Statistical structure of host–phage interactions

Cesar O. Flores¹,², Justin R. Meyer³,¹, Sergi Valverde⁵, Lauren Farr⁴, and Joshua S. Weitz¹,²

¹School of Physics, Georgia Institute of Technology, Atlanta, GA 30332; ²Department of Zoology, Michigan State University, East Lansing, MI 48824; ³Complex Systems Lab and Institute of Evolutionary Biology, University Pompeu Fabra, E-08003 Barcelona, Spain; and ⁴School of Biology, Georgia Institute of Technology, Atlanta, GA 30332

Edited by Simon A. Levin, Princeton University, Princeton, NJ, and approved May 27, 2011 (received for review January 27, 2011)

Interactions between bacteria and the viruses that infect them (i.e., phages) have profound effects on biological processes, but despite their importance, little is known on the general structure of infection and resistance between most phages and bacteria. For example, are bacteria–phage communities characterized by complex patterns of overlapping exploitation networks, or do they conform to a more ordered general pattern across all communities, or are they idiosyncratic and hard to predict from one ecosystem to the next? To answer these questions, we collect and present a detailed metaanalysis of 38 laboratory-veriﬁed studies of host–phage interactions representing almost 12,000 distinct experimental infection assays across a broad spectrum of taxa, habitat, and mode of selection. In so doing, we present evidence that currently available host–phage infection networks are statistically different from random networks and that they possess a characteristic nested structure. This nested structure is typiﬁed by the finding that hard to infect bacteria are infected by generalist phages (and not specialist phages) and that easy to infect bacteria are infected by generalist and specialist phages. Moreover, we ﬁnd that currently available host–phage infection networks do not typically possess a modular structure. We explore possible underlying mechanisms and significance of the observed nested host–phage interaction structure. In addition, given that most of the available host–phage infection networks examined here are composed of taxa separated by short phylogenetic distances, we propose that the lack of modularity is a scale-dependent effect, and then, we describe experimental studies to test whether modular patterns exist at macroevolutionary scales.

Bacteria and their viruses (phages) make up two of the most abundant and genetically diverse groups of organisms (1–3). The extent of this diversity has become increasingly apparent with the advent of community genomics. Microbial DNA isolated from oceans, lakes, soils, and human guts has revealed tremendous taxonomic diversity in a broad range of environmental habitats and conditions (4–11). The ongoing discovery of new taxonomic diversity has, thus far, outpaced gains in understanding the function of specific microbes and their most basic ecology of who interacts with whom. One of the starkest examples of this disparity is the lack of an efﬁcient (bioinformatic or otherwise) approach for determining which viruses can infect which bacteria. Although it is well-known that individual phages do not infect all bacteria, we have little understanding of what the precise host range for any given phage is or whether there are universal patterns or principles governing the set of viruses able to infect a given bacterium and the set of bacteria that a given virus can infect. This deﬁcit is unfortunate given that phage–bacterial interactions are important for both human health and ecosystem function (12–16).

Phages have multifaceted effects on their hosts: they can lyse host cells, thereby releasing new virions, transfer genes between hosts, and form lysogens that can modify host function (17–19). In some cases, phages can transfer genes for pathogenicity between pathogenic and labile strains (e.g., for both Vibrio cholerae and Shigella), facilitating the spread of bacterial infections (20–22). Phages also alter ecosystem functions by the high levels of bacterial mortality that they cause. Bacteria lysed by phage will release their contents, which consequently are scavenged by other bacteria rather than being incorporated into bacterivorous eukaryotes (23, 24). This weakened connection early in the food chain can have effects that ripple throughout the ecosystem. Information on a general pattern of infection by phages on hosts could improve predictions of microbial population dynamics, ecosystem function, and microbial community assembly (25, 26).

What is our expectation for the general pattern of host–phage infection networks? Host–phage infection networks have, in the past, been measured by performing pair-wise infections of hosts by phages isolated from natural ecological communities, evolution experiments, or strain collections. The results of such pair-wise infections can be represented as a network or a matrix, where the rows indicate host isolates, the columns indicate phage isolates, and the cells within the matrix describe whether each combination results in a successful infection. We consider different classes of host–phage interaction networks as alternative hypotheses for an expected pattern (Fig. 1). First, phages may infect a unique host or a limited number of closely related hosts, leading to nearly diagonal matrices (Fig. L4) or block-like matrices that exhibit high degrees of modularity (Fig. 1B). These patterns should occur if host–viral interactions are the result of coevolutionary processes that lead to specialization. Second, diversiﬁcation of hosts and phages may result in nested matrices in which the most specialist phages infect those hosts that are most susceptible to infection rather than infecting those hosts that are most resistant to infection (Fig. 1C). The nested pattern is the predicted outcome of a prominent theory of gene–gene coevolution, where phages evolve so as to broaden host ranges and bacteria evolve so as to increase the number of phages to which they are resistant (27, 28). We should note that these two patterns and hypotheses for the forms of coevolution are not mutually exclusive and in fact, could be scale-dependent. Nested patterns could form within modules if, for instance, microevolutionary changes result in nestedness; however, genetic differences between species or genera that accumulate over macroevolutionary time may limit the exchange of viruses between these phylogenetic groups and create an overall modular structure. Finally, we consider a null model to be that matrices of host–phage infection are statistically indistinguishable from random matrices (Fig. 1D).

Contrary to this null expectation, we show that currently available host–phage interaction matrices are, as a whole, sta-

Author contributions: J.R.M. and J.S.W. designed research; C.O.F., J.R.M., S.V., L.F., and J.S.W. performed research; C.O.F., J.R.M., S.V., and J.S.W. contributed new reagents/analytic tools; C.O.F., J.R.M., L.F., and J.S.W. analyzed data; and J.R.M., S.V., and J.S.W. wrote the paper.

The authors declare no conﬂict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

+Author contributions: J.R.M. and J.S.W. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: jsweitz@gatech.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1101595108/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1101595108
tistically distinguishable from random matrices and possess a characteristic nested structure. We reach this conclusion by performing a metaanalysis on the patterns of host–phage infection matrices collected by a comprehensive search of the literature and supplementing these matrices with an experimental analysis of host–phage infection. The data that we assemble consist of 38 matrices of host–phage infection assays representing the cumulative study of 1,009 bacterial isolates, 502 phage isolates, and almost 12,000 separate attempts to infect a bacteria host with a phage strain (27, 29–64) (SI Appendix, Tables S1 and S2 have more information on the examined studies). This work is an attempt to subject host–phage infection assays to a unified analysis. In doing so, we find a general pattern of host–phage interactions. We discuss biophysical, ecological, and evolutionary mechanisms that could lead to this nested (and not modular) pattern as well as future studies to explore how such a pattern may change as a function of phylogenetic scale.

Results
Compiling a Large-Scale Host–Phage Interaction Dataset. We compiled a set of 37 studies with direct laboratory evidence of host–phage interactions using an extensive literature search supplemented by an experimental study of an evolved Escherichia coli and phage λ-system (SI Appendix, Tables S1 and S2 have complete details of all studies) (27, 29–64). The method of evaluating infection ability in assembling a host–phage infection matrix varies; however, the most commonly used approach is that of spot assays, in which a single virus type is combined with a population of bacteria cells from a single strain. Infection is considered to have occurred given evidence that the phage has infected and lysed (part of) the bacterial population. Hence, the result of each study is a matrix of the infection ability for each phage on each host. The studies included in the host–phage infection assays analyzed here were isolated from one of three sources: co-occurring isolates within natural communities taken directly from the environment and then cultured, coevolutionary laboratory experiments where a single bacterial clone and a single phage clone were allowed to coevolve for a fixed amount of time and then, their evolved progenitors examined, and laboratory stocks of phages and hosts that were artificially combined. Some of the matrices used were composed of bacteria and phage acquired from two separate isolation strategies. For these studies, we classified the matrix by which isolation strategy represented the majority of matrix cells and made a note of the other sources (SI Appendix, Table S2). The criterion by which we searched and cataloged these studies is explained in more detail in SI Appendix, SI Materials and Methods. Overall, we identified and analyzed a wide range of infection networks for organisms that varied in their phylogenetic position, traits, and habitats. For example, the bacterial hosts included Gram-positives and -negatives, heterotrophs, and phototrophs as well as pathogens and non-pathogens.

Some of the assays include graded information about infection (for example, whether a phage simply inhibits bacterial growth or forms regions of complete bacterial mortality like plaques). In other studies, replicate phage populations were used to deduce whether phages always only somewhat cause plaques. Details of the criteria for the interactions can be found in the original works (27, 29–64), and the experimental methods for the experimental study of host–phage infection can be found in Materials and Methods. Because graded information about infection was not uniformly available in all studies, assays were standardized using hand-curved extraction of original data into a single matrix of ones and zeros with H rows (one for every bacterial host) and P columns (one for every phage), where a 1-valued cell represents evidence for infection (either full or partial) and a 0-valued cell represents no evidence for infection (Fig. 2 shows a visual depiction of all host–phage interaction matrices).

Host–Phage Infection Statistics Do Not Vary with Study Type or Show Significant Cross-Correlations. We calculated a variety of global properties of these matrices: number of hosts (H), number of phages (P), number of interactions (I), number of species (S = H + P), size (M = HP), connectance (C = I/M), mean number of interactions across host species (L_H = I/H), and mean number of interactions across phage species (L_P = I/P) (SI Appendix, Tables S1, S2, and S3 show values of each property within each of the 38 studies). Importantly, on a per-study basis, we find that the average number of phages infecting a given host is 4.88 (median = 3.04), whereas the average number of hosts that a phage can infect is 10.91 (median = 6.13). Both results are inconsistent with the hypothesis that phages only infect one host and that hosts are only infected by one phage (Fig. L4).

We first sought to establish whether the source type (natural communities taken directly from the environment and then cultured, coevolutionary laboratory experiments where a single bacterial clone and a single phage clone were allowed to coevolve for a fixed amount of time and then, their evolved progenitors examined, and laboratory stocks of phages and hosts that were artificially combined) had any influence on basic characteristics of the matrices. We performed a principal component analysis (SI Appendix, SI Materials and Methods, SI Appendix, Table S4, and SI Appendix, Fig. S1) using these eight global properties. Despite the significant variation in global properties, we find no statistically significant distinction between the three different types of studies. For example, the distributions of type-specific matrices do not cluster into three groups. We apply a Jaccard clustering validity index (65) and find that the degree of clustering validity is 0.26 (indicating poor separa-
tion of labeled classes into distinct clusters), which is not significantly different from random \((P = 0.33)\) (SI Appendix, SI Materials and Methods and SI Appendix, Figs. S3 and S4).

Not only do we not find evidence for clustering, we also do not find evidence for significant and biologically meaningful correlations among the global properties of all matrices when grouped together. For example, previous work on the analysis of bipartite networks within plant and pollinator systems found inverse relationships between the total number of species in the network and the fraction of interactions that actually occurred (66, 67). We do not find this relationship here. SI Appendix, Fig. S2 plots connectance \((C)\) vs. number of species \((S)\). The observed slope is small and nonsignificant (SI Appendix, Table S5). Moreover, the other correlations between connectance and the size of host–phage infection matrices are not significant (Materials and Methods has details and SI Appendix, Table S5 shows the correlation values).

**Host–Phage Infection Assays Are Typically Nested and Not Modular.**

We measured higher-order properties of the host–phage interaction matrices, specifically modularity and nestedness. In this context, modularity is determined by the occurrence of groups of phages that infect groups of hosts significantly more often than they infect other hosts in the system. Modularity is typically found in biological systems in which groups of organisms preferentially interact with organisms within the group (e.g., plant–pollinator network) (66, 67) and is thought to be an important feature underlying the maintenance of biodiversity (68). Likewise, nestedness is determined by the extent to which phages that infect the most hosts tend to infect bacteria that are infected by the fewest phages (69, 70). Nestedness has been used to characterize species interactions because it is predicted to affect important properties of communities such as stability and extinction potential (67, 71). Both modularity and nestedness may emerge because of coevolutionary adaptation of hosts and phages (28, 72). The individual host–phage infection studies collected here were not subjected to a network analysis with one exception (27). Hence, we examined each study to see if previously unrealized patterns existed within each host–phage interaction network (Fig. 3 and SI Appendix, Fig. S5 have an example of how network properties are extracted from two matrices, Datasets S1 and S2 shows data corresponding to each matrix, and Materials and Methods has additional details on how to calculate modularity and nestedness).

For the 38 matrices shown in Fig. 2, the maximally modular relabeling of each matrix is displayed in Fig. 4 and the maximally nested resorting of each matrix is displayed in Fig. 5. To evaluate the statistical significance of the modularity and nestedness values of observed host–phage matrices, we have to compare the observed values to those values of random matrices. We generate random matrices that have the same size and number of interactions as the original data (SI Appendix, SI Materials and Methods). In that way, we constrain our null model to have exactly the same global properties as detailed
The titles of the study in Fig. 4 (the maximally modular configuration) are red if they are significantly modular, blue if they are significantly antimodular, and black if they are nonsignificantly modular. The majority of studies are significantly antimodular (where we used a P value = 0.05 and $10^5$ random matrices as our null). Our findings stand in contrast to expectations that groups of phages adsorb to nonoverlapping groups of hosts, which would be expected if groups of phages had specialized on groups of hosts within the study systems. The titles of each study in Fig. 5 (the maximally nested configuration) are red if they are significantly nested, blue if they are significantly antinested, and black if they are nonsignificantly nested. The majority of studies are significantly nested ($P < 0.05$), where we used $10^5$ random matrices as our null. Overall, we find 27 of 38 studies to be significantly nested, and when broken down by type, we find significant nestedness in 13 of 19 ecological, 7 of 10 experimental, and 7 of 9 artificial studies. Our findings corroborate, in one case, an earlier effort to characterize nestedness by Poullain et al. (27) using a different nestedness metric. It is also apparent that some matrices are almost perfectly nested [e.g., matrices in the works of Ceyssens et al. (35), McLaughlin and King (49), and Seed and Dennis (57)]. In some cases, like the work of Middelboe et al. (50), the data came from a mix of ecological and experimental studies in that the bacteria were derived from environmental and experimentally evolved isolates, whereas the phages were wild from the same environment as the host. Does the finding of a strongly nested matrix mean, in this case, that in vitro evolution mimics selection in nature, suggesting that there exists robust principles underlying the emergence of nestedness?
Hence, given the number of studies, we ask what evidence is there that host–phage matrices are, as a whole, nested and not modular. We rank all 38 matrices from lowest to largest modularity and lowest to largest nestedness (Fig. 6 A and B). It is evident that matrices tend to be more nested than their random counterparts but not more modular (and apparently, anti-modular) than their random counterparts. How often do we expect to find 27 significantly nested matrices in a sample of 38 random matrices if each of the significantly nested matrices has a P < 0.05? Combinatorically, such a result is highly improbable and given by a binomial distribution with resulting P << 10⁻¹⁰. Likewise, the finding of an excess of antimodular matrices (20 of 38) compared with a small number of modular matrices (6 of 38) is a highly improbable result. Moreover, most of the significantly modular matrices have low values of modularity, suggesting that, although modularity may be deemed significant in a few cases, it is not a driving mechanism underlying the structure of most of these matrices and may be incidental to other patterns. Together, these results imply that currently available host–phage infection networks are typically nested and not modular.

Previously Overlooked Nested Patterns Uncovered. An additional power of subjecting host–phage infection networks to a unified analysis is that, by doing so, we can extract meaningful biological information about the organization of a system that may not have been possible given the original placement of hosts and phages in matrix format. For example, the work by Zinno et al. (64) make no mention of the fact that there are evidently groups of phages that preferentially infect groups of hosts (Fig. 34). Such block-like variability suggests that resistance mechanisms are less haphazard than they seem when network characteristics are not analyzed. Similarly, the work by Holmfeldt et al. (44) highlighted the variability and possibly unique signature of infection for each host and phage. However, reordering hosts according to the number of infecting phages while also reordering phages based on the number of hosts that they can infect leads to a nested pattern, suggesting that specific forms of infection rules may underlie infection variability (Fig. 3 B). To what extent is our finding of nestedness novel? As a reminder, nestedness is a property of a host–phage infection matrix as calculated for a given row and column ordering. Hence, we calculated nestedness for all of the matrices in the format as they were first reported in the literature and then compared these results to the nestedness calculated from our reshuffled matrices. We found that, in 35 of 37 cases of the previously published studies, the reshuffled matrix had a nestedness value higher than that of the original publication, whereas in 2 of 37 studies, the nestedness was equal (47, 50) (SI Appendix, Fig. S6). Hence, our results suggest that, by and large, prior efforts did not identify the extent to which their matrices were nested or whether such nestedness was significant.

Addressing Sample Composition Biases as Potential Drivers of Network Structure. We report a set of analyses to quantify the extent to which potential biases might impact our results. One
The data that we analyzed included phages that did not have the modular structure. We examined 95% Synecococcus that, in fact, still highly nested and marginally significant at a $P = 0.067$ level. We also recalculated modularity for the modified matrices and found that 9 of 38 are modular compared with 6 of 38 in the original analysis (SI Appendix, Fig. S8). Hence, although there are minor changes in the number of significantly nested and modular networks, our finding that matrices have a characteristic nested structure is robust to either of these sources of bias.

Finally, we ask whether there are certain characteristics of matrices that defy the general pattern of nestedness and if it is possible to learn from these outliers? Interestingly, the three matrices with the most significant modular structures (40, 55, 64) were determined for a single bacterial species, *Streptococcus thermophilus*, and its phages. This finding seems robust, because different laboratories performed the studies and the microbes were isolated from three separate continents. Additionally, we did not find an example where a matrix that included *S. thermophilus* did not have the modular structure. We examined bacteria from the same taxonomic order (*Lactobacillales*) and isolated from the environment (dairy products), but these bacteria lacked a modular structure. The consistent modularity observed for this species suggests that species-specific traits may have strong deterministic effects on the form that their interactions with parasites take. We are unsure of which traits produce the modular interactions; however, additional research may help reveal if and what resistance mechanisms determine the shape of microbial interaction networks.

**Possible Scale Dependence of Host–Phage Interactions: From Nestedness to Modularity?** The data that we analyzed included almost 12,000 separate attempts to infect a host isolate with a phage isolate. Although the scale of the current data is beyond the scope of any individual project, it still pales compared with the number of possible interactions in a community at local or regional levels. Scaling up to larger assays presents technical challenges aside from increasing the depth of sampling. Studying many host strains beyond the species (or genus) level often requires distinct culture conditions, a prerequisite for studies that many laboratories cannot or do not want to reach. Here, we present an analysis of what such a hypothesized study may reveal. Consider an experiment in which the hosts from two groups of experiments were combined in a large cross-infection assay with the phages from the same two groups of experiments. If the original matrix sizes were $H_1 \cdot P_1$ and $H_2 \cdot P_2$, then the final matrix size is $(H_1 + H_2) \cdot (P_1 + P_2)$. A total of $H_1 P_2 + H_2 P_1$ new experiments would need to be performed. If the hosts were of sufficiently distant types (e.g., *E. coli* and *Syneceococcus*), we should expect that nearly all of the new cross-infection experiments would lead to no additional infections. Hence, if the original matrices were nested, then the new matrix would have two modules, each of which was nested (Fig. 7 has the results of such a numerical experiment). In other words, we predict that, at larger, possibly macroevolutionary scales, host–phage interaction matrices should be typified by a modular structure, even if there is nested structure at smaller scales.

**Discussion**

**Summary of Major Results.** We have established a unified approach to analyzing host–phage infection matrices. In so doing, we find that a compilation of 38 empirical studies of host–phage interaction networks is nested on average and not modular (Figs. 4 and 5). In most cases, our finding of higher-order structure such as nestedness in the data led to the conclusion that the matrices were modular. However, we have argued that this finding is problematic because it is due to the inclusion of zero rows and columns. In our approach, we recalculate nestedness for the modified matrices and found that 9 of 38 are modular compared with 6 of 38 in the original analysis (SI Appendix, Fig. S8). Hence, although there are minor changes in the number of significantly nested and modular networks, our finding that matrices have a characteristic nested structure is robust to either of these sources of bias.

Potential bias in our study derives from the methods some researchers used for phage isolation. Phages require a bacterial host to reproduce, and therefore, the bacterial host(s) chosen by the researcher can affect the form of the interaction matrix. For instance, if researchers used a single host to isolate phages and included this host in the matrix, then their matrix will necessarily possess a full row of positive infections, thereby introducing the first element of a perfectly nested matrix. We found only six studies that used such an approach (46, 47, 49, 50, 56, 58). To determine if phage isolation strategy biased our results to nestedness, we reanalyzed all six of these matrices after removing the exclusion of zero rows and columns has the potential to bias the consideration of nestedness to a nested pattern. However, such zero rows and columns may be biologically meaningful if hosts or phages have evolved resistance that leads to noninteraction between particular sets of strains. Nonetheless, we performed the entire analysis again by generating alternative matrices such that hosts and phages were only included if they had had at least one nonzero element in their row or column, respectively. Then, we recalculated nestedness for the modified matrices and compared it to the nestedness of appropriately resized null matrices. We found that 26 of 38 studies were nested compared with 27 of 38 using the original analysis (SI Appendix, Fig. S7). Moreover, although the quantitative value of nestedness did decrease in one case, that particular study (39) was, in fact, still highly nested and marginally significant at a $P = 0.067$ level. We also recalculated modularity for the modified matrices and found that 9 of 38 are modular compared with 6 of 38 in the original analysis (SI Appendix, Fig. S8). Hence, although there are minor changes in the number of significantly nested and modular networks, our finding that matrices have a characteristic nested structure is robust to either of these sources of bias.

Finally, we ask whether there are certain characteristics of matrices that defy the general pattern of nestedness and if it is possible to learn from these outliers? Interestingly, the three matrices with the most significant modular structures (40, 55, 64) were determined for a single bacterial species, *Streptococcus thermophilus*, and its phages. This finding seems robust, because different laboratories performed the studies and the microbes were isolated from three separate continents. Additionally, we did not find an example where a matrix that included *S. thermophilus* did not have the modular structure. We examined bacteria from the same taxonomic order (*Lactobacillales*) and isolated from the environment (dairy products), but these bacteria lacked a modular structure. The consistent modularity observed for this species suggests that species-specific traits may have strong deterministic effects on the form that their interactions with parasites take. We are unsure of which traits produce the modular interactions; however, additional research may help reveal if and what resistance mechanisms determine the shape of microbial interaction networks.

**Possible Scale Dependence of Host–Phage Interactions: From Nestedness to Modularity?** The data that we analyzed included almost 12,000 separate attempts to infect a host isolate with a phage isolate. Although the scale of the current data is beyond the scope of any individual project, it still pales compared with the number of possible interactions in a community at local or regional levels. Scaling up to larger assays presents technical challenges aside from increasing the depth of sampling. Studying many host strains beyond the species (or genus) level often requires distinct culture conditions, a prerequisite for studies that many laboratories cannot or do not want to reach. Here, we present an analysis of what such a hypothesized study may reveal. Consider an experiment in which the hosts from two groups of experiments were combined in a large cross-infection assay with the phages from the same two groups of experiments. If the original matrix sizes were $H_1 \cdot P_1$ and $H_2 \cdot P_2$, then the final matrix size is $(H_1 + H_2) \cdot (P_1 + P_2)$. A total of $H_1 P_2 + H_2 P_1$ new experiments would need to be performed. If the hosts were of sufficiently distant types (e.g., *E. coli* and *Syneceococcus*), we should expect that nearly all of the new cross-infection experiments would lead to no additional infections. Hence, if the original matrices were nested, then the new matrix would have two modules, each of which was nested (Fig. 7 has the results of such a numerical experiment). In other words, we predict that, at larger, possibly macroevolutionary scales, host–phage interaction matrices should be typified by a modular structure, even if there is nested structure at smaller scales.

**Discussion**

**Summary of Major Results.** We have established a unified approach to analyzing host–phage infection matrices. In so doing, we find that a compilation of 38 empirical studies of host–phage interaction networks is nested on average and not modular (Figs. 4 and 5). In most cases, our finding of higher-order structure such as nestedness in the data led to the conclusion that the matrices were modular. However, we have argued that this finding is problematic because it is due to the inclusion of zero rows and columns. In our approach, we recalculate nestedness for the modified matrices and found that 9 of 38 are modular compared with 6 of 38 in the original analysis (SI Appendix, Fig. S8). Hence, although there are minor changes in the number of significantly nested and modular networks, our finding that matrices have a characteristic nested structure is robust to either of these sources of bias.
as nestedness within an individual study was not previously observed, in that prior analyses of host–phage interaction matrices usually did not attempt to estimate the network characteristics examined here. We found that host–phage interaction networks are not perfectly nested and that interactions that defy perfect nestedness are typical throughout nearly all of the data. Additionally, we found no significant difference in nestedness or modularity based on taxa, sources, or isolation method. This dataset, although far larger than any individual study, is limited to (largely) microevolutionary scales, an issue that we addressed in Results and will return to later in Discussion. Considering the large range of taxa, habitats, and sampling techniques used to construct the matrices, the repeated sampling of a nested pattern of host–phage infections is salient, although the process driving the nestedness is not obvious. It could result from multiple mechanisms or a single principle. Here, we examine three hypotheses to explain the nestedness pattern based on biochemical, ecological, and evolutionary principles. Note that these hypotheses are not mutually exclusive and that we have only limited ability to test them given our comparative approach. However, each of these hypotheses can be tested with additional laboratory-based or field experiments.

Mechanisms Responsible for Nestedness: Biophysical, Ecological, and Evolutionary. Phage and bacterial infection matrices at microevolutionary scales may be constrained to a nested shape by the nature of their molecular interactions. Phages infect bacteria by using specialized proteins that target and bind to molecules on the outer membranes of bacteria (receptor molecules). Nested infection matrices have been shown for T-phages, which infect strains of E. coli, to be the result of the interactions of the phage proteins and receptor molecules (73). T-phages bind to the lipopolysaccharide (LPS) chains on the cell surface. Mutant E. coli has been observed with shortened LPS chains that confer resistance to some but not all T-phages. There are T-phages that are able to infect these mutants, because they require fewer segments of the LPS molecule to bind. If phage–bacterial molecular interactions are dominated by single traits and variation in these traits is constrained along a single hierarchical dimension such as LPS, then one should expect the nested pattern to arise. There are other examples of traits with physical characteristics that behave similarly: bacteria that evolve a thicker and thicker protective coating (74), phages that evolve increased host range by continually reducing tail length (73), bacteria that reduce their number of receptors, and phages that target fewer receptors (75). Although there are many examples of this type of one-dimensional interaction, the problem with this finding being a universal explanation for the form of bacterial–phage interactions are that host–phage interactions are governed by hundreds of other genes (76), bacteria can use multiple strategies for resistance (74), and phages have complex mechanism to evade bacteria defenses (74, 77). Moreover, a recent discovery of an adaptive immune system, where bacteria acquire targeted sequences to prevent phage infection and phages evolve to evade such immunity, suggests a complex interaction space (78). Given the diversity of host–phage interactions, it seems unlikely that the molecular details alone would constrain the form of their relationship (79). Instead, we turn to the potential guiding forces of community assembly and coevolution to explain this reoccurring pattern.

The nested pattern may be common, because the processes of microbial community assembly select for species with nested relationships. One could imagine that communities may settle into this pattern if this interaction structure is more stable than others (67, 71), noting that the stability of host–phage interaction structures may depend on ecological factors such as resource availability (80). Cohesive interaction structures such as nested patterns have been shown to be more stable than other structures for mutualistic networks (81, 82). The regularity of the interactions and redundancies make these communities less susceptible to the random removal of nodes. However, these networks

Fig. 7. Union of two nested matrices indicates possible host–phage interaction structure at larger, possibly macroevolutionary scales. In this figure, we selected two of the most nested studies and performed a union while assuming that there were no cross-infections of hosts by phages of the other study. In this case, E. coli and cyanobacteria were the host types. (A) Depiction of the original matrices. (B) Randomization of the union matrix. (C) Nested sort of the union matrix. (D) Modularity sort of the union matrix with a nested sort of each module.
are thought to be susceptible to invasion by new species that violate the nested pattern, suggesting that migration of a species would perturb the nestedness. Furthermore, the spatiotemporal complexity of microbial and viral communities suggests that prior theoretical efforts that consider community addition as a process in which invasions occur infrequently may not be widely applicable. Moreover, community assembly models rarely invoke the influence of evolutionary change at similar time scales as ecological change—an issue highly relevant to the study of microbial and viral communities.

Indeed, there may be an evolutionary explanation for nestedness. Most attempts to characterize the form of coevolution with host–phage experiments to date have shown a form of antagonistic evolution called expanded host range (or gene for gene) coevolution (52, 83, 84). Under this model, bacteria evolve ever-increasing resistance to more and more phage genotypes, and phages evolve broader host ranges. If one were to sample a community of bacteria and phages coevolving under this model, they would uncover a diversity of phages and bacteria that exhibit a nested interaction pattern. At any time point, the most-derived bacteria should exist, which is either completely resistant or depending on the timing, sensitive to the most-derived phage. Given that selection by phage may be slow to alleviate the more sensitive ancestral variants or that there may be a tradeoff between resistance and competitiveness, there will exist a diversity of bacteria with ever decreasing sets of phages to which they are resistant. Similarly, the most-derived phages will have the broadest host range, and by the same logic as for the bacteria, its ancestors are likely to persist in the community and display ever-decreasing host ranges. The nested pattern could be a product of taking a snapshot of a dynamically evolving community. Although the majority of experimental results observed in artificial laboratory settings support this hypothesis, there is a single laboratory experiment (85) and models of bacterial host–parasite coevolution that suggest that other forms of coevolution are possible when there are bottom-up costs for modifications to resistance (86, 87). Furthermore, if coevolution provided the only explanation, then the artificially assembled matrices would not have the nested pattern.

**Dispelling and Recognizing Potential Biases.** Three sources of sampling bias challenge the generality of our findings. First, the taxa sampled may poorly represent microbial diversity given that they are subject to both human and methodological biases. If, for instance, only taxa associated with humans were selected or all taxa were cultured similarly, then our results would only be relevant for a small group of microbes. Indeed, the majority of microbial studies were performed on the family Enterobacteriaceae, which lives within human digestive systems; however, the spectrum of bacteria that we examined is much broader and includes both heterotrophic and photosynthetic species. Further, gram-negative and -positive bacteria examined here were isolated from six continents and many disparate environments from the extreme conditions of hot springs, the rich resource conditions of sewage, depauperate marine environments, and the complex matrix of soil to the simplified laboratory environment. Although this study cannot feasibly test the full microbial diversity of the globe, it does include examples from much of it (SI Appendix, Tables S1 and S2).

Second, as previously discussed, the number of hosts used to isolate phages and the inclusion of noninteracting hosts and phages have the potential to alter the nestedness of a matrix. Ideally, the same number of hosts studied in the matrix would be used to isolate phages, or if only a subset of hosts was used, then these hosts would not be included in the matrix. This finding is important to ensure that the pattern of infection is independent of how the parasites were isolated. We found that these biases were not a problem by (i) testing matrices that were created by isolating phages on a single host and (ii) removing hosts and phages that were not interacting. We found that whether the matrices were significantly nested was not affected by including the isolation host in the matrix or by removing noninteracting hosts and phages, which is strong support that the isolation matrix did not enrich for nestedness.

The last category of bias, phylogenetic, is likely to mean that our results define a pattern at relatively narrow taxonomic scales. The majority of our studies was of closely related genotypes and species. As described in Results, we anticipate that more complex patterns of infection may form at larger phylogenetic scales that likely include increasing compartmentalization. Hence, we hypothesize that a multiscale view of host–phage infection networks will reveal nestedness at small scales and modularity at large scales. Our finding of nested interaction matrices is still relevant for characterizing patterns at short phylogenetic distances; they are, arguably, the most relevant for many ecological and evolutionary scenarios, because they likely share the richest connections.

**Prospective View.** Whatever the limitations of this dataset, it is important to point out that viewing host–phage interaction networks through a unifying lens will likely unveil other commonalities of microbial and viral communities. By way of analogy, over 25 y ago, the study of food webs was radically altered by the compilation of many small food webs that were subject to a unified analysis (88–91). The key finding of the earliest food web studies was that the members of a community could be ranked, and that larger species would eat a random fraction of those species smaller than them. From this stage, there were two ways forward. First, by studying larger food webs, the original pattern was refined such that species ranking was found to be correlated with body size (but not equivalent to body size); therefore, individuals eat prey that are smaller, although they are a part of a well-defined size class (92, 93). Second, the topology of food webs was then used as a target and basis for dynamic models of community behavior (i.e., what mechanisms can explain the patterns and how do the patterns influence community function) (94). We hope and envision that a similar process unfolds here in that the finding of a general pattern in the current dataset will stimulate the collection of more and larger host–phage infection networks to continue to provide a fuller picture of who infects whom across an entire community. In so doing, we caution that data completeness can alter the observed patterns of connectivity and refer readers to a number of recent papers that address this topic (95–99).

What do we expect to find when analyzing ever larger host–phage interaction networks collected from an ecological community, evolution experiments, or culture collections? We hypothesize that host–phage interaction matrices are likely characterized by modularity at larger taxonomic scales even if there is structure (e.g., nestedness) at small taxonomic scales (Fig. 7). What would such a multiscale phenomenon inform us about the structure and function of microbiological communities? First, it would suggest the existence of diversifying coevolutionary-induced selection that gave rise to (largely) independent host–phage communities. The molecular basis of such diversification could then be explored. Second, cross-infection assays or similar laboratory-based strategies (100) that test whether phages can infect or at least transmit their genes between phylogenetically divergent hosts have the potential to provide significant advances in understanding patterns of global gene transfer. Such phages (and the bacteria that they infect) may be critical to understanding the direct transfer of genes on a global scale. Instead of phages acting locally (in a taxonomic sense) to shuttle genes between closely related bacteria, a few rare links would permit greater cross-talk between bacterial taxa. Quantifying the frequency of such events may represent the small-
world links that connect distant microbial populations (101), and it is in need of experimental testing.

Furthermore, infections of distantly related groups by the same phages would imply that the bacteria are in indirect competition with one another, even if they do not seem to compete directly for the same set of carbon and nutrient sources. Although whole-genome-based approaches to infer host range and phylogenetic similarity may help provide candidates for such rare links, they are not the only solution. Rather, we suggest that the continued use of laboratory-based assays to catalog the life history traits of cultivable host–phage pairs is essential if we are to improve our understanding of the population dynamics of host–phage communities in the wild. Of course, many (if not most) bacteria and phages are not currently culturable. Hence, in parallel, we recommend attention be given to the development of inverse methods to catalog the life history traits of phages based on community infection assays in those circumstances in which culturing is impossible or yet intractable.

Materials and Methods

Network Statistics. Modularity is estimated by reshuffling the rows and columns of the matrix to find groupings of highly interconnected phages and bacteria, labeling these groups and assessing matrix-wide the ratio of the number of within to outside group connections. This calculation is done using a heuristic called the BRIM algorithm (102) to efficiently find the configuration that maximizes this ratio. We ported the BRIM algorithm to MATLAB from the original code in Octave and used the adaptive BRIM algorithm for all calculations here. By this definition, a perfectly modular matrix is comprised of clusters of completely isolated groups, and modularity decreases as the number of cross-group connections increases. Nestestability is estimated by reordering the rows and columns (103, 104) to determine whether phages that infect fewer hosts are only able to infect a subset of bacteria that are susceptible to many phages. This reordering tries to maximize the position of ones in the matrix such that they clusters above a nullcline (Fig. 1C shows a perfectly nested matrix). The value for nestestability depends on how frequently ones fall above rather than below this nullcline. Complete details are provided in SI Appendix, SI Materials and Methods.

Host–Phage Infection Assay. Matrix 22 is the only dataset not previously published. We constructed the matrix by coevolving an obligately lytic phage–strain with its host E. coli. The E. coli studied was of strain REL606, a derivative of E. coli B acquired from Richard Lenski (Michigan State University, Lansing, MI) and described in ref. 105, and phages were of strain c21 (Ivir) provided by Donald Court (National Cancer Institute, Frederick, MD). The phage lambda bacteria were maintained in 50 mL Erlenmeyer flasks with 10 mL liquid medium, shaken at 120 rpm, and incubated at 37 °C (New Brunswick Innova 4300 Incubator Shaker). This flask was incubated, and the cycle of transfer and incubation was continued one more time. Three 24-h incubations were long enough for the bacteria to evolve resistance and the phages to counter it; however, it was not long enough for a second round of coevolution. We randomly selected 150 bacteria and 150 phage isolates. We determined which of the 150 bacteria isolates were resistant to the 150 phage isolates. To do this task, we performed spot plate assays. All bacterial-phage combinations were replicated five separate times, and a total of 28,125 spots were assayed. To make this process more efficient, we plated up to 96 separate phage stocks onto a single dish (150 mm radius). Phage stock replicates were never placed on the same plate 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 three of five replicates were given an infection titer. Bacteria that had identical infection or resistance profiles as their ancestors were removed from the matrix. Complete details are provided in SI Appendix, SI Materials and Methods.

ACKNOWLEDGMENTS. The authors thank the anonymous reviewers for comments and suggestions that improved the manuscript. J.S.W. acknowledges the support of the James S. McDonnell Foundation and Defense Advanced Projects Research Agency Grant HR0011-09-1-0055. J.S.W. holds a Career Award at the Scientific Interface from the Burroughs Welcome Fund.


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 $k_i = \sum A_{ij}$ (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 them 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
the density of links (2). A more reliable approach uses a null model to assess the quality of a
given network partition. Newman and Girvan(3) defines the modularity as follows:

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

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

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

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

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

Fortunato and Barthélemy have pointed out that, in large networks, modularity
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:

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

where $D_{ij}$ is the diagonal that cross the unexpected cell and $k = 100/U_{\text{max}}$ with $U_{\text{max}} \approx 0.04145$ 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
network. These studies were then divided into one of four sub-classes: aquatic, soil, microbiome, and food items. These sub-classes were based upon the environment from which the hosts and phages were isolated.

Co-evolution (EXP) – 10 studies:
This criterion was applied to studies in which phages and/or hosts were allowed to evolve in the lab. After phages were allowed to evolve, their host ranges were then tested. Sub-classes were based upon methodology of the study, and studies were classified as either serial dilution or chemostat experiments. Importantly, matrices of the EXP class need not be reflective of a given community at a fixed moment in time.

Artificial (ART) – 9 studies:
This criterion was applied to studies in which almost all hosts and phages were either generated within the lab or came from a collection. Sub-classes indicated the origination of the host strains. Host strains were either environmental or pathogenic.

Principal component analysis
The objective of PCA is to find a new coordinate system such that the maximal variance is explained in order of each coordinate (i.e., the principal components). Each variable was normalized to have zero mean and a standard deviation of 1 so that each contributed equally to the PCA. Supplementary Figure 1 shows the projection of each study onto the first two principal axes and Supplementary Table 4 shows the detailed coordinates underlying the principal components. Roughly, principal component 1 (PC1) corresponds to the size of the matrix, and so those studies to the right-side of Supplementary Figure 3 tend to be large matrices and those to the left tend to be small matrices. Roughly, PC2 corresponds to the asymmetry between number of phages and number of hosts, so that the top-most studies of Supplementary Figure 3 have more hosts than phages, whereas the bottom-most studies have more phages than hosts. Finally, the third principal component (not shown) corresponds, roughly, to the connectance of the study.

Statistical analysis of clustering validity using a re-shuffling approach
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 \((x_1, x_1, \ldots, x_n)\), 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:

\[
\arg\min_S = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2
\]

where \(\mu_i\) 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.
Statistical analysis of correlations among global properties using a Bonferroni correction

We study the correlations coefficients among the global properties. These values are show in Supplementary Table 5. In that table is showed also the statistical significance of those values. For evaluate the statistical significance we used a Bonferroni correction, using both, the number of combinations and the number of global properties. This correction is used in statistics when one needs to address multiple comparisons. And comes by the fact that even when there is not statistical significance, we can find just by probability that some of the comparisons are statistically significant. Therefore this correction aims to avoid this problem. We can see in the indicated table that among the statistically significant values there is only a strong correlation between the number of hosts and the number of species. Another interesting result is that there is almost no correlation (no statistical significance) between the connectance and the number of species. This is contrary to the plant-pollinator networks where the relation follows a power law.

Experimental assays of host-phage infection

Conditions and microbial cultures

The phage and bacteria were cocultured in 50ml Erlenmeyer flasks, with 10ml of liquid medium, shaken at 120 rpm, and incubated at 37 °C. The medium was an altered version of Davis Medium (15), in which we added 10 times the magnesium sulfate (1g/L) to improve phage viability and 125 mg/L of maltotriose instead of glucose because E. coli and phage λ are predicted to undergo a coevolutionary arms-race when provided with maltodextrins as its only source of carbon (16-18). The medium was filtered and the magnesium was added just before use in order to stop crystallization of the magnesium during the experiment. 75 separate flasks were initiated with very small populations of bacteria (~1,000 E. coli cells) and phage (~100 phage λ particles) to assure that the initial populations were isogenic and that all mutant bacteria and phage arose de novo, this is important to make sure that each community has the potential to follow its own coevolutionary path. The E. coli studied were of strain REL606, a derivative of E. coli B acquired from Richard Lenski (Michigan State University), described in (19) and phage were of strain cI21 (λvir) provided by Donald Court (National Cancer Institute). Most phage λ strains have two life cycles, lytic and lysogenic, the second includes a latent phase where the phage genome is incorporated into the bacterial chromosome at which time the bacteria acquires 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
infests 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
pleiotropic 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

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 Nestledness 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)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Source Type</th>
<th>H</th>
<th>P</th>
<th>S</th>
<th>I</th>
<th>M</th>
<th>C</th>
<th>Lp</th>
<th>Ls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) Abe (2007)</td>
<td>ecological</td>
<td>11</td>
<td>4</td>
<td>15</td>
<td>22</td>
<td>44</td>
<td>0.50</td>
<td>5.50</td>
</tr>
<tr>
<td>2</td>
<td>(2) Barrangou (2002)</td>
<td>ecological</td>
<td>14</td>
<td>6</td>
<td>20</td>
<td>25</td>
<td>84</td>
<td>0.30</td>
<td>4.17</td>
</tr>
<tr>
<td>3</td>
<td>(3) Braun-Brenton (1981)</td>
<td>experimental</td>
<td>18</td>
<td>3</td>
<td>21</td>
<td>30</td>
<td>54</td>
<td>0.56</td>
<td>10.00</td>
</tr>
<tr>
<td>4</td>
<td>(4) Campbell (1995)</td>
<td>experimental</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>14</td>
<td>45</td>
<td>0.31</td>
<td>2.80</td>
</tr>
<tr>
<td>5</td>
<td>(5) Capparelli (2010)</td>
<td>ecological</td>
<td>18</td>
<td>8</td>
<td>26</td>
<td>54</td>
<td>144</td>
<td>0.38</td>
<td>6.75</td>
</tr>
<tr>
<td>6</td>
<td>(6) Caso (1995)</td>
<td>experimental</td>
<td>23</td>
<td>4</td>
<td>27</td>
<td>17</td>
<td>92</td>
<td>0.18</td>
<td>4.25</td>
</tr>
<tr>
<td>7</td>
<td>(7) Ceyssens (2009)</td>
<td>artificial</td>
<td>5</td>
<td>15</td>
<td>20</td>
<td>29</td>
<td>75</td>
<td>0.39</td>
<td>1.93</td>
</tr>
<tr>
<td>8</td>
<td>(8) Comeau (2005)</td>
<td>experimental</td>
<td>30</td>
<td>13</td>
<td>43</td>
<td>152</td>
<td>390</td>
<td>0.39</td>
<td>11.69</td>
</tr>
<tr>
<td>9</td>
<td>(9) Comeau (2006)</td>
<td>experimental</td>
<td>32</td>
<td>16</td>
<td>48</td>
<td>118</td>
<td>512</td>
<td>0.23</td>
<td>7.38</td>
</tr>
<tr>
<td>10</td>
<td>(10) DePaola (1998)</td>
<td>ecological</td>
<td>5</td>
<td>17</td>
<td>22</td>
<td>39</td>
<td>85</td>
<td>0.46</td>
<td>2.29</td>
</tr>
<tr>
<td>11</td>
<td>(11) Doi (2003)</td>
<td>artificial</td>
<td>15</td>
<td>10</td>
<td>25</td>
<td>41</td>
<td>150</td>
<td>0.27</td>
<td>4.10</td>
</tr>
<tr>
<td>12</td>
<td>(12) Duplessis (2001)</td>
<td>artificial</td>
<td>12</td>
<td>12</td>
<td>24</td>
<td>37</td>
<td>144</td>
<td>0.26</td>
<td>3.08</td>
</tr>
<tr>
<td>13</td>
<td>(13) Gamage (2004)</td>
<td>ecological</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>9</td>
<td>42</td>
<td>0.21</td>
<td>1.29</td>
</tr>
<tr>
<td>14</td>
<td>(14) Goodridge (2003)</td>
<td>ecological</td>
<td>93</td>
<td>2</td>
<td>95</td>
<td>60</td>
<td>186</td>
<td>0.32</td>
<td>30.00</td>
</tr>
<tr>
<td>15</td>
<td>(15) Hansen (2007)</td>
<td>ecological</td>
<td>34</td>
<td>12</td>
<td>46</td>
<td>146</td>
<td>408</td>
<td>0.36</td>
<td>12.17</td>
</tr>
<tr>
<td>16</td>
<td>(16) Holmfield (2007)</td>
<td>artificial</td>
<td>23</td>
<td>46</td>
<td>69</td>
<td>418</td>
<td>1058</td>
<td>0.40</td>
<td>9.09</td>
</tr>
<tr>
<td>17</td>
<td>(17) Kankila (1994)</td>
<td>ecological</td>
<td>32</td>
<td>12</td>
<td>44</td>
<td>346</td>
<td>384</td>
<td>0.90</td>
<td>28.83</td>
</tr>
<tr>
<td>18</td>
<td>(18) Krylov (2006)</td>
<td>ecological</td>
<td>11</td>
<td>10</td>
<td>21</td>
<td>73</td>
<td>110</td>
<td>0.66</td>
<td>7.30</td>
</tr>
<tr>
<td>19</td>
<td>(19) Kudva (1999)</td>
<td>artificial</td>
<td>22</td>
<td>3</td>
<td>25</td>
<td>33</td>
<td>66</td>
<td>0.50</td>
<td>11.00</td>
</tr>
<tr>
<td>20</td>
<td>(20) Langley (2003)</td>
<td>ecological</td>
<td>66</td>
<td>9</td>
<td>75</td>
<td>99</td>
<td>594</td>
<td>0.17</td>
<td>11.00</td>
</tr>
<tr>
<td>21</td>
<td>(21) McLaughlin (2008)</td>
<td>ecological</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td>18</td>
<td>56</td>
<td>0.32</td>
<td>2.57</td>
</tr>
<tr>
<td>22</td>
<td>Meyer (unpub)</td>
<td>experimental</td>
<td>25</td>
<td>27</td>
<td>52</td>
<td>314</td>
<td>675</td>
<td>0.47</td>
<td>11.63</td>
</tr>
<tr>
<td>23</td>
<td>(22) Middelboe (2009)</td>
<td>experimental</td>
<td>11</td>
<td>24</td>
<td>35</td>
<td>202</td>
<td>264</td>
<td>0.77</td>
<td>8.42</td>
</tr>
<tr>
<td>24</td>
<td>(23) Mikic (2003)</td>
<td>ecological</td>
<td>24</td>
<td>14</td>
<td>38</td>
<td>70</td>
<td>336</td>
<td>0.21</td>
<td>5.00</td>
</tr>
<tr>
<td>25</td>
<td>(24) Mizoguchi (2003)</td>
<td>experimental</td>
<td>8</td>
<td>4</td>
<td>12</td>
<td>21</td>
<td>32</td>
<td>0.66</td>
<td>5.25</td>
</tr>
<tr>
<td>26</td>
<td>(25) Pantuczek (1998)</td>
<td>artificial</td>
<td>102</td>
<td>4</td>
<td>106</td>
<td>322</td>
<td>408</td>
<td>0.79</td>
<td>80.50</td>
</tr>
<tr>
<td>27</td>
<td>(26) Paterson (2010)</td>
<td>experimental</td>
<td>100</td>
<td>5</td>
<td>105</td>
<td>267</td>
<td>500</td>
<td>0.53</td>
<td>53.40</td>
</tr>
<tr>
<td>29</td>
<td>(28) Quiberoni (2002)</td>
<td>ecological</td>
<td>20</td>
<td>11</td>
<td>31</td>
<td>89</td>
<td>220</td>
<td>0.40</td>
<td>8.09</td>
</tr>
<tr>
<td>30</td>
<td>(29) Rybnik (2006)</td>
<td>artificial</td>
<td>17</td>
<td>14</td>
<td>31</td>
<td>70</td>
<td>238</td>
<td>0.29</td>
<td>5.00</td>
</tr>
<tr>
<td>31</td>
<td>(30) Seed (2005)</td>
<td>artificial</td>
<td>24</td>
<td>6</td>
<td>30</td>
<td>31</td>
<td>144</td>
<td>0.22</td>
<td>5.17</td>
</tr>
<tr>
<td>32</td>
<td>(31) Stenholm (2008)</td>
<td>ecological</td>
<td>28</td>
<td>22</td>
<td>50</td>
<td>348</td>
<td>616</td>
<td>0.56</td>
<td>15.82</td>
</tr>
<tr>
<td>33</td>
<td>(32) Sullivan (2003)</td>
<td>ecological</td>
<td>21</td>
<td>44</td>
<td>65</td>
<td>148</td>
<td>924</td>
<td>0.16</td>
<td>3.36</td>
</tr>
<tr>
<td>34</td>
<td>(33) Suttle (1993)</td>
<td>artificial</td>
<td>7</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>63</td>
<td>0.17</td>
<td>1.22</td>
</tr>
<tr>
<td>35</td>
<td>(34) Synnott (2009)</td>
<td>ecological</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>207</td>
<td>256</td>
<td>0.81</td>
<td>12.94</td>
</tr>
<tr>
<td>36</td>
<td>(35) Wang (2008)</td>
<td>ecological</td>
<td>18</td>
<td>7</td>
<td>25</td>
<td>11</td>
<td>126</td>
<td>0.09</td>
<td>1.57</td>
</tr>
<tr>
<td>37</td>
<td>(36) Wichels (1998)</td>
<td>ecological</td>
<td>59</td>
<td>23</td>
<td>82</td>
<td>318</td>
<td>1357</td>
<td>0.23</td>
<td>13.83</td>
</tr>
<tr>
<td>38</td>
<td>(37) Zinno (2010)</td>
<td>ecological</td>
<td>18</td>
<td>27</td>
<td>45</td>
<td>49</td>
<td>486</td>
<td>0.10</td>
<td>1.81</td>
</tr>
</tbody>
</table>

| 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 correspond 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).

<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Bacteria</th>
<th>Phage</th>
<th>Majority source</th>
<th>Additional source</th>
<th>Isolation Habitat</th>
<th>Bacterial association</th>
<th>Bacterial trophy</th>
<th>Geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) Abe (2007)</td>
<td>Escherichia coli</td>
<td>T2 and PP01</td>
<td>ecological</td>
<td>artificial</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td></td>
<td>North Carolina, USA</td>
</tr>
<tr>
<td>2</td>
<td>(2) Barrangou (2002)</td>
<td>Leuconostoc</td>
<td>Caudovirales</td>
<td>ecological</td>
<td>artificial</td>
<td>sauerkraut</td>
<td>free</td>
<td></td>
<td>heterotrophic</td>
</tr>
<tr>
<td>3</td>
<td>(3) Braun-Brenton (1981)</td>
<td>Escherichia coli</td>
<td>λ</td>
<td>experimental</td>
<td></td>
<td>lab-agar plates</td>
<td>human symbiont</td>
<td>heterotrophic</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(6) Caso (1995)</td>
<td>Lactobacillus</td>
<td>Siphoviridae</td>
<td>experimental</td>
<td></td>
<td>food, fresh water, soil, sewage</td>
<td>free</td>
<td>heterotrophic</td>
<td>Spain</td>
</tr>
<tr>
<td>7</td>
<td>(7) Ceyssens (2009)</td>
<td>Pseudomonas aeruginosa</td>
<td></td>
<td>artificial</td>
<td></td>
<td>hospital sewage, fresh water</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>global</td>
</tr>
<tr>
<td>8</td>
<td>(8) Comeau (2005)</td>
<td>Vibrio</td>
<td></td>
<td>experimental</td>
<td></td>
<td>marine sewage, fresh water</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>British Columbia, Canada</td>
</tr>
<tr>
<td>10</td>
<td>(10) DePaola (1998)</td>
<td>Vibrio vulnificus</td>
<td>Podoviridae, Siphoviridae, and Myoviridae</td>
<td>ecological</td>
<td></td>
<td>marine</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>Gulf of Mexico</td>
</tr>
<tr>
<td>11</td>
<td>(11) Doi (2003)</td>
<td>Lactobacillus</td>
<td>Siphoviridae and Myoviridae</td>
<td>artificial</td>
<td></td>
<td>silage (fermented bovine feed)</td>
<td>free</td>
<td>heterotrophic</td>
<td>Japan</td>
</tr>
<tr>
<td>12</td>
<td>(12) Duplessis (2001)</td>
<td>Streptococcus thermophillus</td>
<td>Myoviridae and Siphoviridae</td>
<td>artificial</td>
<td></td>
<td>Industrial cheese plants</td>
<td>free</td>
<td>heterotrophic</td>
<td>Quebec, Canada</td>
</tr>
<tr>
<td>13</td>
<td>(13) Gamage (2004)</td>
<td>Escherichia coli</td>
<td></td>
<td>ecological</td>
<td></td>
<td>human and animal fecal isolates</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>Ohio, USA</td>
</tr>
<tr>
<td>17</td>
<td>(17) Kankila (1994)</td>
<td>Rhizobium</td>
<td></td>
<td>ecological</td>
<td></td>
<td>soil</td>
<td>free</td>
<td>heterotrophic</td>
<td>Finland</td>
</tr>
<tr>
<td>19</td>
<td>(19) Kudva (1999)</td>
<td>Enterobacteriaceae</td>
<td></td>
<td>artificial</td>
<td></td>
<td>bovine and ovine feces</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>North West USA</td>
</tr>
<tr>
<td>20</td>
<td>(20) Langley (2003)</td>
<td>Burkholderia</td>
<td>T-even and λ - like</td>
<td>ecological</td>
<td>artificial</td>
<td>soil, freshwater, plant</td>
<td>mycorrhizal</td>
<td>heterotrophic</td>
<td>global</td>
</tr>
<tr>
<td>ID</td>
<td>Reference</td>
<td>Species</td>
<td>Type</td>
<td>Environment</td>
<td>Pathogen Type</td>
<td>Growth Conditions</td>
<td>Location</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>--------------------</td>
<td>--------------------------------</td>
<td>-------------------</td>
<td>------------------------------</td>
<td>--------------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>(21) McLaughlin</td>
<td><em>Salmonella</em></td>
<td>ecological</td>
<td>swine lagoon</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>Mississippi, USA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Meyer (unpub)</td>
<td><em>Escherichia</em> α</td>
<td>experimental</td>
<td>lab - batch culture</td>
<td>human symbiont</td>
<td>heterotrophic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>(22) Middelboe</td>
<td><em>Cellulophaga baltica</em></td>
<td>experimental</td>
<td>ecological</td>
<td>marine</td>
<td>free</td>
<td>free photosynthetic Scandinavia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>(23) Miklic</td>
<td><em>Lactococcus lactis</em></td>
<td>ecological</td>
<td>dairy products</td>
<td>free</td>
<td>heterotrophic</td>
<td>Solvania</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>(24) Mizoguchi</td>
<td><em>Escherichia coli</em> PP01</td>
<td>experimental</td>
<td>lab-chemostat</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>(25) Pantucek</td>
<td><em>Staphylococcus polyvalent</em></td>
<td>artificial</td>
<td>clinical isolates</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>Brno, Czech Republic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>(26) Paterson</td>
<td><em>Pseudomonas fluorescens</em> φ2</td>
<td>experimental</td>
<td>lab - batch culture</td>
<td>plant symbiont</td>
<td>heterotrophic</td>
<td>UK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>(27) Poullain</td>
<td><em>Pseudomonas fluorescens</em> φ2</td>
<td>experimental</td>
<td>lab - batch culture</td>
<td>plant symbiont</td>
<td>heterotrophic</td>
<td>UK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>(28) Quiberoni</td>
<td><em>Streptococcus thermophilus</em></td>
<td>ecological</td>
<td>yogurt industrial plant</td>
<td>free</td>
<td>heterotrophic</td>
<td>Argentina</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>(29) Rybniker</td>
<td><em>Mycobacterium</em></td>
<td>artificial</td>
<td>soil</td>
<td>human pathogen</td>
<td>heterotrophic</td>
<td>global</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>(31) Stenholm</td>
<td><em>Flavobacterium psychrophilum</em></td>
<td>ecological</td>
<td>fresh water</td>
<td>fish pathogen</td>
<td>heterotrophic</td>
<td>Denmark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>(32) Sullivan</td>
<td><em>Prochlorococcus</em></td>
<td>ecological</td>
<td>marine</td>
<td>free</td>
<td>photosynthetic</td>
<td>Atlantic Ocean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>(33) Suttle (1993)</td>
<td><em>Synechococcus and Anacystis</em></td>
<td>ecological</td>
<td>marine</td>
<td>free</td>
<td>photosynthetic</td>
<td>Texas, USA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>(34) Synnott</td>
<td><em>Staphylococcus aureus</em></td>
<td>ecological</td>
<td>sewage, dairy products</td>
<td>bovine pathogen</td>
<td>heterotrophic</td>
<td>Tokyo, Japan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>(35) Wang (2008)</td>
<td><em>Synechococcus and</em></td>
<td>ecological</td>
<td>marine</td>
<td>free</td>
<td>photosynthetic</td>
<td>Chesapeake Bay, USA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>(36) Wichels</td>
<td><em>Pseudofalteromonas</em></td>
<td>ecological</td>
<td>marine</td>
<td>free</td>
<td>heterotrophic</td>
<td>North Sea, Germany</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>(37) Zinno (2010)</td>
<td><em>Streptococcus thermophilus</em></td>
<td>ecological</td>
<td>dairy products</td>
<td>free</td>
<td>heterotrophic</td>
<td>Italy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First column: These ID’s corresponds to indexes in supplementary figures 1-3.
### Supplementary Table 3 Global properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>number of hosts</td>
</tr>
<tr>
<td>$P$</td>
<td>number of phages</td>
</tr>
<tr>
<td>$I$</td>
<td>number of interactions</td>
</tr>
<tr>
<td>$S = H + P$</td>
<td>number of species</td>
</tr>
<tr>
<td>$M = HP$</td>
<td>size</td>
</tr>
<tr>
<td>$C = I/M$</td>
<td>connectance</td>
</tr>
<tr>
<td>$L_H = I/H$</td>
<td>mean number of interactions across host species</td>
</tr>
<tr>
<td>$L_P = I/P$</td>
<td>mean number of interactions across phage species</td>
</tr>
</tbody>
</table>

### Supplementary Table 4 PCA Analysis

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>0.352</td>
<td>0.446</td>
<td>-0.179</td>
<td>0.131</td>
<td>0.389</td>
<td>-0.131</td>
<td>-0.097</td>
<td>0.670</td>
</tr>
<tr>
<td>$P$</td>
<td>0.247</td>
<td>-0.534</td>
<td>-0.203</td>
<td>0.474</td>
<td>-0.461</td>
<td>-0.140</td>
<td>-0.279</td>
<td>0.279</td>
</tr>
<tr>
<td>$I$</td>
<td>0.470</td>
<td>-0.138</td>
<td>0.143</td>
<td>-0.474</td>
<td>0.008</td>
<td>0.517</td>
<td>-0.498</td>
<td>0.000</td>
</tr>
<tr>
<td>$S = H + P$</td>
<td>0.444</td>
<td>0.218</td>
<td>-0.257</td>
<td>0.320</td>
<td>0.192</td>
<td>-0.184</td>
<td>-0.208</td>
<td>-0.688</td>
</tr>
<tr>
<td>$M = HP$</td>
<td>0.397</td>
<td>-0.239</td>
<td>-0.359</td>
<td>-0.524</td>
<td>-0.078</td>
<td>-0.373</td>
<td>0.466</td>
<td>0.000</td>
</tr>
<tr>
<td>$C = I/M$</td>
<td>0.188</td>
<td>0.062</td>
<td>0.743</td>
<td>-0.093</td>
<td>-0.112</td>
<td>-0.601</td>
<td>-0.164</td>
<td>0.000</td>
</tr>
<tr>
<td>$L_H = I/H$</td>
<td>0.281</td>
<td>-0.449</td>
<td>0.359</td>
<td>0.313</td>
<td>0.504</td>
<td>0.224</td>
<td>0.435</td>
<td>0.000</td>
</tr>
<tr>
<td>$L_P = I/P$</td>
<td>0.353</td>
<td>0.431</td>
<td>0.177</td>
<td>0.177</td>
<td>-0.571</td>
<td>0.335</td>
<td>0.434</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### Supplementary Table 5 Correlation analysis

<table>
<thead>
<tr>
<th></th>
<th>$H$</th>
<th>$P$</th>
<th>$S$</th>
<th>$I$</th>
<th>$M$</th>
<th>$C$</th>
<th>$L_P$</th>
<th>$L_H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>1.000</td>
<td>-0.146</td>
<td>0.458</td>
<td>0.394</td>
<td>0.125</td>
<td>0.164</td>
<td>0.000</td>
<td>0.133</td>
</tr>
<tr>
<td>$P$</td>
<td>-0.146</td>
<td>1.000</td>
<td>0.264</td>
<td>0.237</td>
<td>0.244</td>
<td>-0.110</td>
<td>-0.191</td>
<td>0.000</td>
</tr>
<tr>
<td>$S$</td>
<td>0.458</td>
<td>0.264</td>
<td>1.000</td>
<td>0.664</td>
<td>0.686</td>
<td>0.077</td>
<td>0.154</td>
<td>0.000</td>
</tr>
<tr>
<td>$I$</td>
<td>0.394</td>
<td>0.458</td>
<td>0.664</td>
<td>1.000</td>
<td>0.752</td>
<td>0.466</td>
<td>0.553</td>
<td>0.716</td>
</tr>
<tr>
<td>$M$</td>
<td>0.125</td>
<td>-0.110</td>
<td>0.077</td>
<td>0.466</td>
<td>1.000</td>
<td>-0.109</td>
<td>0.204</td>
<td>0.449</td>
</tr>
<tr>
<td>$C$</td>
<td>0.125</td>
<td>-0.110</td>
<td>0.077</td>
<td>0.466</td>
<td>1.000</td>
<td>-0.109</td>
<td>0.204</td>
<td>0.449</td>
</tr>
<tr>
<td>$L_P$</td>
<td>-0.133</td>
<td>-0.191</td>
<td>0.154</td>
<td>0.553</td>
<td>0.304</td>
<td>0.501</td>
<td>0.100</td>
<td>0.035</td>
</tr>
<tr>
<td>$L_H$</td>
<td>-0.133</td>
<td>-0.191</td>
<td>0.154</td>
<td>0.316</td>
<td>0.449</td>
<td>0.517</td>
<td>0.035</td>
<td>1.000</td>
</tr>
</tbody>
</table>

In Green $p$-value $< 0.05/28$  
In Yellow $0.05/28 < p$-value $< 0.05/8$
### Supplementary Table 6 Isolation bias

<table>
<thead>
<tr>
<th>Study</th>
<th>Modularity</th>
<th>Nestedness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Recalculated</td>
</tr>
<tr>
<td>Krylov 2006</td>
<td>0.123</td>
<td>0.136</td>
</tr>
<tr>
<td>Kudva 1999</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>McLaughlin 2008 - Matrix minus TSB control</td>
<td>0.191</td>
<td>0.191</td>
</tr>
<tr>
<td>McLaughlin 2008 - Matrix minus TSB minus isolation host</td>
<td>0.191</td>
<td>0.313</td>
</tr>
<tr>
<td>Middlehoe 2009</td>
<td>-</td>
<td>Matrix minus TSB control</td>
</tr>
<tr>
<td>Rybniker 2006</td>
<td>0.333</td>
<td>0.274</td>
</tr>
<tr>
<td>Stenholm 2009</td>
<td>0.183</td>
<td>0.187</td>
</tr>
</tbody>
</table>

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

### References:

Phages

Bacteria

1 - Abe 2007

$Q = 0.31$

$z_Q = 1.18$

$N = 0.65$

$z_N = -0.16$
Phages

Bacteria

$Q = 0.21$
$z_Q = -4.92$

$N = 0.81$
$z_N = 1.31$

2 - Barrangou 2002
3 - Braun-Breton 1981

Phages

$Q = 0.18$
$z_Q = -2.32$

Bacteria

$N = 1$
$z_N = 4.44$
\[ Q = 0.22 \]
\[ z_Q = -3.23 \]
\[ N = 0.94 \]
\[ z_N = 1.67 \]
Phages

Bacteria

\[ Q = 0.26 \]
\[ z_Q = -2.83 \]

\[ N = 0.95 \]
\[ z_N = 5.61 \]
Phages Bacteria

\[ Q = 0.27 \]
\[ z_Q = -3.65 \]

\[ N = 0.82 \]
\[ z_N = 0.32 \]

6 - Caso 1995
Phages

Bacteria

Q = 0.21

z_Q = -4.03

N = 0.98

z_N = 4.2
8 - Comeau 2005

Phages

$Q = 0.17$

$z_Q = -7.81$

$N = 0.8$

$z_N = 8$
\( Q = 0.22 \)
\( z_Q = -10.05 \)
\( N = 0.86 \)
\( z_N = 6.52 \)
Phages

Bacteria

\[ Q = 0.29 \]
\[ z_Q = 0.48 \]

\[ N = 0.65 \]
\[ z_N = 0.8 \]
Phages

Bacteria

$Q = 0.24$

$z_Q = -6.06$

$N = 0.83$

$z_N = 2.59$
Phages

Bacteria

$Q = 0.48$

$z_Q = 3.03$

$N = 0.88$

$z_N = 2.8$

12 - Duplessis 2001
$Q = 0.32$
$z_Q = -1.88$

$N = 0.92$
$z_N = 0.45$
Phages

$Q = 0.32$

$z_Q = -0.95$

Bacteria

$N = 0.77$

$z_N = 4.21$
Phages

Bacteria

$Q = 0.2$

$z_Q = -6.37$

$N = 0.72$

$z_N = 6.01$

15 - Hansen 2007
**Phages**

**Bacteria**

\[ Q = 0.19 \]

\[ z_Q = -0.83 \]

\[ N = 0.83 \]

\[ z_N = 16.92 \]
Phages

Bacteria

$Q = 0.12$

$z_Q = -4.38$

$N = 0.9$

$z_N = 3.48$
Phages \( Q = 0 \)

Bacteria

\[ z_Q = -8.17 \]

\[ z_N = 0.02 \]

\[ N = 0.63 \]
Phages Bacteria

$Q = 0.31$

$z_Q = -7.24$

$N = 0.86$

$z_N = 4.89$
Phages

Bacteria

\[ Q = 0.19 \]
\[ z_Q = -4.68 \]

\[ N = 0.98 \]
\[ z_N = 2.44 \]
Phages

Bacteria

$Q = 0.22$

$z_Q = 7.54$

$N = 0.98$

$z_N = 18.31$

22 - Meyer unpub
Phages

Bacteria

$Q = 0.08$

$z_Q = -5.22$

$N = 0.99$

$z_N = 5.99$
Bacteria

Phages

$Q = 0.41$

$z_Q = 3.88$

$N = 0.71$

$z_N = 2.35$
Phages

Bacteria

$Q = 0.15$

$z_Q = -2.25$

$N = 0.85$

$z_N = 0.65$
Phages

Bacteria

$Q = 0.05$
$z_Q = -7.98$

$N = 0.88$
$z_N = 5.44$

26 - Pantucek 1998
$Q = 0.37$
$z_Q = -1.52$

$N = 0.81$
$z_N = 4.52$
Phages

Bacteria

Q = 0.38
z_Q = 9.62

N = 0.64
z_N = 2.77

29 - Quiberoni 2003
$Q = 0.33$

$z_Q = -0.16$

$N = 0.93$

$z_N = 5.94$
$Q = 0.23$
$z_Q = -5.87$

$N = 0.98$
$z_N = 3.65$
$Q = 0.18$
$z_Q = 8.77$

$N = 0.93$
$z_N = 15.45$
Phages

Bacteria

33 - Sullivan 2003

\[ Q = 0.39 \]

\[ z_Q = -0.06 \]

\[ N = 0.79 \]

\[ z_N = 4.81 \]
Phages

Bacteria

$Q = 0.48$

$z_Q = -0.62$

$N = 0.91$

$z_N = 0.33$

34 - Suttle 1993
Phages vs. Bacteria

$Q = 0.08$
$z_Q = -0.34$

$N = 0.95$
$z_N = 4.13$
$Q = 0.62$
$z_Q = -0.19$

$N = 0.98$
$z_N = 1.11$
Phages

Bacteria

$Q = 0.22$
$z_Q = -7.42$

$N = 0.89$
$z_N = 14.44$

37 - Wichels 1998
Phages

Bacteria

\[ Q = 0.68 \]
\[ z_Q = 3.27 \]

\[ N = 0.85 \]
\[ z_N = 0.44 \]

38 - Zinno 2010