Environmental Conditions Impact Phage-Bacteria Community Structure Of The Human Microbiome

Geoffrey D Hannigan Melissa B Duhaime Danai Koutra Patrick D Schloss

Contents

Introduction

Viruses and bacteria are crucial components to the human microbiome and play an important role in health and disease. Bacterial communities have been associated diseases including a wide range of skin conditions[1], acute and chronic wound healing[2,3], and gastrointestinal diseases including irritable bowel disease[4,5], *Clostridium difficile* infections[6], and colorectal cancer[7,8]. Altered viromes (consisting primarily of bacteriophages) have also been associated with various diseases and environmental perturbations including irritable bowel disease[5,9], periodontal disease[10], and others[11–16]. The human virome has also been implicated in promoting antibiotic resistance throughout human-associated microbial communities[12,17]. These community shifts are not reflections of the bacterial communities, but rather act in concert with them as a single community[5,18]. Studying these bacterial and viral communities together is essential for more completely understanding the link between the microbiome and human health.

Communities of bacteria and their phages are dynamic and complex. Bacteria and phages act in concert to maintain balanced, efficient ecosystems and their removal can disrupt or even collapse the ecosystem[18–20]. Previous reports of bacterial and viral communities have studied the two separately while forcing bacterial community analyses concepts (e.g. taxonomic classification, alpha diversity, and beta diversity) onto phage communities in an attempt to identify community signatures associated with health and disease. This approach treats the phage and bacterial communities as isolated systems, which is inadequate given the mutual reliance of phages on their bacterial hosts, and vice versa. Some studies have gone further by evaluating correlations between the communities, but often rely on inappropriate linear models and fail to utilize functional links beyond abundance correlations. A more appropriate analysis will attempt to understand the virome and bacterial microbiome as a subset of a greater interacting community by focusing on the relationships between the two, rather than studying them in isolation.

Previous groups have benefited from more sophisticated analyses that focus on the relationships within ecosystems, including those between bacteria and phages. Such approaches allow for the creation of relationship networks that provide unique information into system biodiversity and functionality, including genetic material and resource transfer, as well as ecosystem stability[21–28]. This work has shown that environmental conditions, including resource availability, impact ecological network structure. In isolated phage-bacteria systems, decreased resource availability has been shown to alter virome relationship structures, decreasing connectance and thereby reducing lines of communication (e.g. horizontal gene transfer) and making those communities more vulnerable to network disintegration and extinction events[21]. Approaches such as this are valuable to our understanding of human microbiome ecology, but have yet to be applied to the human bacterial and viral communities. Here we focus our analysis on phage-bacteria relationships with the goal of outlining a foundational understanding of how these relationship networks differ between human environments.

To investigate how networks of bacteria and phages change between environments, we leveraged three published microbiome datasets (one of which was published across two manuscripts) with paired virus and bacterial metagenomic sequence sets[12,13,29,30]. Sites included the human gut and skin. We built off of previous work on large-scale phage-bacteria ecological network analysis by inferring interactions using metagenomic datasets, instead of previous culture-based techniques[23,24]. 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[31].

Just as the human microbiome field has benefited from an understanding of how different conditions impact microbial (primarily bacterial) community composition and diversity, we aim to provide an understanding of how the relationships between bacteria and phages change within the microbiome. Because our findings provide insight into how environmental conditions impact network communication, microbial hubs, and ecosystem stability, they will better inform design of microbiome-based therapeutics (e.g. probiotics) and inform efforts to reduce spread of antibiotic resistance. This will also lay a foundation for incorporating microbiome network analysis in future studies and provide insights into altered network structure in human disease states.

Results

Modeling Phage-Bacteria Infectious Networks

We studied the impact of human system environment (e.g. skin, gut, defined diets) on microbiome phage-bacteria interaction networks by leveraging previously published sequence sets containing purified virome samples paired with bacterial metagenomes from whole metagenomic shotgun sequences. Our study contained datasets for three human virome studies, including a study of the impact of diet on the gut virome[13], the impact of anatomical location on the skin virome[12], and the virome of monozygotic twins and their mothers[29,30]. The viromes associated with these datasets were subjected to virus-like particle (VLP) purification to eliminate other organism DNA including bacteria, fungi, and humans.

Bacterial and viral sequences were quality filtered and assembled into contigs from their respective sample sets. These studies were performed over five years using different methods and technologies, so therefore yielded different sequence abundances **(Supplemental Figure 5 A-B)**. Because contig assembly rarely reconstructs entire genomes, we clustered related bacteria and phage contigs by k-mer frequency and co-abundance using the CONCOCT algorithm. These clusters represent operationally defined clusters of related bacteria and phage genomes that we defined as operational genomic units (OGUs), which are conceptually similar to the operational taxonomic unit (OTU) and operational protein family (OPF) definitions used for grouping highly similar 16S rRNA and open reading frame sequences, respectively[32]. Contigs and OGUs demonstrated high sequence coverage and length **(Supplemental Figure 6 - 7)**.

We predicted which phage OGUs will infect which bacteria using a random forest model trained on experimentally validated infectious relationships from six previous publications[31,33–37]. This training set contained diverse bacteria and phages, with both broad and specific infectious capabilities **(Supplemental Figure 8 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 excluded from the analysis **(Supplemental Figure 8 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 almost exclusively on positive relationships only, 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. The resulting random forest model exhibited high performance with an AUC of 0.853, a sensitivity of 0.851, and a specificity of 0.774 **(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)**.

Upon validation and completion, we used the 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 OGU 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 9)**. 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 200 vertices, less than 20,000 relationships, and diameters of 3, suggesting more sparsely related phages and bacteria **(Figure 1 E - F)**.

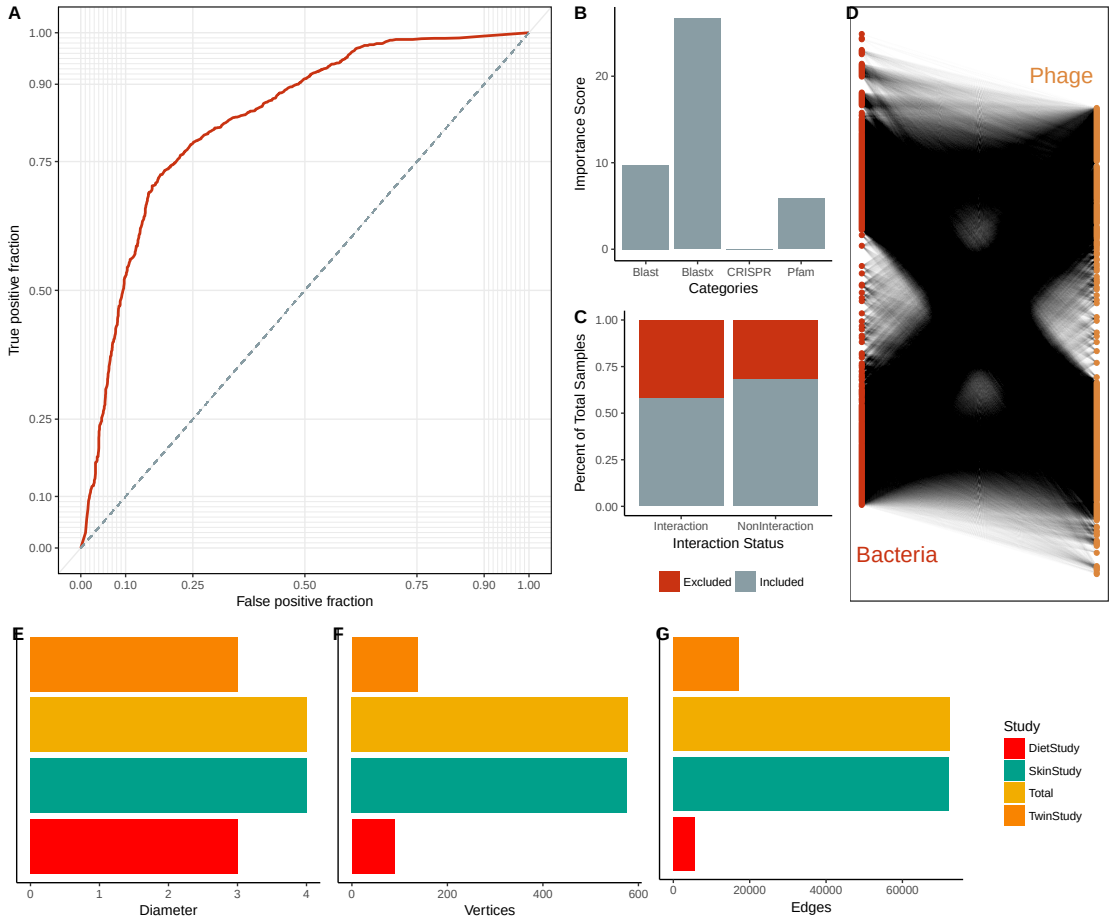

Figure 1: **Summary of Multi-Study Network Model.** *(A) ROC curve resulting from the ten iterations used to create the phage - bacteria infection prediction model. (B) Importance scores associated with the metrics used in the random forest model to predict relationship between bacteria and phages. The importance score is the mean decrease in accuracy of the model when the model is rebuilt without that factor. (C) Proportions of samples excluded from model learning due to a lack of scores. 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. (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).*

Interpersonal and Intrapersonal Diversity of Gut Microbial Networks

Previous work has show reduced intra-personal (within person) diversity in the viromes and bacterial communities of the human gut and skin compared to inter-personal (between people) diversity[12,38,39]. Understanding this conservation of the microbiome has been important for establishing a basic understanding of microbiome individuality and aids in interpreting microbiome study results. We hypothesized that like

other aspects of these microbial communities (e.g. diversity, community composition), there is a strong conservation of network structure in the skin and gut.

We tested this hypothesis by calculating the degree of dissimilarity between subject graphs while incorporating both phage and bacteria abundance, as well as centrality within the node. This was accomplished 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 is defined in the same way. This metric allowed us to identify those bacteria and phage hubs that were capable of broadly disseminating genetic material and most significantly impact community dynamics through fluctuations in abundance. Having more microbial hubs is also associated with greater stability and resilience to perturbations such as extinction events, which could disintegrate a poorly connected network. We compared the centrality profiles between graphs using the Bray-Curtis metric which accounts for shared bacteria and phages, as well as their degree of weighted centrality. The final result was a robust measurement of graph structure dissimilarity that accounted for abundance, centrality, and shared community membership.

We found that gut microbiome network structure was highly individual specific, with networks clustering tightly based on the human host **(Figure 2 A)**. Intrapersonal (within person) network dissimilarity over the 8-10 day sampling period was significantly less than the average interpersonal dissimilarity (between people) associated with each individual **(Figure 2 B)**. The probability of intrapersonal diversity of an individual being less than interpersonal diversity in our dataset was 98% (absolute error $<$ 1.1e-4) **(Figure 2 B)**.

Unlike the strong intrapersonality of the gut, skin network structure was weakly intrapersonally conserved. 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 83% (absolute error $<$ 7.5e-6) **(Figure 2 C)**. This distribution was similar for each separate anatomical site, supporting this as an accurate representation of skin intrapersonal and interpersonal diversity **(Supplemental Figure 10)**.

The individuality of gut network structures did not extend to families. The gut network structures were no more similar between twins and their mothers (intrafamily) compared to other families of twins and families (inter-family) **(Figure 2 D)**. The probability of a family member being more similar to another family member compared to non-family members was 63% (absolute error $<$ 1.1e-4).

Role of Diet & Obesity in Gut Network Structure

Diet is a major environmental factor that influences resource availability and gut microbiome composition and diversity, including bacteria and phages[13,40,41]. Additional 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[21]. We therefore hypothesized that network structure would be altered by changes in diet. We additionally evaluated the potential association between gut network structure and obesity, a disease linked to diet and potentially the microbiome[42].

We evaluated the differences in gut network structure by quantifying graph connectedness as the overall degree of network centrality. 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. Said another way, degree centrality was defined as the number of relationships associated with each bacterium or phage. Because this metric alone offers only minimal insight, we supplemented it with measurements of closeness centrality. **Closeness centrality** is a measure of how close each phage or bacterium is to all of the other phages and bacteria in the network. A higher closeness centrality suggests genetic information or the effects of altered abundance would be more more impactful to all other microbes in the system. 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

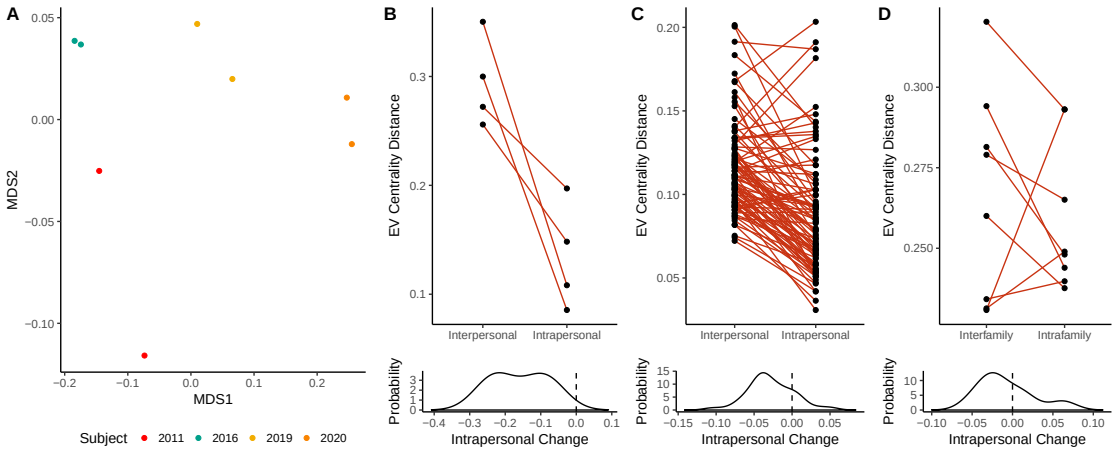

Figure 2: **Intrapersonal vs Interpersonal Network Dissimilarity Across Different Human Systems.** *(A) NMDS ordination illustrating network dissimilarity between subjects over time. Each sample is colored by subject, with each sample pair collected 8-10 days apart. Dissimilarity calculated using the 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). Below is the probability distribution of the changes between intrapersonal and interpersonal diversity, represetning the probability that intrapersonal dissimilarity will be lower than interpersonal dissimilarity (intrapersonal 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 pannel B. (D) Quantification of gut network dissimilarity within subjects from the same family (intrafamily) and the mean dissimilarity between subjects within a familiy and those of other families (interfamily). Probability distribution the same as for pannel B.*

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 information transfer and the impact of altered abundance less likely to impact 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 was associated with a higher degree centrality **(Figure 3 C)** but a less closeness centrality, compared to the healthy controls **(Figure 3 D)**. This means that the obesity network was overall less connected, with more relationships per bacteria and phage, but microbes being further from all other microbes within the community.

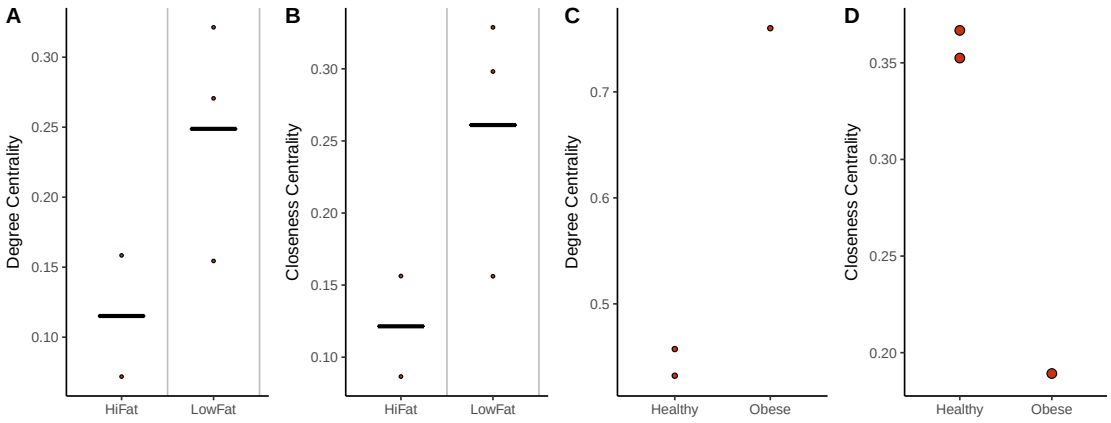

Figure 3: **Impact of Diet and Obesity on Gut Network Structure.** *(A) Quantification of average degree centrality (number of edges per node) and (B) closeness centrality (average distance from each node to every other node) of gut microbiome networks of subjects limited to exclusively high fat or low fat diets. Lines represet the mean degree of centrality for each diet. (C) Quantification of average degree centrality and (D) closeness centrality between obese and healthy adult women. Tests for statistical significance were not performed as they would be misleadingly inappropriate for these small sample sizes.*

Variation of Network Structure Across the Human Skin Landscape

Extensive previous work has shown differences in microbial communities between anatomical sites, including bacteria, viruses, and fungi[12,38,43]. 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 provides a higher value for more abundant phages and bacteria that 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 **(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 intermittently occluded **(Figure 4 B)**.

We supplemented this analysis by comparing the network signatures using the centrality dissimilarity approach described above for our intrapersonal diversity analysis. The dissimilarity between samples was a function of shared relationships, degree of centrality, and bacteria/phage abundance. When using this more sophisticated

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)**.

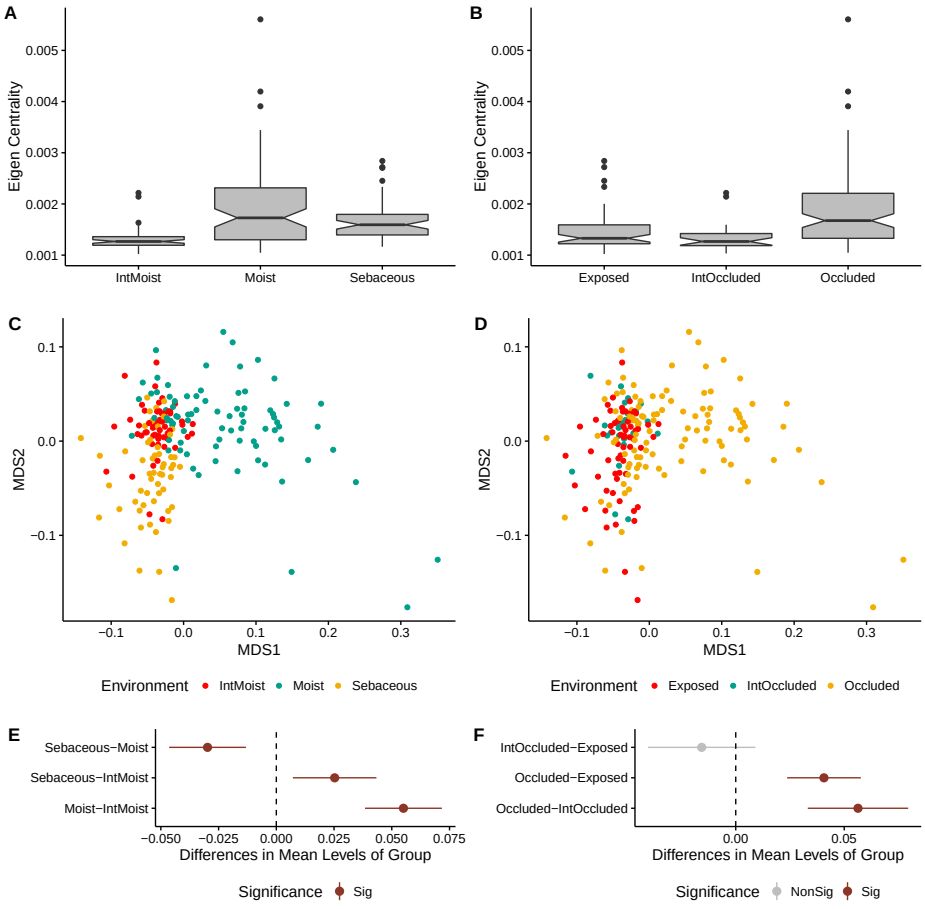

Figure 4: **Impact of Skin Microenvironment on Microbiome Network Structure.** *(A) Notched boxplot depicting differences in average Eigenvector centrality between moist, intermittently moist, and sebaceous skin sites and (B) occluded, intermittently occluded, and exposed sites. Notched boxplots were 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), 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 assocaited 95% confidence intervals. Significant comparisons (p-value < 0.05) are colored in red, and non-significant comparisons are gray.*

Discussion

We developed and implemented a network-based method to understanding the relationships between bacteria and bacteriophages across the human microbiome. This network approach was advantageous over previous

studies because it did not rely on inappropriate linear correlations of abundance, it was trained on positive and negative relationships, and it involved implementation of network theory concepts. By utilizing links between bacteria and phage community members, we gained new insights into communication pathways throughout the communities, including potential routes of horizontal gene transfer, the influence of fluctuating abundance, and overall community stability and resilience to extinction events.

We used this approach to build off of previous human microbiome and virome work and show that the structural properties of bacteria and phage community relationships differ based on environmental conditions. We found that, just as gut virome composition and diversity are conserved in individuals, gut network structure was also conserved within people over time. This conservation did not extend to families, as gut network structures were just as similar to each other as they were to members of other families.

While the gut networks were highly conserved within individuals, the network structure of the skin was weakly conserved. This weaker conservation in the skin compared to the gut could have been due to the greater exposure of the skin to the external environment. Comparison of the networks by skin micro-environment, including moisture levels and environmental exposure, revealed a significant difference in network structures between anatomical micro-environments. The more dynamic regions, including the exposed, intermittently moist, and intermittently occluded sites, exhibited decreased connectedness compared to the more stable micro-environments. Thus the protected and stable environments harbored more stable microbiome networks that were amenable to broad horizontal gene transfer and resilience to network disruption and disintegration. This link between network structure and micro-environment also suggests a role for nutrient availability in defining network structure.

Because diet has been linked to alterations in gut bacterial and viral communities, and because nutrient availability has been associated with altered network structures *in vitro*, we evaluated the association of high and low fat diets with microbiome network structure. Although our sample size was small, we found evidence that there is indeed a difference in network structure between high and low fat diets. We also observed a difference in network structure between healthy and obese individuals, with obesity being a disease associated with altered nutrient availability. Although the obese individual's gut microbiome network had nodes with more relationships, it was associated with reduced connectedness similar to what was observed in high fat diets. Thus the food we eat may not only impact what microbes colonize our guts, but may also impact their ability to communicate and remain stable.

Together these findings suggest that human microbiome network structure and stability depends on the environmental conditions. More specifically, exposed and fluctuating micro-environments seem to be associated with microbial communities less resilient to perturbations and with less capacity for intra-community communication. Nutrient availability also appears to play a role in these community traits.

This work represents an initial step toward understanding the microbiome through its relationships, in addition to its membership. While these findings are informative, there are certainly caveats and future directions that are beyond the scope of this initial study. 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. New and creative scoring metrics can also 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 much many years ago. These archived datasets were limited by the technology and associated 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 certainly be required to validate our observations.

Overall these findings and methodologies are exciting because they represent a new way of understanding the human microbiome. We expect this work will contribute toward understanding the human microbiota as complex interacting communities instead of the reductionist approach often utilized today. Our findings confirm that, in addition to community membership and diversity, microbiome interactions differ between human environments. Some skin sites, diets, and obesity are associated with less connected communities, which could impact gene transfer and community stability. Other highly connected communities are expected

to be more capable of transferring genetic material, such as antibiotic resistance genes. Finally, the link between network structure and nutrient availability could be important for more effective design of probiotics, fecal microbiota transplants, and other microbiome-based therapeutics.

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 within the NCBI SRA Toolkit (version). Sequences were quality trimmed using the Fastx toolkit to exclude bases with quality scores below 33 and shorter than 75 bp. 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 (*version*). 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 (*version*). 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, due to insufficient sequencing depth. In order to reduce redundancy and artificially inflated genomic richness within our dataset, it was important to bin contigs based on their likelihood to be of the same phylogenetic groups. 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).

We clustered contigs using the CONCOCT algorithm (*version*). Because of our large dataset and barriers 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 (version) with the meta mode parameter and default settings.

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[31,33–37]. Associated reference genomes were downloaded from the European Bioinformatics Institute. 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. 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 (version). 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. 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. 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 are predicted to interact using the Pfam interaction information within the Interact database (version). 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 software. 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. 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.

Supplemental Figures

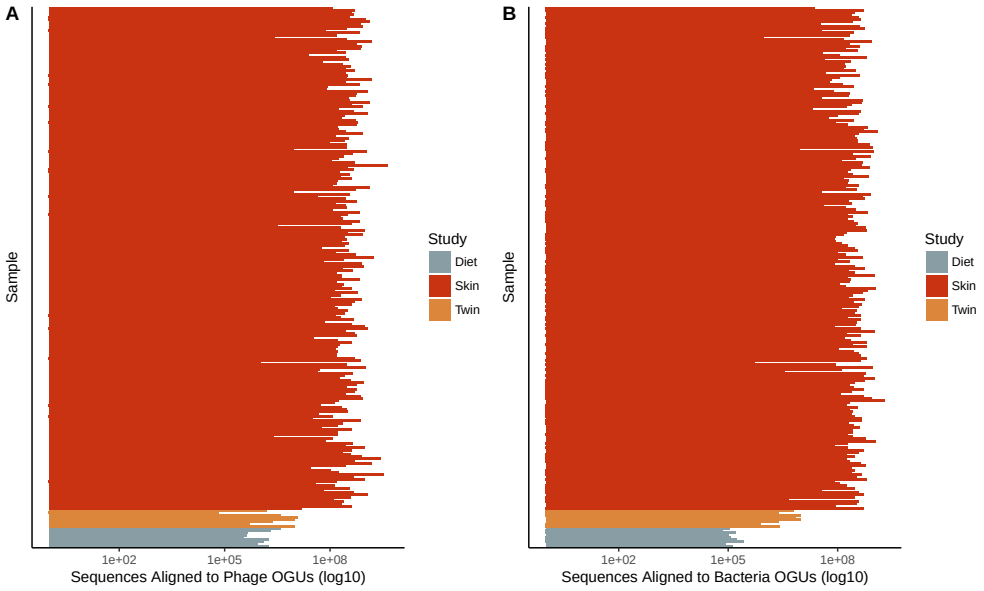

Figure 5: **Sequencing Depth Summary.** *Number of sequences that aligned to (A) Phage and (B) Bacteria operational genomic units. Sequencing count aligned to OGUs was length corrected.*

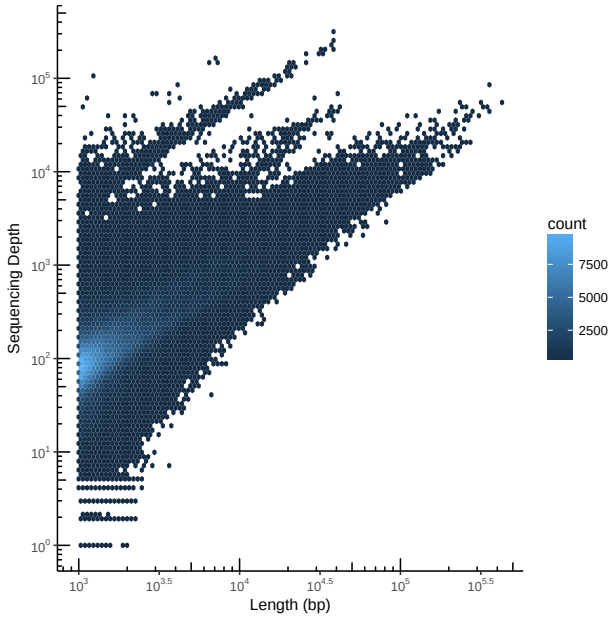

Figure 6: **Contig Summary Statistics.** *Scatter plot heat map with each hexagon representing an abundance of contigs. Contigs are organized by length on the x-axis and the number of aligned sequences on the y-axis.*

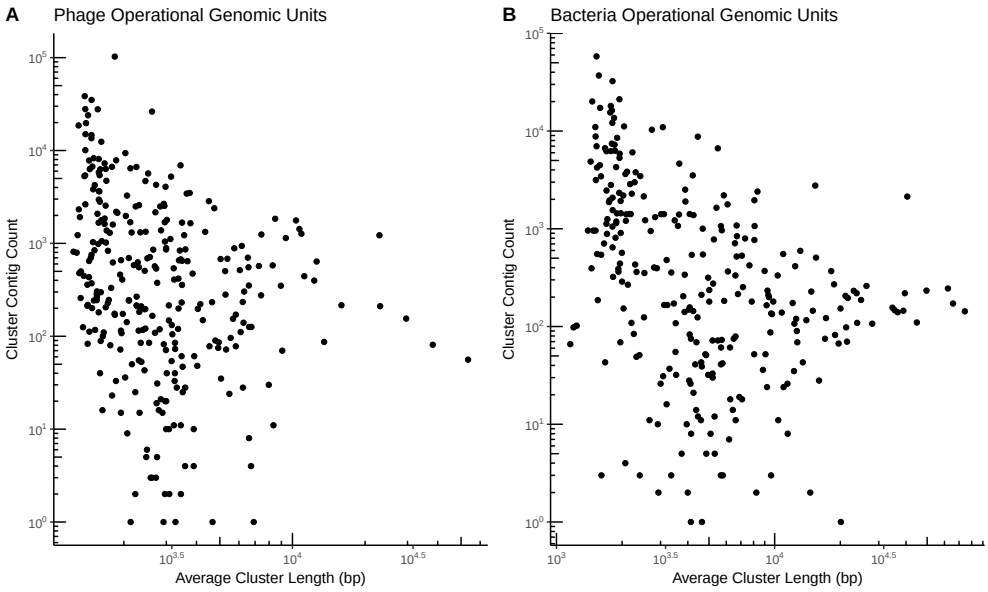

Figure 7: **Operational Genomic Unit Summary Statistics.** *Scatter plot with operational genomic unit clusters organized by average contig length within the cluster on the x-axis and the number of contigs in the cluster on the y-axis. Operational genomic units of (A) bacteriophages and (B) bacteria are shown.*

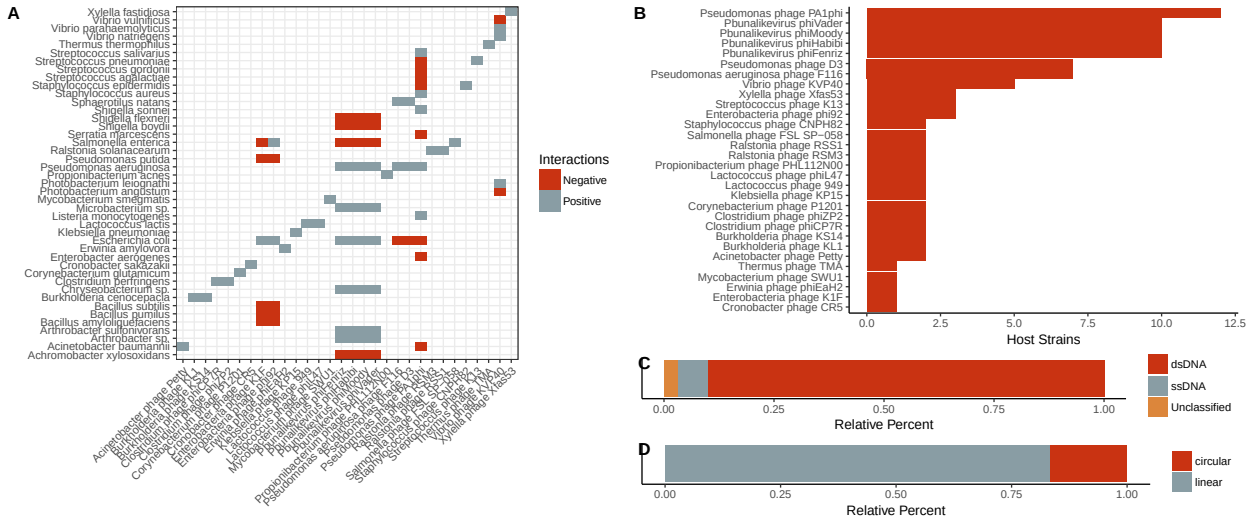

Figure 8: **Summary information of validation dataset used in the interaction predictive model.** *A) Categorical heat-map highlighting the experimentally validated positive and negative interactions. Only bacteria species are shown, which represent multiple reference strains. Phages are labeled on the 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.*

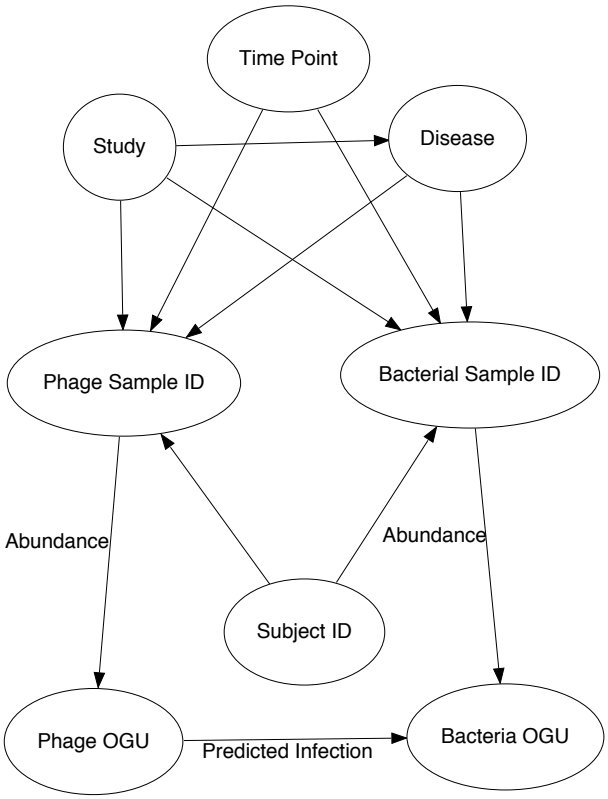

Figure 9: **Structure of the interactive network.** *Metadata relationships to samples (Phage Sample ID and Bacteria Sample ID) included the associated time point, the study, the subject the sample was taken from, and the associated disease. Infectious interactions were recorded between phage and bacteria operational genomic units (OGUs). Sequence count abundance for each OGU within each sample was also recorded.*

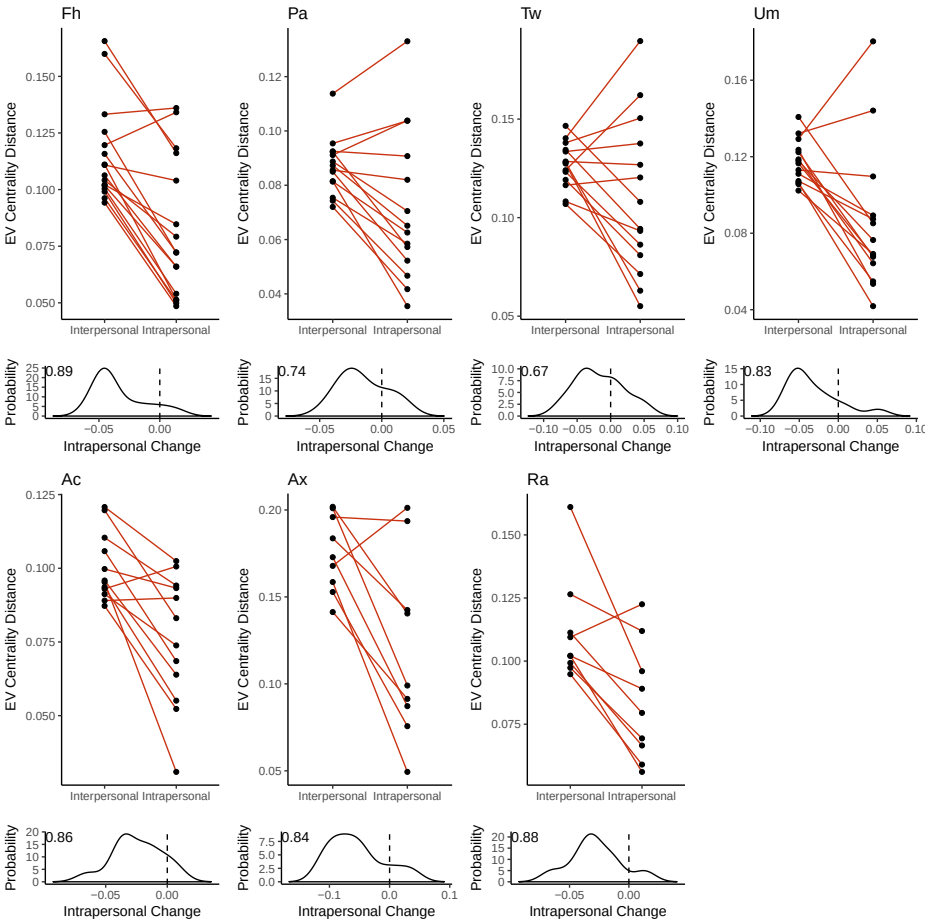

Figure 10: **Intrapersonal vs Interpersonal Dissimilarity of the Skin.** *Quantification of skin network dissimilarity within the same subject and anatomical location over time (intrapersonal) and the mean dissimilarity between the subject of interest and all other subjects at the same time and the same 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

1. Hannigan, G. D. & Grice, E. A. Microbial Ecology of the Skin in the Era of Metagenomics and Molecular Microbiology. *Cold Spring Harbor Perspectives in Medicine* **3,** a015362–a015362 (2013).

2. Hannigan, G. D. *et al.* 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 (2014).

3. Loesche, M. *et al.* Temporal stability in chronic wound microbiota is associated with poor healing. *Journal of Investigative Dermatology* (2016).

4. He, Q. *et al.* Dysbiosis of the fecal microbiota in the TNBS-induced Crohn's disease mouse model. *Applied Microbiology and Biotechnology* 1–10 (2016).

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

6. Seekatz, A. M., Rao, K., Santhosh, K. & Young, V. B. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent Clostridium difficile infection. *Genome medicine* **8,** 47 (2016).

7. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer prevention research (Philadelphia, Pa.)* **7,** 1112–1121 (2014).

8. Baxter, N. T., Zackular, J. P., Chen, G. Y. & Schloss, P. D. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome* **2,** 20 (2014).

9. Manrique, P. *et al.* Healthy human gut phageome. *Proceedings of the National Academy of Sciences of the United States of America* 201601060 (2016).

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

11. 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).

12. 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).

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

14. 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).

15. 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).

16. Abeles, S. R. *et al.* Human oral viruses are personal, persistent and gender-consistent. 1–15 (2014).

17. 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).

18. Haerter, J. O., Mitarai, N. & Sneppen, K. Phage and bacteria support mutual diversity in a narrowing staircase of coexistence. *The ISME Journal* **8,** 2317–2326 (2014).

19. Harcombe, W. R. & Bull, J. J. Impact of phages on two-species bacterial communities. *Applied and Environmental Microbiology* **71,** 5254–5259 (2005).

20. Middelboe, M. *et al.* Effects of Bacteriophages on the Population Dynamics of Four Strains of Pelagic Marine Bacteria. *Microbial Ecology* **42,** 395–406 (2001).

21. 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).

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

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

24. Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME Journal* **7,** 520–532 (2013).

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

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

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

28. Jover, L. F., Flores, C. O., Cortez, M. H. & Weitz, J. S. Multiple regimes of robust patterns between network structure and biodiversity. *Scientific Reports* **5,** 17856 (2015).

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

30. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).

31. 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).

32. Schloss, P. D. & Handelsman, J. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9,** 34–15 (2008).

33. 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).

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

35. 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).

36. 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).

37. 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).

38. Grice, E. A. *et al.* Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324,** 1190–1192 (2009).

39. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 12450–12455 (2013).

40. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine* **1,** 6ra14–6ra14 (2009).

41. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–563 (2014).

42. Sze, M. A. & Schloss, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome.

mBio **7,** e01018–16 (2016).

43. Findley, K. *et al.* Topographic diversity of fungal and bacterial communities in human skin. *Nature* 1–6 (2013).