

1 Biogeography & environmental conditions shape
2 bacteriophage-bacteria networks across the human
3 microbiome

4 Geoffrey D Hannigan¹, Melissa B Duhaime², Danai Koutra³, and Patrick D Schloss^{1,*}

5 ¹Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan, 48109, USA

6 ²Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109, USA

7 ³Department of Computer Science, University of Michigan, Ann Arbor, Michigan, 48109, USA

8 *To whom correspondence may be addressed.

9
10
11
12 **Short Title:** Network Diversity of the Healthy Human Microbiome

13 **Corresponding Author Information**

14 Patrick D Schloss, PhD

15 1150 W Medical Center Dr. 1526 MSRB I

16 Ann Arbor, Michigan 48109

17 Phone: (734) 647-5801

18 Email: pschloss@umich.edu

19 **Abstract**

20 Viruses and bacteria are critical components of the human microbiome and play important roles in health and
21 disease. Most previous work has relied on studying bacteria and viruses independently, thereby reducing
22 them to two separate communities. Such approaches are unable to capture how these microbial communities
23 interact, such as through processes that maintain community robustness or allow phage-host populations
24 to co-evolve. We implemented a network-based analytical approach to describe phage-bacteria network
25 diversity throughout the human body. We built these community networks using a machine learning algorithm
26 to predict which phages could infect which bacteria in a given microbiome. Our algorithm was applied to
27 paired viral and bacterial metagenomic sequence sets from three previously published human cohorts. We
28 organized the predicted interactions into networks that allowed us to evaluate phage-bacteria connectedness
29 across the human body. We observed evidence that gut and skin network structures were person-specific
30 and not conserved among cohabitating family members. High-fat diets appeared to be associated with
31 less connected networks. Network structure differed between skin sites, with those exposed to the external
32 environment being less connected and likely more susceptible to network degradation by microbial extinction
33 events. This study quantified and contrasted the diversity of virome-microbiome networks across the human
34 body and illustrated how environmental factors may influence phage-bacteria interactive dynamics. This
35 work provides a baseline for future studies to better understand system perturbations, such as disease states,
36 through ecological networks.

37 **Author Summary**

38 The human microbiome, the collection of microbial communities that colonize the human body, is a crucial
39 component to health and disease. Two major components to the human microbiome are the bacterial and
40 viral communities. These communities have primarily been studied separately using metrics of community
41 composition and diversity. These approaches have failed to capture the complex dynamics of interacting
42 bacteria and phage communities, which frequently share genetic information and work together to maintain
43 ecosystem homestasis (e.g. kill-the-winner dynamics). Removal of bacteria or phage can disrupt or even
44 collapse those ecosystems. Relationship-based network approaches allow us to capture this interaction
45 information. Using this network-based approach with three independent human cohorts, we were able to
46 present an initial understanding of how phage-bacteria networks differ throughout the human body, so as to
47 provide a baseline for future studies of how and why microbiome networks differ in disease states.

48 Introduction

49 Viruses and bacteria are critical components of the human microbiome and play important roles in health
50 and disease. Bacterial communities have been associated with disease states, including a range of skin
51 conditions [1], acute and chronic wound healing conditions [2,3], and gastrointestinal diseases, such as
52 inflammatory bowel disease [4,5], *Clostridium difficile* infections [6] and colorectal cancer [7,8]. Altered
53 human viromes (virus communities consisting primarily of bacteriophages) also have been associated with
54 diseases and perturbations, including inflammatory bowel disease [5,9], periodontal disease [10], spread of
55 antibiotic resistance [11], and others [12–17]. Viruses act in concert with their microbial hosts as a single
56 ecological community [18]. Viruses influence their living microbial host communities through processes
57 including lysis, host gene expression modulation [19], influence on evolutionary processes such as horizontal
58 gene transfer [22] or antagonistic co-evolution [26], and alteration of ecosystem processes and elemental
59 stoichiometry [27].

60 Previous human microbiome work has focused on bacterial and viral communities, but have reduced them to
61 two separate communities by studying them independently [5,9,10,12–17]. This approach fails to capture the
62 complex dynamics of interacting bacteria and phage communities, which frequently share genetic information
63 and work together to maintain ecosystem structure (e.g. kill-the-winner dynamics that prevent domination
64 by a single bacterium). Removal of bacteria or phages can disrupt or even collapse those ecosystems
65 [18,28–37]. Integrating these datasets as relationship-based networks allow us to capture this complex
66 interaction information. Studying such bacteria-phage interactions through community-wide networks built
67 from inferred relationships begins to provide us with insights into the drivers of human microbiome diversity
68 across body sites and enable the study of human microbiome network dynamics overall.

69 In this study, we characterized human-associated bacterial and phage communities by their inferred
70 relationships using three published paired virus and bacteria-dominated whole community metagenomic
71 datasets [13,14,38,39]. We leveraged machine learning and graph theory techniques to establish
72 and explore the human bacteria-phage network diversity therein. This approach built upon previous
73 large-scale phage-bacteria network analyses by inferring interactions from metagenomic datasets, rather

74 than culture-dependent data [33], which is limited in the scale of possible experiments and analyses.
75 We implemented an adapted metagenomic interaction inference model that made some improvements
76 upon previous phage-host interaction prediction models. Previous approaches have utilized a variety of
77 techniques, such as linear models that were used to predict bacteria-phage co-occurrence using taxonomic
78 assignments [40], and nucleotide similarity models that were applied to both whole virus genomes [41] and
79 clusters of whole and partial virus genomes [42]. Our approach uniquely included protein interaction data
80 and was validated based on experimentally determined positive and negative interactions (i.e. who does
81 and does not infect whom). We built on previous modeling work as a means to our ends, and focused on the
82 biological insights we could gain instead of building a superior model and presenting our work as a toolkit.
83 We therefore did not focus on extensive benchmarking against other existing models [41,43–45]. Through
84 this approach we were able to provide an initial basic understanding of the network dynamics associated
85 with phage and bacterial communities on and in the human body. By building and utilizing a microbiome
86 network, we found that different people, body sites, and anatomical locations not only support distinct
87 microbiome membership and diversity [13,14,38,39,46–48], but also support ecological communities with
88 distinct communication structures and robustness to network degradation by extinction events. Through an
89 improved understanding of network structures across the human body, we aim to empower future studies to
90 investigate how these communities dynamics are influenced by disease states and the overall impact they
91 may have on human health.

92 **Results**

93 **Cohort Curation and Sample Processing**

94 We studied the differences in virus-bacteria interaction networks across healthy human bodies by leveraging
95 previously published shotgun sequence datasets of purified viral metagenomes (viromes) paired with
96 bacteria-dominated whole community metagenomes. Our study contained three datasets that explored
97 the impact of diet on the healthy human gut virome [14], the impact of anatomical location on the healthy

98 human skin virome [13], and the viromes of monozygotic twins and their mothers [38,39]. We selected
99 these datasets because their virome samples were subjected to virus-like particle (VLP) purification,
100 which removed contaminating DNA from human cells, bacteria, etc. To this end, the publishing authors
101 employed combinations of filtration, chloroform/DNase treatment, and cesium chloride gradients to eliminate
102 organismal DNA (e.g. bacteria, human, fungi, etc) and thereby allow for direct assessment of both the
103 extracellular and fully-assembled intracellular virome (**Supplemental Figure S1 A-B**) [14,39]. Each
104 research group reported quality control measures to ensure the purity of the virome sequence datasets,
105 using both computational and molecular techniques (e.g. 16S rRNA gene qPCR) (**Table S1**). These reports
106 confirmed that the virome libraries consisted of highly purified virus genomic DNA.

107 The bacterial and viral sequences from these studies were quality filtered and assembled into contigs. We
108 further grouped the related bacterial and phage contigs into operationally defined units based on their k-mer
109 frequencies and co-abundance patterns, similar to previous reports (**Supplemental Figure S2 - S3**) [42]. This
110 was done both for dimensionality reduction and to prevent inflation of node counts by using contigs which are
111 expected to represent multiple fragments from the same genomes. This was also done to create genome
112 analogs that we could use in our classification model which was built using genome sequences. We referred
113 to these operationally defined groups of related contigs as operational genomic units (OGUs). Each OGU
114 represented a genomically similar sub-population of either bacteria or phages. Contig lengths within clusters
115 ranged between 10^3 and $10^{5.5}$ bp (**Supplemental Figure S2 - S3**).

116 While supplementing the previous virome fraction quality control measures (**Table S1**) we found that, in
117 light of the rigorous purification and quality control during sample collection and preparation, 77% (228 /
118 298 operational genomic units) still had some nucleotide similarity to a given bacterial reference genome
119 (e-value $< 10^{-25}$). As most phages in these communities have been shown to be temperate (they integrate
120 into bacterial genomes) using methods that included nucleotide alignments of phages to bacterial reference
121 genomes, we interpreted this as confirmation that the majority of those phages were temperate and therefore
122 shared elements with bacterial reference genomes – a trend previously reported [14]. We further confirmed
123 that while the majority of these were expected to be temperate phages, there still remained a low level of

124 bacterial sequence noise, which was evident as five (1.7%) OGUs with similar sequence elements to the
125 bacterial 16S rRNA gene (blastn, e-value $< 10^{-25}$, length $> 1,000$ bp). This is in line with previous work
126 which has suggested that bacterial noise is an inevitable technical issue, and we considered this noise while
127 interpreting our findings [50]. We also identified two OGUs as being complete, high confidence phages using
128 the stringent Virsorter phage identification algorithm (class 1 confidence group) [51].

129 The whole metagenomic shotgun sequence samples, which primarily consisted of bacteria, had an average
130 viral relative abundance of 0.4% (**Table S1**) [13,14,38,39]. We found that only 2% (6 / 280 OGUs) of
131 bacterial OGUs had significantly strong nucleotide similarity to phage reference genomes (e-value $< 10^{-25}$)
132 [13,14,38,39]. No OGUs were confidently identified as lytic or temperate phage OGUs in the bacterial
133 dataset using the Virsorter algorithm [51]. Together this suggests minimal bacterial OGU noise that should
134 be considered in the study conclusions.

135 **Implementing Phage-Bacteria Interaction Prediction to Build a Community Network**

136 We predicted which phage OGUs infected which bacterial OGUs using a random forest model trained on
137 experimentally validated infectious relationships from six previous publications [41,52–56]. Only bacteria
138 and bacteriophages were used in the model. The training set contained 43 diverse bacterial species and
139 30 diverse phage strains, including both broad and specific ranges of infection (**Supplemental Figure S4**
140 **A - B**). While it is true that there are more known phages that infect bacteria, we were limited by the
141 information confirming which phages do not infect certain bacteria and attempted to keep the numbers of
142 positive and negative interactions similar. Phages with linear and circular genomes, as well as ssDNA and
143 dsDNA genomes, were included in the analysis. Because we used DNA sequencing studies, RNA phages
144 were not considered (**Supplemental Figure S4 C-D**). This training set included both positive relationships (a
145 phage infects a bacterium) and negative relationships (a phage does not infect a bacterium). This allowed
146 us to validate the false positive and false negative rates associated with our candidate models, thereby
147 building upon previous work that only considered positive relationships [41]. It is also worth noting that while
148 a positive interaction is strong evidence that the interaction exists, we must also be conscious that negative

149 interactions are only weak evidence for a lack of interaction because the finding could be the result of our
150 inability to reproduce conditions in which those interactions occur. Altogether we decided to maintain a
151 balanced dataset at the cost of under-sampling the available positive interaction information because the
152 use of such a severely unbalanced dataset often results in over-fit and uninformative model training.

153 Four phage and bacterial genomic features were used in our random forest model to predict infectious
154 relationships between bacteria and phages: 1) genome nucleotide similarities, 2) gene amino acid
155 sequence similarities, 3) bacterial Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)
156 spacer sequences that target phages, and 4) similarity of protein families associated with experimentally
157 identified protein-protein interactions [57]. These features were calculated using the training set described
158 above. We chose to utilize these metrics that directly compare nucleotide sequences between sample
159 phages and bacteria, instead of relying on alignment to reference genomes or known marker genes, because
160 we are extrapolating our model to highly diverse communities which we expect to diverge significantly from
161 the available reference genomes. The resulting random forest model was assessed using nested cross
162 validation, and the median area under its receiver operating characteristic (ROC) curve was 0.788, the
163 median model sensitivity was 0.905, and median specificity was 0.538 (**Figure 1 A**). This balance of confident
164 true positives at the cost of fewer true negatives is ideal for this type of dataset which consists of primarily
165 positive connections (**Supplemental Figure S5**). Nested cross validation of the model demonstrated that
166 the sensitivity and specificity of the model could vary but the overall model performance (AUC) remained
167 more consistent (**Supplemental Figure S6**). This suggested that our model would perform with a similar
168 overall accuracy despite changes in sensitivity/specificity trade-offs. The most important predictor in the
169 model was amino acid similarity between genes, followed by nucleotide similarity of whole genomes (**Figure**
170 **1 B**). Protein family interactions were moderately important to the model, and CRISPRs were largely
171 uninformative, due to the minimal amount of identifiable CRISPRs in the dataset and their redundancy with
172 the nucleotide similarity methods (**Figure 1 B**). Approximately one third of the training set relationships
173 yielded no score and therefore were unable to be assigned an interaction prediction (**Figure 1 C**).

174 We used our random forest model to classify the relationships between bacteria and phage operational

Figure 1: Summary of Multi-Study Network Model. (A) Median ROC curve (dark red) used to create the microbiome-virome infection prediction model, based on nested cross validation over 25 random iterations. The maximum and minimum performance are shown in light red. (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. Features include the local gene alignments between bacteria and phage genes (denoted *blastx*; the *blastx* algorithm in Diamond aligner), local genome nucleotide alignments between bacteria and phage OGUs, presence of experimentally validated protein family domains (Pfams) between phage and bacteria OGUs, and CRISPR targeting of bacteria toward phages (CRISPR). (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 automatically classified as not having interactions. (D) Network diameter (measure of graph size; the greatest number of traversed vertices required between two vertices), (E) number of vertices, and (F) number of edges (relationships) for the total network (orange) and the individual study sub-networks (diet study = red, skin study = yellow, twin study = green).

175 genomic units, which were then used to build the interactive network. The master network, analogous
176 to the universal microbiome network concept previously described [58], contained the three studies as
177 sub-networks, which themselves each contained sub-networks for each sample (**Supplemental Figure S7**).
178 Metadata including study, sample ID, disease, and OGU abundance within the community were stored in
179 the master network for parsing in downstream analyses (**Supplemental Figure S7**). The phage and bacteria
180 of the master network demonstrated both narrow broad ranges of infectious interactions (**Supplemental**
181 **Figure S8**). Bacterial and phage relative abundance was recorded in each sample for each OGU and
182 the weight of the edge connecting those OGUs was calculated as a function of those relative abundance
183 values. The separate extraction of the phage and bacterial libraries ensured a more accurate measurement
184 of the microbial communities, as has been outlined previously [59,60]. The master network was highly
185 connected and contained 72,287 infectious relationships among 578 nodes, representing 298 phages and
186 280 bacteria. Although the network was highly connected, not all relationships were present in all samples.
187 Relationships were weighted by the relative abundances of their associated bacteria and phages. Like the
188 master network, the skin network exhibited a diameter of 4 (measure of graph size; the greatest number
189 of traversed vertices required between two vertices) and included 576 (297 phages, 279 bacteria, 99.7%
190 total) and 72,127 (99.8%) of the master network nodes and edges, respectively (**Figure 1 E - F**). Additionally,
191 the subnetworks demonstrated narrow ranges of eccentricity across their nodes (**Supplemental Figure S9**).

192 The phages and bacteria in the diet and twin sample sets were more sparsely related, with the diet study
193 consisting of 89 (41 phages, 48 bacteria) nodes and 5,566 relationships, and the twin study containing 137
194 (36 phages, 101 bacteria) nodes and 17,250 relationships (**Figure 1 E - F**). As an interesting validation
195 measure, we identified five (1.7%) examples of phage OGUs which similar genomic elements to the four
196 previously described, broadly infectious phages isolated from Lake Michigan (tblastx; e-value < 10^{-25}) [61].

197 **Role of Diet on Gut Microbiome Connectivity**

198 Diet is a major environmental factor that influences resource availability and gut microbiome composition
199 and diversity, including bacteria and phages [14,62,63]. Previous work in isolated culture-based systems has
200 suggested that changes in nutrient availability are associated with altered phage-bacteria network structures
201 [30], although this has yet to be tested in humans. We therefore hypothesized that a change in diet would
202 also be associated with a change in virome-microbiome network structure in the human gut.

203 We evaluated the diet-associated differences in gut virome-microbiome network structure by quantifying how
204 central each sample's network was on average. We accomplished this by utilizing two common weighted
205 centrality metrics: degree centrality and closeness centrality. Degree centrality, the simplest centrality metric,
206 was defined as the number of connections each phage made with each bacterium. We supplemented
207 measurements of degree centrality with measurements of closeness centrality. Closeness centrality is a
208 metric of how close each phage or bacterium is to all of the other phages and bacteria in the network. A higher
209 closeness centrality suggests that the effects of genetic information or altered abundance would be more
210 impactful to all other microbes in the system. Because these are weighted metrics, the values are functions
211 of both connectivity as well as community composition. A network with higher average closeness centrality
212 also indicates an overall greater degree of connections, which suggests a greater resilience against network
213 degradation by extinction events [30,64]. This is because more highly connected networks are less likely to
214 degrade into multiple smaller networks when bacteria or phages are randomly removed [30,64]. We used
215 this information to calculate the average connectedness per sample, which was corrected for the maximum
216 potential degree of connectedness. Unfortunately our dataset was insufficiently powered to make strong

217 conclusions toward this hypothesis, but this is an interesting observation that warrants further investigation.
218 This observation also serves to illustrate the types of questions we can answer with more comprehensive
219 microbiome sampling and integrative analyses.

220 Using our small sample set, we observed that the gut microbiome network structures associated with high-fat
221 diets appeared less connected than those of low-fat diets, although a greater sample size will be required
222 to more properly evaluate this trend (**Figure 2 A-B**). Five subjects were available for use, all of which had
223 matching bacteria and virome datasets and samples from 8-10 days following the initiation of their diets.
224 High-fat diets appeared to exhibit reduced degree centrality (**Figure 2 A**), suggesting bacteria in high-fat
225 environments were targeted by fewer phages and that phage tropism was more restricted. High-fat diets
226 also appeared to exhibit decreased closeness centrality (**Figure 2 B**), indicating that bacteria and phages
227 were more distant from other bacteria and phages in the community. This would make genetic transfer and
228 altered abundance of a given phage or bacterium less capable of impacting other bacteria and phages within
229 the network.

Figure 2: 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. Each point represents the centrality from a human subject stool sample that was collected 8-10 days following the beginning of their defined diet. There are five samples here, compared to the four in figure 3, because one of the was only sampled post-diet, providing us data for this analysis but not allowing us to compare to a baseline for figure 3. (C) Quantification of average degree centrality and (D) closeness centrality between obese and healthy adult women from the Twin gut study. Each point represents a stool sample taken from one of the three adult woman confirmed as obese or healthy and with matching virus and bacteria data.

230 In addition to diet, we observed a possible trend that obesity influenced network structure. This was done
231 using the three mother samples available from the twin sample set, all of which had matching bacteria and
232 phage samples and confirmed BMI information. The obesity-associated network appeared to have a higher
233 degree centrality (**Figure 2 C**), but less closeness centrality than the healthy-associated networks (**Figure**
234 **2 D**). These results suggested that the obesity-associated networks may be less connected. This again
235 comes with the caveat that this is only an opportunistic observation using an existing sample set with too few
236 samples to make more substantial claims. We included this observation as a point of interest, given the data

237 was available.

238 **Individuality of Microbial Networks**

239 Skin and gut community membership and diversity are highly personal, with people remaining more similar
240 to themselves than to other people over time [13,65,66]. We therefore hypothesized that this personal
241 conservation extended to microbiome network structure. We addressed this hypothesis by calculating
242 the degree of dissimilarity between each subject's network, based on phage and bacteria abundance
243 and centrality. We quantified phage and bacteria centrality within each sample graph using the weighted
244 eigenvector centrality metric. This metric defines central phages as those that are highly abundant (A_O as
245 defined in the methods) and infect many distinct bacteria which themselves are abundant and infected by
246 many other phages. Similarly, bacterial centrality was defined as those bacteria that were both abundant
247 and connected to numerous phages that were themselves connected to many bacteria. We then calculated
248 the similarity of community networks using the weighted eigenvector centrality of all nodes between all
249 samples. Samples with similar network structures were interpreted as having similar capacities for network
250 robustness and transmitting genetic material.

251 We used this network dissimilarity metric to test whether microbiome network structures were more similar
252 within people than between people over time. We found that gut microbiome network structures clustered by
253 person (ANOSIM p-value = 0.005, R = 0.958, **Figure 3 A**). Network dissimilarity within each person over the
254 8-10 day sampling period was less than the average dissimilarity between that person and others, although
255 this difference was not statistically significant (p-value = 0.125, **Figure 3 B**). Four of the five available subjects
256 were used because one of the subjects was not sampled at the initial time point. The lack of statistical
257 confidence was likely due to the small sample size of this dataset.

258 Although there was evidence for gut network conservation among individuals, we found no evidence for
259 conservation of gut network structures within families. The gut network structures were not more similar
260 within families (twins and their mothers; intrafamily) compared to other families (other twins and mothers;
261 inter-family) (p-value = 0.312, **Figure 3 C**). In addition to the gut, skin microbiome network structure was

262 conserved within individuals (p -value < 0.001 , **Figure 3 D**). This distribution was similar when separated by
263 anatomical sites. Most sites were statistically significantly more conserved within individuals (**Supplemental**
264 **Figure S10**).

Figure 3: 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 colored 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. Only four subjects were included here, compared to the five used in figure 2, because one of the subjects was missing the initial sampling time point and therefore lacked temporal sampling.* (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 p -value is provided near the bottom of the figure.* (C) *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). Each point represents the inter-family and intra-family dissimilarity of a twin or mother that was sampled over time.* (D) *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). All p -values were calculated using a paired Wilcoxon test.*

265 **Network Structures Across the Human Skin Landscape**

266 Extensive work has illustrated differences in diversity and composition of the healthy human skin microbiome
267 between anatomical sites, including bacteria, virus, and fungal communities [13,47,65]. These communities
268 vary by degree of skin moisture, oil, and environmental exposure; features which were defined in the original
269 publication [13]. As viruses are known to influence microbial diversity and community composition, we
270 hypothesized that these differences would still be evident after integrating the bacterial and viral datasets
271 and evaluating their microbe-virus network structure between anatomical sites. To test this, we evaluated
272 the changes in network structure between anatomical sites within the skin dataset. The anatomical sites and
273 their features (e.g. moisture & occlusion) were defined in the previous publication through consultation with
274 dermatologists and reference to previous literature [13].

275 The average centrality of each sample was quantified using the weighted eigenvector centrality metric.
276 Intermittently moist skin sites (dynamic sites that fluctuate between being moist and dry) were significantly

277 less connected than the moist and sebaceous environments (p -value < 0.001 , **Figure 4 A**). Also, skin sites
278 that were occluded from the environment were much more highly connected than those that were constantly
279 exposed to the environment or only intermittently occluded (p -value < 0.001 , **Figure 4 B**).

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 $1.58 * IQR / \sqrt{n}$. Sample sizes for each group were: Moist = 81, Sebaceous = 56, IntMoist = 56, Occluded = 106, Exposed = 61, IntOccluded = 26. (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 3. (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.

280 To supplement this analysis, we compared the network signatures using the centrality dissimilarity approach
281 described above. The dissimilarity between samples was a function of shared relationships, degree of
282 centrality, and bacteria/phage abundance. When using this supplementary approach, we found that network
283 structures significantly clustered by moisture, sebaceous, and intermittently moist status (**Figure 4 C,E**).
284 Occluded sites were significantly different from exposed and intermittently occluded sites, but there was no
285 difference between exposed and intermittently occluded sites (**Figure 4 D,F**). These findings provide further
286 support that skin microbiome network structure differs significantly between skin sites.

287 Discussion

288 Foundational work has provided a baseline understanding of the human microbiome by characterizing
289 bacterial and viral diversity across the human body [13,14,46–48,67]. Here we integrated the bacterial and
290 viral sequence sets to offer an initial understanding of how phage-bacteria networks differ throughout the
291 human body, so as to provide a baseline for future studies of how and why microbiome networks differ in
292 disease states. We implemented a network-based analytical model to evaluate the basic properties of the

293 human microbiome through bacteria and phage relationships, instead of membership or diversity alone.
294 This approach enabled the application of network theory to provide a new perspective while analyzing
295 bacterial and viral communities simultaneously. We utilized metrics of connectivity to model the extent to
296 which communities of bacteria and phages interact through mechanisms such as horizontal gene transfer,
297 modulated bacterial gene expression, and alterations in abundance.

298 Just as gut microbiome and virome composition and diversity are conserved in individuals [13,46,47,66], gut
299 and skin microbiome network structures were conserved within individuals over time. Gut network structure
300 was not conserved among family members. These findings suggested that the community properties inferred
301 from microbiome interaction network structures, such as robustness (meaning a more highly connected
302 network is more “robust” to network degradation because a randomly removed bacteria or phage node is less
303 likely to divide or disintegrate [30,64] the overall network), the potential for horizontal gene transfer between
304 members, and co-evolution of populations, were person-specific. These properties may be impacted by
305 personal factors ranging from the body’s immune system to external environmental conditions, such as
306 climate and diet.

307 We observed evidence supporting the ability of environmental conditions to shape gut and skin microbiome
308 interaction network structure by observing that diet and skin location were associated with altered network
309 structures. We observed evidence that diet was sufficient to alter gut microbiome network connectivity,
310 although this needs to be interpreted cautiously as a case observation, due to the small sample size. Although
311 the available sample size was small, our findings provide some preliminary evidence that high-fat diets are
312 less connected than low-fat diets and that high-fat diets may therefore lead to less robust communities
313 with a decreased ability for microbes to directly influence one another. We supported this finding with the
314 observation that obesity may have been associated with decreased network connectivity. Together these
315 findings suggest the food we eat may not only impact which microbes colonize our guts, but may also impact
316 their interactions with infecting phages. Further work will be required to characterize these relationships with
317 a larger cohort.

318 In addition to diet, the skin environment also influenced the microbiome interaction network structure.

319 Network structure differed between environmentally exposed and occluded skin sites. The sites under
320 greater environmental fluctuation and exposure (the exposed and intermittently exposed sites) were less
321 connected and therefore were predicted to have a higher resilience against network degradation when
322 random nodes are removed from the network. Likewise, intermittently moist sites demonstrated less
323 connectedness than the moist and sebaceous sites. These findings agree with previous work that has
324 shown that bacterial community networks differ by skin environment types [58]. Together these data
325 suggested that body sites under greater degrees of fluctuation harbored less connected microbiomes
326 that are potentially less robust to network disruption by extinction events. This points to a link between
327 microbiome and environmental robustness toward network homeostasis and warrants further investigation.

328 While these findings take us an important step closer to understanding the microbiome through interspecies
329 relationships, there are caveats to and considerations regarding our findings. First, as with most classification
330 models, the infection classification model developed and applied is only as good as its training set – in this
331 case, the collection of experimentally-verified positive and negative infection data. Large-scale experimental
332 screens for phage and bacteria infectious interactions that report high-confidence negative interactions (i.e.,
333 no infection) are desperately needed, as they would provide more robust model training and improved model
334 performance. Furthermore, just as we have improved on previous modeling efforts, we expect that new and
335 creative scoring metrics will improve future performance. Other creative and high performing models are
336 currently being developed and the applications of these models to community network creation will continue
337 to move this field forward [43–45].

338 Second, although our analyses utilized the best datasets currently available for our study, this work was done
339 retrospectively and relied on existing data up to seven years old. These archived datasets were limited by
340 the technology and costs of the time. For example, the diet and twin studies, relied on multiple displacement
341 amplification (MDA) in their library preparations—an approach used to overcome the large nucleic acids
342 requirements typical of older sequencing library generation protocols. It is now known that MDA results
343 in biases in microbial community composition [68], as well as toward ssDNA viral genomes [69,70], thus
344 rendering the resulting microbial and viral metagenomes largely non-quantitative. Future work that employs

345 larger sequence datasets and that avoids the use of bias-inducing amplification steps will build on and validate
346 our findings, as well as inform the design and interpretation of further studies.

347 Although our models demonstrated satisfactory accuracy and overall performance, it was important to
348 interpret our findings under the realization that our model was not perfect. This caveat is not new to the
349 microbiome field, with a notable example being the use of 16S rRNA sequencing using the V4 variable
350 region [60]. Use of the V4 variable region excluded detection of major bacterial members, meaning that the
351 findings were not able to completely describe the underlying biological environment. Despite this caveat,
352 skin microbiome studies provided valuable biological insights by focusing on the community differences
353 between groups (e.g. disease and healthy) which were both analyzed the same way. Similarly, here we
354 focused our conclusions on the differences between the groups which were all treated the same, so that we
355 can minimize our dependence on a perfect predictive model.

356 Finally, the networks in this study were built using operational genomic units (OGUs), which represented
357 groups of highly similar bacteria or phage genomes or clustered genome fragments. Similar clustering
358 definition and validation methods, both computational and experimental, have been implemented in other
359 metagenomic sequencing studies, as well [42,71–73]. These approaches could offer yet another level of
360 sophistication to our network-based analyses. While this operationally defined clustering approach allows
361 us to study whole community networks, our ability to make conclusions about interactions among specific
362 phage or bacterial species or populations is inherently limited, compared to more focused, culture-based
363 studies such as the work by Malki *et al* [61]. Future work must address this limitation, e.g., through improved
364 binning methods and deeper metagenomic shotgun sequencing, but most importantly through an improved
365 conceptual framing of what defines ecologically and evolutionarily cohesive units for both phage and bacteria
366 [74]. Defining operational genomic units and their taxonomic underpinnings (e.g., whether OGU clusters
367 represent genera or species) is an active area of work critical to the utility of this approach. As a first
368 step, phylogenomic analyses have been performed to cluster cyanophage isolate genomes into informative
369 groups using shared gene content, average nucleotide identity of shared genes, and pairwise differences
370 between genomes [75]. Such population-genetic assessment of phage evolution, coupled with the ecological

371 implications of genome heterogeneity, will inform how to define nodes in future iterations of the ecological
372 network developed here. Even though we are hesitant to speculate on phage host ranges at low taxonomic
373 levels in our dataset, the data does agree with previous reports of instances of broad phage host range
374 [61,76]. Additionally, visualization of our dataset interactions using the heat map approach previously used
375 in other host range studies, suggests a trend toward modular and nested tropism, but we do not have the
376 strain-level resolution that powered those previous experimental studies.

377 Together our work takes an initial step towards defining bacteria-virus interaction profiles as a characteristic
378 of human-associated microbial communities. This approach revealed the impacts that different human
379 environments (e.g., the skin and gut) can have on microbiome connectivity. By focusing on relationships
380 between bacterial and viral communities, they are studied as the interacting cohorts they are, rather than
381 as independent entities. While our developed bacteria-phage interaction framework is a novel conceptual
382 advance, the microbiome also consists of archaea and small eukaryotes, including fungi and *Demodex* mites
383 [1,77] – all of which can interact with human immune cells and other non-microbial community members [78].
384 Future work will build from our approach and include these additional community members and their diverse
385 interactions and relationships (e.g., beyond phage-bacteria). This will result in a more robust network and a
386 more holistic understanding of the evolutionary and ecological processes that drive the assembly and function
387 of the human-associated microbiome.

388 **Materials & Methods**

389 **Code Availability**

390 A reproducible version of this manuscript written in R markdown and all of the code used to obtain and
391 process the sequencing data is available at the following GitHub repository:

392 https://github.com/SchlossLab/Hannigan_ConjunctisViribus_ploscompbio_2017

393 **Data Acquisition & Quality Control**

394 Raw sequencing data and associated metadata were acquired from the NCBI sequence read archive (SRA).
395 Supplementary metadata were acquired from the same SRA repositories and their associated manuscripts.
396 The gut virome diet study (SRA: SRP002424), twin virome studies (SRA: SRP002523; SRP000319), and
397 skin virome study (SRA: SRP049645) were downloaded as .sra files. For clarity, the sample sizes used
398 for each study subset were described with the data in the results section. Sequencing files were converted
399 to fastq format using the fastq-dump tool of the NCBI SRA Toolkit (v2.2.0). Sequences were quality
400 trimmed using the Fastx toolkit (v0.0.14) to exclude bases with quality scores below 33 and shorter than 75
401 bp [79]. Paired end reads were filtered to exclude sequences missing their corresponding pair using the
402 `get_trimmed_pairs.py` script available in the source code.

403 **Contig Assembly**

404 Contigs were assembled using the Megahit assembly program (v1.0.6) [80]. A minimum contig length of 1
405 kb was used. Iterative k-mer stepping began at a minimum length of 21 and progressed by 20 until 101. All
406 other default parameters were used.

407 **Contig Abundance Calculations**

408 Contigs were concatenated into two master files prior to alignment, one for bacterial contigs and one for
409 phage contigs. Sample sequences were aligned to phage or bacterial contigs using the Bowtie2 global aligner
410 (v2.2.1) [81]. We defined a mismatch threshold of 1 bp and seed length of 25 bp. Sequence abundance was
411 calculated from the Bowtie2 output using the `calculate_abundance_from_sam.pl` script available in
412 the source code.

413 **Operational Genomic Unit Binning**

414 Contigs often represent large fragments of genomes. In order to reduce redundancy and the resulting
415 artificially inflated genomic richness within our dataset, it was important to bin contigs into operational
416 units based on their similarity. This approach is conceptually similar to the clustering of related 16S rRNA
417 sequences into operational taxonomic units (OTUs), although here we are clustering contigs into operational
418 genomic units (OGUs) [67].

419 Contigs were clustered using the CONCOCT algorithm (v0.4.0) [82]. Because of our large dataset and limits
420 in computational efficiency, we randomly subsampled the dataset to include 25% of all samples, and used
421 these to inform contig abundance within the CONCOCT algorithm. CONCOCT was used with a maximum
422 of 500 clusters, a k-mer length of four, a length threshold of 1 kb, 25 iterations, and exclusion of the total
423 coverage variable.

424 OGU abundance (A_O) was obtained as the sum of the abundance of each contig (A_j) associated with that
425 OGU. The abundance values were length corrected such that:

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

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

427 **Operational Genomic Unit Identification**

428 To confirm a lack of phage sequences in the bacterial OGU dataset, we performed blast nucleotide alignment
429 of the bacterial OGU representative sequences using an e-value $< 10^{-25}$, which was stricter than the 10^{-10}
430 threshold used in the random forest model below, against all of the phage reference genomes available in
431 the EMBL database. We used a stricter threshold because we know there are genomic similarities between
432 bacteria and phage OGUs from the interactive model, but we were interested in contigs with high enough
433 similarity to references that they may indeed be from phages. We also performed the converse analysis

434 of aligning phage OGU representative sequences to EMBL bacterial reference genomes. Finally, we ran
435 both the phage and bacteria OGU representative sequences through the Virsorter program (1.0.3) to identify
436 phages (all default parameters were used), using only those in the high confidence identification category
437 “class 1” [51].

438 **Open Reading Frame Prediction**

439 Open reading frames (ORFs) were identified using the Prodigal program (V2.6.2) with the meta mode
440 parameter and default settings [83].

441 **Classification Model Creation and Validation**

442 The classification model for predicting interactions was built using experimentally validated bacteria-phage
443 infections or validated lack of infections from six studies [41,52–56]. No further reference databases were
444 used in our alignment procedures. Associated reference genomes were downloaded from the European
445 Bioinformatics Institute (see details in source code). The model was created based on the four metrics listed
446 below.

447 The four scores were used as parameters in a random forest model to classify bacteria and bacteriophage
448 pairs as either having infectious interactions or not. The classification model was built using the Caret
449 R package (v6.0.73) [84]. The model was trained using five-fold cross validation with ten repeats, and
450 the median model performance was evaluated by training the model on 80% of the dataset and testing
451 performance on the remaining 20%. Pairs without scores were classified as not interacting. The model was
452 optimized using the ROC value. The resulting model performance was plotted using the plotROC R package.

453 **Identify Bacterial CRISPRs Targeting Phages**

454 Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were identified from bacterial
455 genomes using the PilerCR program (v1.06) [85]. Resulting spacer sequences were filtered to exclude

456 spacers shorter than 20 bp and longer than 65 bp. Spacer sequences were aligned to the phage genomes
457 using the nucleotide BLAST algorithm with default parameters (v2.4.0) [86]. The mean percent identity for
458 each matching pair was recorded for use in our classification model.

459 **Detect Matching Prophages within Bacterial Genomes**

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

464 **Identify Shared Genes Between Bacteria and Phages**

465 As a result of gene transfer or phage genome integration during infection, phages may share genes with
466 their bacterial hosts, providing us with evidence of phage-host pairing. We identified shared genes between
467 bacterial and phage genomes by assessing amino acid similarity between the genes using the Diamond
468 protein alignment algorithm (v0.7.11.60) [87]. The mean alignment bitscores for each genome pair were
469 recorded for use in our classification model.

470 **Protein - Protein Interactions**

471 The final method used for predicting infectious interactions between bacteria and phages was the detection
472 of pairs of genes whose proteins are known to interact. We assigned bacterial and phage genes to protein
473 families by aligning them to the Pfam database using the Diamond protein alignment algorithm. We then
474 identified which pairs of proteins were predicted to interact using the Pfam interaction information within the
475 Intact database [57]. The mean bitscores of the matches between each pair were recorded for use in the
476 classification model.

477 Interaction Network Construction

478 The bacteria and phage operational genomic units (OGUs) were scored using the same approach as outlined
479 above. The infectious pairings between bacteria and phage OGUs were classified using the random forest
480 model described above. The predicted infectious pairings and all associated metadata were used to populate
481 a graph database using Neo4j graph database software (v2.3.1) [88]. This network was used for downstream
482 community analysis.

483 Centrality Analysis

484 We quantified the centrality of graph vertices using three different metrics, each of which provided different
485 information graph structure. When calculating these values, let $G(V, E)$ be an undirected, unweighted graph
486 with $|V| = n$ nodes and $|E| = m$ edges. Also, let \mathbf{A} be its corresponding adjacency matrix with entries
487 $a_{ij} = 1$ if nodes V_i and V_j are connected via an edge, and $a_{ij} = 0$ otherwise.

488 Briefly, the **closeness centrality** of node V_i is calculated taking the inverse of the average length of the
489 shortest paths (d) between nodes V_i and all the other nodes V_j . Mathematically, the closeness centrality of
490 node V_i is given as:

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

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

493 Intuitively, the **degree centrality** of node V_i is defined as the number of edges that are incident to that node:

$$C_D(V_i) = \sum_{j=1}^n a_{ij}$$

494 where a_{ij} is the ij^{th} entry in the adjacency matrix \mathbf{A} .

495 The eigenvector centrality of node V_i is defined as the i^{th} value in the first eigenvector of the associated
496 adjacency matrix \mathbf{A} . Conceptually, this function results in a centrality value that reflects the connections of
497 the vertex, as well as the centrality of its neighboring vertices.

498 The **centralization** metric was used to assess the average centrality of each sample graph \mathbf{G} . Centralization
499 was calculated by taking the sum of each vertex V_i 's centrality from the graph maximum centrality C_w , such
500 that:

$$C(G) = \frac{\sum_{i=1}^n C_w - c(V_i)}{T}$$

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

503 Degree and closeness centrality were calculated using the associated functions within the igraph R package
504 (v1.0.1) [89].

505 **Network Relationship Dissimilarity**

506 We assessed similarity between graphs by evaluating the shared centrality of their vertices, as has been
507 done previously. More specifically, we calculated the dissimilarity between graphs G_i and G_j using the
508 Bray-Curtis dissimilarity metric and eigenvector centrality values such that:

$$B(G_i, G_j) = 1 - \frac{2C_{ij}}{C_i + C_j}$$

509 Where C_{ij} is the sum of the lesser centrality values for those vertices shared between graphs, and C_i and
510 C_j are the total number of vertices found in each graph. This allows us to calculate the dissimilarity between
511 graphs based on the shared centrality values between the two graphs.

512 **Statistics and Comparisons**

513 Differences in intrapersonal and interpersonal network structure diversity, based on multivariate data,
514 were calculated using an analysis of similarity (ANOSIM). Statistical significance of univariate Eigenvector
515 centrality differences were calculated using a paired Wilcoxon test.

516 Statistical significance of differences in univariate eigenvector centrality measurements of skin virome-microbiome
517 networks were calculated using a pairwise Wilcoxon test, corrected for multiple hypothesis tests using the
518 Holm correction method. Multivariate eigenvector centrality was measured as the mean differences between
519 cluster centroids, with statistical significance measured using an ANOVA and post hoc Tukey test.

520 **Acknowledgments**

521 We thank the members of the Schloss lab for their underlying contributions. We thank the authors of the
522 original studies for making their data and metadata publicly available and understandable. We also thank
523 the participants in the studies.

524 **Author Contributions**

525 *Conceptualization:* GDH, MBD, DK, PDS. *Data Curation:* GDH. *Formal Analysis:* GDH. *Funding Acquisition:*
526 GDH, PDS. *Writing – Original Draft Preparation:* GDH, PDS. *Writing – Review & Editing:* GDH, MBD, DK,
527 PDS.

528 **Funding Information**

529 GDH was supported in part by the Molecular Mechanisms in Microbial Pathogenesis Training Program (T32
530 AI007528). GDH and PDS were supported in part by funding from the NIH (P30DK034933, U19AI09087,

531 and U01AI124255).

532 **Competing interests**

533 The authors report no conflicts of interest.

534 **Supplemental Figure Captions**

Figure S1: **Sequencing Depth Summary.** *Number of sequences that aligned to (A) Phage and (B) Bacteria operational genomic units per sample and colored by study.*

Figure S2: **Contig Summary Statistics.** *Scatter plot heat map with each hexagon representing the 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) Quantification of bacterial host strains known to exist for each phage. C) Genome strandedness and D) linearity of the phage reference genomes used for the dataset.*

Figure S5: Classification Model Performance By Nested Cross-Validation. *Box plot illustrating the median and variance of phage-bacteria interaction prediction model. Performance was evaluated using nested cross validation, meaning that 20% of the samples were randomly withheld from model training and then used to evaluate performance. The results of 100 random iterations are shown. Metrics include area under the curve (gray), sensitivity (red), and specificity (tan).*

Figure S6: Stable Classification Model Performance Over Random Iterations *In addition to nested cross-validation, here we show the results from the five-fold cross validation, in which 20% of the samples were randomly withheld during the training stage for model evaluation and mtry tuning. The results of 25 random iterations are shown. Metrics include area under the curve (red), sensitivity (green), and specificity (blue). Dashed line highlight the random point of 0.5.*

Figure S7: 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 S8: Heatmap of Phage-Bacteria Interaction Relationships of Master Network. *Heatmap illustrating the ranges of infectious interactions predicted between bacteria and bacteriophages across our three studies. Bacterial OGUs are aligned on the vertical access, and the bacteriophage OGUs are organized on the horizontal access. OGUs are organized near other OGUs with similar infectious profiles, which are further illustrated by the dendrograms. Predicted infections are tan and predicted lacks of interactions are red.*

Figure S9: Distribution of node eccentricity across subnetworks. *Histograms illustrating the distributions of node eccentricity values across the subnetworks, for supplementing the node, edge, and diameter values provided for the networks. Eccentricity of each node is the shortest distance of that node to the furthest other node within the graph.*

Figure S10: 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]). P-value was calculated using a paired Wilcoxon test.*

535 **Supplemental Table Captions**

536 Table S1: Summary of the primary quality control measures reported in the original publications of the viromes
537 used in this current study.

538 **References**

- 539 1. Hannigan GD, Grice EA. Microbial Ecology of the Skin in the Era of Metagenomics and Molecular
540 Microbiology. *Cold Spring Harbor Perspectives in Medicine*. 2013;3: a015362–a015362.
- 541 2. Hannigan GD, Hodkinson BP, McGinnis K, Tyldsley AS, Anari JB, Horan AD, et al. Culture-independent
542 pilot study of microbiota colonizing open fractures and association with severity, mechanism, location, and
543 complication from presentation to early outpatient follow-up. *Journal of Orthopaedic Research*. 2014;32:
544 597–605.
- 545 3. Loesche M, Gardner SE, Kalan L, Horwinski J, Zheng Q, Hodkinson BP, et al. Temporal stability in chronic
546 wound microbiota is associated with poor healing. *Journal of Investigative Dermatology*. 2016;
- 547 4. He Q, Li X, Liu C, Su L, Xia Z, Li X, et al. Dysbiosis of the fecal microbiota in the TNBS-induced Crohn's
548 disease mouse model. *Applied Microbiology and Biotechnology*. 2016; 1–10.
- 549 5. Norman JM, Handley SA, Baldrige MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in
550 the enteric virome in inflammatory bowel disease. *Cell*. 2015;160: 447–460.
- 551 6. Seekatz AM, Rao K, Santhosh K, Young VB. Dynamics of the fecal microbiome in patients with recurrent
552 and nonrecurrent *Clostridium difficile* infection. *Genome medicine*. 2016;8: 47.
- 553 7. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for
554 colorectal cancer. *Cancer prevention research (Philadelphia, Pa)*. 2014;7: 1112–1121.
- 555 8. Baxter NT, Zackular JP, Chen GY, Schloss PD. Structure of the gut microbiome following colonization with
556 human feces determines colonic tumor burden. *Microbiome*. 2014;2: 20.
- 557 9. Manrique P, Bolduc B, Walk ST, Oost J van der, Vos WM de, Young MJ. Healthy human gut phageome.
558 *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 201601060.
- 559 10. Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T, et al. Altered Oral

- 560 Viral Ecology in Association with Periodontal Disease. *mBio*. 2014;5: e01133–14–e01133–14.
- 561 11. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and
562 ecological network of the phage metagenome. *Nature*. 2013;499: 219–222.
- 563 12. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, et al. Altered Virome and
564 Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome.
565 *Cell Host and Microbe*. 2016;19: 311–322.
- 566 13. Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, et al. The Human Skin
567 Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic
568 Associations with the Host Microbiome. *mBio*. 2015;6: e01578–15.
- 569 14. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: Inter-individual
570 variation and dynamic response to diet. *Genome Research*. 2011;21: 1616–1625.
- 571 15. Santiago-Rodriguez TM, Ly M, Bonilla N, Pride DT. The human urine virome in association with urinary
572 tract infections. *Frontiers in Microbiology*. 2015;6: 14.
- 573 16. Abeles SR, Ly M, Santiago-Rodriguez TM, Pride DT. Effects of Long Term Antibiotic Therapy on Human
574 Oral and Fecal Viromes. *PLOS ONE*. 2015;10: e0134941.
- 575 17. Abeles SR, Robles-Sikisaka R, Ly M, Lum AG, Salzman J, Boehm TK, et al. Human oral viruses are
576 personal, persistent and gender-consistent. 2014; 1–15.
- 577 18. Haerter JO, Mitarai N, Sneppen K. Phage and bacteria support mutual diversity in a narrowing staircase
578 of coexistence. *The ISME Journal*. 2014;8: 2317–2326.
- 579 19. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. Photosynthesis genes in marine viruses yield
580 proteins during host infection. *Nature*. 2005;438: 86–89.
- 581 20. Tyler JS, Beerli K, Reynolds JL, Alteri CJ, Skinner KG, Friedman JH, et al. Prophage induction is
582 enhanced and required for renal disease and lethality in an EHEC mouse model. *PLoS Pathogens*. 2013;9:

583 e1003236.

584 21. Hargreaves KR, Kropinski AM, Clokie MR. Bacteriophage behavioral ecology: How phages alter their
585 bacterial host's habits. *Bacteriophage*. 2014;4: e29866.

586 22. Moon BY, Park JY, Hwang SY, Robinson DA, Thomas JC, Fitzgerald JR, et al. Phage-mediated horizontal
587 transfer of a *Staphylococcus aureus* virulence-associated genomic island. *Scientific Reports*. 2015;5: 9784.

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

590 24. Ogg JE, Timme TL, Alemohammad MM. General Transduction in *Vibrio cholerae*. *Infection and Immunity*.
591 1981;31: 737–741.

592 25. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source
593 evolution. *Nature Reviews Microbiology*. 2005;3: 722–732.

594 26. Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and evolutionary
595 processes in microbial communities. *FEMS Microbiology Reviews*. 2014;38: 916–931.

596 27. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. The elemental composition of virus particles:
597 implications for marine biogeochemical cycles. *Nature Reviews Microbiology*. 2014;12: 519–528.

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

600 29. Middelboe M, Hagström A, Blackburn N, Sinn B, Fischer U, Borch NH, et al. Effects of Bacteriophages on
601 the Population Dynamics of Four Strains of Pelagic Marine Bacteria. *Microbial Ecology*. 2001;42: 395–406.

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

604 31. Thompson RM, Brose U, Dunne JA, Hall RO, Hladyz S, Kitching RL, et al. Food webs: reconciling the

- 605 structure and function of biodiversity. *Trends in ecology & evolution*. 2012;27: 689–697.
- 606 32. Moebus K, Nattkemper H. Bacteriophage sensitivity patterns among bacteria isolated from marine waters.
607 *Helgoländer Meeresuntersuchungen*. 1981;34: 375–385.
- 608 33. Flores CO, Valverde S, Weitz JS. Multi-scale structure and geographic drivers of cross-infection within
609 marine bacteria and phages. *The ISME Journal*. 2013;7: 520–532.
- 610 34. Poisot T, Canard E, Mouillot D, Mouquet N, Gravel D. The dissimilarity of species interaction networks.
611 *Ecology letters*. 2012;15: 1353–1361.
- 612 35. Poisot T, Stouffer D. How ecological networks evolve. *bioRxiv*. 2016;
- 613 36. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions.
614 *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108: E288–97.
- 615 37. Jover LF, Flores CO, Cortez MH, Weitz JS. Multiple regimes of robust patterns between network structure
616 and biodiversity. *Scientific Reports*. 2015;5: 17856.
- 617 38. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota
618 of monozygotic twins and their mothers. *Nature*. 2010;466: 334–338.
- 619 39. Turnbaugh PJ, Hamady M, Yatsunencko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome
620 in obese and lean twins. *Nature*. 2009;457: 480–484.
- 621 40. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Ocean plankton. Determinants
622 of community structure in the global plankton interactome. *Science*. 2015;348: 1262073–1262073.
- 623 41. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host
624 relationships. *FEMS Microbiology Reviews*. 2015;40: 258–272.
- 625 42. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential
626 biogeochemical impacts of globally abundant ocean viruses. *Nature*. 2016;537: 689–693.
- 627 43. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, et al. HostPhinder: A Phage Host

628 Prediction Tool. *Viruses*. 2016;8: 116.

629 44. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WlsH: who is the host? Predicting prokaryotic hosts
630 from metagenomic phage contigs. *Bioinformatics*. 2017;33: 3113–3114.

631 45. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity
632 measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids*
633 *Research*. 2017;45: 39–53.

634 46. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and Temporal
635 Diversity of the Human Skin Microbiome. *Science*. 2009;324: 1190–1192.

636 47. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, et al. Topographic diversity of fungal and
637 bacterial communities in human skin. *Nature*. 2013; 1–6.

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

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

641 50. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. Assessment of viral community functional potential
642 from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology*. 2013;3:
643 130160–130160.

644 51. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data.
645 *PeerJ*. 2015;3: e985–20.

646 52. Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, Nickerson KW, et al. Prevalence
647 of broad-host-range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas*
648 *aeruginosa*. *Applied and Environmental Microbiology*. 1998;64: 575–580.

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

651 54. Schwarzer D, Buettner FFR, Browning C, Nazarov S, Rabsch W, Bethe A, et al. A multivalent adsorption

- 652 apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis.
653 *Journal of Virology*. 2012;86: 10384–10398.
- 654 55. Kim S, Rahman M, Seol SY, Yoon SS, Kim J. *Pseudomonas aeruginosa* bacteriophage PA1Ø
655 requires type IV pili for infection and shows broad bactericidal and biofilm removal activities. *Applied and*
656 *Environmental Microbiology*. 2012;78: 6380–6385.
- 657 56. Matsuzaki S, Tanaka S, Koga T, Kawata T. A Broad-Host-Range Vibriophage, KVP40, Isolated from Sea
658 Water. *Microbiology and Immunology*. 1992;36: 93–97.
- 659 57. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct
660 project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids*
661 *Research*. 2014;42: D358–63.
- 662 58. Bashan A, Gibson TE, Friedman J, Carey VJ, Weiss ST, Hohmann EL, et al. Universality of human
663 microbial dynamics. *Nature*. 2016;534: 259–262.
- 664 59. Kleiner M, Hooper LV, Duerkop BA. Evaluation of methods to purify virus-like particles for metagenomic
665 sequencing of intestinal viromes. *BMC Genomics*. 2015;16: 7.
- 666 60. Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q, et al. Skin microbiome
667 surveys are strongly influenced by experimental design. *Journal of Investigative Dermatology*. 2016;
- 668 61. Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, et al. Bacteriophages isolated from Lake
669 Michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal*. 2015;12: 164.
- 670 62. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human
671 gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*.
672 2009;1: 6ra14–6ra14.
- 673 63. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and
674 reproducibly alters the human gut microbiome. *Nature*. 2014;505: 559–563.
- 675 64. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks.

676 Nature. 2000;407: 651–654.

677 65. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and Temporal
678 Diversity of the Human Skin Microbiome. Science. 2009;324: 1190–1192.

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

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

683 68. Yilmaz S, Allgaier M, Hugenholtz P. Multiple displacement amplification compromises quantitative
684 analysis of metagenomes. Nature Methods. 2010;7: 943–944.

685 69. Kim KH, Chang HW, Nam YD, Roh SW. Amplification of uncultured single-stranded DNA viruses from
686 rice paddy soil. Applied and 2008;

687 70. Kim K-H, Bae J-W. Amplification methods bias metagenomic libraries of uncultured single-stranded and
688 double-stranded DNA viruses. Applied and Environmental Microbiology. 2011;77: 7663–7668.

689 71. Minot S, Wu GD, Lewis JD, Bushman FD. Conservation of gene cassettes among diverse viruses of the
690 human gut. PLOS ONE. 2012;7: e42342.

691 72. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging
692 reveals discrete populations in Synechococcus viral genome sequence space. Nature. 2014;513: 242–245.

693 73. Brum JR, Ignacio-Espinoza JC, Roux S, Doucier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns
694 and ecological drivers of ocean viral communities. Science. 2015;348: 1261498–1261498.

695 74. Polz MF, Hunt DE, Preheim SP, Weinreich DM. Patterns and mechanisms of genetic and phenotypic
696 differentiation in marine microbes. Philosophical Transactions of the Royal Society B: Biological Sciences.
697 2006;361: 2009–2021.

698 75. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, et al. Genomic
699 differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics.

700 2016;17: 930.

701 76. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al.

702 Uncovering Earth's virome. *Nature*. 2016;

703 77. Grice EA, Segre JA. The skin microbiome. *Nature Reviews Microbiology*. 2011;9: 244–253.

704 78. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and

705 disease. *Nature reviews Immunology*. 2009;9: 313–323.

706 79. Hannon GJ. FASTX-Toolkit. 2010; GNU Affero General Public License.

707 80. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable

708 metagenome assembler driven by advanced methodologies and community practices. *METHODS*.

709 2016;102: 3–11.

710 81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9:

711 357–359.

712 82. Alneberg J, Bjarnason BS, Ariani A, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic

713 contigs by coverage and composition. *Nature Methods*. 2014; 1–7.

714 83. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in

715 metagenomic sequences. *Bioinformatics*. 2012;28: 2223–2230.

716 84. Kuhn M. caret: Classification and Regression Training. CRAN. 2016;

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

718 18.

719 86. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and

720 applications. *BMC Bioinformatics*. 2009;10: 1.

721 87. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*.

722 2015;12: 59–60.

723 88. Neo Technology, Inc. Neo4j. 2017;

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

725 2006;Complex Systems: 1695.







