Microbial communities are shaped by viral predators. Yet, resolving which viruses (phages) and bacteria are interacting is a major challenge in the context of natural levels of microbial diversity. Thus, fundamental features of how phage-bacteria interactions are structured and evolve in the wild remain poorly resolved. Here we use large-scale isolation of environmental marine *Vibrio* bacteria and their phages to obtain estimates of strain-level phage predator loads, and use all-by-all host range assays to discover how phage and host genomic diversity shape interactions. We show that lytic interactions in environmental interaction networks (as observed in agar overlay) are sparse—with phage predator loads being low for most bacterial strains, and phages being host-strain-specific. Paradoxically, we also find that although overlap in killing is generally rare between tailed phages, recombination is common. Together, these results suggest that recombination during cryptic co-infections is an important mode of phage evolution in microbial communities. In the development of phages for bioengineering and therapeutics it is important to consider that nucleic acids of introduced phages may spread into local phage populations through recombination, and that the likelihood of transfer is not predictable based on lytic host range.
Phages are important predators of bacteria—they shape the structure, function, and evolution of natural microbial communities, and they are potential tools to manipulate microbial communities for industrial, bioengineering, and therapeutic applications.

Key to understanding the roles of phages in natural communities, and to their design and use as efficient and robust tools, is knowledge of their host ranges in the context of the systems in which they exist or will be used. Yet, how phage host ranges are structured in complex microbial communities remains challenging to address because the local genomic diversity of phage and bacterial strains is high, and phage-bacteria interactions are specific. Thus, resolving the structure of interactions at the strain-level requires systematic assays of host ranges of phages against panels of potential host strains. The largest such study in the context of natural microbial communities was performed by Moebus and Nattkemper in the 1970s. Later analyses of the structure of the Moebus-Nattkemper matrix by Flores et al. in 2013 found this network to have a statistically modular structure and numerous singleton interactions. This contrasted predictions made by Flores et al., in their prior large-scale meta-analysis of 38 phage-bacteria interaction networks (PBINs), that whereas interactions in laboratory PBINs were largely nested, larger environmental sampling would reveal modularity in interactions. However, as neither phages nor bacteria of the Moebus-Nattkemper matrix were genome sequenced, the relation of the observed modules to bacterial and phage genomic diversity could not be determined—and thus how phage and bacterial phylogenetic diversity shape the structure of PBINs in natural communities remains unclear.

In this work, we analyze a PBIN for which genomes of the majority of member phages and bacteria have been sequenced to address how environmental PBINs are structured in marine microbial communities. We show that the biological basis of modular structure in large-scale PBINs varies across modules and can be defined by either phage or bacterial phylogenetic boundaries; and we find that whereas overlaps in killing host ranges of phages are rare, local pools of phage genomes are highly recombined. We propose two models that reconcile the contrasting low overlap in killing among paired phages with the prevalence of recombined genomes, and point to cryptic co-infections of bacteria by multiple phages as being important in the ecology and evolution of phage-bacteria interactions in microbial communities.

Results

Co-occurring lytic phage predator loads appear low. To evaluate phage predation on closely related bacteria in the environment, we focused on the well-characterized coastal marine heterotrophic Vibrionaceae bacteria as a model system. We isolated 1440 strains, predominantly in the genus Vibrio, over three days (ordinal day 222, 261, and 286) during the course of the 3-month 2010 Nahant Collection Time Series and sequenced the housekeeping gene hsp60 to initially resolve their phylogenetic relationships. Using these isolates as bait we quantified concentrations of lytic phages present for each strain in seawater collected on the same days. By using direct plating agar overlay methods with virus concentrates, rather than enrichments, we were able to obtain estimates of concentrations of co-occurring plaque-forming phages for each bacterial strain. In previous work we showed that use of oxalate solution in this viral concentration procedure allows initial recovery of 49–55% of infective viruses (see Methods), as well as stable storage - thus, direct and doubled counts provide approximate lower and upper bounds of plaque forming units (PFUs) per ml of seawater concentrate in these assays. Of the 1287 total bacterial strains which both grew for the bait assay and for which we were able to obtain hsp60 gene sequences, 285 (22%) were plaque positive – revealing sensitivity to killing by co-occurring phages.

Our large-scale bait assay revealed that, at the strain level, lytic phage predation pressure on the majority of coastal ocean Vibrio appears low (<134 plaque forming phage L−1; limit of detection based on doubled counts assuming 50% recovery efficiency) compared to total virus-like-particle concentrations (1010 VLP L−1) common in coastal marine environments (Fig. 1, showing undoubled counts). As individual strains of the most abundant Vibrio species in our study typically occur at concentrations of on average <1 cell ml−1, these findings indicate that encounter rates should be very low between these phages and their hosts. These observations are consistent with previous studies of lytic (plaque forming) phage predator loads on heterotrophic marine bacteria (largely in the family Vibrionaceae and genus Vibrio), as well as in the genus Pseudalteromonas by Moebus, which also showed the majority of bacterial strains were subject to 0–10 PFU ml−1 in water samples collected in the same year. These observations suggest that mechanisms that increase encounter rates between Vibrio phages and their hosts—such as host blooms, spatial structure on small scales, and broad host range—should be important features of phage-bacteria interactions in systems where individual host strains are rare.

Lytic phage-bacteria interactions within the Vibrionaceae are overall sparse and modular. To investigate the host ranges of the phages in this system we purified one phage from each plaque-positive host for further study, representing a final set of 248 independent phage isolates (hosts: Supplementary Fig. 1, phages: Supplementary Data 8). In previous work we showed that these phages represent phylogenetically diverse dsDNA viruses ranging in size from 10 kb – 349 kb, including non-tailed members of the recently proposed family Autolykiviridae, as well as representatives of the three morphotypes of the Caudovirales (as predicted by Virfam). Host ranges of each of the phages were assayed against a panel of 294 genome-sequenced bacterial strains, including all plaque-positive hosts and 18 additional Vibrio strains (selected to represent additional populations of Vibrionaceae; for details on these additional strains see Supplementary Data 1 sheet A and filter for all bacterial strains with identifiers without the prefix 10 N). Of these hosts, 279 were lysed by at least one of the 248 phages in the host range assay and were included in subsequent analyses, with genomes of 259 member bacteria and all phages sequenced (Supplementary Data 1).

In this large-scale study of the host ranges of 248 phages on 279 hosts, we found that the majority of bacteria were resistant to the majority of phages and that interactions were overall sparse – with only 1436 lytic interactions observed in agar overlay (hereafter “killing”) out of 69,192 possible interactions. We further found that killing interactions were organized in an overall modular fashion – with groups of phages and bacteria clustering into 89 discrete interconnected sets (“modules”, Fig. 2a, subset shown is 248 phages and 259 hosts with sequenced genomes, details in Supplementary Data 1) using the BiMat modularity evaluation methods developed and employed by Flores et al. to investigate the Moebus-Nattkemper matrix. These features of our matrix are strikingly similar to those of the similarly large matrix generated by Moebus and Nattkemper in the 1980s (Table 1). However, unlike this previous matrix, performed at a time when genome sequencing of all members was not possible, we could now also investigate the structure of phage-bacteria interactions in light of genomic and phylogenetic diversity to understand the biological basis of the modular structure observed.
Diverse processes define membership of different interaction modules. The structure of the phage-bacteria interaction network in this study indicates that both broad and narrow host range strategies are important in the coastal marine environment. Whereas the three largest modules represented the majority of lytic interactions observed in agar overlay (53% of all interactions, 768/1436 total infections), the majority of modules were singletons comprised of only a single phage and bacterial strain interacting exclusively with each other (61/89 modules but only 4% of all interactions).

Central to the organization of each of the three largest modules were phages that were able to kill numerous genomically diverse host strains (Fig. 2b, with the three largest modules shown in blue, green, and red fill, respectively; hosts: Supplementary Fig. 1; Supplementary Data 1). The largest module was organized around killing by members of a new family of recently described phages, the Autolykiviridae, whose members can infect some but not all host strains in up to 6 species20. The second largest module was likewise organized around phages that killed multiple host strains within a single phylogenetically divergent species, Vibrio breoganii. This species is non-motile, lives predominantly attached to macroalgal detritus, and is specialized for degradation of algal polysaccharides23,24 - and thus is also ecologically distinct from other vibrios. The genomically diverse sipho- and podovirus phages infecting V. breoganii hosts were nearly all exclusive to this host species in their infections, suggesting that divergence in bacterial ecology is also reflected in interactions with different groups of phages. The third largest module was organized around
Fig. 2 The nested-modular structure of environmental phage-host interaction networks reflects multiple drivers. 

**a** Network analysis of the Nahant Collection infection matrix shows an overall nested-modular interaction structure and abundance of one-to-one infections. **b** Re-organization of the interactions in light of host phylogenetic and phage genomic diversity reveals that modular structure reflects the influence of host species, phage genera, phage host range strategies, and bloom dynamics. In both panels bacteria are represented as rows and phages as columns; both panels show the same 248 phages and the subset of 259 Nahant Collection hosts which were infected in the host range assay by one of the 248 phages and for which genomes were also available (see details on host subsets in Supplementary Data 1, sheet readme); in both panels a and b, all interactions within each matrix are colored according to BiMat leading eigenvector modules (the five largest groups are shown, for example, as blue, green, red, purple, and light yellow); in panel b bacterial strains are ordered based on phylogeny of concatenated single copy ribosomal protein genes, with leaf colors representing species; in panel b phages are ordered based on manual sorting of VICTOR genus-level trees into groups by morphotype irrespective of their higher order clustering (where VIC-genera of different morphotypes can be intermingled; VICTOR trees represent Genome BLAST Distance Phylogenies (GBDP) based on concatenated protein sequences for each phage genome, with branch lengths representing intergenomic distances scaled in terms of the GBDP distance formula d_6; each of the 49 phage VIC-genera are represented as a distinct group indicated by a circle filled with the color representing the morphotype of the genus (purple: non-tailed; red: myovirus; yellow: podovirus; green: siphovirus); see Supplementary Data 8 for full original VICTOR phylogeny not sorted by morphotype). 

Underlying data are provided in Supplementary Data 1 and Source Data Fig. 2, see Methods for strain sets included in the analyses. Phage icon source: ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics. Source Data Fig. 2.

**Table 1 Comparison of Moebus-Nattkemper and Nahant matrix properties.**

| Matrix Property | Moebus-Nattkemper Atlantic Time Series Matrix | Nahant Collection Matrix |
|-----------------|---------------------------------------------|--------------------------|
| # of Hosts (H) | 286 | 279 |
| # of Phages (P) | 215 | 248 |
| # of Species (S = H + P) | 501 | 527 |
| # of Interactions (I) | 1332 | 1436 |
| Size (M = H * P) | 61,490 | 69,192 |
| Connectance (C = I/M) | 0.02 | 0.02 |
| Host mean interactions (L_H = I/H) | 4.7 | 5.1 |
| Phage mean interactions (L_P = I/P) | 6.2 | 5.8 |
| Modularity | 0.7950* | 0.7306** |

Summary of properties using phageSet248 and baxSet279 (see Supplementary Data 1 for further details on isolates included and on infections). *Calculated using the bipartite recursively induced modules algorithm; **calculated using the leading eigenvector algorithm.
a single broad host range siphovirus (1.215.A) that infected 26 host strains in 6 species in our network, including members of both the Vibionaceae and the Shewanellaceae. All three of these large modules, while organized around broad host range phages that could infect multiple specific host strains, included other phages that were effectively entrained into the module as a result of sharing a host strain with the module-defining broad host range phages. The striking dominance of singleton modules in this network highlights the prevalence of exquisitely narrow host range profiles of phages with respect to their local hosts in the coastal marine environment. This finding parallels that of Moebus, who found that for 200 phages isolated from a coastal marine system nearly half infected only the original strain on which they were isolated. Moebus’ work also suggested that bloom dynamics are likely important in these systems by revealing ephemeral peaks of up to 1500 PFU ml−1 that decayed on the order of days. Such high concentrations and dynamic abundances have also been shown in other marine heterotrophic26 and cyanobacterial27,28 host systems, with observed maxima of up to 36,500 and 35,000 PFU ml−1, respectively. The presence in the Nahant matrix of modules that contain multiple closely related phages and hosts isolated on the same days is consistent with a role for host blooms in driving increases in relative abundances of specific phage types. For example, in the 4th largest module (23% of all infections) the majority of phages (18/19) had an average pairwise average nucleotide identity (ANI) of >99%, and infected largely the same set of closely related host strains (18/19 host strains in the module having >99.95% average pairwise ANI). The potential for Vibrio to form such blooms is well supported as they proliferate rapidly in response to nutrient pulses and have been observed to rapidly undergo large increases in relative abundance in microbial communities in the environment11,29.

Killing host ranges are not clearly defined by phage morphotypes. As host range breadth has previously been shown to be associated with morphotype,30 with myoviruses infecting more broadly than other tailed viruses, we examined whether this was true in this dataset. We found that non-tailed viruses infected significantly more strains than tailed viruses, whereas there were smaller differences between tailed viruses. Student’s t-tests showed significant differences in the number of host strains killed by members of the Autolykiviridae and each of the three tailed morphotypes—the podoviruses (p-value = 4.55e-09), siphoviruses (p-value = 1.22e-08), and myoviruses (p-value = 3.77e-09); and between myoviruses and siphoviruses (p-value = 3.20e-06). By contrast, no significant differences were found between podoviruses and siphoviruses (p-value = 0.02), or between podoviruses and myoviruses (p-value = 0.04). Autolykiviruses killed on average 31.3 strains (standard deviation 11.3), myoviruses on average 2.0 strains (s.d. 1.6), podoviruses on average 3.2 strains (s.d. 3.7), and siphoviruses on average 5.1 strains (s.d. 6.5), yet there were phages of all morphotypes, including the non-tailed autolykiviruses, that killed host strains in only a single species, in two species, in three or more species, and in two genera (Supplementary Data 2). Three siphoviruses killed hosts in both families of bacteria represented in this study, however, the limited representation of potential non-Vibionaceae hosts precludes any conclusion about whether this reflects a broader pattern. Interestingly, there were no myoviruses that killed the ecologically distinctive V. breoganii, though 71/248 phages were of this morphotype and this host species was present on all three isolation days (Fig. 1) and well represented in the host range assay (Supplementary Data 1). Together, these observations indicate that morphotype may not be a reliable indicator of the number of host species a phage will infect but may shape access to hosts with different ecological and habitat associations.

Vibionaceae phages are diverse and under-sampled. To next investigate patterns of phage host range across levels of phage genomic diversity, we operationally clustered phages into species and higher order groups. Because a standard approach has not yet been set, we use two methods for identifying groups of more (−species) and less (-genus) closely related groups of phages, VIRIDIC31 and VICTOR32. VIRIDIC determines intergenomic nucleotide similarities and groups viruses into clusters based on user-defined similarity cut-offs (here defaults of 70% for genera and 95% for species), whereas VICTOR identifies species and genera on the basis of pairwise whole genome distance comparisons followed by clustering benchmarked to previously described viral taxa (here using protein sequences and the d_e distance scaling formula). We find that these two methods largely agree at the species level (171 VICTOR species, 188 VIRIDIC species; both VICTOR and VIRIDIC species and genus clusters for each phage indicated in Supplementary Data 1), yet diverge at the genus level (49 VICTOR genera, 151 VIRIDIC genera; VICTOR taxon sequence similarity thresholds highly variable and reported with respect to VIRIDIC intergenomic similarity values in Supplementary Data 3). We provide comparisons between VICTOR and VIRIDIC as supplementary information (Supplementary Fig. 3 and Supplementary Data 3; and see Fig. 3 for overview of VIC-genera and Supplementary Fig. 2 for representation of genera across sampling days). An overview of all phage genomes organized by VICTOR distances are provided in Supplementary Data 8.

To ask whether any of these phage groups include previously described members, we used vConTACT233 to cluster the Nahant Collection phages with >10,722 previously described phages with available genome sequences in NCBI (details in Supplementary Data 4). We found the VICTOR and vConTACT2 genus-level clusters to be largely concordant and identified 17 Nahant Collection VICTOR genera (hereafter VIC-genera or, for species, VIC-species) that include previously described phages, though none in the same VIC-species as Nahant phages. The majority of previously described phages in these VIC-genera also infected hosts in either the Vibionaceae or Shewanellaceae, consistent with a previous finding that phage genera are largely specific to host families32 (see Supplementary Data 4 for exceptions). We thus find that local phage diversity is overall exceedingly high and under-sampled, even for this well studied host family – with, for example, 51 new VIC-species of phages isolated for one host species (Vibrio lentus) across our 3 sampling days.

Killing host ranges are not defined by phage life history strategy. A small subset of phages in the Nahant Collection show strong evidence for temperate lifestyles (Fig. 3). Phages in 6 of the 49 VIC-genera (VIC-genera 12, 23, 24, 28, 31 and 41; and including all phages in these genera) encode integrases or repli-
cative transposases, and phages in 2 VIC-genera (VIC-genera 31 and 41) were identified as prophages in bacterial genomes. Phages in 20 VIC-genera (including some members of the aforementioned groups) encode transcriptional repressor domains suggestive of potential for temperate life history strategies (see Methods for additional details on annotation of life history strategy and Supplementary Data 5 for read mapping results and summary of phage life history strategy related annotations). A previous study of human microbiome associated coliphages found host ranges of virulent phages to be broader than those of temperate phages34, and to likewise evaluate this here we compare host ranges of phages in species within high confidence
temperate genera (19/248 phages in 12 VIC-species, Supplementary Data 5) with those in high confidence virulent species (75/248 phages in 35 VIC-species, Supplementary Data 5). Overall, we detect no significant difference in the total number of bacterial species or \textit{hsp60}-types killed by a given phage species in relation to predicted life history strategy (two sample Kolmogorov-Smirnov test for species: \(D = 0.14524\), \(p\)-value = 0.9917, two-tailed, and for \textit{hsp60}-types: \(D = 0.2\), \(p\)-value = 0.8671, two-tailed; see Supplementary Data 5 for VIC-species life history assignments).

Overlap in killing host range is generally common only within phage species, yet recombination occurs more broadly. We next considered host range profiles in light of phage species and genera - using VICTOR taxa, as these correspond well with temperate genera (19/248 phages in 12 VIC-species, Supplementary Data 5) with those in high confidence virulent species (75/248 phages in 35 VIC-species, Supplementary Data 5).

Fig. 3 Overview of Nahant Collection phages by VICTOR genus (NCVicG). Features suggestive of temperate life history strategy were evaluated, and findings are highlighted as representing either strong (A and B) or weak (C) evidence, where: A indicates extensive mapping of bacterial genome reads to phage genomes (see Methods); B indicates presence of integrases (PF00239, PF00589) or replicative transposases (PF02914, PF09299); C indicates presence of transcriptional repressors (IPR010982, IPR010744, IPR00387, IPR032499); and D is noted only for reference and indicates sparse mapping of reads from bacterial genomes onto phage genomes (\(\leq 510\) bases total). Note also that: the phages in NCVicG_17 are representatives of the proposed family \textit{Autolykiviridae}; the sole NCVicG_41 phage is in the described subfamily \textit{Peduvirinae}; the phages in NCVicG_41 encode genes for replicative transposition, and the phages in NCVicG_20, _42, and _49 are N4-like in encoding a giant RNA polymerase. All counts reported in the table for Recombinases and Number in Nahant Collection are based on phageSet248 (see Supplementary Data 1 for strain set lists and descriptors); heat map ranges from 1 (green) to 36 (red).
VIRIDIC at the species level but offer greater breadth of diversity at the genus level for systematic supra-species comparisons. We found that whereas overlap in killing is high between phages within VIC-species, it is low between phages in different species. This feature is evident in visual evaluation of host range profiles (infection matrices in Fig. 4a–c), and is consistent with the striking diversity of putative receptor binding proteins (RBPs) among phages of different VIC-species (protein cluster matrices in Fig. 4a–c; Supplementary Data 6). To evaluate these differences quantitatively, we defined a metric of host profile concordance based on Jensen-Shannon distance between host range profiles represented as normalized binary (killed or not-killed) vectors of host strains (see Methods). With this metric, a concordance value of 1 is equivalent to perfect overlap in the host ranges of any two
host range overlap is high within phage VIC-species and low within phage VIC-genera, but recombination occurs both within and between VIC-species. VIC-genus trees for each of a group of podoviruses (a), myoviruses (b), and siphoviruses (c) represent VICTOR Genome BLAST Distance Phylogenies (GBDP) based on concatenated protein sequences for each phage genome, with branch lengths representing intergenomic distances scaled in terms of the GBDP distance formula $d_{gb}$ (complete tree with all phages shown in Supplementary Data 8, with underlying data in Newick format provided in Source Data Fig. 2. Filled in cells in the host range matrices aligned to the right of phage names (in rows) show host strains (in columns) killed by each phage. Protein cluster matrices aligned to the right of the host range matrices show all the MMseqs2 protein sequence clusters present in each genus (columns), ranked based on the number of phages in the VIC-genus in which they occur. Quantified host range profiles for phages across the collection show that: d) overlap in killing profiles (concordance) is high within VIC-species (28 VIC-species with ≥2 phages, 105 phages total) but low within VIC-genera (31 VIC-genera with ≥2 phages, including cases of genera represented by only 1 species, 230 phages total); two-sided Welch’s t-test p-value = 1.45e-07; that e), recombination in conserved regions is commonly a greater contributor to genomic diversity in both species and genera (same phage counts as in panel d); and, f and g, that there is no relationship between concordance in killing and recombination for either VIC-species or VIC-genera, respectively. Underlying data and strain information available in Supplementary Data 1 and 6, and in Source Data Fig. 4, see Methods for description of differences in results when considering only single VIC-species representatives in VIC-genus-level analyses. Boxplot features: central line = median; box limits = 1st and 3rd quartiles; upper whisker = largest value no larger than 1.5 * IQR (inter-quartile range); lower whisker = smallest value no smaller than 1.5 * IQR. Source Data Fig. 4.

phages, and a value of 0 represents no shared hosts. At the VIC-species level we found that concordance values were generally high (Fig. 4d), with phages in 10 VIC-species showing perfect overlap in their host range, including 3 VIC-species with member phages isolated on different days (this analysis included 105 phages, representing the 28 VIC-species with >1 member; see Source Data Fig. 4 sheet A). By contrast, within VIC-genera, concordance in host range among members was generally low (Fig. 4d), even when calculated using a conservative approach that yields higher estimates of concordance when VIC-species in a VIC-genus contain multiple members (see Methods, see Source Data Fig. 4 for concordances calculated using all members of a genus [sheet B] or with only a single representative of each species [sheet C], and see Supplementary Fig. 4 for scaling of concordance value with size of subsampled groups.).

The differences in host range concordance at the phage VIC-species and VIC-genus levels suggested that there should be corresponding differences in levels of recombination between phage genomes within these groups, yet we found recombination to be occurring both within and between phage VIC-species. We observed this qualitatively in the distributions of concordances calculated using all members of a genus [sheet B] or with only a single representative of each species [sheet C], and see Supplementary Fig. 4 for scaling of concordance value with size of subsampled groups.).

The importan of recombination at both the VIC-species and VIC-genus levels indicates that overlap in killing between phages is not predictive of their potential for recombination in the context of natural microbial communities. This is corroborated when both quantitative metrics are considered together, which shows a lack of a positive association between host range concordance and r/m as >1 within VIC-genera is a conservative estimate of the potential extent of homologous recombination occurring between phages in supra-species level taxa.

To understand the potential maximal extent of recent gene flow among all phages in this collection we used a k-mer based approach and found evidence of sequence sharing between phages of different Caudovirales morphotypes with non-overlapping host killing. We first used the liberal metric of OCCurrence of sharing of any 100% identity 25-base pair (25-mer) length sequences between phages, such 25-mers are sufficiently unique in bacterial genomes to be used to recapitulate strain and species level relatedness38, and provide a marker for potential recent horizontal gene transfer events39. Using this approach, we found far fewer potential connectivity of gene flow than suggested by overlaps in host killing (Fig. 5a, b, Source Data Fig. 5). Notably, however, despite extensive overlap in killing host range between autolykiviruses and tailed phages, there were no cases of shared 25-mers between phages in these two groups—a feature consistent with their lack of any shared protein clusters (Supplementary Data 4 and 6).

Considering next the more conservative metric of total numbers of 25-mers shared between any two phage genomes we also found evidence for sequence sharing between divergent tailed phages with non-overlapping host ranges (Fig. 5c, Supplementary Fig. 3); with the maximum number of shared 25-mers between any pair of phages of different morphotypes being 6,169. In the collection, some pairs of phages of different morphotypes show evidence of extensive sequence similarity in non-structural genome regions (e.g. Fig. 5d, similarity between NCVicG45 siphoviruses with the singleton NCVicG25 podovirus), paralleling the observation that the podovirus P22 shares sequence similarity to lambdoid siphoviruses in its non-structural genes; other pairs share only a single 25-mer, for example in a DNA polymerase gene (e.g. in protein cluster mmseq 2145 in podovirus 1.262.O...
and myovirus 1.063.O, cross-reference Source Data Fig. 5 with shared protein cluster in Supplementary Data 6). Finally, an approach designed to detect specific recent gene transfer events (MetaCHIP) from conserved regions of phage genomes also revealed connections between tailed phages in different VIC-genera, without overlapping host ranges, and between which there were often long regions of high genomic sequence similarity (Fig. 5d, Supplementary Fig. 5). Overall, however, when considering only VICTOR distances between phages (based on whole genome concatenated protein sequences), short pairwise distances occur almost exclusively between phages of the same morphotype whereas large pairwise distances can be observed both within and between phage morphotypes (Supplementary Fig. 6).

Altogether, these observations on the extent of recombination among tailed phages despite their lack of overlap in killing suggests that these phages are generally infecting more hosts than they are able to kill. This is supported by a recent study of this collection showing that phages in different VIC-genera can infect the same sets of closely related host strains using different
Nucleotide sharing between phages can lead to cross-mapping of sequencing reads. Considering the potential for widespread sequence sharing to influence mapping of viral reads to reference genomes, we investigated cross-mapping using a recently developed rapid k-mer based pseudo-alignment approach\(^4^3\). We found that cross-mapping of reads can occur between phages of different VIC-species, VIC-genera, and morphotypes within this collection. False positive classifications of reference presence were reduced when using shorter simulated read lengths as a result of overall lower collateral (false positive) sequence coverage when the basis for the mapping was a single 31-mer match in the sequence (Supplementary Fig. 7, see Methods). This observed potential for cross-mapping calls for a cautious approach in using read-mapping to reference genomes in determining whether specific phages are present in metagenomic samples or predicting which hosts virus pools are interacting with.

Recombinases are prevalent in Vibrionaceae phage genomes. The overall prevalence of sequence sharing observed between phage genomes suggests that they harbor homologous recombination systems, and indeed we find that recombinase genes are encoded by the majority of targeted phages in this collection. Low fidelity single strand annealing protein (SSAP) based recombinase systems such as those in the Rad52-superfamily (e.g. Red\(^{B/R}/\)RecT-, ERF-, and Sak-families) are common in temperate phages\(^4^4\), and are thought to play an important role in their extensive genome modularity and mosaicism\(^4^5\). Such recombinases have been shown to be associated with large-scale recombination events of up to 79% genome length between incoming temperate phages and resident prophages\(^4^6\) and are useful tools for in vitro genetic engineering (recombineering)\(^1^7\)–\(^3^0\) as they can facilitate recombination between sequence regions with as little as 23-bp sequence identity\(^5^1\). Noting a number of putative SSAP recombinase genes in our initial annotations, we sought to more systematically evaluate their representation in our diverse collection of phages. Using representative sequences\(^4^4\) as seeds for iterative searches, followed by gene neighborhood analysis, we identified putative recombinases in 196/230 (85%) targeted phages (in phageSet248, see Methods and Supplementary Data 8 showing genome diagrams with recombinases highlighted), with 117 of these resembling low-fidelity Rad52-superfamily and Sak4-like Rad51-superfamily recombinases commonly associated almost exclusively with temperate phages\(^4^4\) (Fig. 6, Supplementary Data 7). That no recombinases were identified among the autolysiviruses, though these viruses also showed evidence for high rates of intraspecies recombination (when calculated using single representatives of each species rather than all members, Source Data Fig. 4, sheet F), indicates distinct pathways underlying observed recombination in this group. These results suggest that just as horizontal gene transfer in microbial communities may allow bacteria to evolve resistance to phages, recombination and genetic exchange between phages may likewise be important in overcoming this resistance.

Discussion

The findings of this work appear at first contradictory: that recombinated phage genomes commonly co-exist in natural microbial communities, yet overlap in phage killing in the context of natural diversity is rare. Previous work has spoken to the importance of recombination in generation of phage diversity globally and locally\(^5^2\), and our observations imply that recombinants are generated in co-infections of shared hosts that are more frequent than are revealed by killing assays. That these largely unobserved implied co-infections are occurring in recent evolutionary and ecological time is indicated by the higher contribution of homologous recombination than mutation to sequence divergence in many phage groups in our dataset. We propose two main complementary model scenarios to unify and reconcile these apparently contradictory observations – one addressing trade-offs between growth and phage defense in bacteria and the other pointing to recombination as an important mechanism for phage survival in the face of selective pressure by bacterial anti-phage defenses.

First (Fig. 7a), recombinant phages may be disproportionately generated during co-infections in broadly sensitive (killed) hosts. Previous work with the marine phototroph Synechococcus by Waterbury and Valois\(^2^7\) showed that the rarest bacterial strains were the fastest growing as well as the most sensitive to phages. As a result of their lower relative abundances in the environment, these ecologically important broadly phage sensitive hosts may thus be underrepresented in cultivated isolate collections such as ours. Recent work\(^4^2\) has also demonstrated that rapid turnover in anti-phage defense systems\(^4^5\) is key to fine-scale differences in phage sensitivity among closely related bacterial strains, and to
what extent this rapid turnover transiently yields rare broadly sensitive hosts susceptible to co-infection and killing by multiple phage types remains to be determined. Second (Fig. 7b), penetrative host ranges of phages are generally substantially broader than the replicative host ranges revealed by killing assays like those used in our study—with narrow host ranges reflecting effective anti-phage defense systems rather than lack of phage adsorption and genome delivery. This has recently been shown to be true for phages in this collection in members of VIC-genera 47 and 48, the “orange” and “purple” phages in Hussain and Dubert et al. 42. Previous studies in marine Synechococcus and Prochlorococcus in E. coli, and in diverse other groups of bacteria, have also shown that phage penetrative host ranges are often broader than replicative host ranges. Findings that highly sensitive indicator strains can yield very high local phage predator loads in both heterotrophic and cyanobacterial host systems further support that local specificity of interactions reflects local defenses against phages, rather than general lack of adsorptive hosts. Altogether, this model is supported by work showing that anti-phage defense systems that effectively abrogate replication are widespread, diverse, and patchily distributed among strains within bacterial species. Indeed, having broad infective host ranges that expose phages to nucleic acid degrading host defense systems may select for carriage of recombinase genes that facilitate rescue in what would otherwise be abrogated infections. In both model scenarios, the potential for co-infections and recombination are expected to be shaped also by both phage and bacterial ecology. Conditions and phage life history strategies enriching for lysogeny or pseudolysogeny will increase potential
phages are able to replicate yet do not readily form plaques. For example, where there is an inverse relationship between phage sensitivity and relative abundance and growth rate, broadly sensitive host strains may be ecologically important but systematically underrepresented in isolation studies. Second, co-infections in commonly observed defended hosts that are only killed by a few phage types (indicated by both filled and empty squares) may also be an important source of recombinant phage progeny, particularly where phage encoded recombinases are expressed in the presence of phage genome fragments generated by host anti-phage defense systems. In both scenarios, where phages are associated with hosts through states of lysogeny or pseudolysogeny, the probability of co-infections, and thus potential for generation of recombinants, will be even greater. Figure created with BioRender.com.

Fig. 7 Cryptic-co-infection models reconcile sparse killing overlap with prevalent recombination in phages. We propose that two classes of cryptic (as in rarely observed in the laboratory) co-infections are key in unifying two at first contradictory observations - that in natural microbial communities overlap in killing is rare among tailed phages, yet recombinant phage genomes are common. First, co-infections in rarely observed broadly sensitive hosts killed by multiple phage types (indicated by filled in squares in the figure) may be an important source of phage recombinants in local microbial communities. Where there is an inverse relationship between phage sensitivity and relative abundance and growth rate, broadly sensitive host strains may be ecologically important but systematically underrepresented in isolation studies. Second, co-infections in commonly observed defended hosts that are only killed by a few phage types (indicated by both filled in and empty squares) may also be an important source of recombinant phage progeny, particularly where phage encoded recombinases are expressed in the presence of phage genome fragments generated by host anti-phage defense systems. In both scenarios, where phages are associated with hosts through states of lysogeny or pseudolysogeny, the probability of co-infections, and thus potential for generation of recombinants, will be even greater. Figure created with BioRender.com.

Methods

Environmental sampling. Samples were collected from the littoral marine zone at Canoe Cove, Nahant, Massachusetts, USA, on 22 August (ordinal day 222), 18 September (261) and 13 October (286) 2010, during the course of the three month Nahant Collection Time Series sampling.

Bacterial isolation and characterization

Bacterial strains were isolated from water samples using a fractionation-based approach as previously described. In brief, seawater was passed first through a 63um plankton net and then sequentially through 3um (Whatman 111113 or Sterlitech PCT5047100), 1um (Whatman 111110 or Sterlitech PCT1047100), and 0.2um (Whatman 111110 or Sterlitech PCT5047100) filters placed directly onto 0.2um polyethersulfone filters (Pall 66234) in a carrier solution of artificial seawater (40 g Sigma Sea Salts, 59883; 0.2um filtered), and filters placed directly onto Vibrio-selective MTCBS plates (BD Difco TCBS Agar 265020, supplemented with 10 g NaCl per liter to 2% w/v). Colonies (96) from each of three replicates of each size fraction were selected from the dilution plates with the fewest numbers of colonies (1,152 isolates per isolation day). Colonies were purified by serial passage, first onto TSB-II (Difco Tryptic Soy Broth, 1.5% BD Difco Bacto Agar 214010, amended with 15 g NaCl to 2% w/v), second onto MTCBS, finally onto TSB-II again. Colonies were inoculated into 1 ml of 2216 Marine Broth (BD Difco 279110) in 96-well 2 ml culture blocks and allowed to grow, shaking at room temperature, for 48 h. Glycerol stocks were prepared by combining 100 ul of culture with 100 ul of 50% glycerol (BDH 1172-4LP) in 96-well microtiter plates and sealed with adhesive aluminum foil for preservation at −80 °C.

Bacterial hsp60 gene sequencing. To obtain hsp60 gene sequences for isolates, Lyse and Go (LNG) (Pierce, Thermo Scientific 78882) treatments of subsamples of the
same overnights cultures used in the bait assay (described below) were used directly as template in PCR amplification reactions. PCR reactions were prepared in 30 µl volumes, as follows: 1 ul 1NG template, 3 ul 10X buffer, 3 of 2 mM dNTPs, 3 ul 2um hsp60-F primer, 3 ul 2um hsp60-R primer, 0.3 ul NEB Taq, 16.7 µl of PCR-grade HOH; with hsp60-F (H279) primer sequence: 5’-GAA TTC GAT III GCI GGI GAY GAI ACI ACI-3’, and hsp60-R (H280) primer sequence: 5’-CCG GCC GGG CTK YTK ITC CIC RAA ICG IGC YTT-3’ (Supplementary Table 1). PCR thermocycling conditions were as follows: initial denaturation at 94 °C for 30 seconds; 35 cycles of 94 °C for 1 min, 37 °C for 1 min, 72 °C for 1 min; final annealing at 72 °C for 6 min; hold at 10 °C. PCR products were cleaned up by isopropyl alcohol (IPA) precipitation, as follows: addition of 50 ul 70% IPA to 30 ul PCR reaction followed by incubation at 25 °C for 20 min, centrifugation at 2800 rcf, inversion on paper to remove IPA, 10 min centrifugation at 700 rcf, air drying in PCR hood for 30 min, resuspension in 30 ul (IPA) precipitation, as follows: addition of 100 ul 75% IPA to 30 ul DNA template. For a subset of strains hsp60 sequences were obtained from subsequently determined whole-genome sequences. Hsp60 sequences were aligned to the hsp60 sequence previously published for Vibrio VS8, and trimmed to 422 bases using Geneious (https://www.geneious.com). Accession numbers for these 1287 strains are provided in Supplementary Data 1, where they are identified as bxset1287.

Bacterial hsp60 phylogenies. A phylogenetic tree of relationships among bacterial isolates screened in the bait assay (described below) was produced based on a 422 bp fragment of the hsp60 gene, derived either from Sanger or whole genome sequences; with E. coli K12 serving as the outgroup. Sequences from each of the three days of isolation were aligned using muscle v.3.8.3162 with default settings. A bacterial RiboTree tool66 to produce a phylogeny based on concatenated single copy ribosomal proteins as in23. We include strains of previously described fibrobacterial isolates in this analysis, with an updated recipe is provided elsewhere67. For the bait assay each of the strains included in these analyses are provided in Supplementary Data 1, where these strains are identified as bxset1287.

Bacterial genome sequencing and assignment to populations. To assign genome-sequenced strains used in the bait assay to species, we used Riboosec66 tool to produce a phylogeny based on concatenated single copy ribosomal proteins as in23. We include strains of previously described Vibrio archaea in preliminary analyses as reference strains and assign species names to new isolates based on clustering with named representatives, as well as provide placeholder names for newly identified clades with no previously described representatives. Trees were reconstructed using FastTree v.2.1.863 (FastTree -gtr-gamma -nt -spr 4 -slow < Seq-ALL.muscleAb > Seq-ALL.muscleAb.fasTree). For presentation in Fig. 1 sub-trees including only nodes from each day were produced using PareTree v.1.0.25 (java -jar PareTree1.0.jar -t 0 -dl -not-Day22 -xt < Seq-ALL.Bound.muscleAb.fasTree-Day22). Trees were visualized using ITOL65 and painted with metadata for each of the strains, including: sensitivity to killing in agar overlay by co-occurring phage predators collected on the same day and, for the subset of strains that were genome sequenced and also included in the host range matrix, the bacterial species, based on concatenated ribosomal protein analysis using Riboosec66 as described below. Isolation days for each of the strains included in these analyses are provided in Supplementary Data 1, where these strains are identified as bxset1289.

Viruses isolation and characterization. We have previously described features of the viruses of the Nahant Collection20, as well as approaches used for the standardization of their genome assemblies24, additional details are provided below.

Virals sample collection. The iron chloride flocculation method was used to generate 1000-fold concentrated viral samples from 0.2 um-filtered seawater, as follows. For each isolation day, triplicate 4L seawater samples were filtered through 0.2 um polycarbonate filter cartriges (Millipore, Sterivex, SVGP0150) into collection bottles, spiked with 400 ul of FeCl3 solution (10 g/l; Fe as 4.83 g FeCl3·4H2O (Mallinkrodt 5029) into 100 ml H2O), and allowed to incubate at room temperature for at least 30 min before transfer to storage at 4 °C. A reagent used in this original formulation (JT-Baker 7501 Mg,EDTA) is no longer available and an updated recipe is provided elsewhere27.

Bait assay and associated viral plaque archival. In order to obtain estimates of co-occurring phage predator loads at bacterial strain level resolution, and generate plaque samples that are to isolate phages, we exposed 1440 purified bacterial isolates to phage concentrates from their same day of isolation (1334 yielded lawns sufficient to evaluate for plaques, and hsp60 sequences could be determined for 1287 of these). Bacterial isolates were screened included up to 400 isolates from each of the 120 strains from each of the 4 size-fractionation classes (0.2 um, 1.0 um, 5.0 um, 63 um) details of isolation origin are provided for each strain in Supplementary Data 1, and description of naming conventions is as previously described27. For the bait assay each strain was mixed in agar overlay with seawater concentrates containing viruses (15 colony forming units equivalent to 15 ml unfiltered seawater assuming 100% recovery efficiency; derived from pooling of three replicate virus concentrates from each day). We note that recoveries were not tested for individual samples and that previous tests11 of recovery efficiency have shown that resuspension of iron flocculates in oxalate solution yields initial recoveries of approximately 50% (49 ± 3% and 55 ± 11% for a marine siphoviridae and myovirus respectively, at 24 h post re-suspension) and shows low decay rate over time (47 ± 5% and 73 ± 16% for a marine siphoviridae and myovirus respectively, at 38 days post-re-suspension). All of our assays were performed approximately 8–9 months post-sampling from oxalate concentrates stored at 4 °C. Agar overlays were performed based on the previously described Tube-free method35, as bait bacterial strains were plated for agar overlays using either purified plating out from glycerol stocks onto 2216MB agar plates with 1.5% agar (Difco, BD Bacto, 214010), and allowed to grow for 2 days at room temperature. Strains were then inoculated into 1 ml 2216MB in a 96-well culture block and incubated 24 h at room temperature shaking at 275 rpm on a VWR D5000 orbital shaker. Immediately prior to use in direct measuring the OD at 600 nm in 96-well plates used for 221SBs were taken for Lyse and Go (LNG) processing for DNA (10 ul culture, 10 ul LNG). Phage concentrates were prepared for plating by pooling 1.2 ml from each of the concentrate replicates into a 7 ml borosilicate scintillation vial. Cultures were transferred from overnight culture blocks to 96-well PCR plates in 100 ul volume and 15 ul of pooled phage concentrate was added to cultures once a time, with each row plated in agar overlay before adding phage concentrate to the next row of bacterial cultures. Mixed samples of 100 ul bacterial overnight culture and 15 ul pooled phage concentrate were transferred to the surface of bottom agar plates (2216MB, 1% agar, 5% glycerol, 125 ml L−1 of chitin supplement [40 g L−1 coarse ground chitin, autoclaved, 0.2 um filtered]). A 2.5 ml volume of 52% molten top agar (2216MB, 1% agar, 5% glycerol, 125 ml L−1) was added to the surface of the bottom agar and swirled around to incorporate and evenly disperse the mixed bacterial and phage sample into an agar overlay lawn. Agar overlay lawns were held at room temperature for 14–16 days and observed for plaque formation. Glycerol was incorporated into this assay to facilitate detection of plaques26. Chitin supplement was incorporated into this assay to facilitate detection of phages interacting with receptors upregulated in response to chitin degradation products. A variety of preliminary tests exploring potential optimizations to agar compositions for direct plating indicated that the addition of chitin did not negatively impact recovery of plaques with control phage strains. Tested after approximately 2 weeks, plaques on agar overlay lawns were cataloged and described with respect to plaque morphology and plaques were picked from plates using the previously described Archiving Bureau31,33. All plaques were archived from plates containing less than 25 plaques, on plates with larger numbers of plaques a random subsample of plaques from each distinct morphology were archived. A polypropylene 96-well PCR plate was filled with 200 ul aliquots of 0.2 um filtered 2216MB, agar plugs were collected from plates using a 1 ml barrier pipette tip and ejected into the 2216MB, skipping one well between each sample to minimize potential for cross-contamination, for a final count of 48 phage plugs per plate. Plate plugs were soaked at 4 °C for several hours to allow elution of phage particles into the media. After soaking, 96-well plates were centrifuged at 2000 rcf for 3 min before proceeding to the next step. Plug soaks were then processed for 2 independent storage treatments. For storage at 4 °C, plates were processed by transferring 150 ul of eluate from each well to a 0.2 um filtration plate (Millipore, Multiscreen HTS GV 0.22um Filter Plate MSGB522) and gently filtered under vacuum to remove bacteria, the cell-free filtrates containing eluted phage particles from each plate plug were stored at 4 °C. For storage at −20 °C, 50 ul of glycerol was added to the residue of the plug eluting the agar plug. In this way all plaques were characterized and many plaques from each strain were archived in two independent sets of conditions. Total plaque counts for all strains included in the bait assay are represented in Fig. 1, and provided in Supplementary Data 1, where they are identified as bxset1287. Notes on limitations to the associated viral genomes on the previous section is provided elsewhere33. A subset of plaques archived during the bait assay was selected as baxSet1287. Membrane filters (Millipore, 3M TLR10-50), 5.25 µM Tris-HCl (Promega PH8142), 25 µl 1 M oxalic acid (Mallinkrodt 2725), adjusted to pH 6 with 10 M NaOH (I.T. Baker, 3722-01); final volume 60 ul. After 7 days of preparation and maintained at room temperature in the dark was added to the vial and the sample allowed to dissolve at room temperature for at least 30 min before transfer to storage at 4 °C. A reagent used in this final preparation (JT-Baker 7501 Mg,EDTA) is no longer available and an updated recipe is provided elsewhere27.
subset included single randomly-selected representatives from each plaque-positive bacterial strain. Minor details of the purification and lysis preparation varied accordingly. Plates were prepared as follows: Plates were purified from inocula derived primarily from ~20 °C plaque cultures, and secondarily from 4 °C cultures when primary attempts with ~20 °C stocks failed to produce plaques. Three serial passages were performed using Molten Streaking for Singles [13] method. Agar overlay lawns for passages were prepared by aliquoting 100 ul of host overnight culture onto the surface of Nonidet P40 (NP-40) mucedo agar (BD Difco Agar, shaken overnight at RT at 250 rpm on VWR DS500E orbital shaker) onto a standard size bottom agar plate and adding 2.5 ml of molten 52 °C top agar as in the bait assay, swirling to disperse the host into the top agar and form a lawn, and streaking with barrier tip 1 ml pipettes to include archi-fimbrial host phages. Following plaque formation on the third serial passage plate plaque plugs were picked using barrier tip 1 ml pipettes and ejected into 250 ml of 2216MB to elute overnight at 4 °C. Plaque eluates were spiked with 20 ul of host culture and grown with shaking for several hours to generate a primary small-scale lysate. Small scale primary lysates were centrifuged to pellet cells and filtered by drop spot assay to estimate optimal inoculum volume to achieve confluent lysis in a 150 mm agar overlay plate lysate. Plate lysates were generated by mixing 250 ul of overnight host culture with primary lysate and plating in 7.5 ml agar overlay. After development of confluent lysis of lawns, plates were compared against negative control without addition, the lysates were harvested by addition of 25 ml of 2216MB, shredding of the agar overlay with a dowl, and collection of the broth and top agar. Freshly harvested lysates were stored at 4 °C overnight for elution of phage particles, the following day lysates were centrifuged at 5,000 rcf for 20 min and the supernatant filtered through a 0.2 um Sterivex filter into a 50 ml tube and stored at 4 °C.

Viral genome sequencing. Sequencing of Nahant Collection viruses was described in previous work [19], and was performed as follows. For DNA extraction approximately 18 ml of phage lysate was concentrated using a 30 kD centrifugal filtration device (Millipore, Amicon Ultra Centrifugal Filters, Ultracel 30 K, UFC903024) and washed with 1:100 2216MB to reduce salt concentrations inhibitory to downstream nuclease treatments. Concentrates were brought to approximately 500 ul using 1:100 diluted 2216MB to reduce salt concentrations inhibitory to downstream nuclease treatments. Concentrates were centrifuged at 5,000 rcf for 20 min and the supernatant filtered through a 0.2 um Sterivex filter into a 50 ml tube and stored at 4 °C.

\[ \text{Illumina constructs were prepared from sheared DNA as follows: end repair of sheared DNA (NEB, Quick Blunting Kit, E1021L), 0.75x/0.21x dSPRI (AMPure XP SPRI Beads) size selection to enrich for ~300 bp sized fragments, ligation (NEB, Quick Ligation Kit, M2200L) of Illumina adapters and unique pairs of forward primers for each sample with a M2200L AMPure XP SPRI bead clean up, nick translation (NEB, Rsa DNA polymerase, M2072L), and final SPRI (AMPure XP SPRI Beads) clean up (Rodrigue et al., 2010). Controls were enriched by PCR using PE primers following qPCR-based normalization of template concentrations. Enrichment PCR were performed in octuplicate 25 μl volumes, with the result of 1 Illumina construct primer set 5x 5 μl of each SPRI bead clean up buffer (NEB, 5X SPRI Binding Buffer), 0.5 μl of 10 mM dNTPs (NEB, dNTP Mix (1 mM: 0.5 μl), N1201AA), 0.25 μl 0.1 μM IGA-PCR-PE-F primer, 0.25 μl 0.1 μM IGA-PCR-PE-R primer, 0.25 μl Phusion polymerase (NEB, Phusion High Fidelity DNA Pol, M0530L), 17.75 μl 1X PE-grade PCR water. PCR thermocycling conditions were as follows: initial denaturation at 98 °C for 20 sec; batch dependent number of cycles (range of 12–26) of 98 °C for 15 sec, 60 °C for 20 sec, 72 °C for 20 sec; final annealing at 72 °C for 5 min; hold at 10 °C. For each sample 8 replicate enrichment PCR reactions were pooled and purified by 0.8x SPRI beads (AMPure XP PE) clean up. Each sample was then checked by Bioanalyzer (2100 expert High Sensitivity DNA Assay) to confirm the presence of a unique monodisperse plasmid. Concentrates were brought to ~500 ml. Sequencing of phage genomes was distributed over 4 paired-end sequencing runs as follows: HiSeq library of 18 samples pooled with 18 external samples, 3 MiSeq libraries each containing ~100 multiplexed plasmid genomes. Accession numbers for all sequenced phage genomes are provided in Supplementary Data 1, where they are identified as pha-ce283-4. Several subsets of phage used in the portfolio analyses in this manuscript are identified as phageSet248 and exclude non-independent isolates derived from the same plaque, as well as additional phages isolated from multiple independent plaques from the same host strain in the bait assay.

Viral protein clustering. To characterize and annotate groups of proteins in assembly, the Nahant Collection viral proteins were clustered using MSeqs2 v. 2.2369 [20] with default parameter settings, the 21,937 proteins reported in the GenBank files associated with each of the 283 Nahant Collection phage genomes were clustered into 5,929 clusters including 2,978 singletons. MSeqs2 cluster assignments for each protein sequence are provided in Supplementary Data 6.
superfamily and family, as well as all associated annotations, are provided in Supplementary Data 6 (sheet A.proteins_overview column anno_Recombinase_manual). Additional details regarding seed sequences and MCL cluster assignments associated with recombinase analyses are provided in Supplementary Data 7 which contains a main description sheet (00.readline), an overview of the 224 Nahant phages with recombinases (sheet 01.NahantPhageRecombinases_224), a table of InterPro domains associated with each of the reference and Nahant recombinases, with specific mmseqs and MCL annotations. Logos from MCL analysis of the annotation entries were visualized (sheet 03.List1_LOPes_ALL_noTTuision), the output of the iterative jackhammer search with seeds against all Nahant Collection proteins (sheet 04.List1_vs_NahantProt), the output of the all-by-all jackhammer search for 194 references and 224 Nahant recombinases (sheet 05.Lopese +Nahant224_v_self2iter), and information on the assignment of all Nahant and reference proteins to MCL clusters as shown in Fig. 6 (06.Recombinase_assign_by_MCL).

All proteins were assigned to one of three broad categories - structural, other (non-structural), or no prediction - based on manual review of annotations derived from: NCBI product ID, Virfam73, PfANNs74, PHOG33, eggNOG-mapper33, Phyre272, the MPI Bioinformatics implementation of HHpred78, and targeted annotations of predicted receptor binding proteins and recombinases (see descriptions for targeted annotations in Methods, above). Protein clusters (mmseq groups) were reviewed for conflicting calls and ultimately all proteins within each protein (mmseq) ID were reviewed to a single category. All assignments, and annotations on which they were based, are provided in Supplementary Data 6.

The approach for assigning annotations to these broad categories was as follows: Step 1) All genes identified as putative recombinases through targeted annotations were assigned as "other". Step 2) All genes identified as putative receptor binding proteins through targeted annotations were assigned as "structural". Step 3) Genes not assigned to a category in steps 1 and 2, and which were identified by Virfam as "head-neck-tail" associated were assigned as follows: Genes annotated by Virfam as a terminase (TerI) were assigned as "other"; genes annotated by Virfam as a major capsid protein (MCP), portal (portal), adaptors (Ad1, Ad2, Ad3), DNA-assembly proteins (Helic, Hel2, Tc1, Tc2), major tail protein (MTP), neck (Ne), or sheath (sheath) were assigned as "structural". Step 4) Genes not assigned to a category in steps 1–3, were assigned as "structural" or "other" (non-structural) if identified as such by PfANNs with a confidence of ≥95%. Cases where conflicting annotations were observed between PfANNs and other annotations were flagged for review in subsequent steps. Step 5) Genes with annotations of VOG0263 (DNA transfer protein), terminal protein, any reference to internal virion protein, DNA circulation protein, and Mu-like proteins were assigned as "other"; in the case of conflict the Step 5 annotation superseded the prior annotation. Step 6) Genes with annotation as a terminase (large subunit, small subunit, and unspecified) by any of the tools (requiring ≥90% confidence if based on Phyre2) were assigned as "other". Step 7) All genes lacking support across annotations were assigned as "no prediction", high confidence Phyre2 predictions qualitatively judged as inappropriate were disregarded. Step 8) Genes flagged in Step 4 were reviewed and as assigned as "structural" when containing any structural related genes (i.e. those listed in Step 3 and any others identifiable as structural based on words in the annotations and consensus across tools, e.g. containing the word baseplate, capsid, coat, head, spike, tail, whisker, fibrin). Additional targeted annotation by HHpred was used to facilitate assignment to "structural" (known structural proteins as described for Step 3 and in the aforementioned list), "other" (non-structural) and "no prediction" (e.g. no assignable function based on available annotations and a PfANNs confidence of <95% for its category of "other"). Step 9) All protein clusters (genes with the same mmseqID) were reviewed for consistency of annotation among member genes, and additional targeted annotation by HHpred was used to facilitate assignment to "structural" (known structural proteins as described for Step 3 and Step 8), "other" (non-structural), "no prediction" (e.g. no assignable function based on available annotations, a PfANNs confidence of <95% for its category of "other"). In cases where existing assignments of genes within the protein cluster contained both "no prediction" and "other" calls, the "no prediction" call prevailed where these represented more than ~30% of the calls across all genes in the cluster.

Annotation of viral potential for temperate lifestyle Overview. We identified 6 genera of phages as likely representing temperate phages (indicated in Fig. 2).

Bacterial genome read mapping. In order to evaluate the possibility that phages closely related to the Nahant Collection phages reside in the bacterial hosts in this study as prophages we used a read mapping approach. Briefly, reads from each of 276 bacterial genomes isolated from Nahant were mapped against each of the 248 phages (276 bacterial genomes x 248 phages) covering a total of bases 125,675,138 (7 phages, NCVICG-41), both genera of which show extensive recruitment of bacterial reads, covering 100 and 93% of their genomes, respectively. In addition to these 2 phages, 58 phages recruited bacterial genome reads covering up to 510 bases (range: 30–510 bases covered). Investigation of the genes to which these reads mapped revealed that the only one coding a gene for a functional integrase was in phage NVICG-43 (a group which includes 20 members). Where there was any mapping of reads the patterns were bimodal, with either extensive coverage (which we describe in Fig. 2 as strong evidence, in category A) or very limited coverage (which we note in Fig. 2 as weak evidence, in category B). Results of the mapping analyses are consistent with our other approaches for assessing potential for lysogeny in these phages. Con
phage), were identified as encoding integrases based on annotations with iterative jackhammer searches with PF00589 phage integrase seed alignment, and by EggNOG-Mapper and InterProScan annotation. The phages in NCVICG_23 and NCVICG_28 were identified as encoding PF00239 family integrases based on InterProScan annotation. The phages in NCVICG_41 (7 phages) were identified as encoding a Mu-like transposase, PF02914, on the basis of InterProScan annotations. Iterative jackhammer searches with N15 phage linear plasmid maintenance sequences (NCVC_227_4911) and E. coli ParA (AA002930.1) retrieved hits to any of the Nahant Collection phages. Finally, phages in 19 genera in the collection encode genes with transcriptional repressor domains represented by IPR010982, IPR010744, IPR001387, IPR032499.

**Viral genome clustering.** To understand how the diversity of viral genomes in the Nahant Collection is organized, we use the VICTOR classifier, which determines genome to genome distances between concatenated amino acid sequences of viral proteomes using the Genome-BLAST Distance Phylogeny method and clusters these using OPTSIL17 and criteria optimized by benchmarking to reference ICTV prokaryotic virus taxonomic units available at the time of the development of the tool12, with the fraction of links required for cluster fusion of 0.58. Average support values for the phylogenomic trees using the d0, d4, and d6 metrics were: kmer size of 16, fragment length of 100, minimum fraction of shorter hits to any of the phage genomes generously provided as a public resource by the laboratory of the Nahant Collection phages. Finally, phages in 19 genera in the collection encode genes with transcriptional repressor domains represented by IPR010982, IPR010744, IPR001387, IPR032499.

**Phage-host interaction features.** To characterize large scale features of the infection network we use the BiMat MatLab package22 as described in5,6,90. Modularity was quantified using the leading eigenvector method, with a Kernighan-Lin tuning step performed using overlap and decreasing fill (NODF). Statistical significance of modularity and nestedness was tested against 1000 random matrices generated using the equiprobable method, preserving the overall matrix connectivity. Modularity values were as follows: Qb value 0.7306, mean 0.4362, std 0.0047, z-score 63.2774, t-score 2001.0077, percentile 100; Qr value 0.9318, mean 0.1004, std 0.0219, z-score 37.9184, t-score 1199.0848, percentile 100. Nestedness values were as follows: Nestedness value: 0.0300, mean 0.0230, std 0.0025, z-score 14.0350, t-score 434.6833, percentile 100. The 248 phages included in the analysis (phageSet248 in Supplementary Data 1) are genome sequenced phages purified from the same plaque and excluding duplicate phages purified from different plaques on the same host; the 279 bacteria included in the analyses (bactSet248 in Supplementary Data 1) include all bacterial strains screened in the host range assay for which there was a positive interaction with a phage in phageSet248 (ie. host strains that were assayed but not killed by any phages were not included). These representations 1.43% of the 16,919 high-quality interactions yielding a connectance or fill of 0.021. To facilitate visual comparisons between the matrix of interactions between phages and bacteria with known species assignments (Fig. 2b) and the BiMat results, the representation of the BiMat analysis as shown in the main text (Fig.2a) includes only the subset of interactions between phages in phageSet248 and bacterial strains (bactSet248) that is the intersection of the 279 bacterial strains (bactSet248) included in the BiMat analysis and the 294 bacterial genomes (bactSet248) for which genomes were available. BiMat module assignments for all phages and hosts are provided in Supplementary Data 1 sheet C.

**Characterization of phage-host interaction features.** BiMat modularity analysis. To characterize large scale features of the infection network we use the BiMat MatLab package22 as described in5,6,90. Modularity was quantified using the leading eigenvector method, with a Kernighan-Lin tuning step performed using overlap and decreasing fill (NODF). Statistical significance of modularity and nestedness was tested against 1000 random matrices generated using the equiprobable method, preserving the overall matrix connectivity. Modularity values were as follows: Qb value 0.7306, mean 0.4362, std 0.0047, z-score 63.2774, t-score 2001.0077, percentile 100; Qr value 0.9318, mean 0.1004, std 0.0219, z-score 37.9184, t-score 1199.0848, percentile 100. Nestedness values were as follows: Nestedness value: 0.0300, mean 0.0230, std 0.0025, z-score 14.0350, t-score 434.6833, percentile 100. The 248 phages included in the analysis (phageSet248 in Supplementary Data 1) are genome sequenced phages purified from the same plaque and excluding duplicate phages purified from different plaques on the same host; the 279 bacteria included in the analyses (bactSet248 in Supplementary Data 1) include all bacterial strains screened in the host range assay for which there was a positive interaction with a phage in phageSet248 (ie. host strains that were assayed but not killed by any phages were not included). These representations 1.43% of the 16,919 high-quality interactions yielding a connectance or fill of 0.021. To facilitate visual comparisons between the matrix of interactions between phages and bacteria with known species assignments (Fig. 2b) and the BiMat results, the representation of the BiMat analysis as shown in the main text (Fig.2a) includes only the subset of interactions between phages in phageSet248 and bacterial strains (bactSet248) that is the intersection of the 279 bacterial strains (bactSet248) included in the BiMat analysis and the 294 bacterial genomes (bactSet248) for which genomes were available. BiMat module assignments for all phages and hosts are provided in Supplementary Data 1 sheet C.

**Average nucleotide identity.** FastANI v1.32 was used to determine average nucleotide identity (ANI) for phages and hosts as follows. For phages, run parameters were: kmer size of 16, fragment length of 100, minimum fraction of shorter genome coverage of 75% (fastANI -kmer 16 -fragLen 100 -minFraction 0.75 -matrix

---

**Host range.** Host range assay. Host range of viruses was determined as follows, and as also previously described1,2,9. Cell-free phage lysates were stamped onto host agar overlay laws and observed for changes in lawn morphology proximal to each stamp. Phage application to host lawns was performed using a 96-well blower (BelArt, Bel-blower 96-tip replicator 376760002) that was set into a microtiter plate containing arrayed phage lysates transferred to the surface of the host lawn, and allowed to remain in contact for several minutes. Each 96-stamp contained 3 replicates of each phage lysate, distributed across three panels (columns 1–4, 5–8, 9–12) each with a unique array of the 32 samples (including one negative control). 96-well blowers were microtomed to steam sterilized (Tomodex, Tippee. Closer to Nature Microwave Steam Sterilizer) and used in batches for continuing to contact the plating lawn. Bacillus strains were prepared for the infection assay by inoculating 1 ml volumes of 2216MB in 2 ml 96-well culture blocks directly from glycerol stocks and shaking them at RT for approximately 48 h. Agar overlays were prepared by transferring 250 ul aliquots of host culture to bottom agar plates (2216MB, 1% Bacto Agar, 5% glycerol; in 150 mm diameter Petri dishes) and adding 7 ml of top agar (1% Bacto Agar, 0.4% Bacto Album, 5% glycerol). Phages were prepared by distributing lysates into a 2 ml 96-well culture block in panels as described above, aliquots of <200 ul were then transferred into shallow microtiter plates so that the blower could phage lysate by capillary action. Host lawns were stamped with phage lysates within 5-6 h of plating laws. Agar overlay plates were assessed for changes in lawn morphology associated with phage treatment and scored blind with respect to phage identity and arrangement of replicates. Plates were scored for the presence of interactions on days 1, 2, 3, 7, 14, 21, and 30, and the outer bands of the interaction zones were marked with a different color for each time point. After 30 days the interactions for each strain were recorded and a connectance diameter (time to first interaction at each overlay) was recorded. During recording of the interactions for each plate an additional qualitative measure of confidence in the projected positive or negative call of the interaction was made. For example, where 2 of 3 replicates were positive for a phage on a lawn with no other positive interactions such an interaction would be called by the qualitative measure as “positive”, already, where 2 of 3 replicates were positive for a phage containing several other positive interactions the qualitative measure might call these replicates “contam” if they were high-titer interactions and occurred in close proximity to other positive interactions. Evaluation of changes in clearing sizes showed that the majority of clearings (>90%) increased in size over the course of the observation periods and with these replicates, the small size markers used for titration or killing from without. If any of the clearings that do not increase in size represent non-replicative interactions this would reduce the total true killing interactions further and underscore our finding of the general sparsity of interactions. A list of all infection pairs and information regarding plaque size and increases are provided in Supplementary Data 1 sheet C and represent interactions between phages identified in the Supplemental figure as phageSet248 and hosts identified as bactSet279.
Host range divergence analyses within VIC-species and VIC-genera. To quantify overlap in host range profiles between phages we develop a metric of host range divergence (represented as concordance (1-divergence) in the main text and Fig. 4). We normalize the binary vector \( x = (x_1, x_2, \ldots, x_m) \) representing the killing host range of a phage across all m hosts so that it sums to 1, and interpret the probability \( p_i = x_i / \sum x_i \) as a probability distribution of killing across all m hosts for a single phage. We then define the scaled host range divergence of a given genus consisting of \( n \) phages as the generalized Jensen-Shannon divergence (gJSD) of their infection probability vectors, \( D_g = \text{gJSD}(\{p_1, \ldots, p_n\}) / \log(n) \). This has the property that \( 0 \leq D_g \leq 1 \), where \( D_g = 0 \) means the host ranges of all phages overlap exactly, and \( D_g = 1 \) means none of the host ranges have any overlap with each other. We note that we took a conservative approach in determining genus-level concordances as presented in Fig. 3 by including all phages within each VIC-genus in the calculation of concordance rather than collapsing intra-species blooms, which are largely comprised of phages with overlapping host ranges. Because high overlap of host range within VIC-species increases the value of the overall genus-level concordance metric (because of the higher number of pairwise comparisons with high values of concordance within species), this approach will be affected by the evenness of species abundances within a genus and including all members yields up to 25% higher values of the concordance metric than when only including single representatives of each species in this dataset. Calculations of concordance are provided in Source Data Fig. 4 for VIC-genera using both the approach of using all members (Source Data Fig. 4 sheet B, with phages identified as phageSet248) and that of using only single representatives of each VIC-species (Source Data Fig. 4 sheet C, with phages identified as phageSet171); values for VIC-species are provided in Source Data Fig. 4 sheet A. A Welch’s t-test between the set of VIC-genus level concordances and the set of VIC-species level concordances yielded a p-value of 1.45e-07, suggesting that concordance differs significantly between the VIC-genus and VIC-species levels. Analyses and visualizations were performed in R 3.6.1 with the following packages: ggrepel 0.8.1, ape 5.3, combinat 0.0.8, Infoheiro 1.2.0, phylotools 0.4.0, igraph 1.2.4.1, igraph 2.0.0, cowplot 1.0.0, data.table 1.12.2, ggplot2 3.2.1, tidyverse 1.3.0, gtuee 2.0.4, and patchwork 1.1.1.

Characterization of sequence sharing

Homologous recombination within species and genera. To assess the extent of recombination between closely related viruses we used viral species and genera as operationally defined by VICTOR as the framework and estimated effective relative contribution of recombination over mutation (r/m) as follows. HomBlocks v.1.07 was used with default parameter settings to identify, extract, and trim conserved regions within genera based on alignments with progressiveMauve (build February 13, 2015)[24] and trimAl v.1.2.493. Phylogenetic relationships between sequences were predicted using IQ-TREE v.1.6.129[25]. ClonalFrameML v.1.1.269[26] was used to visualize evidence for recombination within species and genera. Estimates of relative contribution of recombination (r/m) were based on the formula \( r / m = R / \theta \delta + \delta R R \), where \( R / \theta \) is the ratio of recombination to mutation rate, \( \delta \) is the importa
Received: 10 August 2020; Accepted: 12 November 2021; 
Published online: 18 January 2022

References
1. Sheth, R. U., Cabral, V., Chen, S. P. & Wang, H. H. Manipulating bacterial communities by in situ microbiome engineering. Trends Genet. 32, 189–200 (2016).
2. Koskella, B. & Meaden, S. Understanding bacteriophage specificity. Nat. Rev. Microbiol. 17, 126–139 (2015).
3. Gilbert, J. A. et al. Defining seasonal marine microbial community dynamics. ISME J. 6, 298–308 (2012).
4. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanobacteria infecting the oceanic cyanobacterium Prochlorococcus. Nature 424, 1047–1051 (2003).
5. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33, 3396–3404 (2017).
6. Bin Jiang, H. et al. Taxonomic assignment of uncultivated prokaryotic viruses genome is enabled by gene-sharing networks. Nat. Biotechnol. 37, 632–639 (2019).
7. Mathieu, A. et al. Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages. Nat. Commun. 11, 378 (2020).
8. Preheim, S. P. et al. Metapopulation structure of bacterioplankton. Science 357, 1147–1150 (2017).
9. Swanstrom, M. & Adams, M. H. Agar layer method for production of high titre bacteriophage. Appl. Environ. Microbiol. 34, 375–385 (1981).
10. Martin-Platero, A. M. et al. High resolution time series reveals cohesive but heterogeneous bacterioplankton communities by in situ microbiome engineering. Microbiome 7, 806–823 (2019).
11. Thompson, J. R. et al. Genotypic diversity within a natural coastal bacterioplankton population. Science 307, 1311–1313 (2005).
12. Lopes, A., Tavares, P., Petit, M.-A., Guérois, R. & Zinn-Justin, S. Automated concentration technique for concentration of ocean viruses. MethodsX 5, 139–172 (2018).
13. Kupczok, A. et al. Rates of mutation and recombination in siphoviridae phage genome evolution over three decades. Mol. Biol. Evol. 35, 1147–1159 (2018).
14. Babenko, V. V. et al. Phages associated with horses provide new insights into the dominance of lateral gene transfer in virulent bacteriophages evolution in natural systems. 542787. https://doi.org/10.1101/542787.
15. Bernard, G., Greenfield, P., Ragan, M. A. & Chau, C. X. K-mer similarity, networks of microbial genomes, and taxonomic rank. mSystems 3, e00257–18 (2018).
16. Corzett, C. H. et al. Evolution of a vegetarian vibrio: metabolic specialization of Vibrio breogami to macroialgal substrates. J. Bacteriol. 190,0020-18. https://doi.org/10.1128/JB.00020-18 (2016).
17. Moebus, K. & Nattkemper, H. Bacteriophage sensitivity patterns among bacteria isolated from marine waters. Helgoländer Meeresunters 34, 375–385 (1981).
18. Flores, C. O., Valderle, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. ISME J. 7, 520–532 (2013).
19. Flores, C. O., Meyer, J. R., Valderle, S., Fatt, L. & Weitz, J. S. Statistical structure of host–phage interactions. Proc. Natl. Acad. Sci. USA 108, E288–E297 (2011).
20. Hunt, D. E. et al. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320, 1081–1085 (2008).
21. Preheim, S. P. et al. Metapopulation structure of Vibrioaceae among coastal marine invertebrates. Environ. Microbiol. 13, 265–275 (2011).
22. Szabo, G. et al. Reproducibility of Vibrioaceae population structure in coastal bacterioplankton. ISME J. 7, 509–519 (2013).
23. Arevalo, P., Vaninssberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A reverse ecology approach based on a biological definition of microbial populations. Cell 178, 830–834.e4 (2019).
24. Martin-Platero, A. M. et al. High resolution time series reveals cohesive but short-lived communities in coastal plankton. Nat. Commun. 9, 266 (2018).
25. Duhaime, M. B., Wichels, A., Waldmann, J., Teeling, H. & Glöckner, F. O. Ecogenomics and genome landscapes of marine Pseudoalteromonas phage H1051. ISME J. 5, 107–121 (2011).
26. Kaufman, K. M. et al. Viruses of the Nahant Collection, characterization of 251 marine Vibrioaceae viruses. Sci. Data 5, 180114 (2018).
27. Kaufman, K. M. et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature. https://doi.org/10.1038/s41586-018-0355-y (2018).
28. Dekel-Bird, N. P., Sabehi, G., Mosevitzyk, B. & Lindell, D. Host-dependent differences in abundance, composition and host range of cyanophages from the Red Sea. Environ. Microbiol. 17, 1268–1279 (2015).
29. Lopes, A., Amariz-Bouharm, J., Faure, G., Petit, M.-A. & Guerois, R. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res. 38, 3952–3962 (2010).
30. Martinsohn, J. T., Radman, M. & Petit, M.-A. The λ red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. PLOS Genet. 4, e1000605 (2008).
31. Kupczok, A. et al. Mutations and recombination in siphoviridae phage genome evolution over three decades. Mol. Biol. Evol. 35, 1147–1159 (2018).
32. Murphy, K. C. Use of bacteriophage λ recombination functions to promote gene replacement in Escherichia coli. J. Bacteriol. 180, 2063–2071 (1998).
33. Zhang, Y., Buchholz, F., Muyrers, J. P. & Stewart, A. F. A new logic for DNA engineering using recombination in Escherichia coli. Nat. Genet. 20, 123–128 (1998).
34. Copeland, N. G., Jenkins, N. A. & Court, D. L. Recombineering: a powerful new tool for mouse functional genomics. Nat. Rev. Genet. 2, 769–779 (2001).
35. Fislinger, G. T. et al. Characterizing the portability of phage-encoded homologous recombination proteins, Nat. Chem. Biol. 1–9 https://doi.org/10.1038/nchembio.2018-0357 (2021).
36. Sawitzke, J. A. et al. Probing cellular processes with oligo-mediated recombination and using the knowledge gained to optimize recombineering. J. Mol. Biol. 407, 45–59 (2011).
37. Zborowsky, S. & Lindell, D. Resistance in marine cyanobacteria differs against specialist and generalist cyanophages. Proc. Natl. Acad. Sci. USA 116, 16899–16908 (2019).
38. Maffei, E. et al. Systematic exploration of Escherichia coli phage-host interactions with the BASEL phage collection. PLoS Biol. https://doi.org/10.1101/2021.03.08.434281 (2021).
39. Mayer, O. et al. Fluorescent reporter D56A mycobacteriophages reveal unique variations in infectibility and phage production in mycobacteria. J. Bacteriol. 198, 3220–3223 (2016).
57. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pan-genome. *Science* **359**, eaar1120 (2018).

58. Berdahl, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).

59. Paepe, D. M. et al. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. *PLoS Genet.* **10**, e1004181 (2014).

60. Liu, J. M., Gole, P., Wegrzyn, A. & Loi, S. Simple method for plating Escherichia coli bacteriophages forming very small plaques or no plaques under standard conditions. *Appl. Environ. Microbiol.* **74**, 5113–5120 (2008).

61. Goh, S. H. et al. HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. *J. Clin. Microbiol.* **34**, 818–821 (1996).

62. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

63. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—a maximum-likelihood tree for large alignments. *PLoS One* **5**, e9490 (2010).

64. Hodcroft, E. PareTree 1.0: Remove Sequences, Bootstraps, and Branch Lengths From Your Tree! http://emmahodcroft.com/PareTree.html (2017).

65. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOl): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

66. Arevalo, P. Philarevalo/RiboTree (2017).

67. Pouls, B. T., John, S. G. & Sullivan, M. B. Iron Chloride Flocculation of Bacteriophages from Seawater. in Bacteriophages: Methods and Protocols (eds Clokie, M. R. J., Kroinski, A. M. & Lavigne, R.) 49–57 (Springer, 2018). https://doi.org/10.1007/978-1-4939-7343-9_4

68. Santos, S. B. et al. The use of antibiotics to improve phage detection and enumeration by the double-layer agar technique. *BMC Microbiol.* **9**, 148 (2009).

69. Stümpel, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

70. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

71. Huerta-Cepas, J. et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msx148 (2017).

72. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

73. Charoenkwan, P., Nantasenamat, C., Hasan, M. D. & Shoombuthawong, W. Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *J. Comput. Aided Mol. Des.* https://doi.org/10.1007/s10822-020-00323-z (2020).

74. Graziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus PareTree 1.0: Remove Sequences, Bootstraps, and Branch Lengths From Your Tree! http://emmahodcroft.com/PareTree.html (2017).

75. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sterberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 455–458 (2015).

76. Davidson, A. R., Cardarelli, L., Pell, L. G., Radford, D. R. & Maxwell, K. L. Birds in the RNA world? A pattern scanning tool with a new HHpred server at its core. *Bioinformatics* **31**, 142–148 (2015).

77. Suyama, M., Haji, S. & Baldi, P. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

78. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

79. Alkhamis, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).

80. Homer, N. nh13/DWGSIM. https://github.com/nh13/DWGSIM (2019).

81. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

Acknowledgements

We are grateful to Jan Meier-Kolthoff and Markus Göker, for running their VICTOR tool analysis on the Nahant Collection phage genomes on their cluster. We thank Michael Cutler, Philip Arevalo, and Joseph Elsberini for support in sequencing, assembly, and annotation of bacterial genomes. We thank David Varnslnberge for support in early phage genome analyses. We thank all members of the 2010 Polz lab for support with field sampling and bacterial isolation, and in particular Michael Cutler, Alison Takemura, Hong Xue, Tara Soni, and Gitta Szabo. We thank the Edward Ruby lab for sharing *Vibrio fischeri* strains for inclusion in the host range assay. This work was supported by grants from the National Science Foundation OCE 1435993 and 1435868 to M.P. and L.K., respectively, the Simons Foundation (LIFE ID 572792) to M.P., the NSF GRFP to F.H., and the WHOI Ocean Ventures Fund to K.K. I.Y. was supported by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308.

Author contributions

K.R. and W.C. designed the study, performed analyses, and wrote the manuscript with substantial contributions throughout from J.B., F.H., J.Y., M.P. and L.K.

Competing interests

The authors declare no competing interests.

Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27583-z.

**Correspondence** and requests for materials should be addressed to Martin F. Polz or Lisabel Basha.

**Peer review information** *Nature Communications* thanks Nathan Ahlgren, Karin Holmblad and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
