

1 Gene networks provide a high-resolution view of bacteriophage ecology

2 Jason W. Shapiro^{1,2,3,*}, and Catherine Putonti^{1,2,3,4}

3 ¹ Department of Biology, Loyola University of Chicago, Chicago, IL, United States of America

4 ² Department of Computer Science, Loyola University of Chicago, Chicago, IL, United States of
5 America

6 ³ Bioinformatics Program, Loyola University of Chicago, Chicago, IL, United States of America

7 ⁴ Department of Microbiology and Immunology, Loyola University of Chicago, Maywood, IL, United
8 States of America

9 *corresponding author email: jshapiro2@luc.edu

10

11 **Abstract**

12 Bacteriophages are the most abundant and diverse biological entities on the planet, and new phage
13 genomes are being discovered at a rapid pace from metagenomes. As more novel, uncultured phage
14 genomes are published, new tools are needed for placing these genomes in an ecological and
15 evolutionary context. Phages are difficult to study with phylogenetic methods, because they exchange
16 genes regularly, and no single gene is conserved across all phages. Instead, genome-level networks
17 have been used to group similar viruses into clusters for taxonomy. Here, we show that gene-level
18 networks provide a high-resolution view of phage genetic diversity and offer a novel perspective on
19 virus ecology. To that end, we developed a method that identifies informative associations between a
20 phage's annotated host and clusters of genes in the network. Given these associations, we were able to
21 predict a phage's host with 86% accuracy at the genus level, while also identifying genes that underlie
22 these virus-host interactions. This approach, thus, provides one of the most accurate means of host
23 prediction while also pointing to directions for future empirical work.

24

25 **Introduction**

26 Bacteriophages (phages) are viruses that infect bacteria, and with over 10^{31} estimated on the planet, are
27 often the most abundant and diverse members of any ecosystem (Edwards & Rohwer 2005). Phages act
28 as predators, drivers of biogeochemical cycles (Wilhelm & Suttle 1999), industrial contaminants
29 (McGrath et al. 2007), and as important mutualists within bacterial pathogens that cause disease in
30 plants and animals (e.g. Addy et al. 2012, Waldor & Mekalanos 1996). Phages have also been used as
31 therapeutics in agriculture (Greer 2005) and for treating antibiotic-resistant bacterial infections (Chan et
32 al. 2016, Biswas et al 2002). Similar to bacteria, the majority of phages cannot be propagated in the lab,
33 either because their host cannot be grown or because their host is not known. Nonetheless, new
34 metagenomes from diverse environments are being published regularly, and the rate of uncultured,
35 novel virus discovery has increased rapidly in the past decade (e.g., Simmonds et al. 2017, Paez-Espino
36 et al 2016, Bruder et al. 2016, Roux et al. 2015, Dutilh et al 2014). Coping with this deluge of data
37 requires new computational methods for both classifying virus diversity and for inferring key features
38 of virus ecology and evolution.

39 Except for strain-level variation of a particular virus, traditional phylogenetic methods cannot
40 be applied to derive a “species” tree for phages. There are no universal genes shared by all phages, and
41 horizontal gene transfer (HGT) between viruses is common. In essence, every phage genome is a
42 mosaic that reflects the often disparate evolutionary histories of its genes (Pedulla et al. 2003, Hendrix
43 et al. 1999), and genome-level classification is, therefore, difficult. To overcome these challenges,
44 network-based approaches have been used to depict the relationship between phage genomes on the
45 basis of the similarity of their genic content or overall sequence identity (Cresawn et al. 2011, Halary et
46 al. 2010, Lima-Mendez et al. 2008, Roux et al. 2015, Paez-Espino et al. 2016).

47 Genome-level network analyses are appealing, because they make it possible to visualize phage
48 relationships in place of traditional phylogenies (e.g. Paez-Espino et al 2016, Lima-Mendez et al 2008).

49 At the same time, these whole-genome analyses continue to ignore the mosaic architecture of phage
50 genomes and take the focus away from the actual targets of selection: genes. As a result, it is unclear
51 how to apply these genome networks to questions beyond taxonomy. In the present work, we instead
52 build a network of genes, where genes are connected if they are ever found within the same genome.
53 By extending network analyses from genomes to genes, it is possible to address questions directly
54 related to virus ecology and evolution, such as how particular genes affect the mode of infection,
55 virulence, and host range of a virus.

56 Host range, in particular, constrains viral ecology and evolution, and predicting a virus' host is a
57 key challenge when characterizing novel, uncultured genomes. Host range typically depends on
58 individual virus-host gene interactions (Labrie et al. 2010), and both phages and their hosts can acquire
59 genes that alter these interactions through HGT (Meyer et al. 2016, Sachs & Bull 2005, Tzipilevich et
60 al. 2016). Methods for predicting virus host range from genomes commonly rely on comparing
61 genomic properties such as k-mer frequencies, codon usage, or, when possible, host CRISPR content.
62 The best of these methods, however, are rarely better than 80% accurate for predicting a phage's host at
63 the genus level (Ahlgren et al. 2016, Villaroel et al. 2016, Edwards et al. 2016). Here, we build a gene-
64 level network representing the co-occurrence of genes across phage genomes. In addition to providing
65 a robust view of virus genetic diversity, clusters within this network can be associated with virus host
66 range. By identifying genes that increase the correspondence between phages and their hosts, we are
67 able to predict virus host range at the genus level with over 85% accuracy for many host genera.

68

69 **Building Genome- and Gene-Level Networks**

70 We built genome- and gene-level networks for a set of 945 phage RefSeq genomes, consisting of
71 92,801 gene sequences. In the genome network (Figure 1a), nodes represent virus genomes, and two
72 nodes are connected if they share at least one gene. In the gene network (Figure 1b), nodes represent

73 homologous phage protein sequences, and two nodes are connected if these genes are found in the
74 same genome. Homologous genes were identified with as low as 35% identity via clustering by usearch
75 (Edgar 2010). Singleton and doubleton clusters were removed from consideration to increase the
76 reliability of connections between genes. This filter yielded a final set of 8,847 gene clusters from
77 across 913 phage genomes, dropping 32 phage genomes from primarily under-sampled, tailless phage
78 families, which are often underrepresented in metaviromes (Steward et al. 2013).

79 In each network, there exist subsets of nodes that form subgraphs in which members have more
80 connections in common with each other than with the rest of the network. We formally identified these
81 subsets of interconnected nodes using the Markov Clustering Algorithm (MCL) (Enright et al. 2002).
82 MCL relies on an inflation parameter that transforms the adjacency matrix of the underlying network.
83 Higher inflation values generally yield more clusters from a network, and others have previously used a
84 measure of cohesion within subgraphs, the “intracluster clustering coefficient” (ICCC), to optimize this
85 parameter choice for virus taxonomy (Roux et al. 2015, Lima-Mendez et al. 2008). Using this metric,
86 we chose an inflation factor of 6 for the genome network and 4.1 for the gene network (see Figure S1).
87 These values correspond to 209 clusters in the genome network and 135 clusters in the gene network.
88 As seen in Figure 2, the MCL clusters in the gene network appear to provide a cleaner visualization of
89 virus diversity than clusters in the genome network.

90

91 **Clusters of phage genes are associated with phage host genera**

92 Given the gene and genome networks, we then recolored the nodes according to the phage host genus
93 (Figure 3). In the gene network, each node represents a set of homologous genes, and only the most
94 common host associated with these homologs is indicated for each node. As can be seen in Figure 3,
95 phage host was poorly associated with graphical clustering in the genome network but maps closely to
96 graphical clusters in the gene network. In fact, for several hosts, distinct clusters could be identified in

97 the gene network that correspond at the species or strain-level of the phage host (see Figure 4).

98 In the case of *Bacillus* phages, genes are found in clusters corresponding to their annotated host
99 species: *B. anthracis*, *B. subtilis*, *B. thuringiensis*, *B. pumilus*, or *B. cereus*. Further, overlap exists
100 between *B. anthracis* and *B. thuringiensis*, closely related pathogens belonging to the *B. cereus* group
101 (Priest et al. 2004). Host associations at the species level are also visible within the genera
102 *Prochlorococcus* and *Streptococcus*.

103 Not all graphical clusters, however, correspond to a specific host species or strain. *Lactococcus*
104 *lactis*, for instance, is frequently used in dairy starter cultures as *L. lactis* subsp. *lactis* and *L. lactis*
105 subsp. *cremoris*, and phages have been well-sampled from both hosts (Deveau et al. 2006). Genes from
106 these diverse phages occur across three clusters of phage genetic diversity in the gene network, with no
107 clear associations with either host subspecies. Notably, these phages often are found to infect multiple
108 strains of *L. lactis* (Mahony et al. 2013), and recombination between dairy phages may be frequent
109 (Brüssow and Desiere 2001). Interestingly, one cluster of *Lactococcus*-associated genes shares many
110 connections with a cluster of *Streptococcus thermophilus*, another common member of dairy
111 fermentations.

112 The largest and most distinct cluster of phage genes corresponds to phages infecting
113 *Mycobacterium smegmatis*, a non-pathogenic and more readily-cultured relative of *M. tuberculosis*.
114 These phages have been heavily sampled compared to other hosts because of the SEA-PHAGES
115 program, in which undergraduates isolate and sequence phage genomes (Jordan et al. 2014). Though
116 phages of other species of *Mycobacterium* have not been thoroughly studied, genes from phages
117 infecting *M. tuberculosis* are also present across MCL clusters found within this subgraph. This
118 observation suggests it may be worthwhile, if technically difficult, to test more of the phages of *M.*
119 *smegmatis* on *M. tuberculosis*, as has been previously suggested (Hatfull 2014).

120 Though not as well-sampled as phages of *Mycobacterium*, genes from phages infecting

121 *Escherichia coli* and *Pseudomonas* species appear across the network, often more closely related to
122 phages infecting other genera. Genes from phages of *Salmonella*, *Shigella*, *Acinetobacter*, and
123 generically-identified *Enterobacteria* can all be found within clusters that are largely associated with *E.*
124 *coli*. There is a distinct cluster of phage genes affiliated with *Pseudomonas fluorescens*, but other
125 species-specific designations are not readily-observed. Iranzo *et al.* (2016) recently introduced a
126 bipartite network connecting phage genes to phage genomes, which may provide further insight into
127 how recombination events have structured phage host range.

128

129 **Quantifying and optimizing associations between network clusters and phage hosts**

130 We next sought to quantify the reliability of these visible associations and to ask if subsets of genes
131 could be used to predict a phage's host. We estimated the degree of overlap between graphical clusters
132 and host associations in each network by determining their mutual information (see Supplemental
133 Methods). This metric suggested that clusters in the genome network may, in fact, be more closely
134 associated with host annotations than clusters in the gene network ($MI_{\text{genome}} = 2.18$, $MI_{\text{gene}} = 1.42$). This
135 effect likely arises, however, because each node in the genome network corresponds to exactly one
136 host, and each MCL cluster in the genome network has, on average, only 4.36 members. In contrast,
137 there are an average of 65.5 genes within each MCL cluster in the gene network, and each node within
138 these clusters corresponds to at least 3 homologous genes from different phage genomes. More
139 importantly, many genes are not directly linked to host specificity, and homologs represented by a
140 single node in the gene network may come from phages that infect different hosts. Thus, graphical
141 clusters built from the gene network will contain many genes with variable host associations, whereas
142 those within the genome network are buffered from this noise. In the gene network, this variation
143 reduces the mutual information between cluster membership and host. This effect would also imply that
144 there exists a subset of genes within the gene network that would provide greater correspondence with

145 host associations.

146 To address this hypothesis, we developed an evolutionary algorithm, *mimax*, to identify the
147 subset of genes that maximizes the mutual information of MCL clusters and hosts. The *mimax*
148 algorithm works as follows: in each iteration, an MCL cluster in the gene network is removed from a
149 matrix of cluster-host associations at random. If doing so would result in removing a phage genome
150 from the dataset, the deletion is rejected. If no genomes are lost, then the mutual information of the new
151 matrix is calculated. If this value exceeds the value from the previous iteration, the deletion is retained,
152 otherwise it is rejected. Because the *mimax* algorithm depends on removing uninformative clusters of
153 genes, it should be more effective when there are more clusters from which to choose. When applied to
154 the 135 clusters previously found in the gene network, *mimax* removed 47 clusters containing 1375
155 genes (~15% of the dataset), resulting in a modest improvement in mutual information but still falling
156 short of the value observed in the genome network.

157 Three methods have been suggested to increase the granularity of MCL clusters (see
158 <https://micans.org>): increasing the inflation factor, removing highly connected nodes before finding
159 clusters, and introducing noise to the network. Initially, we chose an inflation factor of 4.1 to optimize
160 the ICCC, a measure of within-cluster cohesion. The ICCC, though, is largely of interest when clusters
161 represent naturally distinct sets of nodes, such as for taxonomic classification using genome-level
162 networks. Here, we are more interested in subdividing genes into co-occurring subsets, and optimizing
163 ICCC comes at the cost of sensitivity for the *mimax* algorithm. We tested each of the three methods
164 described above (see Supplemental Methods and Figure S2), finding the best results, 1355 clusters,
165 with an inflation factor of 15 and adding 5 random edges per node. Given this new set of clusters, we
166 ran *mimax* 10 times and retained the resulting matrix with the highest mutual information. In each
167 replicate, the mutual information between MCL membership and host associations converged to a
168 higher value than found in the genome network (Figure S3). On average, *mimax* reduced the number of

169 MCL clusters and associated genes within the gene network to 483.5 and 4070.6, respectively. These
170 deletions suggest that over half of the genes in the gene network are uninformative with respect to host
171 range.

172 Two questions emerge from maximizing the mutual information between graphical clusters and
173 host associations: 1) Are the retained genes more closely associated with functions characteristic of
174 phage-host interactions? and 2) can the resulting gene network be used as a tool for predicting the
175 primary host of phages?

176 To address the first question, we annotated the complete and *mimax*-reduced sets of genes using
177 RAST (Aziz et al. 2008). We then compared the frequency of common annotations of non-hypothetical
178 proteins for each set of genes (see Table S1 and Figure S3). Phage baseplate, neck, replication, and
179 DNA synthesis genes are over-represented following *mimax*, whereas phage packaging and regulatory
180 genes are under-represented. Phage baseplate proteins directly affect virus adsorption to host receptors
181 (Mahony and van Sinderen 2015), suggesting that gene function does affect *mimax* results.

182 The cluster-host correspondence in the *mimax*-reduced gene network offers a novel means to
183 predict a phage's host. Given a phage's genome, we identified all genes that belong to the *mimax*-
184 reduced set. We then recorded how often each potential host was associated with a homolog of one of
185 these remaining genes (excluding a phage's own contribution if already within the network). Finally,
186 we chose the most frequent host affiliated with this subset of genes as the predicted host. (See the
187 Supplemental Methods for additional details of the procedure.) When applied to all phage in the
188 network, this approach predicted the host genus with 86% accuracy. If the full gene network is used in
189 place of the *mimax*-reduced network, accuracy declines to 72%. This difference confirms that the
190 *mimax* procedure reduces the gene network to a set of genes with stronger ties to phage host
191 determination.

192 We deconstructed the host prediction accuracy (from *mimax*-improved predictions) for each

193 host genus in order to account for uneven sampling of phages across hosts (Table 1). Doing so indicates
194 that accuracy varied with host genus. Predictions for phages of *Mycobacterium* were nearly 100%
195 accurate, and this reflects the large, unique space occupied by their genes in the network. In contrast,
196 host predictions were less accurate for hosts with few representatives in the dataset, such as
197 *Clostridium* and *Yersinia*. Accuracy also declined for well-sampled hosts, such as *Escherichia*, where
198 phages have been sampled from closely-related genera (e.g. *Salmonella*, *Shigella*, and *Yersinia*). As has
199 been seen for other host prediction methods (e.g. Villaroel et al. 2016), incorrect host predictions
200 tended to predict that phage infect closely-related hosts (see Table 1). Improving the accuracy of
201 predictions within these groups requires additional sampling and wet lab characterization of phages
202 from across host genera. We should also be careful when assessing the quality of negative predictions.
203 While phage host range can be exceptionally specific, many phages infect multiple genera (Hamdi et
204 al. 2017, Jensen et al. 1998) or even across phyla (Malki et al. 2015), and additional lab work is
205 required to confirm that putatively incorrect predictions are not, in fact, false negative results.

206 We next tested this approach with a set of novel phages not included in the original gene
207 network. Over 1000 new phage genomes have been published since we built our original network. We
208 chose 500 phage genomes at random from this new set. Of these, 185 were annotated as infecting hosts
209 already included in our network. The genes in these phages were assigned to the *mimax*-reduced set of
210 MCL clusters identified previously. While 52 of these phages shared no genes in the *mimax* set with
211 any phages in our original dataset, for the remaining 133 phage, our procedure predicted the host genus
212 67.7% of the time (see Table S2). Moreover, accuracy remained high for well-sampled hosts, such as
213 *Mycobacterium* and *Escherichia*, but was low for others, such as *Bacillus*, that we could previously
214 predict with over 90% accuracy. This discrepancy suggests that a gene network approach to host
215 prediction should be updated regularly to account for the frequent addition of new virus genomes to
216 repositories.

217

218 **Conclusion**

219 In this work, we have shown that gene-level networks provide both a high-resolution view of viral
220 genetic diversity and a means to connect specific groups of genes to broad patterns in viral ecology.
221 When applied to virus host range, phage gene clusters correlated with a phage's annotated host, and
222 proximity of clusters in the network reflected the evolutionary relatedness of these hosts. Using an
223 evolutionary algorithm, *mimax*, we were then able to identify specific groups of genes with the
224 strongest correlation to virus host range. The *mimax*-reduced dataset was enriched for genes known to
225 affect host recognition, and the enhanced network offers one of the most accurate means of host
226 prediction to date.

227 This approach should be extensible to aspects of viral ecology beyond host range, including
228 isolation source (e.g. freshwater, marine, soil, leaf, gut, hospital, etc.) and abiotic or biotic factors that
229 vary across locations (e.g. temperature, pH, O₂, nutrient concentrations, and available host diversity).
230 Moreover, phage have a direct impact on the growth of their host bacteria, and knowing a phage's
231 ecological and evolutionary history is critical to understanding how that phage affects an ecosystem.
232 Gene network analysis should facilitate new discoveries in any environment, be it a dairy vat, a
233 freshwater lake, or the human gut.

234

235 **Acknowledgments**

236 We are grateful to past and present members of the Putonti lab for thoughtful feedback on this work.

237

238 **Funding**

239 This work was supported by the NSF (NSF awards 119387 and 1661357).

240 **References**

241 Addy HS, Askora A, Kawasaki T, Fujie M, & Yamada T (2012). The Filamentous Phage phi RSS1
242 Enhances Virulence of Phytopathogenic *Ralstonia solanacearum* on Tomato. *Phytopathology*
243 102(3):244-251.

244

245 Ahlgren, Nathan A., Jie Ren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. "Alignment-free
246 d_2 oligonucleotide frequency dissimilarity measure improves prediction of hosts from
247 metagenomically-derived viral sequences." *Nucleic Acids Research* (2016): gkw1002.

248

249 Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards,
250 Kevin Formsma et al. "The RAST Server: rapid annotations using subsystems technology." *BMC*
251 *genomics* 9, no. 1 (2008): 75.

252

253 Biswas, Biswajit, Sankar Adhya, Paul Washart, Brian Paul, Andrei N. Trostel, Bradford Powell,
254 Richard Carlton, and Carl R. Merrill. "Bacteriophage therapy rescues mice bacteremic from a clinical
255 isolate of vancomycin-resistant *Enterococcus faecium*." *Infection and immunity* 70, no. 1 (2002):
256 204-210.

257

258 Bruder, Katherine, Kema Malki, Alexandria Cooper, Emily Sible, Jason W. Shapiro, Siobhan C.
259 Watkins, and Catherine Putonti. "Freshwater Metaviromics and Bacteriophages: A Current
260 Assessment of the State of the Art in Relation to Bioinformatic Challenges." *Evolutionary*
261 *bioinformatics online* 12, no. Suppl 1 (2016): 25.

262

263

- 264 Brüssow, Harald, and Frank Desiere. "Comparative phage genomics and the evolution of Siphoviridae:
265 insights from dairy phages." *Molecular microbiology* 39, no. 2 (2001): 213-223.
266
- 267 Chan, Benjamin K., Mark Siström, John E. Wertz, Kaitlyn E. Kortright, Deepak Narayan, and Paul E.
268 Turner. "Phage selection restores antibiotic sensitivity in MDR *Pseudomonas aeruginosa*." *Scientific*
269 *reports* 6 (2016).
270
- 271 Chatteraj, Partho, Tridib Ganguly, Ranjan Kumar Nandy, and Subrata Sau. "Overexpression of a
272 delayed early gene hlg1 of temperate mycobacteriophage L1 is lethal to both *M. smegmatis* and *E.*
273 *coli*." *BMB reports* 41, no. 5 (2008): 363-368.
274
- 275 Cresawn, Steven G., Matt Bogel, Nathan Day, Deborah Jacobs-Sera, Roger W. Hendrix, and Graham F.
276 Hatfull. "Phamerator: a bioinformatic tool for comparative bacteriophage genomics." *BMC*
277 *bioinformatics* 12, no. 1 (2011): 395.
278
- 279 Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal,
280 Complex Systems 1695. 2006. <http://igraph.org>
281
- 282 Deveau, H el ene, Simon J. Labrie, Marie-Christine Chopin, and Sylvain Moineau. "Biodiversity and
283 classification of lactococcal phages." *Applied and environmental microbiology* 72, no. 6 (2006):
284 4338-4346.
285
- 286 Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo GZ Silva, Lance
287 Boling, Jeremy J. Barr et al. "A highly abundant bacteriophage discovered in the unknown

- 288 sequences of human faecal metagenomes." *Nature communications* 5 (2014).
- 289
- 290 Edgar, Robert C. "Search and clustering orders of magnitude faster than BLAST." *Bioinformatics* 26,
- 291 no. 19 (2010): 2460-2461.
- 292
- 293 Edwards, Robert A., Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. "Computational
- 294 approaches to predict bacteriophage–host relationships." *FEMS microbiology reviews* 40, no. 2
- 295 (2016): 258-272.
- 296
- 297 Edwards, Robert A., and Forest Rohwer. "Viral metagenomics." *Nature Reviews Microbiology* 3, no. 6
- 298 (2005): 504-510.
- 299
- 300 Enright, Anton J., Stijn Van Dongen, and Christos A. Ouzounis. "An efficient algorithm for large-scale
- 301 detection of protein families." *Nucleic acids research* 30, no. 7 (2002): 1575-1584.
- 302
- 303 Greer, G. Gordon. "Bacteriophage control of foodborne bacteria." *Journal of food protection* 68, no. 5
- 304 (2005): 1102-1111.
- 305 Halary, Sébastien, Jessica W. Leigh, Bachar Cheaib, Philippe Lopez, and Eric Bapteste. "Network
- 306 analyses structure genetic diversity in independent genetic worlds." *Proceedings of the National*
- 307 *Academy of Sciences* 107, no. 1 (2010): 127-132.
- 308
- 309 Hamdi, Sana, Geneviève M. Rousseau, Simon J. Labrie, Denise M. Tremblay, Rim Saïed Kourda,
- 310 Karim Ben Slama, and Sylvain Moineau. "Characterization of two polyvalent phages infecting
- 311 Enterobacteriaceae." *Scientific reports* 7 (2017).

312

313 Hatfull, Graham F. "Mycobacteriophages: windows into tuberculosis." *PLoS Pathogens* 10, no. 3
314 (2014): e1003953.

315

316 Hendrix, Roger W., Margaret CM Smith, R. Neil Burns, Michael E. Ford, and Graham F. Hatfull.
317 "Evolutionary relationships among diverse bacteriophages and prophages: all the world's a
318 phage." *Proceedings of the National Academy of Sciences* 96, no. 5 (1999): 2192-2197.

319

320 Iranzo, Jaime, Mart Krupovic, and Eugene V. Koonin. "The double-stranded DNA virosphere as a
321 modular hierarchical network of gene sharing." *MBio* 7, no. 4 (2016): e00978-16.

322

323 Iranzo, Jaime, Mart Krupovic, and Eugene V. Koonin. "A network perspective on the virus
324 world." *Communicative & Integrative Biology* 7, no. just-accepted (2017): 00-00.

325

326 Jensen, Ellen C., Holly S. Schrader, Brenda Rieland, Thomas L. Thompson, Kit W. Lee, Kenneth W.
327 Nickerson, and Tyler A. Kokjohn. "Prevalence of Broad-Host-Range Lytic Bacteriophages of
328 *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*." *Applied and environmental*
329 *microbiology* 64, no. 2 (1998): 575-580.

330

331 Jordan, Tuajuanda C., Sandra H. Burnett, Susan Carson, Steven M. Caruso, Kari Clase, Randall J.
332 DeJong, John J. Dennehy et al. "A broadly implementable research course in phage discovery and
333 genomics for first-year undergraduate students." *MBio* 5, no. 1 (2014): e01051-13.

334

335 Labrie, Simon J., Julie E. Samson, and Sylvain Moineau. "Bacteriophage resistance

- 336 mechanisms." *Nature Reviews Microbiology* 8, no. 5 (2010): 317-327.
- 337
- 338 Li, Li, Christian J. Stoeckert, and David S. Roos. "OrthoMCL: identification of ortholog groups for
339 eukaryotic genomes." *Genome research* 13, no. 9 (2003): 2178-2189.
- 340
- 341 Lima-Mendez, Gipsi, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae. "Reticulate
342 representation of evolutionary and functional relationships between phage genomes." *Molecular
343 biology and evolution* 25, no. 4 (2008): 762-777.
- 344
- 345 Mahony, Jennifer, Witold Kot, James Murphy, Stuart Ainsworth, Horst Neve, Lars H. Hansen, Knut J.
346 Heller et al. "Investigation of the relationship between lactococcal host cell wall polysaccharide
347 genotype and 936 phage receptor binding protein phylogeny." *Applied and environmental
348 microbiology* 79, no. 14 (2013): 4385-4392.
- 349
- 350 Mahony, Jennifer, and Douwe van Sinderen. "Novel strategies to prevent or exploit phages in
351 fermentations, insights from phage–host interactions." *Current opinion in biotechnology* 32 (2015):
352 8-13.
- 353
- 354 Malki, Kema, Alex Kula, Katherine Bruder, Emily Sible, Thomas Hatzopoulos, Stephanie Steidel,
355 Siobhan C. Watkins, and Catherine Putonti. "Bacteriophages isolated from Lake Michigan
356 demonstrate broad host-range across several bacterial phyla." *Virology journal* 12, no. 1 (2015): 164.
- 357
- 358 Mc Grath, Stephen, Gerald F. Fitzgerald, and Douwe van Sinderen. "Bacteriophages in dairy products:
359 pros and cons." *Biotechnology journal* 2, no. 4 (2007): 450-455.

360

361 Meyer, Justin R., Devin T. Dobias, Joshua S. Weitz, Jeffrey E. Barrick, Ryan T. Quick, and Richard E.

362 Lenski. "Repeatability and contingency in the evolution of a key innovation in phage

363 lambda." *Science* 335, no. 6067 (2012): 428-432.

364

365 Meyer, Justin R., Devin T. Dobias, Sarah J. Medina, Lisa Servilio, Animesh Gupta, and Richard E.

366 Lenski. "Ecological speciation of bacteriophage lambda in allopatry and sympatry." *Science* (2016):

367 aai8446.

368

369 Paez-Espino, David, Emiley A. Eloë-Fadrosch, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel

370 Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides.

371 "Uncovering Earth's virome." *Nature* 536, no. 7617 (2016): 425-430.

372

373 Pedulla, Marisa L., Michael E. Ford, Jennifer M. Houtz, Tharun Karthikeyan, Curtis Wadsworth, John

374 A. Lewis, Debbie Jacobs-Sera et al. "Origins of highly mosaic mycobacteriophage

375 genomes." *Cell* 113, no. 2 (2003): 171-182.

376

377 Priest, Fergus G., Margaret Barker, Les WJ Baillie, Edward C. Holmes, and Martin CJ Maiden.

378 "Population structure and evolution of the *Bacillus cereus* group." *Journal of bacteriology* 186, no.

379 23 (2004): 7959-7970.

380

381 R Core Team (2015). R: A language and environment for statistical computing. R Foundation for

382 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

383

- 384 Roux, Simon, Steven J. Hallam, Tanja Woyke, and Matthew B. Sullivan. "Viral dark matter and virus–
385 host interactions resolved from publicly available microbial genomes." *Elife* 4 (2015): e08490.
386
- 387 Sachs, Joel L., and James J. Bull. "Experimental evolution of conflict mediation between
388 genomes." *Proceedings of the National Academy of Sciences of the United States of America* 102,
389 no. 2 (2005): 390-395.
390
- 391 Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage,
392 Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: a software environment for
393 integrated models of biomolecular interaction networks." *Genome research* 13, no. 11 (2003): 2498-
394 2504.
395
- 396 Simmonds, Peter, Mike J. Adams, Mária Benkő, Mya Breitbart, J. Rodney Brister, Eric B. Carstens,
397 Andrew J. Davison et al. "Consensus statement: Virus taxonomy in the age of
398 metagenomics." *Nature Reviews Microbiology* (2017).
399
- 400 Steward, Grieg F., Alexander I. Culley, Jaclyn A. Mueller, Elisha M. Wood-Charlson, Mahdi Belcaid,
401 and Guylaine Poisson. "Are we missing half of the viruses in the ocean?." *The ISME journal* 7, no. 3
402 (2013): 672-679.
403
- 404 Tzipilevich, Elhanan, Michal Habusha, and Sigal Ben-Yehuda. "Acquisition of Phage Sensitivity by
405 Bacteria through Exchange of Phage Receptors." *Cell* (2016).
406
- 407 van Dongen, Stijn Marinus. "Graph clustering by flow simulation." PhD diss., 2001.

408

409 Villarroel, Julia, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, Henrike Zschach, Ole Lund, Morten

410 Nielsen, and Mette Voldby Larsen. "HostPhinder: A Phage Host Prediction Tool." *Viruses* 8, no. 5

411 (2016): 116.

412

413 Waldor MK & Mekalanos JJ (1996). Lysogenic conversion by a filamentous phage encoding cholera

414 toxin. *Science* 272(5270):1910-1914.

415

416 Wilhelm SW & Suttle CA (1999). Viruses and Nutrient Cycles in the Sea - Viruses play critical roles in

417 the structure and function of aquatic food webs. *Bioscience* 49(10):781-788.

418

419 **Table 1: Host accuracy varies with genus and sampling**

420

Host Genus	Accuracy	Top Mistake	Total Count
<i>Chlamydia</i>	1	N/A	4
<i>Lactococcus</i>	1	N/A	36
<i>Mycobacterium</i>	0.991	<i>Lactococcus</i>	226
<i>Bacillus</i>	0.97	<i>Chlamydia</i>	66
<i>Streptococcus</i>	0.947	<i>Bacillus</i>	38
<i>Escherichia</i>	0.906	<i>Salmonella</i>	138
<i>Prochlorococcus</i>	0.905	<i>Synechococcus</i>	21
<i>Staphylococcus</i>	0.897	<i>Bacillus</i>	87
<i>Pseudomonas</i>	0.847	<i>Escherichia</i>	85
<i>Burkholderia</i>	0.833	<i>Pseudomonas</i>	30
<i>Salmonella</i>	0.804	<i>Escherichia</i>	56
<i>Vibrio</i>	0.686	<i>Escherichia</i>	51
<i>Clostridium</i>	0.667	<i>Streptococcus</i>	21
<i>Acinetobacter</i>	0.583	<i>Escherichia</i>	12
<i>Shigella</i>	0.273	<i>Escherichia</i>	11
<i>Yersinia</i>	0.273	<i>Escherichia</i>	11
<i>Anabaena</i>	0	<i>Escherichia</i>	1
<i>Microcystis</i>	0	<i>Escherichia</i>	1
<i>Chlamydophila</i>	0	<i>Chlamydia</i>	1
<i>Synechococcus</i>	0	<i>Prochlorococcus</i>	15
<i>Bdellovibrio</i>	0	<i>Escherichia</i>	2

421

422

423 **Figure Captions**

424

425 **Figure 1**

426 Genome-level (a) and gene-level (b) networks for a set of 913 phage. In the genome network, nodes are
427 genomes, and two nodes are connected by an edge if they share any genes. Inversely, in the gene
428 network, nodes are genes, and two nodes are connected if they are found in the same genome.

429

430 **Figure 2**

431 The genome (a) and gene (b) networks are identical to those in Figure 1, except nodes have been
432 colored based on their membership in graphical clusters identified using MCL with inflation set to 6 for
433 the genome network and to 4.1 for the gene network.

434

435 **Figure 3**

436 The genome (a) and gene (b) networks are identical to those in Figures 1 and 2, except nodes have now
437 been colored to reflect the host genus associated with each phage. In the gene network, each node
438 signifies a set of homologous sequences, and colors match the most common host for the genomes
439 containing these homologs.

440

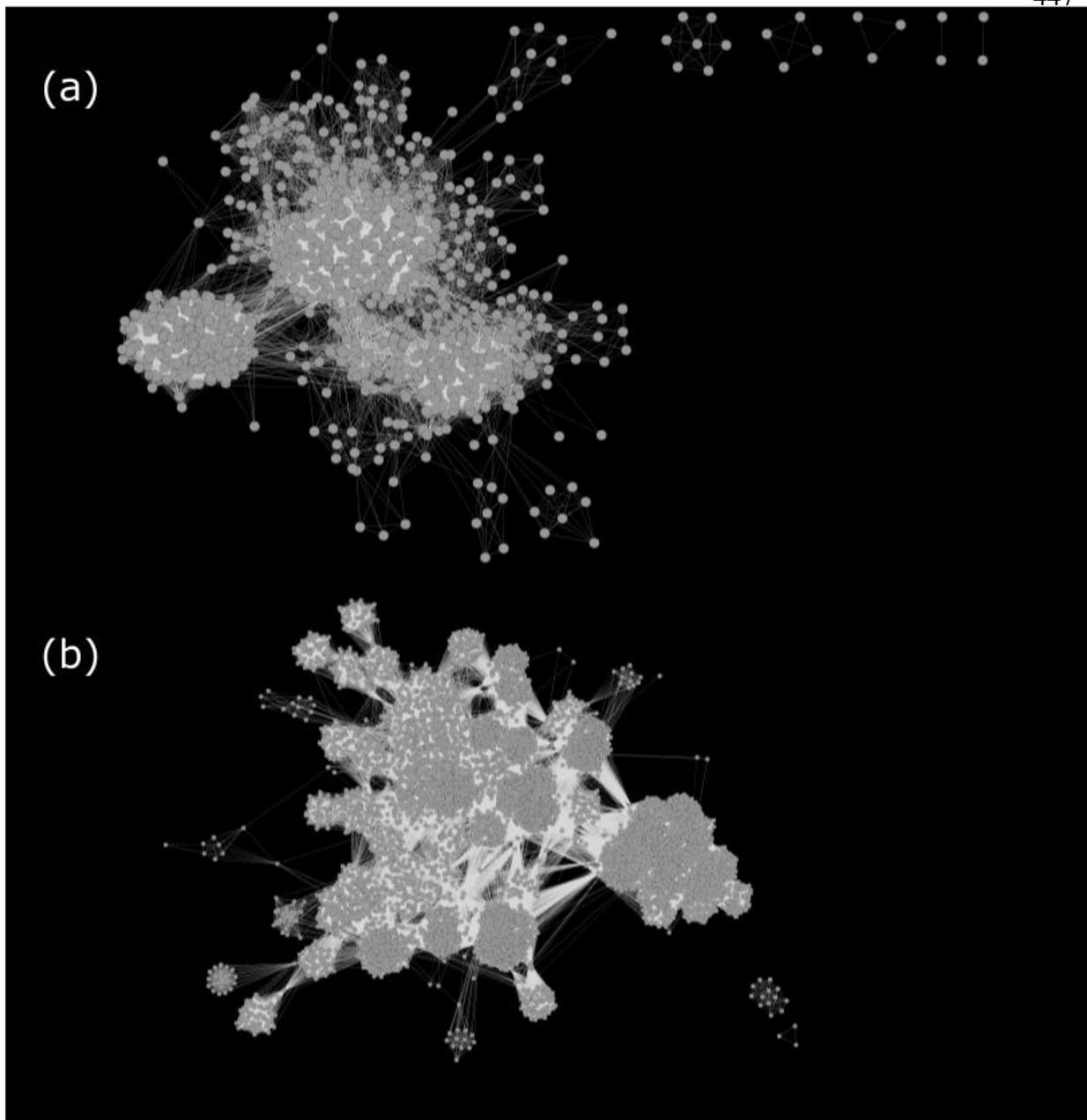
441 **Figure 4**

442 The gene network shown is identical to the network in Figures 1b and 2b, but with nodes recolored
443 according to the host species, where annotation was available. Labels and arrows indicate specific cases
444 highlighted in the main text.

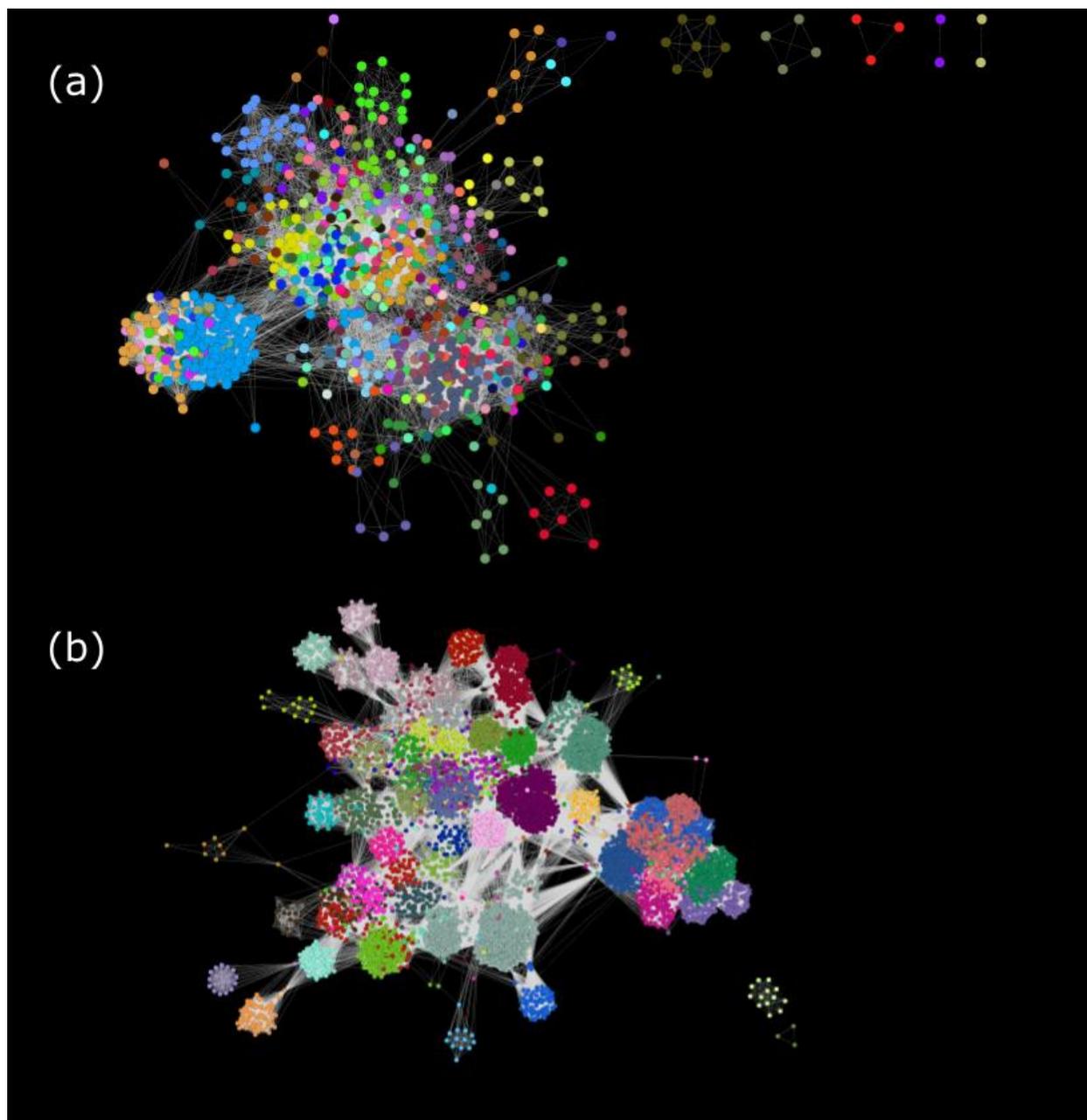
445

446 **Figure 1: Uncolored genome (a) and gene (b) networks**

447

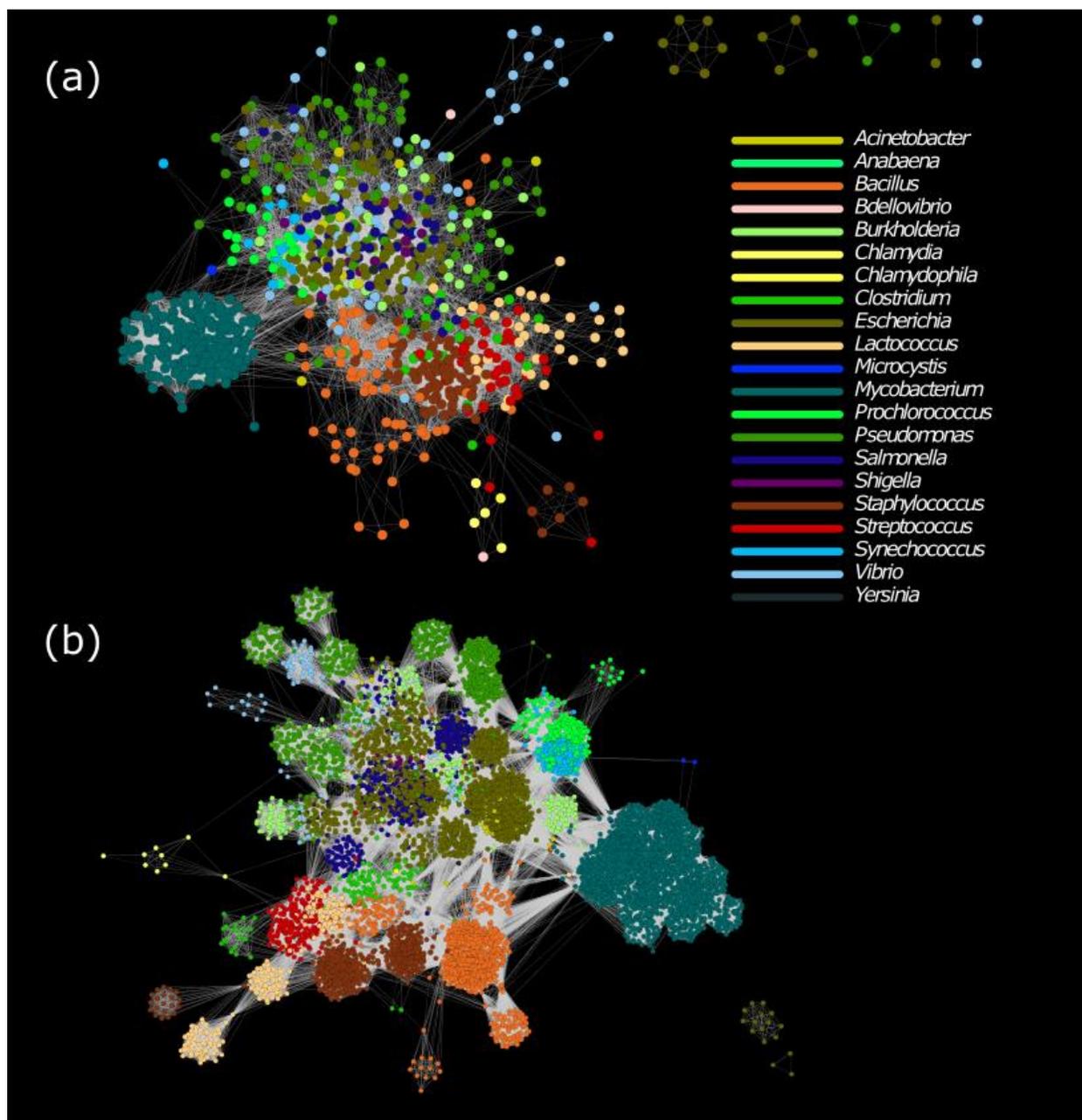


449 **Figure 2: Genome (a) and gene (b) networks colored by MCL clustering**
450



452
453

Figure 3: Genome (a) and gene (b) networks colored by annotated host genus



454 **Figure 4: Gene network highlighting clusters that vary by host species**
 455

