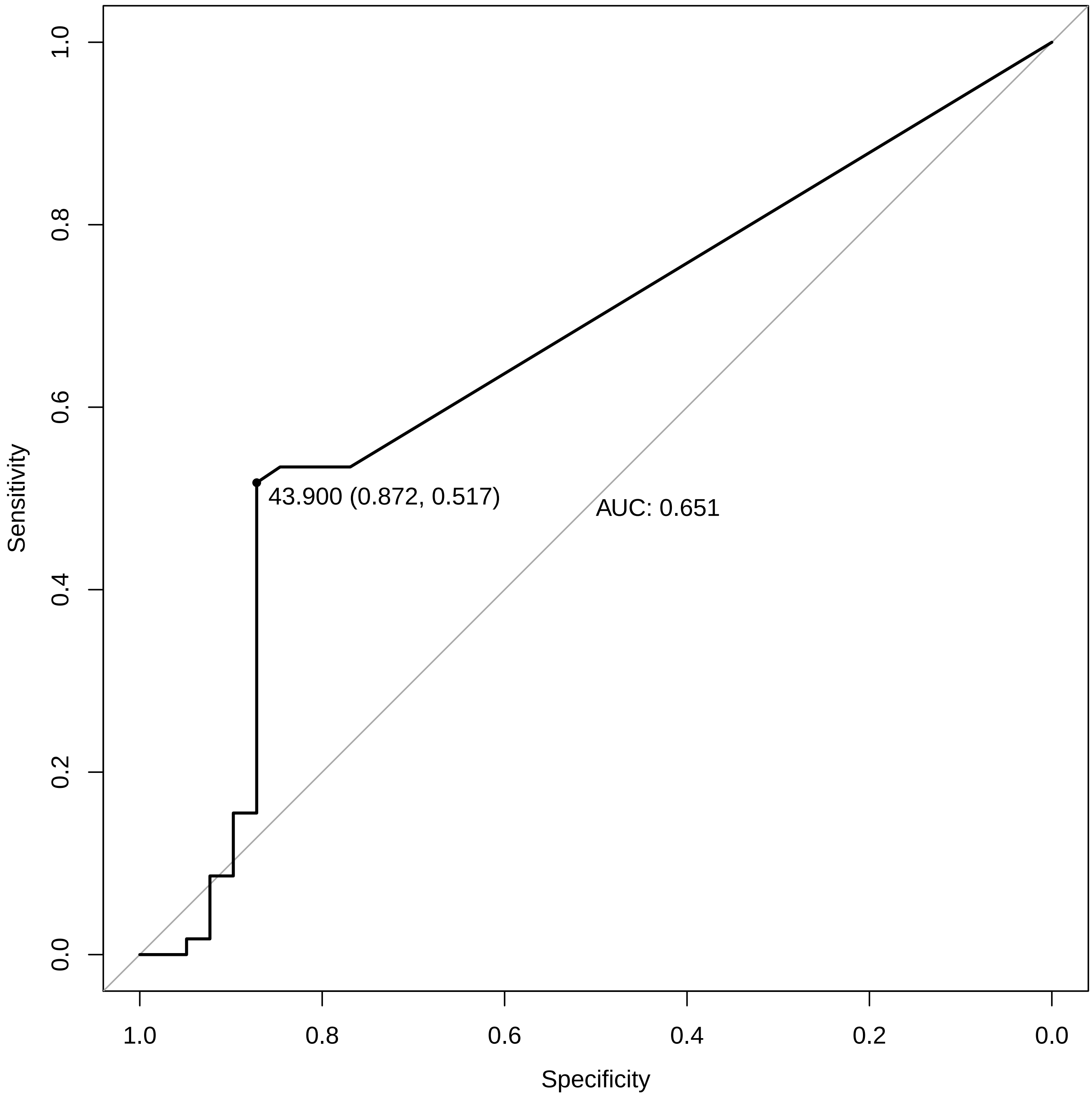

Figure 3: ROC curve used to validate the graph model of phage-bacteria interactions.

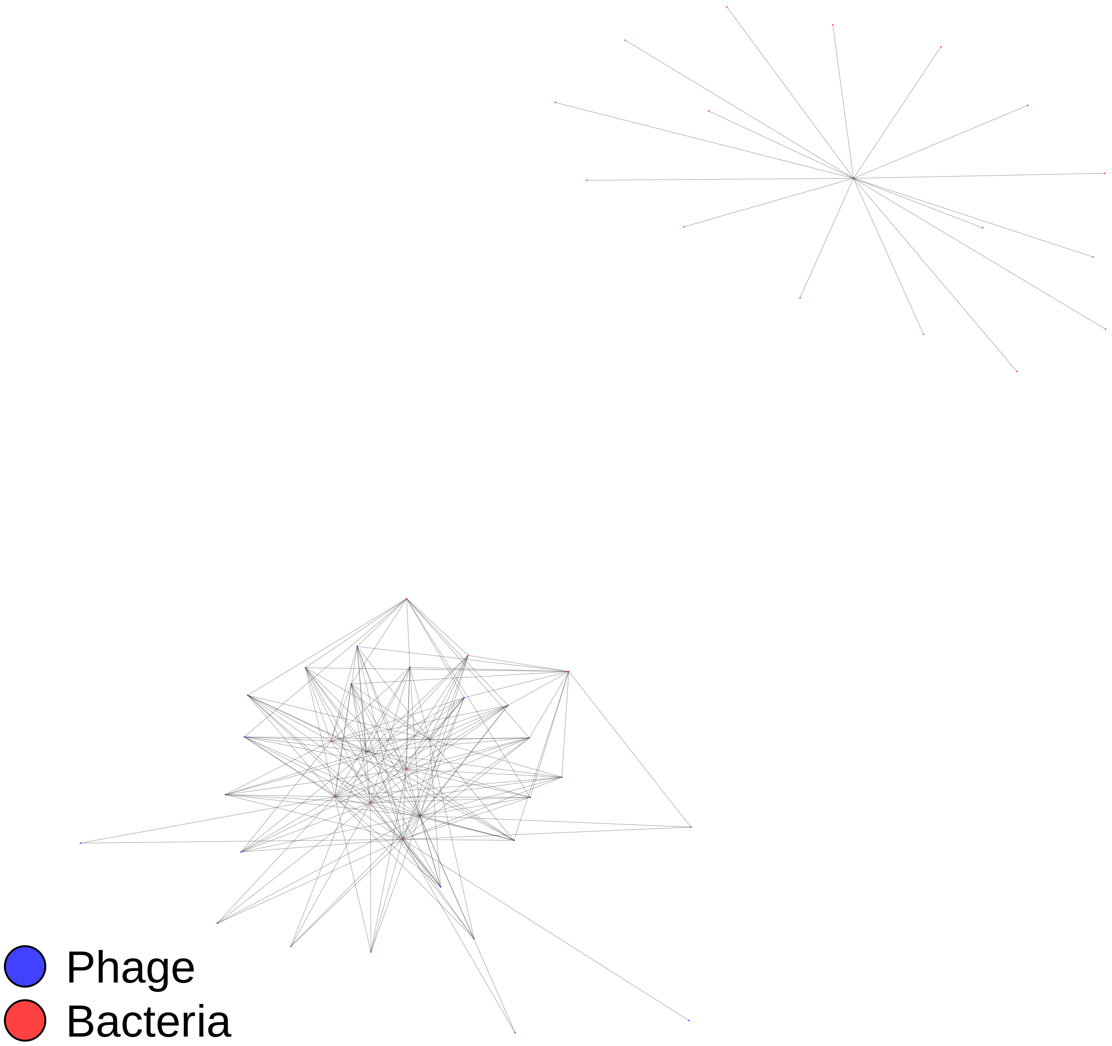

Figure 4: Network diagram of the phage - bacteria network.

Bibliography

1. 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).

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

3. 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).

4. 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).

5. 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).

6. 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).