Biogeography and Environmental Conditions Shape Phage and Bacteria Interaction Networks Across the Healthy Human Microbiome

Geoffrey D Hannigan[1], Melissa B Duhaime[2], Danai Koutra[3], and Patrick D Schloss[1,*]

[1]Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan, 48108

[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48108

[3]Department of Computer Science, University of Michigan, Ann Arbor, Michigan, 48108

[*]To whom correspondence may be addressed.

Corresponding Author Information

Patrick D Schloss, PhD
1150 W Medical Center Dr. 1526 MSRB I
Ann Arbor, Michigan 48109
Phone: (734) 647-5801
Email: pschloss@umich.edu

Running Title: Network Diversity of the Healthy Human Microbiome

Journal: Genome Research (*Preparation Details*)

Keywords: Virome, Microbiome, Graph Theory, Machine Learning

Text Length: 35,640 / 50,000 Characters

* *Figures at the end of the document for internal review only.*

Abstract

Viruses and bacteria are critical components to the human microbiome and play important roles in health and disease. Previous work has relied on studying bacteria and phages in isolation (e.g. taxonomic classification, alpha diversity, and beta diversity), reducing them to two separate communities. This approach cannot capture how these communities interact to share information (e.g. horizontal gene transfer) and maintain stability. We developed and implemented a network-based analytical approach to provide an initial understanding of phage-bacteria network diversity across the human body and over time. We built a machine learning algorithm to predict which phages will infect which bacteria in each microbiome. This model was applied to paired viral and bacterial metagenomic sequence sets from three previously published human cohorts. We organized the predicted interactions into networks that allowed us to evaluate the diversity of phage-bacteria connectedness and network structure across the human body. We found that gut and skin network structure was person-specific, and was only weakly conserved among cohabitating family members. High fat diets and obesity were associated with less connected networks. There were significant differences in network structure between skin sites, with those more exposed to the external environment being less connected and more prone to instability. This study characterizes the diversity of microbiome networks across the human body, and provides a baseline for future studies to investigate the role of ecological networks in disease states.

Word Count: 231 / 250

Introduction

Viruses and bacteria are critical components to the human microbiome and play important roles in health and disease. Bacterial communities have been associated with diseases including a wide range of skin conditions (Hannigan and Grice 2013), acute and chronic wound healing conditions (Hannigan et al. 2014; Loesche et al. 2016), and gastrointestinal diseases including inflammatory bowel disease (He et al. 2016; Norman et al. 2015), *Clostridium difficile* infections (Seekatz et al. 2016), and colorectal cancer (Zackular et al. 2014; Baxter et al. 2014). Altered viromes (virus communities consisting primarily of bacteriophages in humans) have also been associated with various diseases and environmental perturbations including inflammatory bowel disease (Norman et al. 2015; Manrique et al. 2016), periodontal disease (Ly et al. 2014), and others (Monaco et al. 2016; Hannigan et al. 2015; Minot et al. 2011; Santiago-Rodriguez et al. 2015; Abeles et al. 2015, 2014). Individual phages within these communities are capable of lysing their bacterial hosts and modulating their functionality through horizontal gene transfer (i.e. transduction) and altering bacterial host gene expression. At the community level, the human virome has begun to be implicated in promoting antibiotic resistance throughout human-associated microbial communities (Modi et al. 2013; Hannigan et al. 2015). These virus community shifts are not reflections of the bacterial communities, but rather act in concert with bacteria as one single overall community (Norman et al. 2015; Haerter et al. 2014).

Previous work has relied on studying bacteria and phages in isolation (e.g. taxonomic classification, alpha diversity, and beta diversity), reducing them to two separate communities. In reality, bacteria and phage communities are dynamic and complex, acting in concert to share genetic information (e.g. horizontal gene transfer) and to maintain stable ecosystems, and removal of members can disrupt or even collapse those ecosystems (Haerter et al. 2014; Harcombe and Bull 2005; Middelboe et al. 2001; Poisot et al. 2011, 2012; Thompson et al. 2012; Moebus and Nattkemper 1981; Flores et al. 2013, 2011; Poisot and Stouffer 2016; Jover et al. 2015). Relationship-based network approaches allow us to capture this information, including community resilience to instability and the potential for information flow. The conventional approach to understanding the human microbiome is unable to capture how these communities interact and influence each other. We bridge this knowledge gap by characterizing the human bacterial and phage communities by their *relationships*, leveraging machine learning and graph theory techniques. We focus on characterizing the network diversity of the healthy human microbiome (including the skin and gut), so as to provide a foundation on which we and others can build further studies into the network dynamics associated with disease states.

To begin characterizing the networks of bacteria and phages within the human microbiome, we leveraged three published microbiome datasets with paired virus and bacterial metagenomic sequence sets (Hannigan et al. 2015; Minot et al. 2011; Reyes et al. 2010; Turnbaugh et al. 2009a). These datasets included gut and skin samples, allowing us to gain broader insights into human microbiome network diversity across different body sites. Our approach built off of previous large-scale phage-bacteria microbiome network analyses by inferring interactions using metagenomic datasets, instead of using culture-based techniques (Moebus and Nattkemper 1981; Flores et al. 2013). Our metagenomic interaction inference model is powered beyond previous models by its inclusion of protein interaction data, inclusion of negative interactions as well as positive, and the use of a more sophisticated machine learning algorithm instead of the linear models used by some groups (Edwards et al. 2015).

Using this approach, we were able to go beyond contemporary isolated methods to uncover a basic understanding of the community network dynamics associated with healthy human phage and bacteria. By building and utilizing a microbiome network, we showed that different people, body sites, and anatomical locations not only support distinct microbiome membership and diversity (Hannigan et al. 2015; Minot et al. 2011; Reyes et al. 2010; Turnbaugh et al. 2009a; Grice et al. 2009a; Findley et al. 2013; Costello et al. 2009, Consortium (2012)), but also support communities with distinct communication structures and propensities toward community instability. By understanding the healthy state of network structures across the human body, we empower future studies to begin investigating how these community structures change in disease or otherwise altered states.

Results

Establishing a Model of Phage-Bacteria Infectious Networks

We studied the differences in microbiome phage-bacteria interaction networks across healthy human bodies by leveraging previously published sequence sets containing purified virome samples paired with bacterial metagenomes from whole metagenomic shotgun sequences. Our study contained three datasets, including a study of the impact of diet on the healthy human gut virome (Minot et al. 2011), the impact of anatomical location on the healthy human skin virome (Hannigan et al. 2015), and the viromes of monozygotic twins and their mothers (Reyes et al. 2010; Turnbaugh et al. 2009a). The viromes associated with these datasets were subjected to virus-like particle (VLP) purification to eliminate other organism DNA including bacteria, fungi, and humans, and allowed us to assess the actively replicating virome.

Bacterial and viral sequences were quality filtered and assembled into contigs. The published datasets we used were conducted over the span of five years using different methods and technologies, and therefore yielded different sequence abundances **(Supplemental Figure S1 A-B)**. Because contig assembly most commonly returns genome fragments, we clustered related bacteria and phage contigs by k-mer frequency and co-abundance using the CONCOCT algorithm. These clusters represent operationally defined units of related bacteria and phage genomes that we defined as operational genomic units (OGUs). These OGUs are conceptually similar to the operational taxonomic unit (OTU) and operational protein family (OPF) definitions used for grouping highly similar 16S rRNA gene and open reading frame sequences, respectively (Schloss and Handelsman 2008). The resulting contigs and OGUs demonstrated high sequence coverage and length **(Supplemental Figure S2 - S3)**.

We predicted which phage OGUs infected which bacteria using a random forest model trained on experimentally validated infectious relationships from six previous publications (Jensen et al. 1998; Malki et al. 2015; Schwarzer et al. 2012; Kim et al. 2012; Matsuzaki et al. 1992; Edwards et al. 2015). This training set contained diverse bacteria and phages, with both broad and specific infectious ranges **(Supplemental Figure S4 A - B)**. Phages with linear and circular genomes, as well as ssDNA and dsDNA genomes, were included in the analysis. Because this was a DNA sequencing study, RNA phages were considered in the analysis **(Supplemental Figure S4 C-D)**. This training set included both positive relationships (a phage infects a bacterium) and negative relationships (a phage does not infect a bacterium). This built on previous work, which focused only on positive relationships, by allowing us to validate the false positive *and* false negative rates associated with our candidate models.

Four phage and bacterial genomic markers were used to predict infectious relationships between bacteria and phages: 1) genome nucleotide similarities, 2) gene amino acid sequence similarities, 3) CRISPR targeting of phages by bacterial CRISPR spacer sequences, and 4) similarity of protein families known to be associated with known protein-protein interactions (Orchard et al. 2014). The resulting random forest model exhibited high performance with an AUC of 0.846, a sensitivity of 0.829, and a specificity of 0.767 **(Figure 1 A)**. The most important predictor in the model was nucleotide similarity between genes, followed by nucleotide similarity of whole genomes **(Figure 1 B)**. Protein family interactions were moderately important to the model, while CRISPRs were minimally important, likely because they were redundant with the blast information. Approximately one third of the relationships yielded no score and were automatically classified as being non-infectious **(Figure 1 C)**.

We used our random forest model to classify the relationships between bacteria and phages in the experimental datasets. The relationships within the three studies were used to construct one master network, containing the three study sub-networks, which themselves each contain sub-networks for each sample **(Figure 1 D)**.

Metadata including study, sample ID, disease, and abundance within the community (based on sequence count) were also stored in the multi-study master network to allow for effective parsing and downstream analysis **(Supplemental Figure S5)**. The resulting master network was highly connected and contained 72,287 infectious relationships among 578 nodes, 298 of which represented phages and 280 that represented bacteria. Although the network was highly connected, not all relationships were present in all samples. Furthermore, relationships were weighted by relative abundance of their associated bacteria and phage, meaning that lowly abundant relationships could be present but insignificant compared to those that were more highly abundant. Like the master network, the skin network exhibited a diameter of 4 (measure of graph size; the greatest number of traversed vertices required between two vertices) and included almost all of the master network nodes and edges **(Figure 1 E - F)**. The gut diet and twin sample sets each contained less than 150 vertices, less than 20,000 relationships, and diameters of 3, suggesting more sparsely related phages and bacteria **(Figure 1 E - F)**.

Individuality of Microbial Networks

Previous work has show community membership and diversity are highly personal, with signatures remaining more similar to themselves over time compared to other people (Grice et al. 2009b; Hannigan et al. 2015; Minot et al. 2013). Using our microbiome network model, we determined whether this personal conservation extended to network structure. We calculated the degree of dissimilarity between each subject's network, based on phage and bacteria abundance, as well as centrality. Centrality was evaluated by first calculating the weighted eigenvector centrality of all bacteria and phages within each sample graph. Conceptually, this metric defines central phages as those that are highly abundant and infect many bacteria which themselves are abundant and infected by many other phages. Bacterial centrality was defined in the same way. We then calculated the similarity of community networks using the overall weighted eigenvector centrality of all nodes between all samples. Biologically, samples with similar network structures are interpreted as having similar capacities to influence other microbes within the community, transmit genetic material (e.g. horizontal gene transfer), and maintain stability.

We found that gut microbiome network structure was highly individual specific, with networks clustering significantly by person (anosim p-value = 0.01, anosim statistic R = 0.979, **Figure 2 A)**. Network dissimilarity within people over the 8-10 day sampling period was significantly less than the average dissimilarity between that person and others **(Figure 2 B)**. We evaluated the significance of an individual's network being more similar to itself than others by fitting a probability distribution to the reductions in intrapersonal dissimilarity compared to interpersonal dissimilarity (Lines between points in Figure 2 B - D). By taking the integral of the probability distribution from negative infinity to zero, we were able to calculate the probability that an individual's network structure was more similar to itself than others, in a given population. Thus a probability of 1 would mean that all members in a given population will be expected to be more similar to themselves compared to others, and 0.5 means they are no more likely to be more similar to themselves than dissimilar. This is a Bayesian approach that tells us how prevalent we expect this trend to be. The probability of intrapersonal diversity of an individual being less than interpersonal diversity in our dataset was 0.972 (absolute error = 3.59 × 10-5, **Figure 2 B)**.

Skin microbiome network structure was also conserved within individuals, although not as strongly as the gut. The probability of a given skin site being more similar to itself after a month compared to a different person at the same time was 0.888 (absolute error = 7.39 × 10-5, **Figure 2 C)**. This distribution was similar when separated by anatomical sites, suggesting this was an accurate representation of overall skin network dissimilarity **(Supplemental Figure S6)**.

The conservation of gut network structures was only weakly associated with families. The gut network structures were more similar between twins and their mothers (intrafamily) compared to other families of twins and mothers (inter-family) **(Figure 2 D)**. The probability of a family member being more similar to another family member compared to non-family members was 0.724 (absolute error = 7.27 × 10-5).

Role of Diet & Obesity in Gut Microbiome Conectivity

Diet is a major environmental factor that influences resource availability and gut microbiome composition and diversity, including bacteria and phages (Minot et al. 2011; Turnbaugh et al. 2009b; David et al. 2014). Previous work in isolated culture-based systems has suggested that changes in nutrient availability are associated with altered phage - bacteria network structures, although this has yet to be applied to humans (Poisot et al. 2011). Understanding the impact diet might have on gut microbiome networks is a valuable contribution to the baseline understanding we aim to achieve here. We therefore hypothesized that network structure would be altered by changes in diet. We also described a potential association between gut network structure and obesity, a disease linked to diet and potentially the microbiome (Sze and Schloss 2016).

We evaluated the differences in gut network structure by quantifying how central each sample's network was on average across all microbes. We accomplished this by utilizing two common centrality metrics: degree centrality and closeness centrality. **Degree centrality**, the simplest centrality metric, was defined as the number of connections each phage made to bacteria, or each bacterium made to phages. Because this metric alone offers only minimal insight, we supplemented it with measurements of closeness centrality. **Closeness centrality** is a metric of how close each phage or bacterium is to all of the other phages and bacteria in the network. A higher closeness centrality suggests that genetic information or the effects of altered abundance would be more impactful to all other microbes in the system. Networks with higher closeness centrality also indicates an overall greater degree of connections, which indicates a greater resilience to instability. Because these values are assigned to each phage and bacterium within each network, we calculated the average connectedness and corrected for the maximum potential degree of connectedness to obtain a single value for the connectedness of each graph.

We found that gut microbiome network structures associated with high fat diets were less connected than those of low fat diets **(Figure 3 A-B)**. Tests for statistical differences were not performed, so as to prevent misleading interpretations from the small sample size. High fat diets exhibited less degree centrality **(Figure 3 A)**, meaning bacteria were overall targeted by less phages and phage tropism was more specific. High fat diets also exhibited decreased closeness centrality **(Figure 3 B)**, meaning the microbes were more distant from other microbes in the community, making genetic information transfer and the impact of altered abundance less capable of impacting other bacteria and phages within the network.

In addition to diet, there was also an association between obesity and network structure **(Figure 3 C-D)**. The obesity-associated network demonstrated a higher degree centrality **(Figure 3 C)** but a less closeness centrality, compared to the healthy controls **(Figure 3 D)**. This meant that the obesity network was overall less connected, having microbes further from all other microbes within the community.

Variation of Network Structure Across the Human Skin Landscape

Extensive work has illustrated differences in healthy human skin microbiome between anatomical sites, including bacteria, viruses, and fungi (Grice et al. 2009b; Findley et al. 2013; Hannigan et al. 2015). These communities vary by degree of skin moisture, oil, and environmental exposure. We hypothesized that like microbial composition and diversity, microbial network structure differs between anatomical sites. We addressed this hypothesis by evaluating the changes in network structure between anatomical sites within our skin dataset.

We quantified the average centrality of each sample using the weighted eigenvector centrality metric, which yields a larger value for phages and bacteria that are more abundant and have more relationships to other bacteria and phages that themselves have more relationships. We found that intermittently moist skin sites (dynamic sites that fluctuate between being moist and dry) were significantly less connected than the more stable moist and sebaceous environments (p-value < 0.001, **Figure 4 A)**. We also found that skin sites that were protected from the environment (occluded) were much more highly connected than those that were constantly exposed to the environment or only intermittently occluded (p-value < 0.001, **Figure 4 B)**.

We supplemented this analysis by comparing the network signatures using the centrality dissimilarity approach described above. The dissimilarity between samples was a function of shared relationships, degree of centrality, and bacteria/phage abundance. When using this supplementary approach, we found that network structures significantly clustered by moisture, sebaceous, and intermittently moist status **(Figure 4 C,E)**. We also found that occluded sites were significantly different from exposed and intermittently occluded sites, but there was no difference between exposed and intermittently occluded sites **(Figure 4 D,F)**.

Discussion

We developed and implemented a network-based analytical approach that allowed us to begin evaluating the basic properties of the human microbiome through bacteria and phage relationships, instead of membership or diversity of separate communities. The goal of this study was to provide an initial understanding of how phage-bacteria networks differ throughout the human body, so as to provide a baseline for future studies of how microbiome networks differ in disease states. This goal is similar to the initial studies of bacteria and viral diversity across the human body, as well as other environments (Grice et al. 2009a; Findley et al. 2013; Hannigan et al. 2015; Costello et al. 2009, Consortium (2012); Schloss and Handelsman 2005; Minot et al. 2011). Our approach was advantageous over previous microbiome network analyses because it did not rely on inappropriate linear correlations of abundance, it was trained on both positive and negative relationships, and it involved utilization of network theory concepts.

The principles of network theory offer extensive analytical opportunities which can be applied to understand complex ecological communities. In this study we focused on utilizing metrics of connectivity to understand the extent to which communities of bacteria and phages interact (e.g. horizontal gene transfer, modulated bacterial gene expression, alterations in abundance), and how resilient they are to instability. Like earlier work laid a foundational understanding of the healthy human microbiome using membership and diversity measurements (Grice et al. 2009a; Findley et al. 2013; Hannigan et al. 2015; Costello et al. 2009, Consortium (2012)), we provide a similar understanding of healthy human microbiome network structures. These properties included variations between individuals and families, different diet types, and different anatomical skin sites.

One major overarching principle we observed was that, just as gut microbiome and virome composition and diversity are conserved in individuals (Hannigan et al. 2015; Grice et al. 2009a; Findley et al. 2013; Minot et al. 2013), gut and skin microbiome network structures were also conserved within individuals over time, with the gut being more strongly similar than the skin. Network structure was weakly conserved among family members. One explanation for the weaker intrapersonal network conservation of skin compared to the gut communities is that the skin is in more direct contact with the external environment. This was further supported by the observation that network structure differed between environmentally exposed and occluded skin sites. The exposed and intermittently exposed sites were less connected and therefore are expected to have a higher propensity for instability. Likewise, intermittently moist sites demonstrated less connectedness than the more stable moist and sebaceous sites. This suggests that body sites under greater degrees of fluctuation harbored less connected, potentially less stable microbiomes.

We also observed a link between diet and microbiome network structure. Although our sample size was small, we found that high fat diets appeared to be less connected than low fat diets. This suggested that high fat diets may lead to less stable communities with a decreased ability to influence each other. We supported this finding with the observation that obesity is associated with decreased network connectivity. Further work will be required to further characterize these relationships beyond our small study, but our results allow us to speculate that the food we eat may not only impact what microbes colonize our guts, but may also impact their ability to communicate and maintain stability.

Together this work represents an initial step toward understanding the microbiome through its relationships. While these findings are informative, there are certainly caveats that should be noted. *First*, while our infection classification model is advantageous over existing models, we recognize that, like most classification models, there remains opportunity for improvement. For example, such a model is only as good as its training set, and future large-scale endeavors into infectious relationships (and the associated genomes) will provide more robust training and higher model accuracy. Just as we have improved on previous modeling efforts, we expect that new and creative scoring metrics will likely be integrated into this model to further improve model performance.

Second, while informative, this work was done retrospectively and relied on published research from as many as seven years ago. These archived datasets were limited by the technology and costs of the time, meaning the datasets are poorly powered for the statistical analysis we strive for today. While we were able to present initial observations, follow-up studies will be required to validate our observations.

Despite their limitations, these results will be important for informing design and interpretation of future studies of the human microbiome. By showing that microbiome relationship structure, and therefore the community potential for dissemination of information and influence, as well as the potential for maintaining stability, differs significantly between body sites and environmental conditions we demonstrate the diversity of relationships across the human body. We found that other environmental factors, such as diet, also influence relationship structures. This information builds on previous work by suggesting that bacteria and phages not only preferentially colonize different body sites, but also interact differently, allowing for different communication structures and capacities for maintaining stability. This work will also provide a foundation for future studies to begin evaluating how microbiome network dynamics change in disease states, and how the information can be leveraged therapeutically.

Materials & Methods

Data Availability

All associated source code is available on GitHub at the following repository:

https://github.com/SchlossLab/Hannigan-2016-ConjunctisViribus.

Data Acquisition & Quality Control

Raw sequencing data and associated metadata was acquired from the NCBI sequence read archive (SRA). Supplementary metadata was acquired from the same SRA repositories and their associated manuscripts. The gut virome diet study (SRA: `SRP002424`), twin virome studies (SRA: `SRP002523`; `SRP000319`), and skin virome study (SRA: `SRP049645`) were downloaded as `.sra` files. Sequencing files were converted to `fastq` format using the `fastq-dump` tool of the NCBI SRA Toolkit (v2.2.0). Sequences were quality trimmed using the Fastx toolkit (v0.0.14) to exclude bases with quality scores below 33 and shorter than 75 bp (Hannon). Paired end reads were filtered to exclude and sequences missing their corresponding pair using the `get_trimmed_pairs.py` available in the source code.

Contig Assembly

Contigs were assembled using the Megahit assembly program (v1.0.6) (Li et al. 2016). A minimum contig length of 1 kb was used. Iterative k-mer stepping began at a minimum length of 21 and progressed by 20 until 101. All other default parameters were used.

Contig Abundance Calculations

Contigs were concatenated into two master files prior to alignment, one for bacterial contigs and one for phage contigs. Sample sequences were aligned to phage or bacteria contigs using the Bowtie2 global aligner (v2.2.1) (Langmead and Salzberg 2012). We defined a mismatch threshold of 1 bp and seed length of 25 bp. Sequence abundance was calculated from the Bowtie2 output using the `calculate_abundance_from_sam.pl` script available in the source code.

Operational Genomic Unit Binning

Contigs often represent large fragments of genomes. In order to reduce redundancy, and the resulting artificially inflated genomic richness within our dataset, it was important to bin contigs into operational units based on their similarity. This approach is conceptually similar to the clustering of related 16S rRNA sequences into operational taxonomic units (OTUs), although here we are clustering contigs into operational genomic units (OGUs) (Schloss and Handelsman 2005).

We clustered contigs using the CONCOCT algorithm (v0.4.0) (Alneberg et al. 2014). Because of our large dataset and limits in computational efficiency, we randomly subsampled the dataset to include 25% of all samples, and used these to inform contig abundance within the CONCOCT algorithm. CONCOCT was used with a maximum of 500 clusters, a k-mer length of four, a length threshold of 1 kb, 25 iterations, and exclusion of the total coverage variable.

OGU abundance (`Ao`) was obtained as the sum of the abundance of each contig (`Aj`) associated with that OGU. The abundance values were length corrected such that:

$$A_O = \frac{10^7 \sum_{j=1}^{k} A_j}{\sum_{j=1}^{k} L_j}$$

Where `L` is the length of each contig `j` within the OGU.

Open Reading Frame Prediction

Open reading frames (ORFs) were identified using the Prodigal program (V2.6.2) with the meta mode parameter and default settings (Hyatt et al. 2012).

Classification Model Creation and Validation

The classification model for predicting interactions was built using experimentally validated bacteria-phage infections or validated lack of infections from six studies (Jensen et al. 1998; Malki et al. 2015; Schwarzer et al. 2012; Kim et al. 2012; Matsuzaki et al. 1992; Edwards et al. 2015). Associated reference genomes were downloaded from the European Bioinformatics Institute (see details in source code). The model was created based on the four metrics listed below.

The four scores were used as parameters in a random forest model to classify bacteria and bacteriophage pairs as either having infectious interactions or not. The classification model was built using the Caret R package (v6.0.73) (Kuhn). The model was trained using five-fold cross validation with ten repeats. Pairs without scores were classified as not interacting. The model was optimized using the ROC value. The resulting model performance was plotted using the plotROC R package.

Identify Bacterial CRISPRs Targeting Phages

CRISPRs were identified from bacterial genomes using the PilerCR program (v1.06) (Edgar 2007). Resulting spacer sequences were filtered to exclude spacers shorter than 20 bp and longer than 65 bp. Spacer sequences were aligned to the phage genomes using the nucleotide blast algorithm with default parameters (v2.4.0) (Camacho et al. 2009). The mean percent identity for each matching pair was recorded for use in our classification model.

Detect Matching Prophages within Bacterial Genomes

Temperate bacteriophages infect and integrate into their bacterial host's genome. We detected integrated phage elements within bacterial genomes by aligning phage genomes to bacterial genomes using the nucleotide blast algorithm and a minimum e-value of 1e-10. The resulting bitscore of each alignment was recorded for use in our classification model.

Identify Shared Genes Between Bacteria and Phages

Phages may share genes with their bacterial hosts, providing us with evidence of phage-host infectious pairs. We identified shared genes between bacterial and phage genomes by assessing amino acid similarity between the genes using the Diamond protein alignment algorithm (v0.7.11.60) (Buchfink et al. 2015). The mean alignment bitscores for each genome pair was recorded for use in our classification model.

Protein - Protein Interactions

The final method we used for predicting infectious interactions between bacteria and phages was by detecting pairs of genes whose proteins are known to interact. We assigned bacterial and phage genes to protein families by aligning them to the Pfam database using the Diamond protein alignment algorithm. We then identified which pairs of proteins were predicted to interact using the Pfam interaction information within the Intact database (Orchard et al. 2014). The mean bitscores of the matches between each pair were recorded for use in our classification model.

Virome Network Construction

The bacteria and phage operational genomic units (OGUs) were scored using the same approach as outlined above. The infectious pairings between bacteria and phage OGUs were classified using the random forest model described above. The predicted infectious pairings and all associated metadata was saved as a graph database using Neo4j graph database software (v2.3.1) (). This network was used for downstream community analysis.

Centrality Analysis

Degree and closeness centrality were calculated using the associated functions within the igraph R package (v1.0.1) (Csardi and Nepusz). Briefly, the **closeness centrality** of Vi was calculated taking the inverse of the average length of the shortest path (`d`) between nodes Vi and the `k` other nodes Vj, such that closeness centrality of node Vi is:

$$C_C\left(V_i\right) = \left(\sum_{j=1}^{k} d\left(V_i, V_j\right)\right)^{-1}$$

The distance between nodes (`d`) was calculated as the shortest number of edges required to be traversed to move from one node to another.

The simple metric of **degree centrality** of node Vi was defined as the sum of the number of k edges Ei associated with that node, such that:

$$C_D\left(V_i\right) = \sum_{i=1}^{k} E_i$$

The Eigenvector centrality was calculated using the values of the first eigen vector of the associated adjacency matrix that are associated with each vertex `Vi`. Conceptually, this function results in a centrality value that reflects the connections of the vertex, as well as the centrality of the connected vertices.

The **centralization** metric was used to assess the average centrality of each sample graph `G`. Centralization was calculated by taking the sum of each vertex `Vi` centrality from the graph maximum centrality `Cw`, such that:

$$C\left(G\right) = \frac{\sum_{i=1}^{k} \max_{w} c\left(w\right) - c\left(v_{i}\right)}{T}$$

The values were corrected for uneven graph sizes by dividing the centralization score by the maximum theoretical centralization (`T`) for a graph with the same number of vertices.

Network Relationship Dissimilarity

We assessed similarity between graphs by evaluating the shared centrality of their vertices, as has been done previously. More specifically, we calculated the dissimilarity between graphs `Gi` and `Gj` using the Bray-Curtis dissimilarity metric and Eigenvector centrality values such that:

$$B\left(G_{i}, G_{j}\right) = 1 - \frac{2C_{ij}}{C_{i} + C_{j}}$$

Where `Cij` is the sum of the lesser centrality values for those vertices shared between graphs, and `Ci` and `Cj` are are the total number of vertices found in each graph. This allows us calculate the dissimilarity between graphs based on the shared centrality values between the two graphs.

Acknowledgments

We thank the members of the Schloss lab for their underlying contributions. GDH is supported in part by the University of Michigan Molecular Mechanisms of Microbial Pathogenesis Fellowship.

Disclosure Declaration

The authors report no conflicts of interest.

Figures

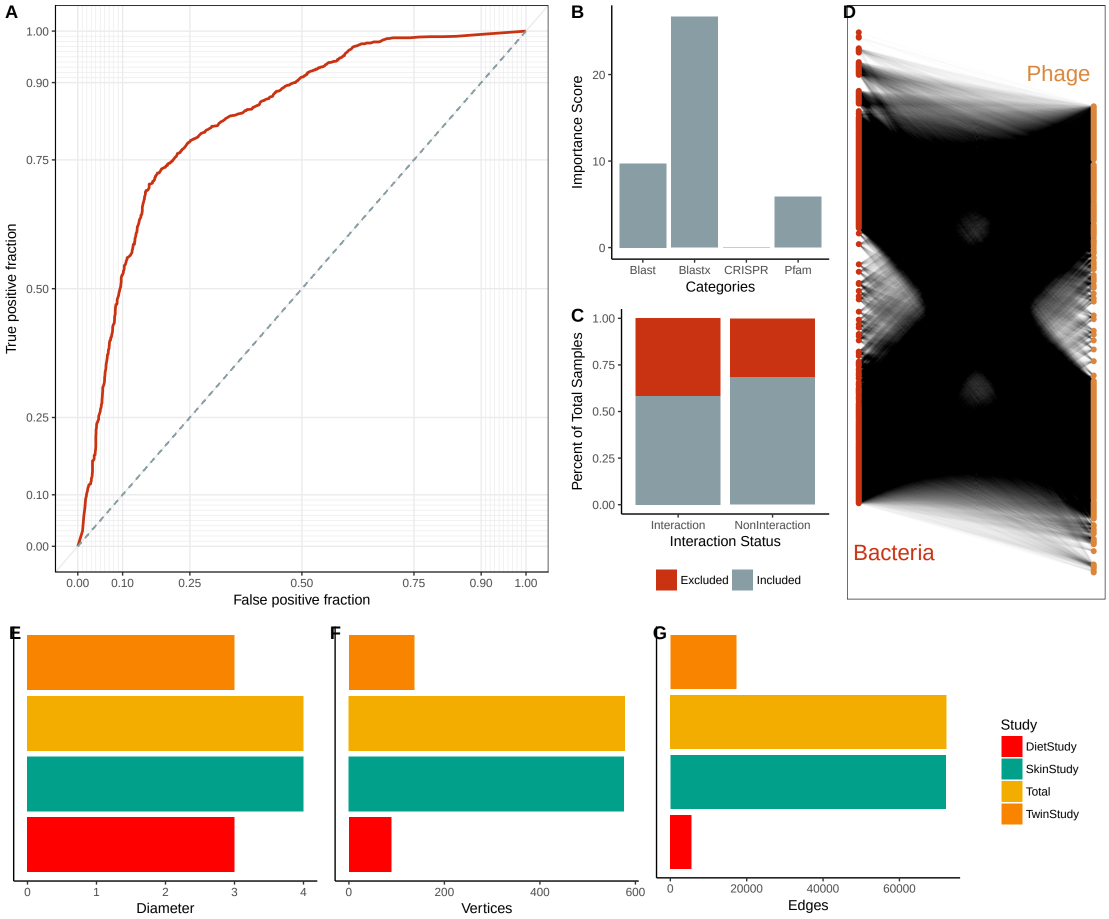

Figure 1: **Summary of Multi-Study Network Model.** *(A) ROC curve resulting from ten iterations used to create the phage - bacteria infection prediction model. (B) Importance scores associated with the metrics used in the random forest model to predict relationships between bacteria and phages. The importance score is defined as the mean decrease in accuracy of the model when a feature (e.g. Pfam) is excluded. (C) Proportions of samples included (gray) and excluded (red) in the model. Samples were excluded from the model because they did not yield any scores. Those interactions without scores were defined as not having interactions. (D) Bipartite visualization of the resulting phage-bacteria network. This network includes information from all three published studies. (E) Network diameter (measure of graph size; the greatest number of traversed vertices required between two vertices), (F) number of vertices, and (G) number of edges (relationships) for the total network (yellow) and the individual study sub-networks (diet study = red, skin study = green, twin study = orange).*

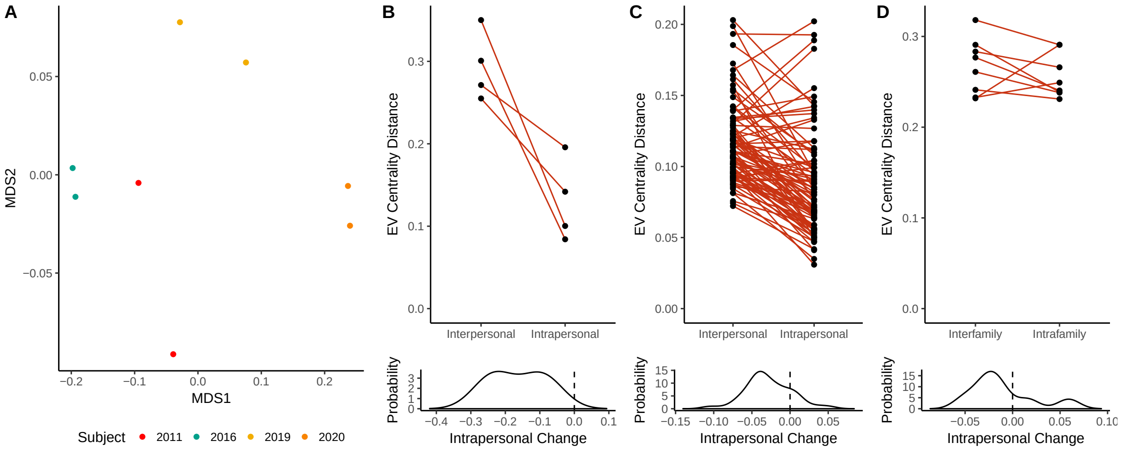

Figure 2: **Intrapersonal vs Interpersonal Network Dissimilarity Across Different Human Systems.** *(A) NMDS ordination illustrating network dissimilarity between subjects over time. Each sample is colored by subject, with each sample pair collected 8-10 days apart. Dissimilarity was calculated using the Bray-Curtis metric based on abundance weighted Eigenvector centrality signatures, with a greater distance representing greater dissimilarity in bacteria and phage centrality and abundance. (B) Quantification of gut network dissimilarity within the same subject over time (intrapersonal) and the mean dissimilarity between the subject of interest and all other subjects (interpersonal). The density curve is the probability distribution of the changes between intrapersonal and interpersonal diversity, representing the probability that intrapersonal dissimilarity will be lower than interpersonal dissimilarity (lower intrapersonal dissimilarity defined as change less than zero). (C) Quantification of skin network dissimilarity within the same subject and anatomical location over time (intrapersonal) and the mean dissimilarity between the subject of interest and all other subjects at the same time and the same anatomical location (interpersonal). Probability distribution the same as for panel B. (D) Quantification of gut network dissimilarity within subjects from the same family (intrafamily) and the mean dissimilarity between subjects within a family and those of other families (interfamily). Probability distribution the same as for panel B.*

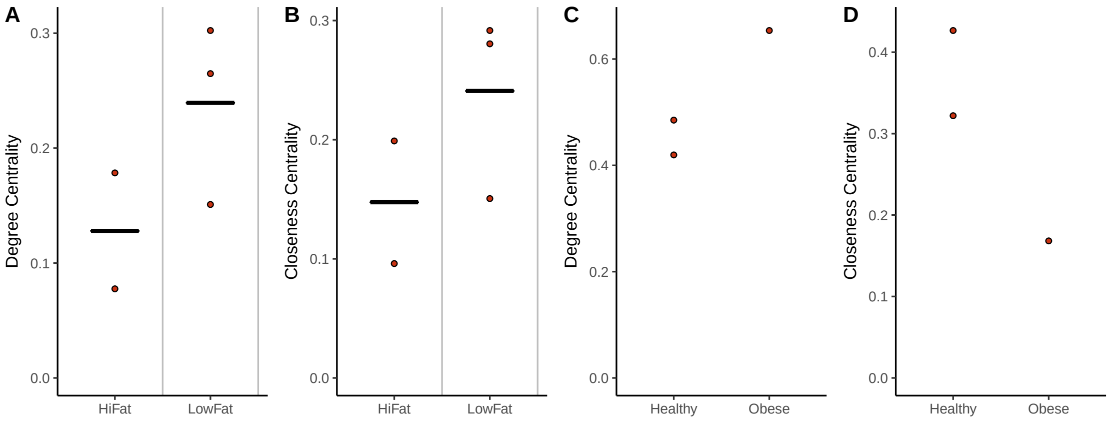

Figure 3: **Impact of Diet and Obesity on Gut Network Structure.** *(A) Quantification of average degree centrality (number of edges per node) and (B) closeness centrality (average distance from each node to every other node) of gut microbiome networks of subjects limited to exclusively high fat or low fat diets. Lines represent the mean degree of centrality for each diet. (C) Quantification of average degree centrality and (D) closeness centrality between obese and healthy adult women.*

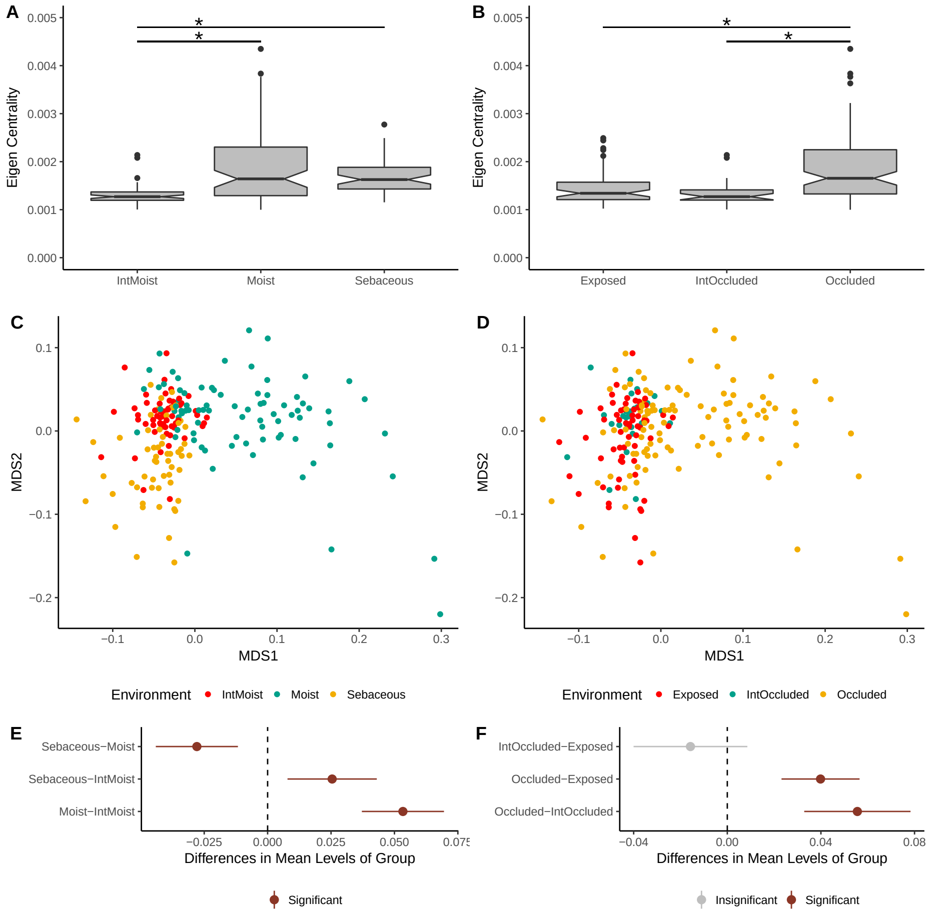

Figure 4: **Impact of Skin Micro-Environment on Microbiome Network Structure.** *(A) Notched box-plot depicting differences in average Eigenvector centrality between moist, intermittently moist, and sebaceous skin sites and (B) occluded, intermittently occluded, and exposed sites. Notched box-plots were created using ggplot2 and show the median (center line), the inter-quartile range (IQR; upper and lower boxes), the highest and lowest value within 1.5* IQR (whiskers), outliers (dots), and the notch which provides an approximate 95% confidence interval as defined by 1.58 * IQR / sqrt(n). (C) NMDS ordination depicting the differences in skin microbiome network structure between skin moisture levels and (D) occlusion. Samples are colored by their environment and their dissimilarity to other samples was calculated as described in figure 2. (E) The statistical differences of networks between moisture and (F) occlusion status were quantified with an anova and post hoc Tukey test. Cluster centroids are represented by dots and the extended lines represent the associated 95% confidence intervals. Significant comparisons (p-value < 0.05) are colored in red, and non-significant comparisons are gray.*

Supplemental Figures

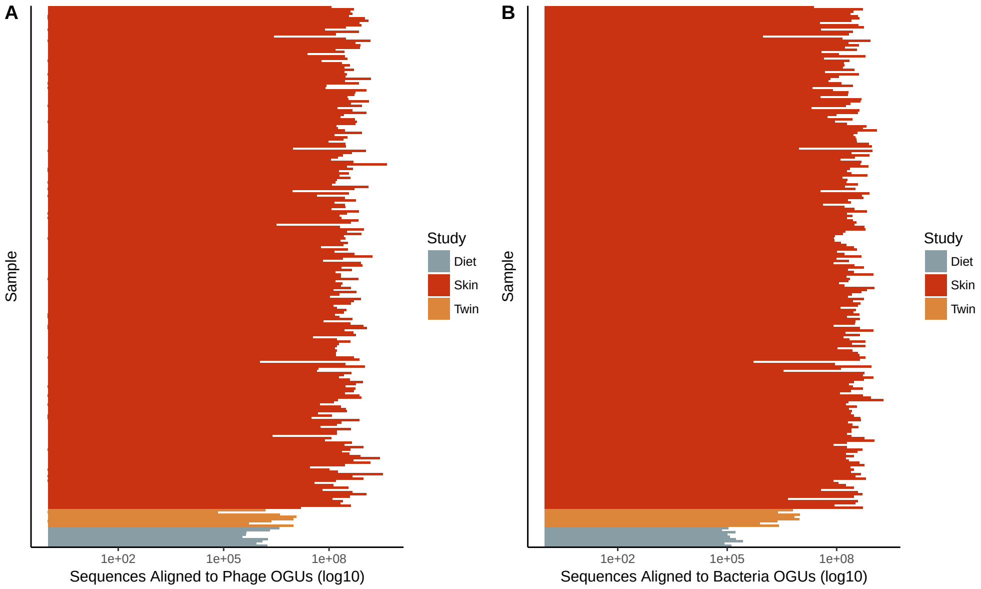

Figure S1: **Sequencing Depth Summary.** *Number of sequences that aligned to (A) Phage and (B) Bacteria operational genomic units. Sequencing count aligned to OGUs was length corrected.*

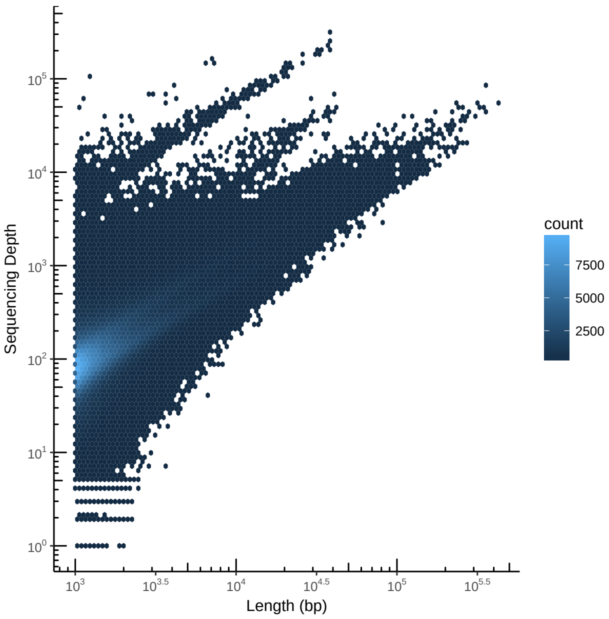

Figure S2: **Contig Summary Statistics.** *Scatter plot heat map with each hexagon representing an abundance of contigs. Contigs are organized by length on the x-axis and the number of aligned sequences on the y-axis.*

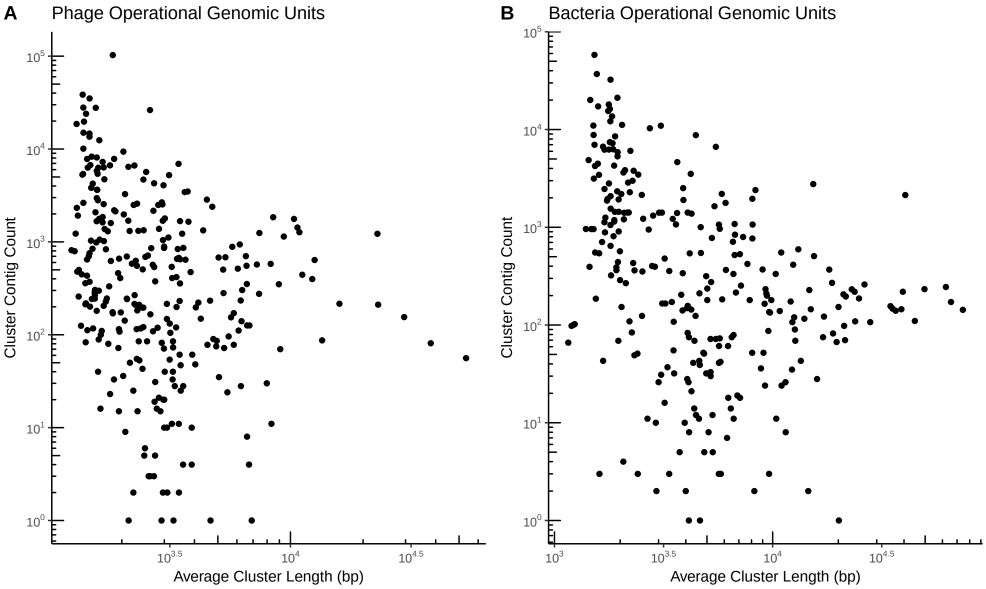

Figure S3: **Operational Genomic Unit Summary Statistics.** *Scatter plot with operational genomic unit clusters organized by average contig length within the cluster on the x-axis and the number of contigs in the cluster on the y-axis. Operational genomic units of (A) bacteriophages and (B) bacteria are shown.*

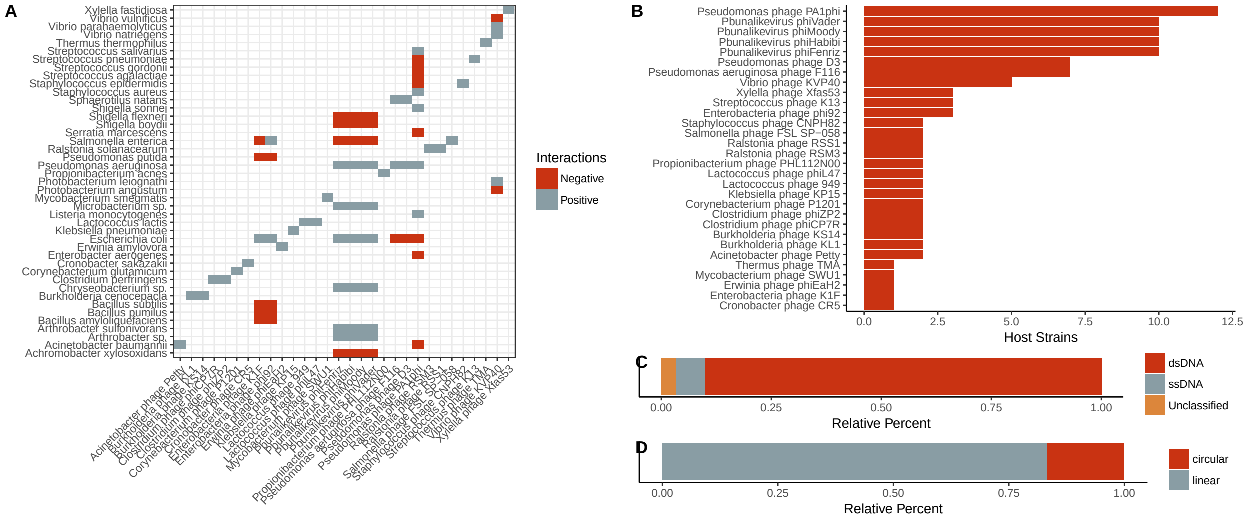

Figure S4: **Summary information of validation dataset used in the interaction predictive model.** *A) Categorical heat-map highlighting the experimentally validated positive and negative interactions. Only bacteria species are shown, which represent multiple reference strains. Phages are labeled on 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) Quantification 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.*

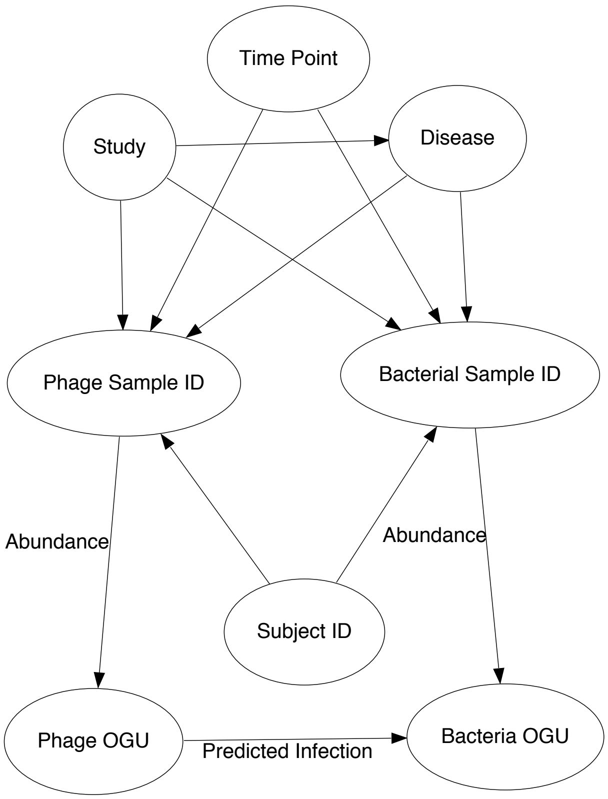

Figure S5: **Structure of the interactive network.** *Metadata relationships to samples (Phage Sample ID and Bacteria Sample ID) included the associated time point, the study, the subject the sample was taken from, and the associated disease. Infectious interactions were recorded between phage and bacteria operational genomic units (OGUs). Sequence count abundance for each OGU within each sample was also recorded.*

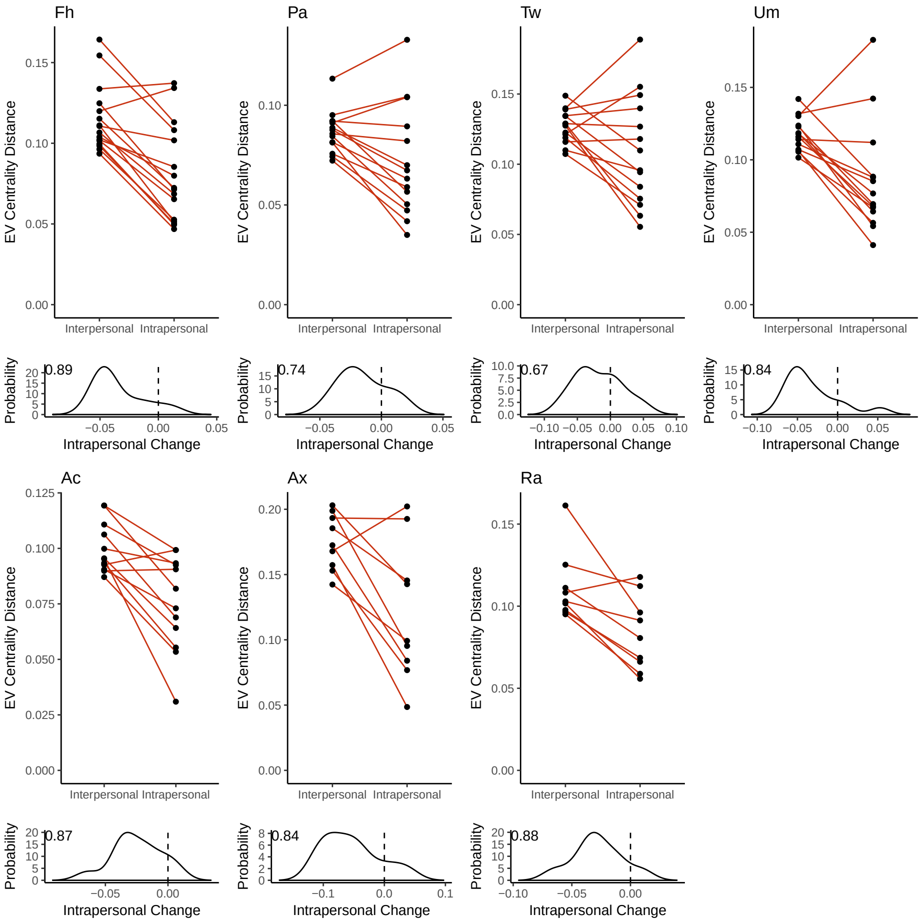

Figure S6: **Intrapersonal vs Interpersonal Dissimilarity of the Skin.** *Quantification of skin network dissimilarity within the same subject and anatomical location over time (intrapersonal) and the mean dissimilarity between the subject of interest and all other subjects at the same time and the same anatomical location (interpersonal), separated by each anatomical site (forehead [Fh], palm [Pa], toe web [Tw], umbilicus [Um], antecubital fossa [Ac], axilla [Ax], and retroauricular crease [Ra]). Below is the probability distribution of the changes between intrapersonal and interpersonal diversity, representing the probability that intrapersonal dissimilarity will be lower than interpersonal dissimilarity (intrapersonal change less than zero). The probability that the slope will be less than zero (integral from negative infinity to zero) is provided in the top left corner.*

References

Abeles SR, Ly M, Santiago-Rodriguez TM, Pride DT. 2015. Effects of Long Term Antibiotic Therapy on Human Oral and Fecal Viromes. *PLOS ONE* **10**: e0134941.

Abeles SR, Robles-Sikisaka R, Ly M, Lum AG, Salzman J, Boehm TK, Pride DT. 2014. Human oral viruses are personal, persistent and gender-consistent. 1–15.

Alneberg J, Bjarnason BS aacute ri, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 1–7.

Baxter NT, Zackular JP, Chen GY, Schloss PD. 2014. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome* **2**: 20.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59–60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 1.

Consortium THMP. 2012. A framework for human microbiome research. *Nature* **486**: 215–221.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

Csardi G, Nepusz T. The igraph software package for complex network research.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, et al. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**: 559–563.

Edgar RC. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**: 18.

Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2015. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews* **40**: 258–272.

Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, NIH Intramural Sequencing Center Comparative Sequencing Program, et al. 2013. Topographic diversity of fungal and bacterial communities in human skin. *Nature* 1–6.

Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. 2011. Statistical structure of host-phage interactions.

Proceedings of the National Academy of Sciences of the United States of America **108**: E288–97.

Flores CO, Valverde S, Weitz JS. 2013. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME Journal* **7**: 520–532.

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, NISC Comparative Sequencing Program, Bouffard GG, Blakesley RW, Murray PR, et al. 2009a. Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324**: 1190–1192.

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, NISC Comparative Sequencing Program, Bouffard GG, Blakesley RW, Murray PR, et al. 2009b. Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324**: 1190–1192.

Haerter JO, Mitarai N, Sneppen K. 2014. Phage and bacteria support mutual diversity in a narrowing staircase of coexistence. *The ISME Journal* **8**: 2317–2326.

Hannigan GD, Grice EA. 2013. Microbial Ecology of the Skin in the Era of Metagenomics and Molecular Microbiology. *Cold Spring Harbor Perspectives in Medicine* **3**: a015362–a015362.

Hannigan GD, Hodkinson BP, McGinnis K, Tyldsley AS, Anari JB, Horan AD, Grice EA, Mehta S. 2014. Culture-independent pilot study of microbiota colonizing open fractures and association with severity, mechanism, location, and complication from presentation to early outpatient follow-up. *Journal of Orthopaedic Research* **32**: 597–605.

Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, Minot S, Bushman FD, Grice EA. 2015. The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *mBio* **6**: e01578–15.

Hannon GJ. FASTX-Toolkit. GNU Affero General Public License.

Harcombe WR, Bull JJ. 2005. Impact of phages on two-species bacterial communities. *Applied and Environmental Microbiology* **71**: 5254–5259.

He Q, Li X, Liu C, Su L, Xia Z, Li X, Li Y, Li L, Yan T, Feng Q, et al. 2016. Dysbiosis of the fecal microbiota in the TNBS-induced Crohn's disease mouse model. *Applied Microbiology and Biotechnology* 1–10.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223–2230.

Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, Nickerson KW, Kokjohn TA. 1998. Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa.

Applied and Environmental Microbiology **64**: 575–580.

Jover LF, Flores CO, Cortez MH, Weitz JS. 2015. Multiple regimes of robust patterns between network structure and biodiversity. *Scientific Reports* **5**: 17856.

Kim S, Rahman M, Seol SY, Yoon SS, Kim J. 2012. 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.

Kuhn M. caret: Classification and Regression Training.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *METHODS* **102**: 3–11.

Loesche M, Gardner SE, Kalan L, Horwinski J, Zheng Q, Hodkinson BP, Tyldsley AS, Franciscus CL, Hillis SL, Mehta S, et al. 2016. Temporal stability in chronic wound microbiota is associated with poor healing. *Journal of Investigative Dermatology.*

Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T, Pride DT. 2014. Altered Oral Viral Ecology in Association with Periodontal Disease. *mBio* **5**: e01133–14–e01133–14.

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

Manrique P, Bolduc B, Walk ST, Oost J van der, Vos WM de, Young MJ. 2016. Healthy human gut phageome. *Proceedings of the National Academy of Sciences of the United States of America* 201601060.

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

Middelboe M, Hagström A, Blackburn N, Sinn B, Fischer U, Borch NH, Pinhassi J, Simu K, Lorenz MG. 2001. Effects of Bacteriophages on the Population Dynamics of Four Strains of Pelagic Marine Bacteria. *Microbial Ecology* **42**: 395–406.

Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 12450–12455.

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* **21**: 1616–1625.

Modi SR, Lee HH, Spina CS, Collins JJ. 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**: 219–222.

Moebus K, Nattkemper H. 1981. Bacteriophage sensitivity patterns among bacteria isolated from marine waters. *Helgoländer Meeresuntersuchungen* **34**: 375–385.

Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge MT, Wilen CB, Flagg M, et al. 2016. Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host and Microbe* **19**: 311–322.

Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, et al. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**: 447–460.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, et al. 2014. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**: D358–63.

Poisot T, Canard E, Mouillot D, Mouquet N, Gravel D. 2012. The dissimilarity of species interaction networks. *Ecology letters* **15**: 1353–1361.

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

Poisot T, Stouffer D. 2016. How ecological networks evolve. *bioRxiv*.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**: 334–338.

Santiago-Rodriguez TM, Ly M, Bonilla N, Pride DT. 2015. The human urine virome in association with urinary tract infections. *Frontiers in Microbiology* **6**: 14.

Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9**: 34–15.

Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71**: 1501–1506.

Schwarzer D, Buettner FFR, Browning C, Nazarov S, Rabsch W, Bethe A, Oberbeck A, Bowman VD, Stummeyer K, Mühlenhoff M, et al. 2012. A multivalent adsorption apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis. *Journal of Virology* **86**: 10384–10398.

Seekatz AM, Rao K, Santhosh K, Young VB. 2016. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent Clostridium difficile infection. *Genome medicine* **8**: 47.

Sze MA, Schloss PD. 2016. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio* **7**: e01018–16.

Thompson RM, Brose U, Dunne JA, Hall RO, Hladyz S, Kitching RL, Martinez ND, Rantala H, Romanuk TN, Stouffer DB, et al. 2012. Food webs: reconciling the structure and function of biodiversity. *Trends in ecology & evolution* **27**: 689–697.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. 2009a. A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. 2009b. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine* **1**: 6ra14–6ra14.

Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. 2014. The human gut microbiome as a screening tool for colorectal cancer. *Cancer prevention research (Philadelphia, Pa)* **7**: 1112–1121.

Neo4j.