Supplementary Materials and Methods

for

The Statistical Structure of Host-Phage Interactions

Cesar Flores[1,x], Justin R. Meyer[2,x], Sergi Valverde[3] Lauren Farr[4], and Joshua S. Weitz[4,1,*]

[1]School of Physics, Georgia Tech, Atlanta GA 30332, USA
[2]Department of Zoology, MSU, East Lansing, MI 48824, USA
[3]Complex Systems Lab, University Pompeu Fabra, Barcelona, E-08003, Spain
[4]School of Biology, Georgia Tech, Atlanta GA 30332, USA
x These authors contributed equally to this work
* Corresponding author: jsweitz@gatech.edu

Quantitative estimation of nestedness and modularity

We represent the host-phage network with a *bipartite network* consisting of three sets $G = (U,V,E)$, where U and V are disjoint sets of nodes and $E = \{\{u_i, v_j\}\}$ is the set of edges connecting nodes of different type. For example, Supplementary Figure 5A shows the host-phage network described in Quiberoni (1). Define $P = |U|$ the number of phages and $H = |V|$ the number of hosts. The *adjacency matrix* of the bipartite network is $A_{ij} = 1$ if there is an edge $\{u_i, v_j\}$ \in E or $A_{ij} = 0$ otherwise (see Supplementary Figure 5b-c). The number of links attached to node u_i is the so-called *degree* $k_i = \sum_j A_{ij}$ (similarly, we can define the degree for v_j). Distinct colors indicate whether the node is a host (blue) or a phage (yellow) and bright (dark) shading depicts high (low) degree. Visual inspection of the network reveals significant structure, which can be rigorously detected by means of standard network measurements.

We have examined different properties of host-phage networks. Many real networks have a natural community structure, where disjoint subgroups of nodes exchange many internal connections among then than with the rest of nodes. Formally, we want to compute the optimal division of the network that minimizes the number of links between subgroups (also called communities). The raw number of links at the boundary does not give a good partition of the network. For example, the community structure can be a consequence of random variations in

1 the density of links (2). A more reliable approach uses a null model to assess the quality of a
2 given network partition. Newman and Girvan(3) defines the *modularity* as follows:

3
$$Q = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(g_i, g_j)$$

4 where $2m = \sum_{ij} A_{ij}$ is the number of links and g_i gives the label of the community the node i
5 belongs to. Notice that maximizing the above function yields a partition that minimizes the
6 expected number of links falling between different communities, i.e., when $\delta(g_i, g_j) = 0$.
7 Modularity Q takes values between 0 and 1: low modularity indicates the number of links
8 between distinct communities is not significantly different from the random distribution and high
9 modularity indicates there is a strong community structure.

10 Our networks are different from the networks studied with the standard modularity
11 measure Q (see above). Here, we study bipartite networks, i.e., networks having two distinct
12 types of nodes and there are no links between nodes of the same type. Barber defines a new
13 modularity quantity $Q_{bipartite}$ using a specific null model for bipartite networks:

14
$$Q_{bipartite} = \frac{1}{4m}\sum_{ij}\left(A_{ij} - b_{ij}\frac{k_i k_j}{2m}\right)\delta(g_i, g_j)$$

15 where $b_{ij} = 1$ if nodes i and j are of different type and 0 otherwise. Related studies of modularity
16 in plant-pollinator networks have used the standard modularity Q *(4)*. Empirical analyses of
17 bipartite networks have shown that $Q_{bipartite} > Q$, that is, the bipartite modularity can often find
18 better community divisions than the standard modularity when we do not consider the possibility
19 to have links between nodes of the same type (5). We use the BRIM (5) (Bipartite Recursive
20 Induced Modules) algorithm to maximize this bipartite modularity in our host-phage networks
21 (see the paper by Barber for full details on the BRIM algorithm). For example Supplementary
22 Figure 5A and 5D show the matrix and network representations of the optimal community
23 structure found in a host-phage network. Figure 5B maps the four network communities found
24 with BRIM into coherent matrix blocks of the (sorted) adjacency matrix. Alternatively, the
25 network representation of community structure in Figure 7d suggests a geometrical interpretation
26 of the maximization of bipartite modularity in terms of link crossing minimization, a hard
27 problem that has been extensively studied in literature (6).

28 Fortunato and Barthélemy have pointed out that, in large networks, modularity
29 optimization may fail to identify modules smaller than a characteristic size-dependent scale (7).

A check of the modularity obtained through modularity optimization is thus necessary. When modularity optimization finds a module S with l_s internal links, it may be that the latter is a combination of two or more smaller modules. In this case:

$$l_s < \sqrt{2L}$$

where L is the number of links in the full network (see the paper by Fortunato and Barthélemy (7) for full details on the derivation). Modules close to this resolution limit can result from the random merging of two or more sub-modules. Then, modularity optimization might fail to detect the fine modularity structure in these situations.

An important measurement of ecological networks determines to what extent they form a nested network, i.e., when the specialist species only interact with proper subsets of the species interacting with the generalists (8). The computation of the degree of nestedness involves three steps: (i) computing the isocline of perfect order, which is the curve that separates all the non-zero entries in the adjacency matrix (above the isocline) from the absence of interactions (below the isocline) in a perfectly nested network, (ii) re-arrange all the rows and columns of the adjacency matrix in a way that maximizes the nestedness and (iii) compute the temperature T as the sum of distances d_{ij} between the expected and unexpected matrix entries and the isocline:

$$T = \frac{k}{HP} \sum_{ij \in \text{unexpected cells}} \left(\frac{d_{ij}}{D_{ij}} \right)^2$$

where D_{ij} is the diagonal that cross the unexpected cell and $k = 100/U_{max}$ with $U_{max} \approx 0.04145$ is a normalization factor that makes $0 \leq T \leq 100$ (9, 10). Finally, we have normalized the temperature T in such a way that the new range is $0 \leq N \leq 1$:

$$N = \frac{100 - T}{100}$$

Supplementary Figure 5C shows the sorted matrix corresponding to the optimal nestedness temperature. This matrix ordering indicates the network is highly nested. There are a few unexpected interactions below the isocline of perfect order, which correspond to the links of the right side of Supplementary Figure 5E.

Criterion for cataloging studies as Co-evolution (EXP), Natural Communities (NAT) or Host-phage typing (TYP):

Representative host-phage studies were found using a literature search using ISI Web of Science and tracking references (both to and from the original article). Productive search terms were as follows:

- (phage or bacteriophage) and host and range
- (phage or bacteriophage) and host and typing
- (phage or bacteriophage) and host and infectivity
- (phage or bacteriophage) and characterization

Searching cross-references were also a useful means of collecting infectivity matrices. Web of Science also generated the BibTex reference information for each article. The criteria of inclusion of a study was as follows:

1) Data is available in a matrix/table format in the paper
2) The matrix included interpretable quantitative information on infection
3) The matrix had no missing values
4) The matrix could be manually verified at each cell.
5) The matrix included at least 2 hosts and 2 phages.

Thirty-eight matrices were included in the analysis. Infectivity was indicated either with shading or a (+/-) system. Different amounts of shading would indicate the degree of infection. In the (+/-) system, a '+' generally indicated a positive infection, while a '-' indicated no infection. According to these criterion, we excluded three datasets because of missing data (11-13).

The criterion for cataloging studies was as follows:

Natural communities (NAT) – 19 studies:

This criterion was applied to studies in which both phages and hosts were isolated from the environment. These types of studies are indicative of community interactions within a natural

1 network. These studies were then divided into one of four sub-classes: aquatic, soil, microbiome,
2 and food items. These sub-classes were based upon the environment from which the hosts and
3 phages were isolated.
4
5 *Co-evolution (EXP) – 10 studies:*
6 This criterion was applie to studies in which phages and/or hosts were allowed to evolve in the
7 lab. After phages were allowed to evolve, their host ranges were then tested. Sub-classes were
8 based upon methodology of the study, and studies were classified as either serial dilution or
9 chemostat experiments. Importantly, matrices of the EXP class need not be reflective of a given
10 community at a fixed moment in time.
11
12 *Artificial (ART) – 9 studies:*
13 This criterion was applied to studies in which almost all hosts and phages were either generated
14 within the lab or came from a collection. Sub-classes indicated the origination of the host strains.
15 Host strains were either environmental or pathogenic.
16
17 **Principal component analysis**
18 The objective of PCA is to find a new coordinate system such that the maximal variance is
19 explained in order of each coordinate (i.e., the principal components). Each variable was
20 normalized to have zero mean and a standard deviation of 1 so that each contributed equally to
21 the PCA. Supplementary Figure 1 shows the projection of each study onto the first two principal
22 axes and Supplementary Table 4 shows the detailed coordinates underlying the principal
23 components. Roughly, principal component 1 (PC1) corresponds to the size of the matrix, and so
24 those studies to the right-side of Supplementary Figure 3 tend to be large matrices and those to
25 the left tend to be small matrices. Roughly, PC2 corresponds to the asymmetry between number
26 of phages and number of hosts, so that the top-most studies of Supplementary Figure 3 have
27 more hosts than phages, whereas the bottom-most studies have more phages than hosts. Finally,
28 the third principal component (not shown) corresponds, roughly, to the connectance of the study.
29
30 **Statistical analysis of clustering validity using a re-shuffling approach**
31 In order to find clusters the K-means algorithm (14) (with $k=3$) has been applied to the two main

components of the PCA analysis output. This output is used as benchmark for study the subdivision of the studies and compare with those of random labels. The way in which this algorithm works is the next.

Given a set of observations ($\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_n$), where each observation in our case represents a point in the PCA-analysis output, the k-means aims to partition the n observations into k sets ($k \leq n$) S = {$S_1, S_2, \ldots, S_k$} so as to minimize the within-cluster sum of squares:

$$\arg\min_{\mathbf{S}} = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \mu_i \right\|^2$$

where μ_i is the mean of the points in S_i. In our case $n=38$ and $k=3$. See Supplementary Figure 3 for the output of this algorithm.

In order to compare the three clusters found in this algorithm with the three real categories (NAT, EXP, ART) of our studies we used the *Jaccard Index* defined as:

$$J(C,K) = \frac{a}{a+b+c}$$

Where C represents the real labels and K the labels of the output in the k-means algorithm. a denotes the number of pairs of points with the same label in C and assigned to the same cluster in K, b denotes the number of pairs with the same label, but in different clusters and c denotes the number of pairs in the same cluster, but with different class labels. The index produces a result in the range [0,1], where a value of 1 indicates that C and K are identical.

We find that the three real categories when compared with the output of the k-means algorithm share a Jaccard Index of 0.26. This value indicates that there exist a poor clustering of labels of the studies with the labels of the k-means algorithm. And by consequence we can say (assuming that the k-means output is the perfect subdivision) that there is not significant subdivision in the three real categories (EXP, NAT and ART).

We subjected this index to a randomization test. We generated 10,000 trials where we relabeled the studies while retaining the number of each class (EXP, NAT and ART). The distribution of the Jaccard index of these random trials is showed in Supplementary Figure 4. We found a p-value = 0.34 in the Jaccard index of the real labels. This indicates that there is not a statistically significant difference between the real subdivision of the studies and those that are labeled randomly.

1 **Statistical analysis of correlations among global properties using a Bonferroni correction**
2 We study the correlations coefficients among the global properties. These values are show in
3 Supplementary Table 5. In that table is showed also the statistical significance of those values.
4 For evaluate the statistical significance we used a Bonferroni correction, using both, the number
5 of combinations and the number of global properties. This correction is used in statistics when
6 one needs to address multiple comparisons. And comes by the fact that even when there is not
7 statistical significance, we can find just by probability that some of the comparisons are
8 statistically significant. Therefore this correction aims to avoid this problem. We can see in the
9 indicated table that among the statistically significant values there is only a strong correlation
10 between the number of hosts and the number of species. Another interesting result is that there is
11 almost no correlation (no statistical significance) between the connectance and the number of
12 species. This is contrary to the plant-pollinator networks where the relation follows a power law.
13
14 **Experimental assays of host-phage infection**
15 ***Conditions and microbial cultures***
16 The phage and bacteria were cocultured in 50ml Erlenmeyer flasks, with 10ml of liquid medium,
17 shaken at 120 rpm, and incubated at 37 °C. The medium was an altered version of Davis
18 Medium (15), in which we added 10 times the magnesium sulfate (1g/L) to improve phage
19 viability and 125 mg/L of maltotriose instead of glucose because *E. coli* and phage λ are
20 predicted to undergo a coevolutionary arms-race when provided with maltodextrins as its only
21 source of carbon (16-18). The medium was filtered and the magnesium was added just before
22 use in order to stop crystallization of the magnesium during the experiment. 75 separate flasks
23 were initiated with very small populations of bacteria(~1,000 *E. coli* cells) and phage (~100
24 phage λ particles) to assure that the initial populations were isogenic and that all mutant bacteria
25 and phage arose *de novo,* this is important to make sure that each community has the potential to
26 follow its own coevolutionary path. The *E. coli* studied were of strain REL606, a derivative of
27 *E. coli* B acquired from Richard Lenski (Michigan State University), described in (19) and phage
28 were of strain cI21 (λvir) provided by Donald Court (National Cancer Institute). Most phage λ
29 strains have two life cycles, lytic and lysogenic, the second includes a latent phase where the
30 phage genome is incorporated into the bacterial chromosome at which time the bacteria acquires
31 immunity to phage infection. Because the goal of this study was to characterize evolved phage

resistance instead of acquired resistance, we used a phage that was unable to create the resistant lysogenic bacteria. cI21 is only able to reproduce through the lytic phase because it has a chemically induced mutation in the cI gene which is a repressor protein required for lysogeny. Each flask was cultured for 24 hours and then a random subsample of 100µl of the culture was removed and transferred to 9.9ml of fresh medium. This flask was incubated and the cycle of transfer and incubation was continued once more. Three 24 hour incubations were long enough for the bacteria to evolve resistance and the phage to counter it, however not long enough for a second round of coevolution.

Isolation strategies

After 72 hours of coculturing, two bacterial clones were isolated from each flask by streaking on LB (Luria Burtani medium, recipe found in (20)) agar plates and picking single colonies. These colonies were restreaked twice more to assure the bacteria was separated from the phage. A mixed phage stock of all coevolved genotypes was created from each flask by adding 500 µl of chloroform to the remaining culture in order to kill the bacterial cells, which were removed by centrifugation (21). Two phage clones were isolated from each of these mixed phage stocks by applying an aliquot of diluted stocks onto soft agar plates and picking isogenic 'plaques'. Soft agar plates are created by suspending an isogenic population of bacteria combined with the diluted phage stock in a thin agar matrix on top of a petri dish. When a single phage particle infects a bacterial cell trapped in the agar, the phage reproduces and spreads to nearby bacteria, this continues for a number of rounds and a clearing known as a plaque is produced in the 'lawn' of viable bacteria after 24 hours of incubations at 37 °C. This plaque contains an isogenic population of phage that can be removed to create a clonal stock of phage. We made three plates for each coevolved viral population; one from each bacterial clone isolated from the same population and then one of the ancestral bacteria REL606. Clonal phage cultures were created by isolating single plaques from the soft-agar plates and following the procedure given by (21). Plaques on the coevolved bacteria were chosen over ones grown on REL606 to increase the chance of isolating phage that had evolve specialized counter-resistance strategies that have the plietropic consequence of losing the ability to infect the ancestral REL606. Despite this effort, none of the phage isolated lost the ability to infect REL606. Besides favoring plaques on the

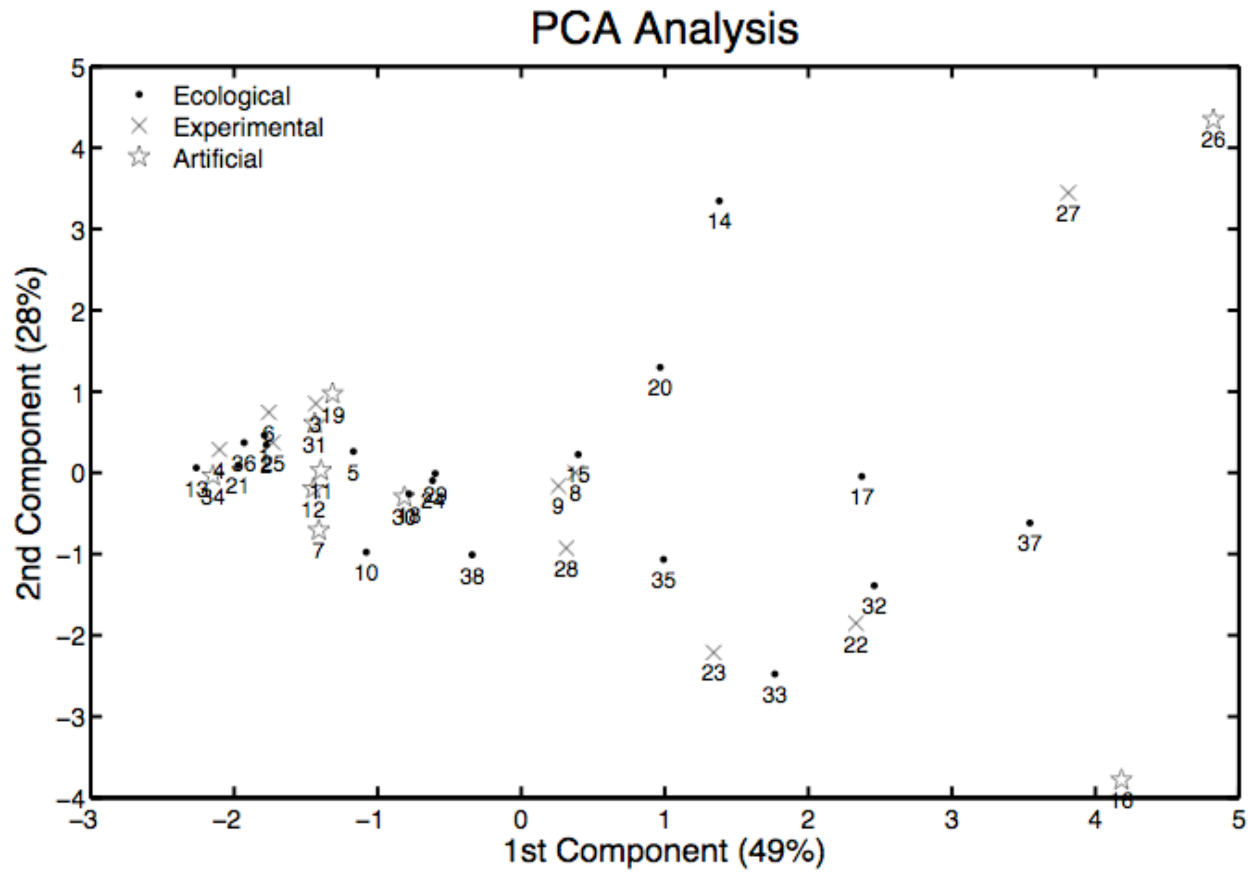
evolved bacterial plates, we tried to choose plaques from separate plates to improve our chances of picking different phage genotypes.

Evaluating patterns of infection and cross-resistance

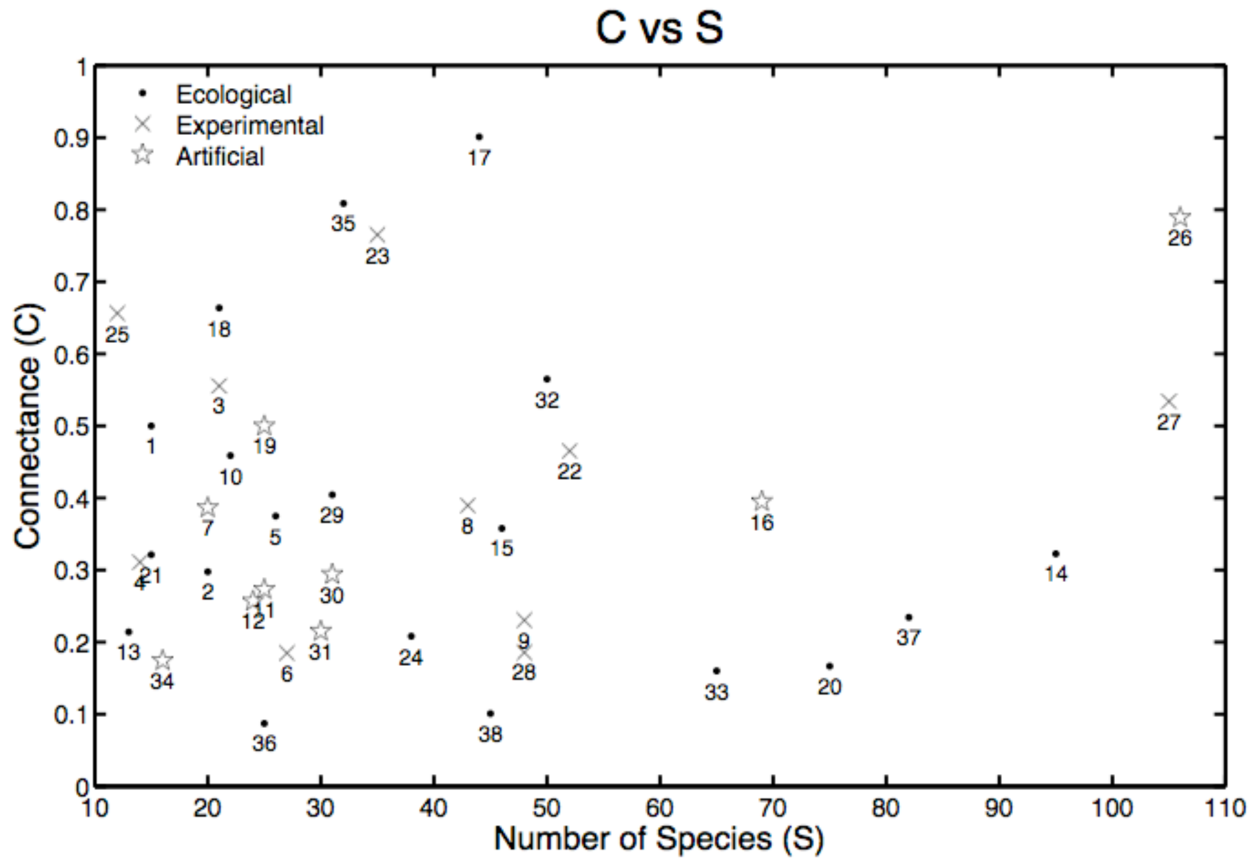
We determined which of the 150 bacteria isolates were resistant to the 150 phage isolates. To do this we preformed 'spot' plate assays. Spot plates are created just as the soft agar plates above were, except instead of combining dilute samples of phage into the agar, one drops 2 µl of concentrated phage stock on top of the bacterial-agar matrix. If the phage is able to infect and reproduce on the bacterium, then a clearing or 'spot' larger than a single plaque will form in the bacterial lawn after 24 hours of incubations at 37 °C. If any clearing or inhibition of bacterial growth larger than a single plaque was observed a '1' was recorded. Plaque-sized clearings were excluded because they likely represent cross-contamination or a mutant phage that has a broader host-range than the originally isolated phage. All bacterial-phage combinations without '1's were given '0's. All bacterial phage combinations were replicated five separate times, a total of 28,125 spots were assayed. To make this processes more efficient, we placed up to 96 separate phage stocks onto a single dish (150mm radius). Phage stock replicates were never placed on the same plate in order to reduce the signal of any stochastic plating effects. The five replicates were combined and a phage was only determined to be able to infect a bacterium if 3 of 5 replicates were given '1's. Lastly, phage or bacteria that had identical infection or resistance profiles as their ancestors were removed from the matrix.

References

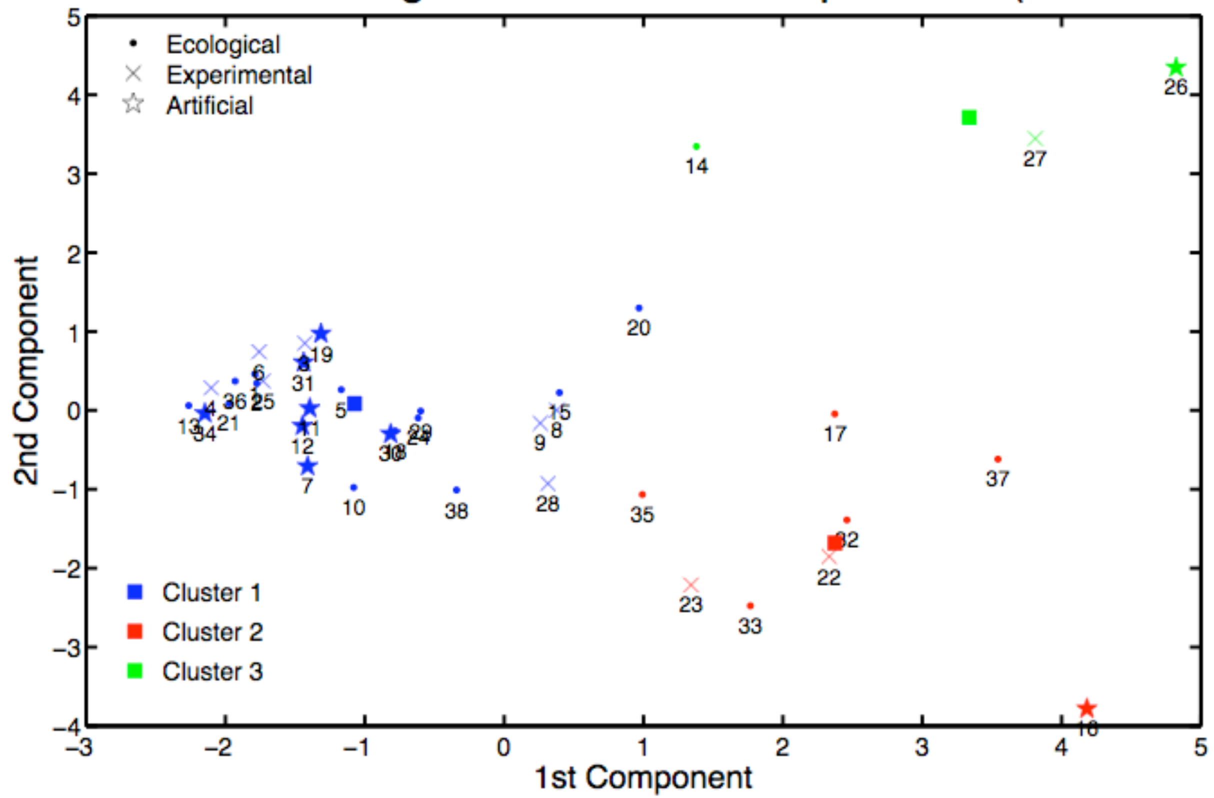
1. Quiberoni A (2003) Comparative analysis of Streptococcus thermophilus bacteriophages isolated from a yogurt industrial plant. *Food Microbiol.* 20(4):461-469.
2. Guimera R, Sales-Pardo M, & Amaral L (2004) Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70(2):25101.
3. Newman M & Girvan M (2004) Finding and evaluating community structure in networks. *Physical Review E* 69(2):26113.
4. Fortuna M*, et al.* (2010) Nestedness versus modularity in ecological networks: two sides of the same coin? *Journal of Animal Ecology* 79(4):811-817.
5. Barber M (2007) Modularity and community detection in bipartite networks. *Physical Review E* 76(6):1-9.
6. Garey M & Johnson D (1983) Crossing number is NP-complete. *SIAM Journal on Algebraic and Discrete Methods* 4:312.
7. Fortunato S & Barthélemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 104(1):36-41.
8. Bascompte J, Jordano P, Melián CJ, & Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. U. S. A.* 100(16):9383-9387.
9. Atmar W & Patterson BD (1993) The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* 96(3):373-382.
10. Rodriguez-Girones Ma & Santamaria L (2006) A new algorithm to calculate the nestedness temperature of presence-absence matrices. *Journal of Biogeography* 33(5):924-935.
11. Jensen EC*, et al.* (1998) Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* 64(2):575-580.
12. Sillankorva S, Neubauer P, & Azeredo J (2008) Isolation and characterization of a T7-like lytic phage for Pseudomonas fluorescens. *BMC Biotechnol.* 8:80-80.
13. Yu MX, Slater MR, & Ackermann HW (2006) Isolation and characterization of Thermus bacteriophages. *Arch. Virol.* 151(4):663-679.
14. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. (California, USA), p 14.
15. Carlton BC & Brown BJ (1981) Gene mutation. *Manual of methods for general bacteriology*, ed Gerhard P (American Society of Microbiology, Washington DC), pp 222-242.
16. Lenski RE & Levin BR (1985) Constraints on the coevolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities. *Am. Nat.* 125(4):585-602.
17. Spanakis E & Horne MT (1987) Co-adaptation of *Escherichia coli* and Coliphage λvir in Continuous Culture *J. Gen. Microbiol.* 133:353-360.
18. Weitz JS, Hartman H, & Levin SA (2005) Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl. Acad. Sci. U. S. A.* 102(27):9535-9535.
19. Daegelen P, Studier FW, Lenski RE, Cure S, & Kim JF (2009) Tracing ancestors and relatives of Escherichia coli B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* 394(4):634-643.
20. Sambrook J, Fritsch EF, & Maniatis T (1989) *Molecular Cloning* (Cold Spring Harbor Laboratory Press, Plainview, NY).
21. Adams MH (1959) *Bacteriophages* (Interscience, New York).
22. Doi K*, et al.* (2003) A comparative study and phage typing of silage-making Lactobacillus bacteriophages. *J. Biosci. Bioeng.* 95(5):518-525.



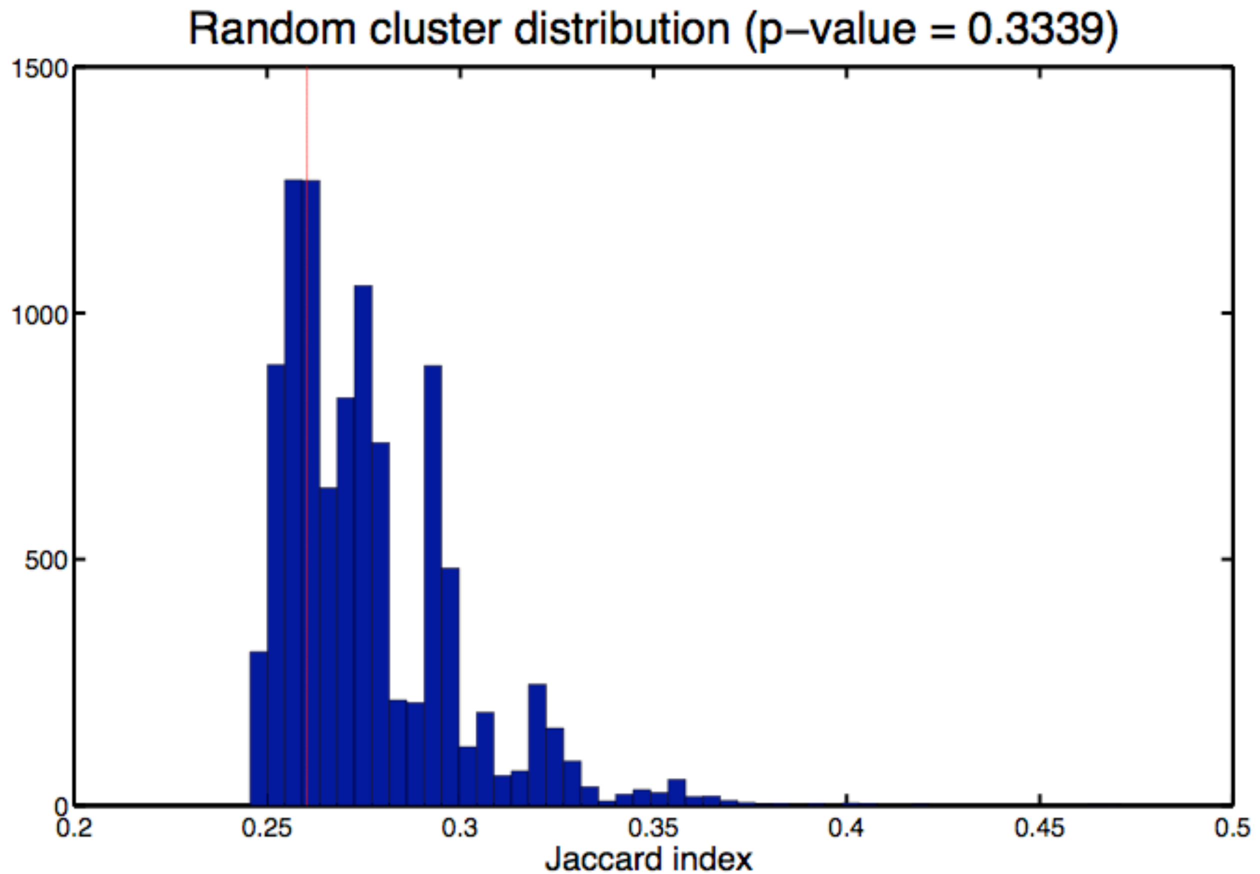
Supplementary Figure 1 PCA Analysis in the global properties of the collected studies. Only the two main components are showed. There is no distinction between the three different type of studies.



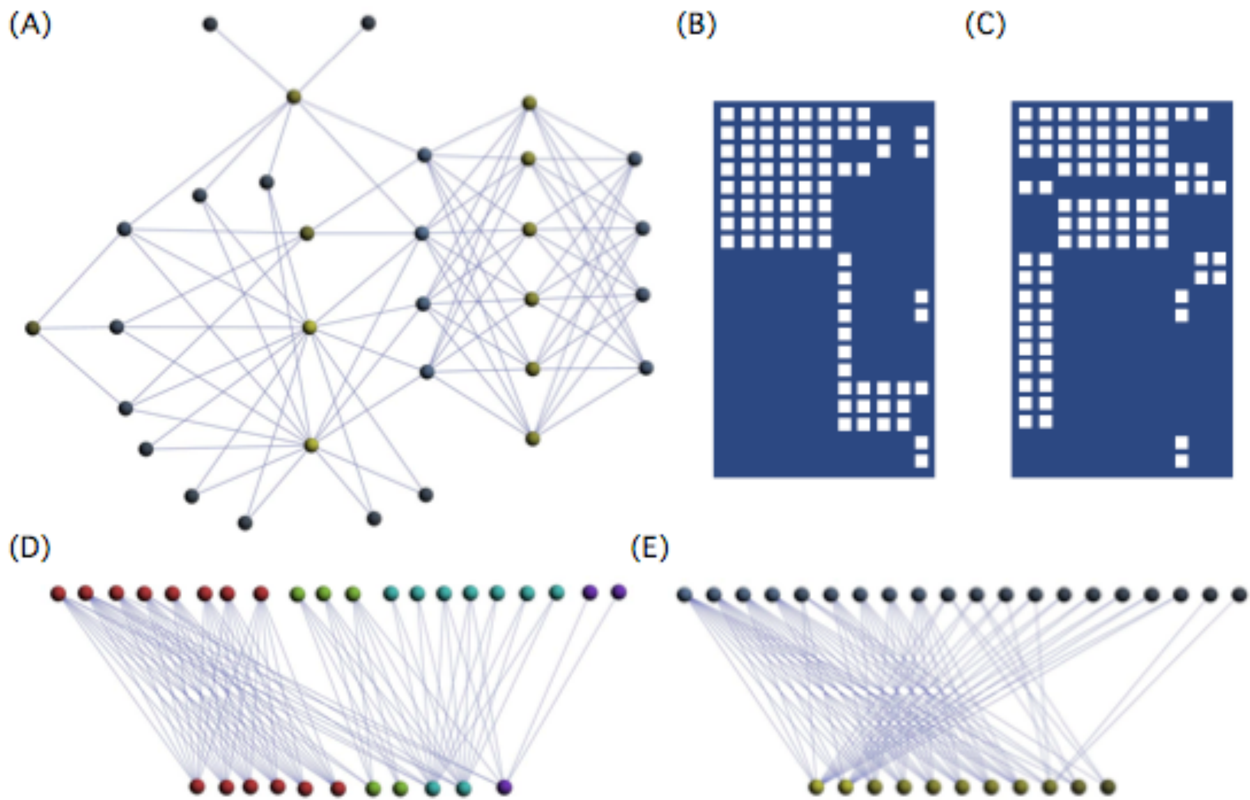
Supplementary Figure 2 Correlation between connectance (*C*) and number of species (*S*). This plot shows that there is no relation between the connectance and the number of species. Numbers in both plots indicate the study id that can be consulted in the appendix



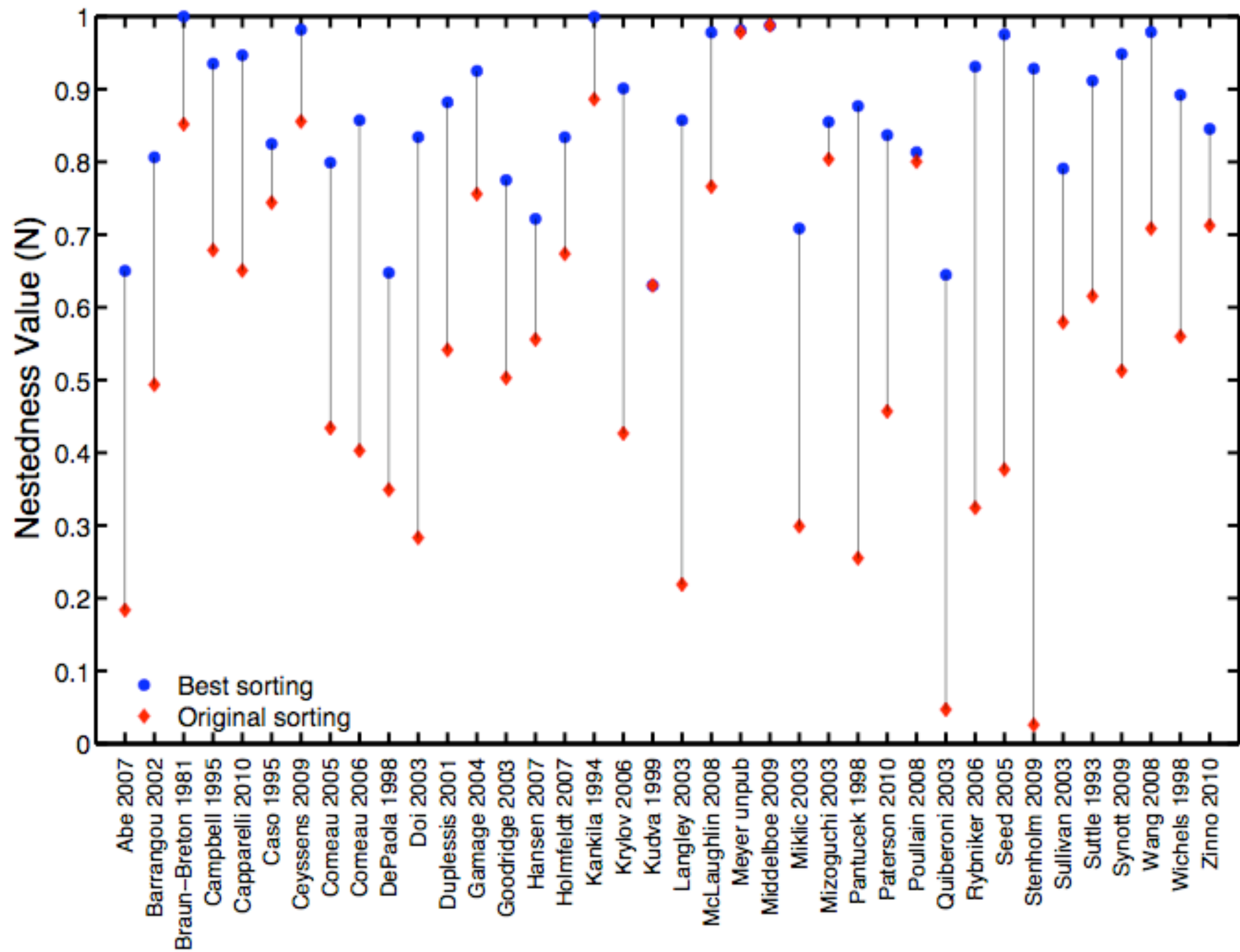
Supplementary Figure 3 Output of the k-means (with $k = 3$) algorithm when applied to the two main components of the PCA-analysis output.



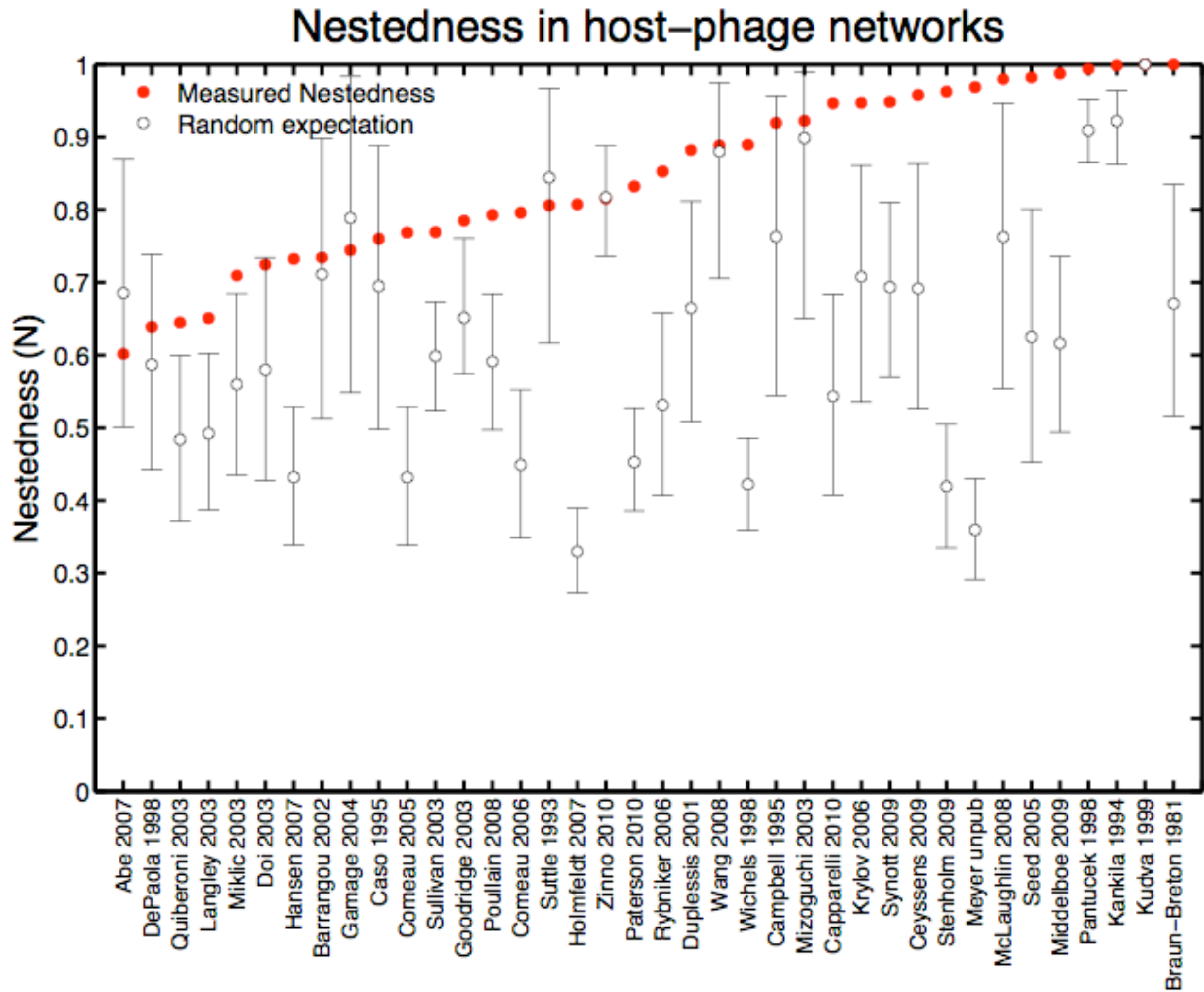
Supplementary Figure 4 Distribution of clustering validity of source types (EXP, NAT and ART) based on global properties. The histogram denotes 10,000 randomization trials in which the labels of each study were relabeled while retaining the total number of each class (EXP, NAT and ART). The value on the x-axis is the Jaccard index of clustering validity (see Supplementary Materials and Methods). The red line denotes the observed clustering validity for the data set which is non-significant, $p = 0.34$.



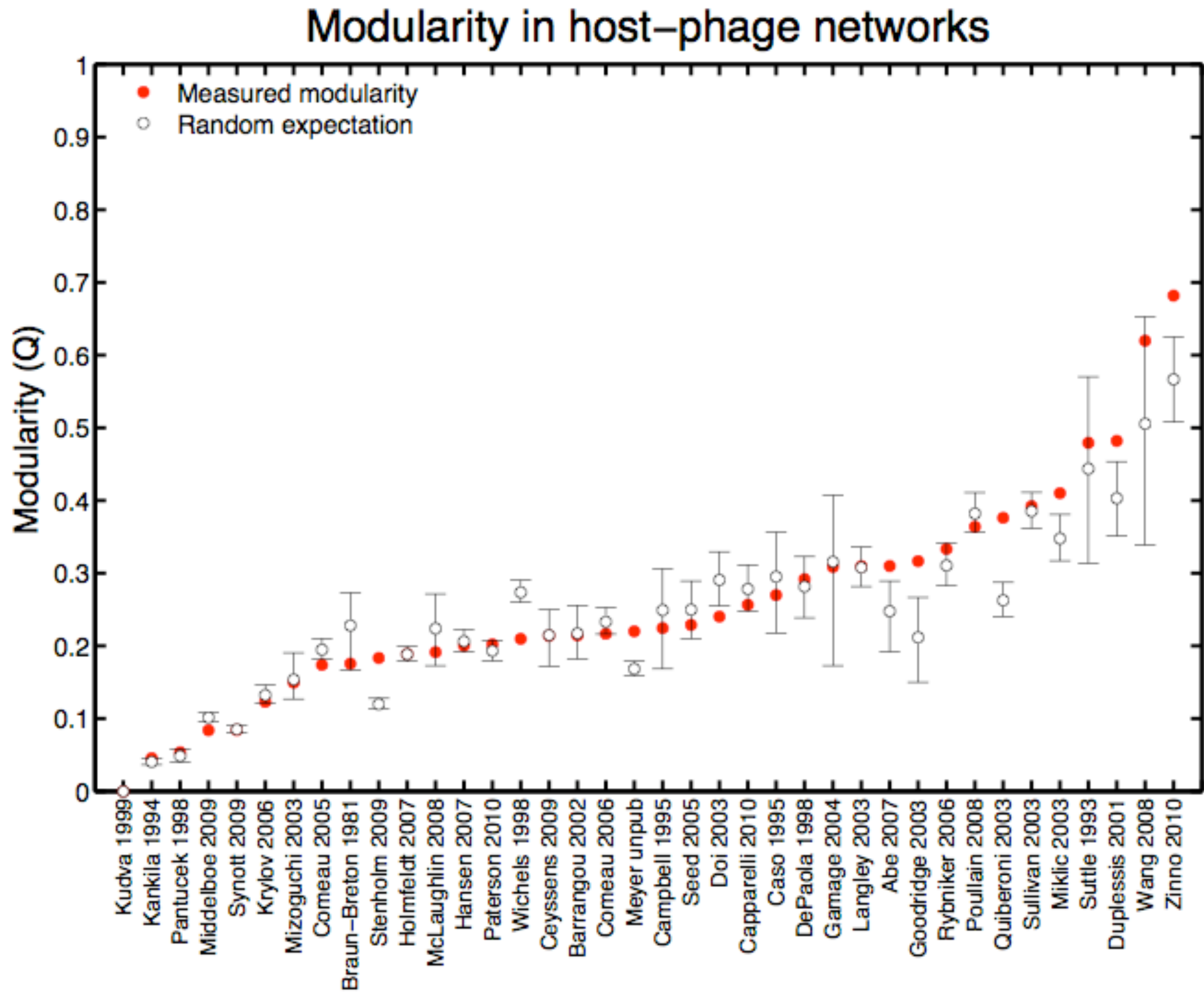
Supplementary Figure 5 Matrix and network representations reveal non-random patterns in host-phage networks. (A) Force-directed layout of the host-phage network where yellow and blue nodes represent phages and hosts, respectively. Shading represents the number of node connections, or degree (see text). We can re-arrange the rows and columns of the adjacency matrix according to optimal network modularity (B) and degree of nestedness (C). (D) Strong modularity indicates the presence of subsets of nodes with the same color (communities) having many more internal links than external links (i.e., less crossings across different modules). (E) Network representation evidences a high degree of nestedness overall, with a few unexpected interactions between specialist species (on the right). Notice that generalist species have more connections and they are located on the left.



Supplementary Figure 6 Nestedness value compared for the original publication format of the matrix (red diamonds) vs. the value found in this study (blue circles). X-axis lists all studies in alphabetical order. Y-axis denotes the value of nestedness. Lines connect the points for ease of comparison. Note that in all cases the current value exceeded that of the original publication.



Supplementary Figure 7 Statistical distribution of nestedness for random matrices compared to that of the original data. Here, empty rows/columns from all matrices were removed so that matrices only contain hosts that were infected by at least one phage and phages that infected at least one host. Error bars denote *95 %* confidence intervals based on 10^5 randomizations of appropriately randomized null networks. Here 26/38 are significantly nested, where Doi et al.(22) is the only study to no longer be significant at the 0.05 level compared to the original data, yet it remains highly nested (p = 0.067).



Supplementary Figure 8 Statistical distribution of modularity for random matrices compared to that of the original data. Here, empty rows/columns from all matrices were removed so that matrices only contain hosts that were infected by at least one phage and phages that infected at least one host. Error bars denote *95 %* confidence intervals based on 10^5 randomizations of appropriately randomized null networks. Here 9/38 are significantly modular as opposed to 6/38 which were significantly modular in the original data.

Supplementary Table 1 Characteristics of complete host-phage networks included in the present study (1-37)

	Reference	*Source Type*	*H*	*P*	*S*	*I*	*M*	*C*	*Lp*	L_h
1	(1) Abe (2007)	ecological	11	4	15	22	44	0.50	5.50	2.00
2	(2) Barrangou (2002)	ecological	14	6	20	25	84	0.30	4.17	1.79
3	(3) Braun-Brenton (1981)	experimental	18	3	21	30	54	0.56	10.00	1.67
4	(4) Campbell (1995)	experimental	9	5	14	14	45	0.31	2.80	1.56
5	(5) Capparelli (2010)	ecological	18	8	26	54	144	0.38	6.75	3.00
6	(6) Caso (1995)	experimental	23	4	27	17	92	0.18	4.25	0.74
7	(7) Ceyssens (2009)	artificial	5	15	20	29	75	0.39	1.93	5.80
8	(8) Comeau (2005)	experimental	30	13	43	152	390	0.39	11.69	5.07
9	(9) Comeau (2006)	experimental	32	16	48	118	512	0.23	7.38	3.69
10	(10) DePaola (1998)	ecological	5	17	22	39	85	0.46	2.29	7.80
11	(11) Doi (2003)	artificial	15	10	25	41	150	0.27	4.10	2.73
12	(12) Duplessis (2001)	artificial	12	12	24	37	144	0.26	3.08	3.08
13	(13) Gamage (2004)	ecological	6	7	13	9	42	0.21	1.29	1.50
14	(14) Goodridge (2003)	ecological	93	2	95	60	186	0.32	30.00	0.65
15	(15) Hansen (2007)	ecological	34	12	46	146	408	0.36	12.17	4.29
16	(16) Holmfeldt (2007)	artificial	23	46	69	418	1058	0.40	9.09	18.17
17	(17) Kankila (1994)	ecological	32	12	44	346	384	0.90	28.83	10.81
18	(18) Krylov (2006)	ecological	11	10	21	73	110	0.66	7.30	6.64
19	(19) Kudva (1999)	artificial	22	3	25	33	66	0.50	11.00	1.50
20	(20) Langley (2003)	ecological	66	9	75	99	594	0.17	11.00	1.50
21	(21) McLaughlin (2008)	ecological	8	7	15	18	56	0.32	2.57	2.25
22	Meyer (unpub)	experimental	25	27	52	314	675	0.47	11.63	12.56
23	(22) Middelboe (2009)	experimental	11	24	35	202	264	0.77	8.42	18.36
24	(23) Miklic (2003)	ecological	24	14	38	70	336	0.21	5.00	2.92
25	(24) Mizoguchi (2003)	experimental	8	4	12	21	32	0.66	5.25	2.63
26	(25) Pantucek (1998)	artificial	102	4	106	322	408	0.79	80.50	3.16
27	(26) Paterson (2010)	experimental	100	5	105	267	500	0.53	53.40	2.67
28	(27) Poullain (2008)	experimental	24	24	48	107	576	0.19	4.46	4.46
29	(28) Quiberoni (2002)	ecological	20	11	31	89	220	0.40	8.09	4.45
30	(29) Rybniker (2006)	artificial	17	14	31	70	238	0.29	5.00	4.12
31	(30) Seed (2005)	artificial	24	6	30	31	144	0.22	5.17	1.29
32	(31) Stenholm (2008)	ecological	28	22	50	348	616	0.56	15.82	12.43
33	(32) Sullivan (2003)	ecological	21	44	65	148	924	0.16	3.36	7.05
34	(33) Suttle (1993)	artificial	7	9	16	11	63	0.17	1.22	1.57
35	(34) Synnott (2009)	ecological	16	16	32	207	256	0.81	12.94	12.94
36	(35) Wang (2008)	ecological	18	7	25	11	126	0.09	1.57	0.61
37	(36) Wichels (1998)	ecological	59	23	82	318	1357	0.23	13.83	5.39
38	(37) Zinno (2010)	ecological	18	27	45	49	486	0.10	1.81	2.72
		Average	26.55	13.21	39.76	114.87	314.32	0.39	10.91	4.88
		Median	19.00	10.50	31.00	65.00	203.00	0.34	6.13	3.04
		Total	1009	502	1511	4365	11944			

First column: These ID's corresponds to indexes in supplementary figures 1-3.

Supplementary Table 2 Characteristics of complete host-phage networks included in the present study, including additional information on biological context of each study (1-37).

ID	*Reference*	*Bacteria*	*Phage*	*Majority source*	*Additional source*	*Isolation Habitat*	*Bacterial association*	*Bacterial trophy*	*Geography*
1	(1) Abe (2007)	*Escherichia coli*	*T2 and PP01*	ecological	artificial		human pathogen	heterotrophic	
2	(2) Barrangou (2002)	*Leuconostoc*	*Caudovirales*	ecological	artificial	sauerkraut	free	heterotrophic	North Carolina, USA
3	(3) Braun-Brenton (1981)	*Escherichia coli*	*λ*	experimental		lab-agar plates	human symbiont	heterotrophic	
4	(4) Campbell (1995)	*Pseudomonas*	*Myoviridae*	experimental	ecological	barley roots	plant symbiont	heterotrophic	Hojbakkegaard, Denmark
5	(5) Capparelli (2010)	*Salmonella*		ecological		gastroenteritis patients	human pathogen	heterotrophic	Europe
6	(6) Caso (1995)	*Lactobacillus*	*Siphoviridae*	experimental		food, fresh water, soil, sewage	free	heterotrophic	Spain
7	(7) Ceyssens (2009)	*Pseudomonas aeruginosa*		artificial		hospital sewage, fresh water	human pathogen	heterotrophic	global
8	(8) Comeau (2005)	*Vibrio*		experimental		marine	human pathogen / oysters	heterotrophic	British Columbia, Canada
9	(9) Comeau (2006)	*Vibrio*	*Siphoviridae and Podoviridae*	experimental		marine	human pathogen	heterotrophic	British Columbia, Canada
10	(10) DePaola (1998)	*Vibrio vulnificus*	*Podoviridae, Styloviridae, and Myoviridae*	ecological		marine	human pathogen / oysters	heterotrophic	Gulf of Mexico
11	(11) Doi (2003)	*Lactobacillus*	*Siphoviridae and Myoviridae*	artificial		silage (fermented bovine feed)	free	heterotrophic	Japan
12	(12) Duplessis (2001)	*Streptococcus thermophilus*	*Myoviridae and Siphoviridae*	artificial		Industrial cheese plants	free	heterotrophic	Quebec, Canada
13	(13) Gamage (2004)	*Escherichia coli*		ecological		human and animal fecal isolates	human pathogen	heterotrophic	Ohio, USA
14	(14) Goodridge (2003)	*Enterobacteriaceae*	*Myoviridae*	ecological		human and animal	human pathogen	heterotrophic	global
15	(15) Hansen (2007)	*Campylobacter*	*Myoviridae*	ecological		poultry intestine	human pathogen	heterotrophic	Denmark
16	(16) Holmfeldt (2007)	*Flavobacteriaceae*	*Myoviridae, Siphoviridae, and Podoviridae*	artificial	ecological	marine	free	heterotrophic	Scandinavia
17	(17) Kankila (1994)	*Rhizobium*		ecological		soil	free	heterotrophic	Finland
18	(18) Krylov (2006)	*Escherichia and Salmonella*	*T-even superfamily*	ecological		sewage	human pathogen	heterotrophic	
19	(19) Kudva (1999)	*Enterobacteriaceae*		artificial		bovine and ovine feces	human pathogen	heterotrophic	North West USA
20	(20) Langley (2003)	*Burkholderia*	*T-even and λ - like*	ecological	artificial	soil, freshwater, plant	mycorrhizal	heterotrophic	global

21	(21) McLaughlin (2008)	*Salmonella*		ecological	artificial	swine lagoon	human pathogen	heterotrophic	Mississippi, USA
22	Meyer (unpub)	*Escherichia*	*λ*	experimental		lab - batch culture	human symbiont	heterotrophic	
23	(22) Middelboe (2009)	*Cellulophaga baltica*	*Myoviridae, Siphoviridae, and Podoviridae*	experimental	ecological	marine	free	photosynthetic	Scandinavia
24	(23) Miklic (2003)	*Lactococcus lactis*	*Siphoviridae*	ecological		dairy products	free	heterotrophic	Solvania
25	(24) Mizoguchi (2003)	*Escherichia coli*	*PP01*	experimental		lab-chemostat	human pathogen	heterotrophic	
26	(25) Pantucek (1998)	*Staphylococcus*	*polyvalent staphylophage*	artificial		clinical isolates	human pathogen	heterotrophic	Brno, Czech Republic
27	(26) Paterson (2010)	*Pseudomonas fluorescens*	*ϕ2*	experimental		lab - batch culture	plant symbiont	heterotrophic	UK
28	(27) Poullain (2008)	*Pseudomonas fluorescens*	*ϕ2*	experimental		lab - batch culture	plant symbiont	heterotrophic	UK
29	(28) Quiberoni (2002)	*Streptococcus thermophilus*	*Siphoviridae*	ecological		yogurt industrial plant	free	heterotrophic	Argentina
30	(29) Rybniker (2006)	*Mycobacterium*		artificial		soil	human pathogen	heterotrophic	global
31	(30) Seed (2005)	*Burkholderia*	*Myoviridae*	artificial		soil, freshwater, plant	human pathogen	heterotrophic	
32	(31) Stenholm (2008)	*Flavobacterium psychrophilum*	*Siphoviridae, Myoviridae, and Podoviridae*	ecological		fresh water	fish pathogen	heterotrophic	Denmark
33	(32) Sullivan (2003)	*Prochlorococcus Synechococcus*	*Myoviridae and Podoviridae*	ecological		marine	free	photosynthetic	Atlantic Ocean
34	(33) Suttle (1993)	*Synechococcus and Anacystis*	*Siphoviridae, Myoviridae, and Podoviridae*	artificial	ecological	marine	free	photosynthetic	Texas, USA
35	(34) Synnott (2009)	*Staphylococcus aureus*	*Myoviridae*	ecological		sewage, dairy products	bovine pathogen	heterotrophic	Tokyo, Japan
36	(35) Wang (2008)	*Synechococcus and Prochlorococcus*	*Myoviridae and Podoviridae*	ecological		marine	free	photosynthetic	Chesapeake Bay, USA
37	(36) Wichels (1998)	*Pseudoalteromonas*	*Siphoviridae, Myoviridae, and Podoviridae*	ecological		marine	free	heterotrophic	North Sea, Germany
38	(37) Zinno (2010)	*Streptococcus thermophilus*		ecological		dairy products	free	heterotrophic	Italy

First column: These ID's corresponds to indexes in supplementary figures 1-3.

Supplementary Table 3 Global properties

Property	*Definition*
H	number of hosts
P	number of phages
I	number of interactions
$S = H + P$	number of species
$M = HP$	size
$C = I/M$	connectance
$L_H = I/H$	mean number of interactions across host species
$L_P = I/P$	mean number of interactions across phage species

Supplementary Table 4 PCA Analysis

	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*
H	0.352	0.446	-0.179	0.131	0.389	-0.131	-0.097	0.670
P	0.247	-0.534	-0.203	0.474	-0.461	-0.140	-0.279	0.279
I	0.470	-0.138	0.143	-0.474	0.008	0.517	-0.498	0.000
$S = H + P$	0.444	0.218	-0.257	0.320	0.192	-0.184	-0.208	-0.688
$M = HP$	0.397	-0.239	-0.359	-0.542	-0.078	-0.373	0.466	0.000
$C = I/M$	0.188	0.062	0.743	-0.093	-0.112	-0.601	-0.164	0.000
$L_H = I/H$	0.281	-0.449	0.359	0.313	0.504	0.224	0.435	0.000
$L_P = I/P$	0.353	0.431	0.177	0.177	-0.571	0.335	0.434	0.000
	48.95%	*27.98%*	*18.55%*	*2.03%*	*1.30%*	*1.07%*	*0.11%*	*0.00*

Supplementary Table 5 Correlation analysis

	H	***P***	***S***	***I***	***M***	***C***	***Lp***	***Lh***
H	1.000	-0.146	0.916	0.458	0.394	0.125	0.847	-0.133
P	-0.146	1.000	0.264	0.535	0.744	-0.110	-0.191	0.697
S	0.916	0.264	1.000	0.664	0.686	0.077	0.748	0.154
I	0.458	0.535	0.664	1.000	0.752	0.466	0.553	0.716
M	0.394	0.744	0.686	0.752	1.000	-0.109	0.204	0.449
C	0.125	-0.110	0.077	0.466	-0.109	1.000	0.501	0.517
Lp	0.847	-0.191	0.748	0.553	0.204	0.501	1.000	0.035
Lh	-0.133	0.697	0.154	0.716	0.449	0.517	0.035	1.000

In Green p-value < 0.05/28

In Yellow 0.05/28 < p-value < 0.05/8

Supplementary Table 6 Isolation bias

Study	*Modularity*		*Nestedness*	
	Original	*Recalculated*	*Original*	*Recalculated*
Krylov 2006	0.123	0.136	0.901	0.839
Kudva 1999	0	0	0.630	0.630
McLaughlin 2008 - Matrix minus TSB control	0.191	0.191	0.978	0.951
McLaughlin 2008 - Matrix minus TSB minus isolation host	0.191	0.313	0.978	1.000
Middleboe 2009	0.084	0.079	0.988	0.980
Rybniker 2006	0.333	0.274	0.931	0.908
Stenholm 2009	0.183	0.187	0.928	0.931

In Red significant modular/nested studies
In Blue significant anti- modular/nested studies

References:

1. Abe M, Izumoji Y, & Tanji Y (2007) Phenotypic transformation including host-range transition through superinfection of T-even phages. *FEMS Microbiol. Lett.* 269(1):145-152.
2. Barrangou R, Yoon SS, F., Fleming HP, & Klaenhammer TR (2002) Characterization of six Leuconostoc fallax bacteriophages isolated from an industrial sauerkraut fermentation. *Appl. Environ. Microbiol.* 68(11):5452-5452.
3. Braun-Breton C & Hofnung M (1981) Invivo and invitro functional alterations of the bacteriophage-lambda receptor in LamB missense mutants of Escherichia coli K-12. *J. Bacteriol.* 148(3):845-852.
4. Campbell JIA, Albrechtsen M, & Sørensen J (2006) Large Pseudomonas phages isolated from barley rhizosphere. *FEMS Microbiol. Ecol.* 18(1):63-74.
5. Capparelli R*, et al.* (2010) Bacteriophage therapy of Salmonella enterica: a fresh appraisal of bacteriophage therapy. *The Journal of Infectious Diseases* 201(1):52-61.
6. Caso JL*, et al.* (1995) Isolation and characterization of temperate and virulent bacteriophages of Lactobacillus plantarum. *J. Dairy Sci.* 78(4):741-750.
7. Ceyssens P-J*, et al.* (2009) Survey of Pseudomonas aeruginosa and its phages: de novo peptide sequencing as a novel tool to assess the diversity of worldwide collected viruses. *Environ. Microbiol.* 11(5):1303-1313.
8. Comeau AM, Buenaventura E, & Suttle CA (2005) A persistent, productive, and seasonally dynamic vibriophage population within Pacific oysters (Crassostrea gigas). *Appl. Environ. Microbiol.* 71(9):5324-5324.
9. Comeau AM, Chan AM, & Suttle CA (2006) Genetic richness of vibriophages isolated in a coastal environment. *Environ. Microbiol.* 8(7):1164-1176.
10. DePaola A, Motes ML, Chan AM, & Suttle CA (1998) Phages infecting Vibrio vulnificus are abundant and diverse in oysters (Crassostrea virginica) collected from the Gulf of Mexico. *Appl. Environ. Microbiol.* 64(1):346-351.
11. Doi K*, et al.* (2003) A comparative study and phage typing of silage-making Lactobacillus bacteriophages. *J. Biosci. Bioeng.* 95(5):518-525.
12. Duplessis M & Moineau S (2001) Identification of a genetic determinant responsible for host specificity in Streptococcus thermophilus bacteriophages. *Mol. Microbiol.* 41(2):325-336.
13. Gamage SD, Patton AK, Hanson JF, & Weiss AA (2004) Diversity and host range of Shiga toxin-encoding phage. *Infect. Immun.* 72(12):7131-7131.
14. Goodridge L, Gallaccio A, & Griffiths MW (2003) Morphological, host range, and genetic characterization of two coliphages. *Appl. Environ. Microbiol.* 69(9):5364-5364.
15. Hansen VM, Rosenquist H, Baggesen DL, Brown S, & Christensen BB (2007) Characterization of Campylobacter phages including analysis of host range by selected Campylobacter Penner serotypes. *BMC Microbiol.* 7:90-90.

16. Holmfeldt K, Middelboe M, Nybroe O, & Riemann L (2007) Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. *Appl. Environ. Microbiol.* 73(21):6730-6739.
17. Kankila J & Lindstrom K (1994) Host range, morphology and DNA restriction patterns of bacteriophage isolates infecting Rhizobium leguminosarum bv. trifolii. *Soil Biol. Biochem.* 26(4):429-437.
18. Krylov VN*, et al.* (2006) Ambivalent bacteriophages of different species active on Escherichia coli K12 and Salmonella sp. strains. *Russ. J. Genet.* 42(2):106-114.
19. Kudva IT, Jelacic S, Tarr PI, Youderian P, & Hovde CJ (1999) Biocontrol of Escherichia coli O157 with O157-specific bacteriophages. *Appl. Environ. Microbiol.* 65(9):3767-3773.
20. Langley R (2003) Lysogeny and bacteriophage host range within the Burkholderia cepacia complex. *J. Med. Microbiol.* 52(6):483-490.
21. McLaughlin MR & King RA (2008) Characterization of Salmonella bacteriophages isolated from swine lagoon effluent. *Curr. Microbiol.* 56(3):208-213.
22. Middelboe M, Holmfeldt K, Riemann L, Nybroe O, & Haaber J (2009) Bacteriophages drive strain diversification in a marine Flavobacterium: implications for phage resistance and physiological properties. *Environ. Microbiol.* 11(8):1971-1982.
23. Miklic A & Rogelj I (2003) Characterization of lactococcal bacteriophages isolated from Slovenian dairies. *Int. J. Food Sci. Technol.* 38(3):305-311.
24. Mizoguchi K*, et al.* (2003) Coevolution of bacteriophage PP01 and Escherichia coli O157: H7 in continuous culture. *Appl. Environ. Microbiol.* 69(1):170-176.
25. Pantůcek R*, et al.* (1998) The polyvalent staphylococcal phage phi 812: its host-range mutants and related phages. *Virology* 246(2):241-252.
26. Paterson S*, et al.* (2010) Antagonistic coevolution accelerates molecular evolution. *Nature* 464(7286):275-U154.
27. Poullain V, Gandon S, Brockhurst MA, Buckling A, & Hochberg ME (2008) The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. *Evolution* 62(1):1-11.
28. Quiberoni A (2003) Comparative analysis of Streptococcus thermophilus bacteriophages isolated from a yogurt industrial plant. *Food Microbiol.* 20(4):461-469.
29. Rybniker J, Kramme S, & Small PL (2006) Host range of 14 mycobacteriophages in Mycobacterium ulcerans and seven other mycobacteria including Mycobacterium tuberculosis--application for identification and susceptibility testing. *J. Med. Microbiol.* 55(Pt 1):37-42.
30. Seed KD & Dennis JJ (2005) Isolation and characterization of bacteriophages of the Burkholderia cepacia complex. *FEMS Microbiol. Lett.* 251(2):273-280.
31. Stenholm AR, Dalsgaard I, & Middelboe M (2008) Isolation and characterization of bacteriophages infecting the fish pathogen Flavobacterium psychrophilum. *Appl. Environ. Microbiol.* 74(13):4070-4078.
32. Sullivan MB, Waterbury JB, & Chisholm SW (2003) Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature* 424(6952):1047-1051.
33. Suttle C & Chan A (1993) Marine cyanophages infecting oceanic and coastal strains of Synechococcus: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* 92:99-109.
34. Synnott AJ*, et al.* (2009) Isolation from sewage influent and characterization of novel Staphylococcus aureus bacteriophages with wide host ranges and potent lytic capabilities. *Appl. Environ. Microbiol.* 75(13):4483-4490.
35. Wang K & Chen F (2008) Prevalence of highly host-specific cyanophages in the estuarine environment. *Environ. Microbiol.* 10(2):300-312.
36. Wichels A*, et al.* (1998) Bacteriophage diversity in the North Sea. *Appl. Environ. Microbiol.* 64(11):4128-4133.
37. Zinno P, Janzen T, Bennedsen M, Ercolini D, & Mauriello G (2010) Characterization of Streptococcus thermophilus lytic bacteriophages from mozzarella cheese plants. *Int. J. Food Microbiol.* 138(1-2):137-144.