Twelve previously unknown phage genera are ubiquitous in global oceans

Karin Holmfeldt[a,1,2], Natalie Solonenko[a], Manesh Shah[b], Kristen Corrier[b], Lasse Riemann[c], Nathan C. VerBerkmoes[b,3], and Matthew B. Sullivan[a,1]

[a]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [b]Chemical Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; and [c]Department of Biology, University of Copenhagen, 3000 Helsingor, Denmark

Edited by James L. Van Etten, University of Nebraska–Lincoln, Lincoln, NE, and approved June 17, 2013 (received for review April 2, 2013)

Viruses are fundamental to ecosystems ranging from oceans to humans, yet our ability to study them is bottlenecked by the lack of ecologically relevant isolates, resulting in "unknowns" dominating culture-independent surveys. Here we present genomes from 31 phages infecting multiple strains of the aquatic bacterium *Cellulophaga baltica* (*Bacteroidetes*) to provide data for an underrepresented and environmentally abundant bacterial lineage. Comparative genomics delineated 12 phage groups that (*i*) each represent a new genus, and (*ii*) represent one novel and four well-known viral families. This diversity contrasts the few well-studied marine phage systems, but parallels the diversity of phages infecting human-associated bacteria. Although all 12 *Cellulophaga* phages represent new genera, the podoviruses and icosahedral, nontailed ssDNA phages were exceptional, with genomes up to twice as large as those previously observed for each phage type. Structural novelty was also substantial, requiring experimental phage proteomics to identify 83% of the structural proteins. The presence of uncommon nucleotide metabolism genes in four genera likely underscores the importance of scavenging nutrient-rich molecules as previously seen for phages in marine environments. Metagenomic recruitment analyses suggest that these particular *Cellulophaga* phages are rare and may represent a first glimpse into the phage side of the rare biosphere. However, these analyses also revealed that these phage genera are widespread, occurring in 94% of 137 investigated metagenomes. Together, this diverse and novel collection of phages identifies a small but ubiquitous fraction of unknown marine viral diversity and provides numerous environmentally relevant phage–host systems for experimental hypothesis testing.

model systems | phage genomics | phage taxonomy | queuosine biosynthesis | prophage

Microbes are the main drivers of the planet's biogeochemical cycles (1), and their viruses (phages) play important roles, ranging from mortality and nutrient cycling to gene transfer (reviewed in ref. 2). However, our knowledge of phage biology, ecology, and evolution is biased toward phages that infect only a few hosts. For example, 85% of 1,100 sequenced phage genomes in GenBank are isolated by using bacteria from only three of 45 known bacterial phyla (*Actinobacteria*, *Firmicutes*, and *Proteobacteria* of the class Gammaproteobacteria), predominantly involved in human diseases and food processing. In contrast, with the exception of phages infecting cyanobacteria (cyanophages), phages that infect environmental microbes are largely unstudied and unknown. This lack of genomic representation results in unidentified DNA sequences accounting for ∼90% of the sequences in nearly any viral metagenome, even when using technologies that provide longer read lengths (3), which hinders inference power in viral ecology.

One bacterial phylum whose phages are underexplored is *Bacteroidetes*, which is currently represented by the genomes of only six phages in GenBank, isolated from both aquatic and human-related environments (e.g., refs. 4, 5). *Bacteroidetes* bacteria are abundant and active members of bacterial communities in various habitats ranging from Antarctic soil (6) to surface (7) and deep (8) oceans and even the human gut. In humans, *Bacteroidetes* comprise 30% of the gut microbiota and play important roles for fat storage (9) and the immune system (10). In the oceans, *Bacteroidetes* is the third most abundant bacterial phylum (7, 8), and there these bacteria are active in degrading biopolymers (11) and involved in recycling of phytoplankton bloom-related organic matter (12).

Here we present 31 genomes and 13 representative structural proteomes of *Bacteroidetes* phages isolated by using 17 *Cellulophaga* host strains. This genomic and proteomic information helps define *Cellulophaga* phage taxonomy, diversity, and functional potential, and metagenomic comparisons elucidate their distribution in natural aquatic systems.

Results and Discussion

Cellulophaga Phages Represent at Least 12 Novel Phage Genera. Genomes were sequenced from 31 phages isolated from the strait of Öresund, between Sweden and Denmark, using 17 closely related *Cellulophaga baltica* host strains [99.5–100% 16S rRNA gene identity (13)]. The genomes ranged in size from 6.5 to 145 kb with a G+C content of 29% to 38% (summarized in Table 1 and detailed in Table S1). As is common for environmental phages (e.g., refs. 4, 14), few (average, 39%; range, 23–53%) predicted ORFs matched anything in databases, with only 3% to 39% (average, 22%) functionally annotated beyond "hypothetical" protein (Table 1 and Table S1; full annotation details in Dataset S1). Few structural proteins could be identified based on sequence data alone: 83% of the 192 proteins identified through virion structural MS-based proteomics lacked any sequence-based similarity to known structural proteins (Dataset S1). Per genome, this allowed 8 to 27 ORFs to be annotated as "structural," compared with zero to nine ORFs based on sequence similarity alone (Dataset S1). This clearly delineated each genome's structural module (example in Fig. S1), as observed for other environmental phages (e.g., ref. 4). MS also identified that proteins with lytic activity (based on sequence similarity) were present in the structural particle of seven of the investigated phages (Dataset S1), whose function could be to aid penetration of the bacterial cell wall upon entry (e.g., ref. 15).

Comparative genomics delineated 12 groups (Fig. 1*A*), which, using current taxonomic metrics whereby phages within a genus share 40% of their genes (16), represent 12 new viral genera with >40% of the genes shared within genera and 0% to 18% of the

Author contributions: K.H., L.R., and M.B.S. designed research; K.H., N.S., and K.C. performed research; M.S. and N.C.V. contributed new reagents/analytic tools; K.H. and M.S. analyzed data; and K.H., L.R., N.C.V., and M.B.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. KC821604–KC821634).

[1]To whom correspondence may be addressed. E-mail: k.holmfeldt@gmail.com or mbsulli@e-mail.arizona.edu.

[2]Present address: School of Natural Sciences, Linnaeus University, 39182 Kalmar, Sweden.

[3]Present address: Chemical Biology Division, New England Biolabs, Ipswich, MA 01938.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1305956110/-/DCSupplemental.

Table 1. General features of *Cellulophaga* phage genera

Group	Type phage	Family	Putative genus	DNA type	Genome size, kb	G+C, %	ORFs	tRNA	ORFs with hits to databases, %	ORFs with function, %	ORFs with proteomic data
1	ϕ40:1	*Podoviridae*	Cba401likevirus	dsDNA	72.5	38	101	16	35	15	14
2	ϕ18:3	*Podoviridae*	Cba183likevirus	dsDNA	73.2	33	125	1	41	14	12
3	ϕ14:2	*Podoviridae*	Cba142likevirus	dsDNA	100.4	30	133	—	26	15	19
4	ϕ4:1	*Podoviridae*	Cba41likevirus	dsDNA	145.7	33	198	24	32	14	27
5	ϕS_M	*Myoviridae*	CbaSMlikevirus	dsDNA	54.2	33	80	—	41	29	22
6	ϕ39:1	*Siphoviridae*	Cba391likevirus	dsDNA	28.8	31	49	—	43	22	14
7	ϕ46:1	*Siphoviridae*	Cba461likevirus	dsDNA	34.8	38	54	—	52	28	8
8	ϕ18:1	*Siphoviridae*	Cba181likevirus	dsDNA	38.9	37	64	—	52	37	10
9	ϕ10:1	*Siphoviridae*	Cba101likevirus	dsDNA	55.6	31	112	3	33	16	12
10	ϕ13:1	*Siphoviridae*	Cba131likevirus	dsDNA	77.8	30	107	—	40	24	15
11	ϕ18:4	*Microviridae*	Cba184likevirus	ssDNA	6.5	34	13	—	23	15	10
12	ϕ48:2	Novel	Cba482likevirus	ssDNA	11.5	29	30	—	27	3	17

Each row represents the average data for each genus. Group numbers refer to numbers used for each genus in Figs. 1 and 5.

genes shared between genera. The genera are named following International Committee on Taxonomy of Viruses conventions that use the name of the host bacterium (Cba for *C. baltica*) and type phage, where we remove the ":" from the type phage name (Table 1 and Table S1). Although most genera shared >65% of their genes, the two phages in genus Cba101 were more divergent, sharing only 41% of their genes and at lower percent identity (averages of 97% vs. 82% aa identity, respectively), and are tracked specifically here as Cba101a (ϕ10:1) and Cba101b (ϕ19:1; Fig. S1*B*).

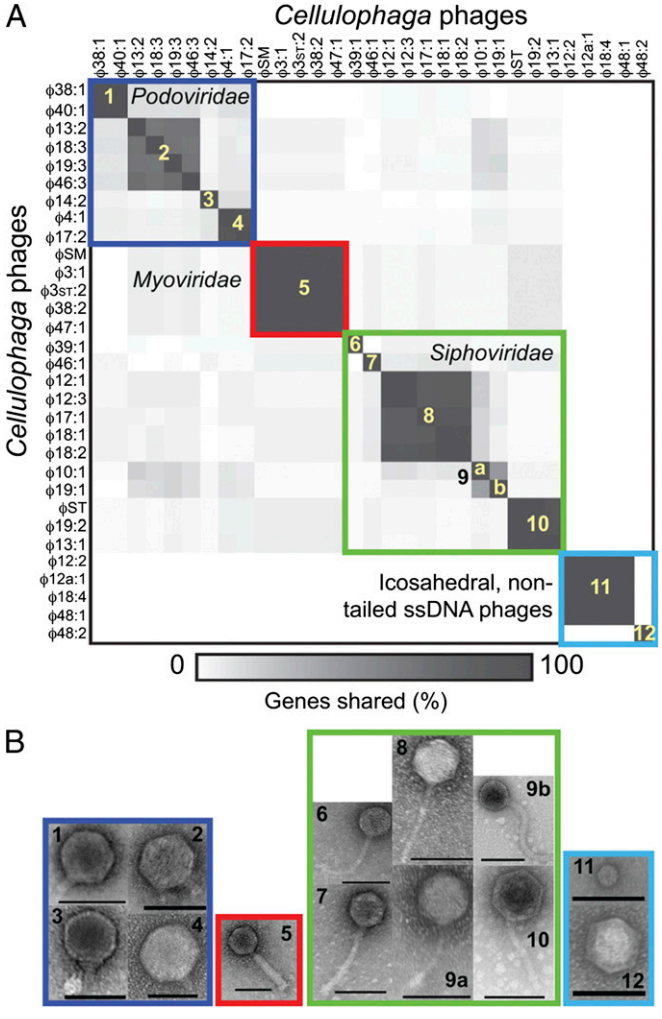

Fig. 1. (*A*) Heat map showing percentage of shared genes between the 31 *Cellulophaga* phages. Numbers in the boxes indicate the 12 genera delineated by this genome comparison. (*B*) EM images of representative phages from each genus (affiliation designated by number in the micrograph) delineated from gene comparisons. (Scale bars: 100 nm.)

Morphologically, these 12 genera derive from at least four viral families according to current International Committee on Taxonomy of Viruses taxonomy (17): 10 belonged to dsDNA tailed phage families [order Caudovirales, families Myoviridae ($n = 1$ genus), Podoviridae ($n = 4$ genera), and Siphoviridae ($n = 5$ genera); Fig. 1*B*], whereas two were previously described as nontailed ssDNA phages (18).

High Diversity and Novelty: Breaking Marine Paradigms. *Cellulophaga* phage diversity starkly contrasts the only other extensively sequenced aquatic phage collections, but parallels collections of phages that infect heterotrophic bacteria (Fig. 2). The marine systems, *Prochlorococcus* and *Synechococcus* cyanophages, represent fewer ($n = 4$ and $n = 5$, respectively) groups per host genus (Fig. 2*A*), despite more diverse hosts and water samples used to isolate cyanophages compared with *Cellulophaga* phages. In contrast, the nonmarine *Escherichia coli* and *Pseudomonas aeruginosa* phages were as diverse as the *Cellulophaga* phages (Fig. 2 *B* and *C*). Curiously, the nonmarine *Mycobacterium* phages also show large genomic diversity [15 clusters and seven singletons (19)], but more limited morphological diversity [*Siphoviridae and Myoviridae* (20)]. Although no genomes were available, similarly high diversity was also reported among 22 marine *Pseudoalteromonas* phages isolated from the North Sea, where morphology, DNA hybridization, and host range analysis delineated 13 groups (21). One possible explanation for the apparent reduced diversity in cyanophages might be that K-strategist hosts (e.g., cyanobacteria) have relatively invariant host physiology compared with r-strategists [e.g., *C. baltica* and *E. coli* (22)]. This could reduce phage diversity via fewer niche opportunities on the former and increase diversity via host-state–dependent phage genera cycling on the latter. Notably, some marine phages that infect heterotrophic bacteria appear less diverse, but this could be an artifact, as these phage collections suffer from a paucity of isolates (23); the use of broad, nonquantitative metrics for delineating groupings [60 phages for *Vibrio parahaemolyticus* (24)]; or ascertainment bias [e.g., selection in isolation procedures to recover near-identical roseophages (25)].

Evolutionarily, signatures of horizontal gene transfer were observed in entero- and cyanophages. Among enterophages, phages of different morphotypes (podo- and siphoviruses) shared a large number of genes (Fig. 2*B*), which complicates phage taxonomy (e.g., refs. 16, 26). Cyanophages, on the contrary, shared

MICROBIOLOGY

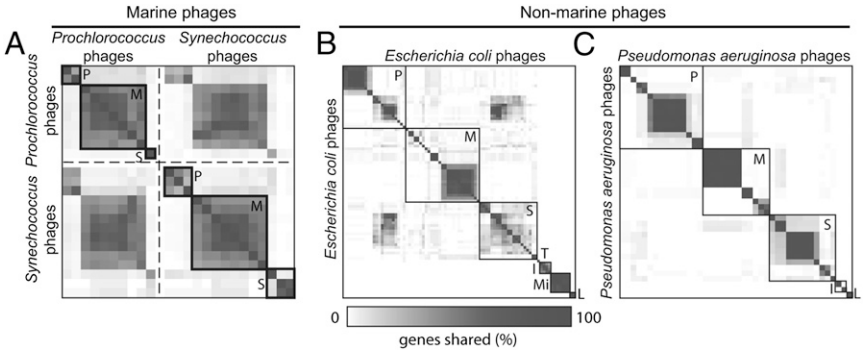

Fig. 2. Heat maps showing percentage of shared genes between phages infecting the same bacterial host species. (*A*) Cyanophages isolated on *Prochlorococcus* or *Synechococcus*, (*B*) *E. coli* phages, and (*C*) *P. aeruginosa* phages. Dashed lines separate *Prochlorococcus* and *Synechococcus* phages. Squares enclose phages belonging to the same phage family: I, *Inoviridae*; L, *Leviviridae*; M, *Myoviridae*; Mi, *Microviridae*; P, *Podoviridae*; S, *Siphoviridae*; T, *Tectiviridae*. Dark areas (large proportion of genes shared) outside of family squares indicate putatively horizontally transferred genes.

genes between phages that were isolated by using different host genera (Fig. 2*A*). For example, cyanomyoviruses commonly share half of their genes whether isolated within one genus or from two genera. Indeed, exchange of genes, even core genes, between cyanophages that infect across generic boundaries is well known (e.g., ref. 27). However, no such signature, either sharing >50% of their genes between phages of different families or between phages infecting different hosts, was observed for *Cellulophaga* phages.

Marine cyanophages and roseophages also share genes with nonmarine coliphages (e.g., refs. 28, 29), and again *Cellulophaga* phages do not. Marine cyanophages share 16 (podoviruses) and 38 (myoviruses) genes with nonmarine enterophages (28), and roseophage DSS3ϕ2 shares 26 genes with enterophage N4 (29). In contrast, most *Cellulophaga* phages share few genes (average, n = 3; range, n = 1–6) with any non-*Cellulophaga* phage. The exceptions are *Cellulophaga* myovirus genus CbaSM (14 shared genes with *Vibrio* phage VP16T, GenBank accession no. AY328852; Fig. 1*A*) and siphovirus genera Cba181 and Cba101a (10 and 18 genes shared, respectively, with *Flavobacterium* phage 11b, GenBank accession no. NC_006356; Fig. 1*B*). These are both phages of aquatic origin (4, 14) and the host of the latter (*Flavobacterium* phage 11b) belongs to the same family as *C. baltica*. In all cases, these shared genes are highly divergent (26–40% aa identity). This highlights that, even though a few *Cellulophaga* phages share genes with other phages, they are still exceptionally different from known phages, even when comparing the siphoviruses, which represent the bulk (57% per GenBank, accessed December 2012) of sequenced *Caudovirales* phages.

Phage Giants: Exceptionally Large Cellulophaga Podo- and Nontailed Phages. Six *Cellulophaga* phage genera (the myo- and siphoviruses) had genome sizes within the known range, whereas the genome sizes of the other six genera (the podoviruses and nontailed icosahedral phages) were quite different from known representatives (Fig. 3).

Podoviruses. The four *Cellulophaga* podovirus genera defined through sequence similarity fell into three groups by genome size: 71 to 76 kb (genera Cba401 and Cba183), 100 kb (genus Cba142), and 145 to 146 kb (genus Cba41). Considering that only 7% of previously sequenced podovirus genomes are >70 kb and the largest was only 79 kb (Fig. 3), the *Cellulophaga* podoviruses are clearly exceptionally large. Not surprisingly, their genomes differed from known podoviruses: only 4% to 10% of genes were similar to podoviruses and, instead, often, more genes were similar to those from sipho- and myoviruses—although, in all cases with relatively low similarities (averaging 32–33% aa identity).

Numerous other traits highlight the novelty of the *Cellulophaga* podoviruses. First, although tRNAs occur in ∼20% of previously sequenced podovirus genomes (GenBank, accessed June 2012), their frequency is one to three per genome, which contrasts the 16 to 24 observed for genera Cba401 and Cba41 *Cellulophaga* podovirus genomes (Table 1 and Table S1). Such phage-encoded tRNA abundances are more characteristic of myoviruses, for which as many as 33 tRNAs are used to expand the phage's codon use capabilities during infection of diverse hosts (30). Consistent with this, the *Cellulophaga* podoviruses with several tRNAs (genera Cba401 and Cba41) infect 9 to 15 *Cellulophaga* strains, whereas those with fewer tRNAs (genera Cba183 and Cba142, 0–1 tRNA) infect only one to four *Cellulophaga* strains (13). Second, the protein-folding genes chaperonin GroEL or chaperonin Cpn10 occur in all four *Cellulophaga* podovirus genera (Dataset S1). Such genes (GroEL and Cpn10) have previously been reported in myoviruses (four GroEL and 56 Cpn10 of 284 investigated myoviruses) but are lacking in siphoviruses and podoviruses (GenBank, accessed December 2012). Chaperonins are critical for protein folding, perhaps here involved in folding of phage capsid proteins (31). Why these are shared among all these *Cellulophaga* podovirus genera is unknown, but may reflect larger-genome podoviruses requiring larger and possibly more complex capsid structures. This would be in agreement with chaperonin-containing myoviruses, all 60 of which have genome sizes >100 kb (GenBank, accessed December 2012). Finally, relatively unique combinations of nucleotide metabolism genes occur in the large podovirus genera Cba142 and Cba41 (details provided later).

Nontailed ssDNA phages. The nontailed, ssDNA *Cellulophaga* phages (18) are either slightly larger (genus Cba184 exceeds the largest *Microviridae* genome by 300 bp) or nearly double the size of known nontailed, icosahedral ssDNA phages (11.7 kb, genus

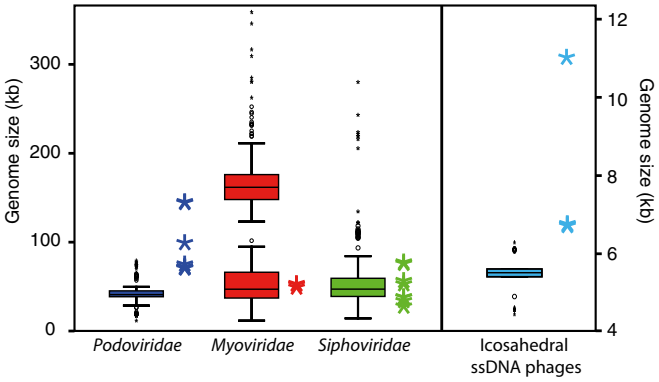

Fig. 3. Genome size comparison of the *Cellulophaga* phages (colored asterisks) to genome-sequenced phages available in GenBank (accessed December 2012; box plots). Phages within the family *Myoviridae* have been divided into two groups (larger and smaller than 100 kb) in view of the large range of genome sizes. [Note: a 498-kb myovirus, *Bacillus* phage G (JN63751), was not included here to minimize white space in the figure.] The box plot of icosahedral ssDNA phages represents phages belonging to *Microviridae*, the only known nontailed, icosahedral ssDNA phage family. The box represents the lower and upper quartiles with the median marked. The whiskers present 1.5 interquartile range (IQR) from the lower and upper quartiles, respectively; circles are outliers (1.5–3 IQR from the end of the box) and black asterisks are extremes (>3 IQR from the end of the box).

Cba482; Fig. 3). For phages of genus Cba184, none of their 14 predicted genes were similar to *Microviridae* phage genes or *Microviridae*-like genes found integrated in *Bacteroidetes* genomes (32), but functional parallels emerged as follows. First, the largest gene (gp 2) is a DNA replication protein (Dataset S1), as in the ssDNA model phage ϕX174 [the A protein (33)]. Second, the gene gp 8 is in the viral particle (Dataset S1) and annotated as a mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase. This is likely an analogue to the ϕX174 pilot protein (H) (33) and adapted to recognize and penetrate the polysaccharides present in the *C. baltica* host cell wall (34). Third, the second largest gene (gp 4) garnered the most spectral counts in proteomic analyses (Dataset S1), which parallels the ϕX174 major capsid protein (F) (33). Thus, we posit that, despite being half the size of F, gp 4 is the major capsid protein for the genus Cba184 *Cellulophaga* phages. Together, our data suggest that these phages represent a divergent subfamily within *Microviridae*.

In contrast, the genus Cba482 phage represents an entirely new phage family. At 11.7 kb, its genome is the largest among known ssDNA bacteriophages and icosahedral ssDNA viruses but is only half the genome size of the largest known ssDNA virus—a rod-shaped archaeal virus, ACV, with a 25 kb genome (35). Functionally, one gene (gp 19) encoded a peptidase (Dataset S1), whereas another 23% were similar to nonviral database sequences. Among these, four were exclusively similar to *Zunongwangia profunda* (phylum *Bacteroidetes*) genes, and, along with a highly divergent but syntenic peptidase, a remnant prophage could be identified in this microbial genome (Fig. 4). Together with turbid plaque morphology (Fig. S2), this suggests that the genus Cba482 phage is temperate (i.e., capable of lysogeny). Although common in filamentous ssDNA phages (e.g., ref. 36), lysogeny is rare in icosahedral ssDNA phages. The only evidence available is that *Microviridae*-like phage genomes occur in bacterial genomes [*Chlamydia* (37) and *Bacteroidetes* (32)], and a nontailed ssDNA phage has been induced from *Synechococcus* cultures [no phage genomes available (38)]. Interestingly, the latter has a capsid size similar to the genus Cba482 phage.

Although nontailed phages are thought uncommon as they are underrepresented in culture collections [e.g., <4% are nontailed among 5,500 isolated phages (39)], recent work shows that they dominate marine viral communities in the global surface oceans (40). Thus, both nontailed phage genera described here offer windows into a dominant marine phage type as they represent the first in culture infecting a *Bacteroidetes* bacterium, and join only four other marine nontailed phages in culture [two infecting *Synechococcus* (38, 41), one infecting *Pseudoalteromonas* (only sequenced; ref. 42), and one infecting a host of unknown taxonomy (43)].

Unusual Phage-Encoded Nucleotide Metabolisms. Three *Cellulophaga* phage genera (large podovirus genera Cba142 and Cba41 and large siphovirus genus Cba131) contained genes for de novo nucleotide synthesis including ribonucleoside-diphosphate reductase (RNR) and thymidylate synthase or thymidylate synthase complementing protein (Thy; Tables S2 and S3). These genes are common among myoviruses (>50% of 284 genomes), but less common in siphoviruses (70 [RNR] and 86 [Thy] of 625 genomes; Table S2) and podoviruses (only 12 of 186 genomes have either; GenBank, accessed December 2012; Table S3). Siphoviruses have any of three classes of RNRs, whereas published podoviruses are restricted to class II (11 roseo- and cyanophages) or III (1 *Cronobacter* phage) RNRs. All *Cellulophaga* phage RNRs are class I, regardless of morphology. Previous phylogenetic work showed that podovirus-encoded RNR type reflects that of the host, whereas myovirus-encoded RNRs might not (44). Our data, along with additional RNR-containing roseo- and cyanopodovirus and host RNR data obtained from GenBank, support this hypothesis (Table S3). Notably, only marine (roseo- or cyanophages) or large (N4- or phiEco32-like) phages have RNR or Thy genes (Table S4). This suggests that smaller genome phages gain little from the ability to convert available RNA pools into DNA for phage replication (45) except in predominantly phosphorous-limited marine environments.

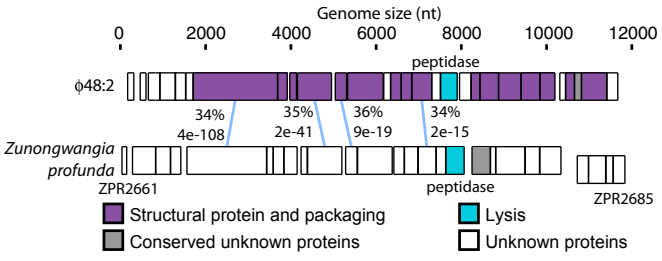

Fig. 4. Comparison of *Cellulophaga* phage ϕ48:2 to the region in *Z. profunda* (phylum *Bacteroidetes*) possibly containing a temperate phage (ORF 2,661–2,685). Lines drawn between the genomes represent shared sequence similarity, which is given next to each line as percentage amino acid identity and e-value.

Besides genes for RNR and Thy, all *Cellulophaga* phages in genera CbaSM (myovirus) and Cba131 (siphovirus) have queuosine (Que) biosynthesis genes (Dataset S1). Que is a hypermodified guanosine derivative in tRNAs specific for four amino acids (Asp, Asn, His, or Tyr) that increase translation efficiency and is found across all domains of life (46). Que de novo synthesis in prokaryotes requires five genes in the bacteria preQ1 pathway (46), of which five (genus CbaSM) and three (genus Cba131) are present in the *Cellulophaga* phages (Table 2 and Fig. S3). However, Que is uncommon among phages: (*i*) only 16 phages have Que genes similar to those in *Cellulophaga* (Table 2), and (*ii*) only one phage, *Streptococcus* phage Dp-1, has a larger number of Que genes [also involved in Que insertion (47)]. Analyses of viral metagenomes, however, suggest that virus-encoded Que genes are broadly distributed in aquatic environments [occurring in 55 of 137 viral metagenomes available as Broad Phage Metagenomes at Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA); Table 2], indicating that Que might be more important for aquatic phages than for phages in other environments.

Global Distribution of the Cellulophaga Phage Types. Given *Cellulophaga* phage novelty, their distributions were investigated using recruitment to 137 Broad Phage Metagenomes. Reads were recruited from 17 to 115 (average, 87) metagenomes per phage (details in Fig. 5) with representation from environments ranging from freshwater to marine, coastal to open ocean, and surface water to deep sediments. Curiously, 35% to 52% of genes from the tailed, dsDNA *Cellulophaga* phages recruited metagenomic reads, whereas only 11% to 14% of the nontailed, ssDNA phage genes did (Fig. 5 and Dataset S1). We posit that nontailed phage recruitment is repressed as a result of artifacts: (*i*) nontailed phages are commonly lost during CsCl purification as their low buoyant density differs from tailed phages (48), and (*ii*) DNA extraction methods are not optimized for the *Cellulophaga* nontailed phages (*Methods*) (18). Consistent with this hypothesis, a metagenome that targeted ssDNA phages (CAM_SMPL_000841; available on CAMERA) was responsible for most (58%) of the recruited ssDNA, nontailed phage reads.

The types of genes that recruited reads and the quality of recruitment were used to better interpret their meaning. First, reads recruited to conserved proteins (e.g., DNA replication or nucleotide metabolism) as well as phage-group–specific, experimentally identified structural proteins and predicted ORFs of unknown function. Although the former may be spurious recruits, indicative of conserved proteins found across many phage groups, the latter are likely bona fide recruits because phage structural proteins rarely share sequence similarity across groups (49). Second, on average, many of the recruited reads were exclusive to the *Cellulophaga* phages (average, 15%) or had higher bitscore to the *Cellulophaga* phages (average, 14%) than to anything in GenBank (details in Fig. 5 and Dataset S1). Third,

Table 2. Summary of Que biosynthesis genes occurring in all phages in genera CbaSM (g5) and Cba131 (g10)

	Phage homologues (CAMERA)		Recruited metagenome reads of phage origin, %		Metagenomes containing phage recruits	
Function	g5	g10	g5	g10	g5	g10
GTP cyc I*	16	0	0.3	84	1	32
QueD	1	0	46	90	33	23
QueC	16	NA	3	NA	6	NA
QueE	16	0	8	88	14	34
QueF	3	NA	0.9	NA	1	NA

The table includes data concerning phage homologues to the *Cellulophaga* phages Que genes in CAMERA and recruitment of metagenome reads from Broad Phage Metagenome to the *Cellulophaga* phages Que genes. cyc, cyclohydrolase; NA, not applicable, i.e., gene did not occur in genome.

sensitivity analysis with metagenomic recruitment to the T4-like phages suggested that these *Cellulophaga* phage genera are indeed ubiquitous (Fig. 5). Here, cyanophage P-SSM4, described as abundant in the oceans (50), recruited reads to 84% of its genes, averaging 57% aa identity. In contrast, reads recruited to only 38% of *Vibrio* phage KVP40 and *Enterobacteria* phage T4 genes and at a lower average percent identity (42% aa identity), which was similar to that observed for *Cellulophaga* phages. Many (51% for KVP40 and 68% for T4) of these recruited reads did so redundantly across the T4 phages, likely because of nonspecific recruitment to core T4 genes (Table S4), where the cyanobacterial T4 phage (P-SSM4) was likely the most representative reference genome for these metagenomes. Given such hierarchical recruitment results, we propose that, although the particular *Cellulophaga* phages investigated here are not abundant in these metagenomes, the phage genera they represent are ubiquitous, as they occur in 94% of 137 investigated aquatic viral metagenomes. However, caution is warranted when making quantitative statements from available metagenomic datasets (reviewed in ref. 51). Future work with the use of emerging viral metagenomic methods (3, 51, 52) to generate quantitative datasets (e.g., ref. 3) will enable better quantification of these and other new viral genera in the global oceans. Given these caveats, however, the novel *Cellulophaga* phages presented here help identify a small (average maximum of 0.04%; details in Fig. 5) but ubiquitous portion of the vast unknown sequence space that dominates (~90%) (3) marine viral metagenomes.

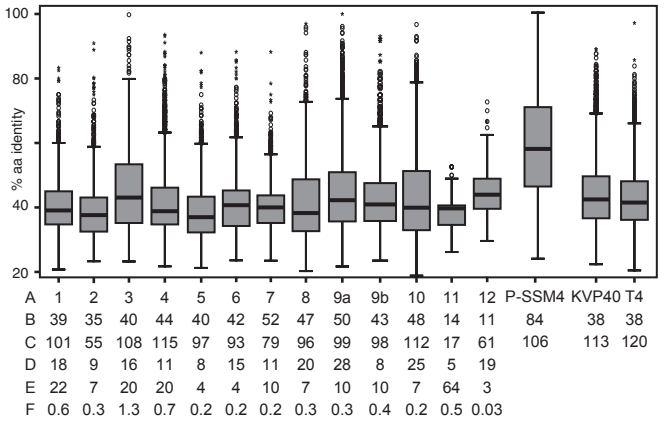

Fig. 5. Box plots show the percent amino acid identity for metagenomic reads (all 137 metagenomes available at CAMERA, Broad Phage Metagenome, January 2012) recruiting to predicted genes from *C. baltica* phages (designated as genera 1–12), as well as three T4-like phages: marine *Prochlorococcus* phage P-SSM4 (GenBank accession no. NC_006884), marine *Vibrio* phage KVP40 (GenBank accession no. NC_005083), and nonmarine *Enterobacteria* phage T4 (GenBank accession no. NC_000866). (*A*) Cellulophaga phage group (Table 1) or T4-like phage isolate. (*B*) Gene products to which the metagenomic reads were recruited (as percentages). (*C*) Number of metagenomes from which the recruited reads originated. (*D*) Reads with higher bitscore to a *Cellulophaga* phage than NCBI (as percentages). (*E*) Reads exclusively recruiting to a *Cellulophaga* phage (as percentages). (*F*) Maximum proportion of novel reads identified in a single metagenome (as per mils). The box represents the lower and upper quartiles with the median marked. The whiskers present 1.5 IQR from the lower and upper quartiles, respectively; circles are outliers (1.5–3 IQR from the end of the box) and asterisks are extremes (>3 IQR from the end of the box).

Conclusions

The 31 *Cellulophaga* phages represent 12 novel genera and together comprise the largest collection of genome-sequenced aquatic phages that infect a single host species. Their novelty, diversity, and ubiquity in the oceans are striking and warrant future structural, ecological, and evolutionary investigations. Although less abundant than the marine pelagi-, roseo-, or cyanophages (23), *Cellulophaga* phages likely present a first glimpse into the phage side of the "rare biosphere" (reviewed in ref. 53). Although not very abundant, such rare biosphere bacteria are thought to impact nutrient cycling during blooms conditions, with low abundances leading to cryptic escape from virus infections (53). Given the large diversity of *Cellulophaga* phages, bloom events must be frequent, with viral and microbial population structure likely shaped by microscale heterogeneity (54). These novel phage genera may also be critical in other environments in which *Bacteroidetes* phylum hosts are abundant, including the human gut (9). Mechanistic and discovery-based exploration of these new viral types will help elucidate yet another aspect of the large virus–host diversity that is increasingly being recognized as important in natural and manmade ecosystems.

Methods

Bacterial strains and phages were isolated as described previously (13). DNA for dsDNA phages was extracted by using the Wizard Genomic DNA Purification Kit (Promega) per the manufacturer's recommendations. DNA for ssDNA phages was extracted as a dsDNA replicative intermediate during infection. Phages were sequenced by using a combination of 454, Illumina, and Ion Torrent sequencing, and closed as needed by using Sanger sequencing. Because ssDNA phage genomes were nearly completely novel, they were also fully Sanger-sequenced by using CsCl-purified viral particles as PCR template. ORFs were predicted by using GeneMark, then annotated by using BLASTP against the National Center for Biotechnology Information (NCBI) nonredundant (as of November 2012), Conserved Domain, and Pfam databases (e-value cutoff $<1 \times 10^{-3}$); tRNAs were identified by using tRNAscan-SE. Percentage of shared genes between the *Cellulophaga* phages was calculated from BLASTP comparison (e-value cutoff $<1 \times 10^{-3}$). Metagenomic reads were recruited to predicted ORFs by using the TBLASTN workflow (cutoffs of e-value $<1 \times 10^{-3}$, >20% aa identity, alignment length >45 nt, and 300 alignments per query) in CAMERA against Broad Phage Metagenomes (https://portal.camera.calit2.net/; January 2012). Recruited reads were compared with NCBI nonredundant (e-value cutoff $<1 \times 10^{-3}$) to

calculate novelty of *Cellulophaga* phage contributions to each metagenome. Transmission EM was conducted as described previously (18). For proteomics, CsCl-purified viral particles were tryptically digested by using an optimized Filter-Aided Sample Preparation kit protocol (Protein Discovery; now Expedion) (55) and analyzed via 2D nano-LC-MS/MS (56). Resultant MS/MS spectra were searched against a compiled viral predicted protein database with SEQUEST and conservatively filtered with DTASelect (56). For proteomics, databases, peptide and protein results, MS/MS spectra and Tables S1–S4 are archived and available at https://compbio.ornl.gov/Cellulophaga_phages_proteome. MS .raw files or other extracted formats are available upon request. Phage genome GenBank accession numbers are KC821604 to KC821634. A complete description of materials and methods is provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank J. Cesar Ignacio-Espinoza and Melissa Duhaime for help with phage genome analyses; Jarl Haggerty and Kenth Holmfeldt for development of in-house bioinformatics programs; and Forest Rohwer for launching M.B.S. and K.H. in phage genomics and supporting preliminary genome sequencing of isolates ϕ4:1, ϕ13:1, and ϕ39:1. This study was supported by the Gordon and Betty Moore Foundation (M.B.S.) and postdoctoral fellowships from the Sweden–America Foundation and the Swedish Research Council (to K.H.).

1. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320(5879):1034–1039.
2. Breitbart M (2012) Marine viruses: Truth or dare. *Annu Rev Mar Sci* 4:5.1–5.24.
3. Hurwitz BL, Sullivan MB (2013) The Pacific Ocean virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* 8(2):e57355.
4. Borriss M, et al. (2007) Genome and proteome characterization of the psychrophilic *Flavobacterium* bacteriophage 11b. *Extremophiles* 11(1):95–104.
5. Hawkins SA, Layton AC, Ripp S, Williams D, Sayler GS (2008) Genome sequence of the *Bacteroides fragilis* phage ATCC 51477-B1. *Virol J* 5:97.
6. Aislabie JM, et al. (2006) Dominant bacteria in soils of Marble Point and Wright Valley, Victoria Land, Antarctica. *Soil Biol Biochem* 38:3041–3056.
7. Kirchman DL, Dittel AI, Malmstrom RR, Cottrell MT (2005) Biogeography of major bacterial groups in the Delaware estuary. *Limnol Oceanogr* 50(5):1697–1706.
8. Baltar F, Aristegui J, Gasol JM, Hernandez-Leon S, Herndl GJ (2007) Strong coast-ocean and surface-depth gradients in prokaryotic assemblage structure and activity in a coastal transition zone region. *Aquat Microb Ecol* 50(1):63–74.
9. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307(5717):1915–1920.
10. Yanagibashi T, et al. (2009) *Bacteroides* induce higher IgA production than *Lactobacillus* by increasing activation-induced cytidine deaminase expression in B cells in murine Peyer's patches. *Biosci Biotechnol Biochem* 73(2):372–377.
11. Fernández-Gómez B, et al. (2013) Ecology of marine *Bacteroidetes*: A comparative genomics approach. *ISME J* 7(5):1026–1037, 10.1038/ismej.2012.1169.
12. Pinhassi J, et al. (2004) Changes in bacterioplankton composition under different phytoplankton regimens. *Appl Environ Microbiol* 70(11):6753–6766.
13. Holmfeldt K, Middelboe M, Nybroe O, Riemann L (2007) Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their *Flavobacterium* hosts. *Appl Environ Microbiol* 73(21):6730–6739.
14. Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM (2003) Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. *J Bacteriol* 185(21):6434–6447.
15. Nakagawa H, Arisaka F, Ishii S (1985) Isolation and characterization of the bacteriophage T4 tail-associated lysozyme. *J Virol* 54(2):460–466.
16. Lavigne R, Seto D, Mahadevan P, Ackermann HW, Kropinski AM (2008) Unifying classical and molecular taxonomic classification: Analysis of the *Podoviridae* using BLASTP-based tools. *Res Microbiol* 159(5):406–414.
17. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (2012) *Virus Taxonomy; Eighth Report of the International Committee on Taxonomy of Viruses* (Academic, San Diego).
18. Holmfeldt K, Odić D, Sullivan MB, Middelboe M, Riemann L (2012) Cultivated single-stranded DNA phages that infect marine *Bacteroidetes* prove difficult to detect with DNA-binding stains. *Appl Environ Microbiol* 78(3):892–894.
19. Jacobs-Sera D, et al.; Science Education Alliance Phage Hunters Advancing Genomics And Evolutionary Science Sea-Phages Program (2012) On the nature of mycobacteriophage diversity and host preference. *Virology* 434(2):187–201.
20. Hatfull GF (2010) Mycobacteriophages: Genes and genomes. *Annu Rev Microbiol* 64:331–356.
21. Wichels A, et al. (1998) Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* 64(11):4128–4133.
22. Lauro FM, et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106(37):15527–15533.
23. Zhao Y, et al. (2013) Abundant SAR11 viruses in the ocean. *Nature* 494(7437):357–360.
24. Kellogg CA, Rose JB, Jiang SC, Thurmond JM, Paul JH (1995) Genetic diversity of related vibrophages isolated from marine environments around Florida and Hawaii, USA. *Mar Ecol Prog Ser* 120:89–98.
25. Angly F, et al. (2009) Genomic analysis of multiple Roseophage SIO1 strains. *Environ Microbiol* 11(11):2863–2873.
26. Lawrence JG, Hatfull GF, Hendrix RW (2002) Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J Bacteriol* 184(17):4891–4905.
27. Ignacio-Espinoza JC, Sullivan MB (2012) Phylogenomics of T4 cyanophages: Lateral gene transfer in the 'core' and origins of host genes. *Environ Microbiol* 14(8):2113–2126.
28. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol* 3(5):e144.
29. Zhao Y, Wang K, Jiao N, Chen F (2009) Genome sequences of two novel phages infecting marine roseobacters. *Environ Microbiol* 11(8):2055–2064.
30. Enav H, Béjà O, Mandel-Gutfreund Y (2012) Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. *ISME J* 6(3):619–628.
31. Hildenbrand ZL, Bernal RA (2012) Chaperonin-mediated folding of viral proteins. *Advances in Experimental Medicine and Biology*, eds Rossmann MG, Rao VB (Springer, New York), Vol 726, pp 307–324.
32. Krupovic M, Forterre P (2011) *Microviridae* goes temperate: Microvirus-related proviruses reside in the genomes of *Bacteroidetes*. *PLoS ONE* 6(5):e19893.
33. Cherwa JE, Fane BA (2011) Microviridae: Microviruses and gokushoviruses. *Encyclopedia of Life Sciences* (Wiley, Chichester, UK).
34. Tomshich SV, et al. (2007) Structure of acidic O-specific polysaccharide from the marine bacterium Cellulophaga baltica. *Bioorg Khim* 33(1):91–95.
35. Mochizuki T, et al. (2012) Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proc Natl Acad Sci USA* 109(33):13386–13391.
36. Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270):1910–1914.
37. Read TD, et al. (2003) Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): Examining the role of niche-specific genes in the evolution of the *Chlamydiaceae*. *Nucleic Acids Res* 31(8):2134–2147.
38. McDaniel LD, delaRosa M, Paul JH (2006) Temperate and lytic cyanophages from the Gulf of Mexico. *J Mar Biol Assoc U K* 86:517–527.
39. Ackermann HW (2007) 5500 Phages examined in the electron microscope. *Arch Virol* 152(2):227–243.
40. Brum JR, Schenck RO, Sullivan MB (2013) Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J*, 10.1038/ismej.2013.1067.
41. Kuznetsov YG, Chang SC, Credaroli A, Martiny J, McPherson A (2012) An atomic force microscopy investigation of cyanophage structure. *Micron* 43(12):1336–1342.
42. Männistö RH, Kivelä HM, Paulin L, Bamford DH, Bamford JK (1999) The complete genome sequence of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* 262(2):355–363.
43. Hidaka T, Ichida K-i (1976) Properties of a marine RNA-containing bacteriophage. *Mem Fac Fish Kagoshima Univ* 25:77–89.
44. Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M (2013) A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evol Biol* 13(33):33, 10.1186/1471-2148-1113-1133.
45. Horton HR, Moran LA, Ochs RS, Rawn JD, Scrimgeour KG (1996) *Principles of Biochemistry* (Prentice-Hall, Upper Saddle River, NJ), 2nd Ed.
46. El Yacoubi B, Bailly M, de Crécy-Lagard V (2012) Biosynthesis and function of post-transcriptional modifications of transfer RNAs. *Annu Rev Genet* 46:69–95.
47. Sabri M, et al. (2011) Genome annotation and intraviral interactome for the *Streptococcus pneumoniae* virulent phage Dp-1. *J Bacteriol* 193(2):551–562.
48. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4(4):470–483.
49. Abrescia NG, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. *Annu Rev Biochem* 81:795–822.
50. Williamson SJ, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3(1):e1456.
51. Duhaime MB, Sullivan MB (2012) Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 434(2):181–186.
52. Solonenko SA, et al. (2013) Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* 14:320.
53. Pedrós-Alió C (2012) The rare bacterial biosphere. *Annu Rev Mar Sci* 4:449–466.
54. Stocker R (2012) Marine microbes see a sea of gradients. *Science* 338(6107):628–633.
55. Wiśniewski JR, Zougman A, Mann M (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J Proteome Res* 8(12):5674–5678.
56. Verberkmoes NC, et al. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 3(2):179–189.

Supporting Information

Holmfeldt et al. 10.1073/pnas.1305956110

SI Methods

Bacteria and Phage Growth Conditions. *Cellulophaga baltica* host strains were grown at room temperature (RT) on agar plates (1 g yeast extract, 5 g Bacto Peptone, and 15 g of agar per liter sea salt solution; 15 psu). Single colonies were inoculated into MLB liquid media (0.5 g yeast extract, 0.5 g Bacto Peptone, 0.5 g casamino acids, and 3 mL glycerol per liter sea salt solution; 15 psu) and grown without agitation overnight. Phages were grown by using the top-agar plating technique whereby phages diluted in MSM buffer (450 mM NaCl, 50 mM $MgSO_4 \times 7H_2O$, 50 mM Tris base, pH 8), were mixed with 300 mL of bacterial overnight culture, and 3.5 mL molten soft agar (MSM buffer containing 0.5% Low Melting Point Agarose; Fisher), and dispersed on agar plates. Plates were grown at RT in the dark, and plaques were visible after 1 to 2 d. For phage lysate, 5 mL of MSM buffer was added to fully lysed plates, the plates were shaken for 30 min at RT, and the liquid was collected and 0.2-μm filtered. Lysate was stored in the dark at 4 °C until further investigation.

DNA Extraction and Sequencing. Large-scale phage concentrates (n = 3–10 fully lysed plates) were precipitated with PEG 8000 (Fisher) to use for DNA extraction and proteomics (as detailed later). Briefly, 6.5 g NaCl was added per 100 mL of phage lysate, incubated on ice for 1 h to overnight at 4 °C, and centrifuged (11,000 × g, 10 min). To the supernatant, 10 g PEG was added per 100 mL, incubated 1 h to overnight at 4 °C, and centrifuged (10,000 × g, 10 min). After the pellet had air-dried, it was resuspended in MSM buffer. DNA from the dsDNA phages was extracted by using the Wizard DNA purification kit (Promega) according to the manufacturer's recommendations. As no DNA could be obtained from the ssDNA phages when extracted from viral particles, this DNA was extracted as dsDNA plasmids during replication. Here, phages at a multiplicity of infection of 3 were added to overnight bacterial culture, and cells were harvested through centrifugation (3,220 × g, 15 min) after one-third latent period. Phage DNA was extracted by using a plasmid extraction kit (BioRad).

Phages were sequenced commercially by using Illumina HiSEq.2000 (ϕ4:1, ϕSM, ϕ3:1, ϕ47:1; http://uagc.arl.arizona.edu), Roche 454 (all other dsDNA phages; www.genome.duke.edu/cores/sequencing/ and http://med.emory.edu/main/research/core_labs/genomics/index.html), or IonTorrent (ssDNA; http://uagc.arl.arizona.edu) sequencing.

Genome Assembly and Annotation. Phage genomes were assembled by using the Newbler assembly software package (454 Life Sciences) with all settings set to default. For genomes not fully closed or only assembled into two or three contigs, primers were designed (Primer3; http://frodo.wi.mit.edu) to perform PCR across the missing parts. PCR was performed by using GoTaq Green (Promega) with 0.2 μM primer and a thermal program of 94 °C for 5 min initial denaturation; 30 cycles of 94 °C for 30 s denaturation, various temperatures (depending on primer pair) for 30 s annealing, and 72 °C for 30 to 90 s (depending on product size) elongation; and a final elongation of 5 min at 72 °C. PCR products were sequenced commercially through the University of Arizona Genetics Core and aligned to the genomes by using SeqMan (Lasergene). As a result of their novel appearance and the large risk of bacterial contamination because of the extraction method, the ssDNA phages genomes were fully Sanger-sequenced as described earlier by using heat-treated (95 °C, 10 min), CsCl-purified viral particles as template. For the CsCl treatment, ∼5 mL of phage lysate was loaded on top of a CsCl gradient consisting of 1 mL 1.5 $g \cdot cm^{-3}$, 1 mL 1.4 $g \cdot cm^{-3}$, 3 mL 1.3 $g \cdot cm^{-3}$, and 4 mL 1.2 $g \cdot cm^{-3}$ from bottom to top. The gradients were centrifuged at 102,000 × g for 3 h at 4 °C (SW40; Beckman). If the phage-containing fraction could be visualized as a band, it was pulled out with a needle; otherwise, fractions (500 μL) were collected from the bottom of the tubes (through a needle) and the phage-containing fraction was determined by plaque assay as described earlier.

The assembled genomes were annotated by using a pseudoautomated pipeline. ORFs were predicted by using GeneMark Heuristic (1), followed by refinement through synteny and maximizing ORF size where alternative start sites were present. Functional annotations for predicted ORFs were assigned by using BLASTP (e-value cutoff $<1 \times 10^{-3}$) against the National Center for Biotechnology Information (NCBI) nonredundant database (as of November 2012). Further, conserved regions and protein families were identified through searches against the Conserved Domain Database (2) and Pfam (3) (e-value cutoff $<1 \times 10^{-3}$; as of November 2012). Identification of tRNAs was done by using tRNAscan-SE (4) (cover score cutoff >25).

Genomic and Metagenomic Comparison. By using BLASTP, all predicted ORFs were compared (e-value cutoff $<1 \times 10^{-3}$) vs. a database created from the same ORFs (MAKEBLASTDB), and the percentage of shared genes per genome was calculated. Besides the *Cellulophaga* phages, the same analyses were performed on complete genomes available for *Prochlorococcus*, *Synechococcus*, *Escherichia*, and *Pseudomonas* phages (NCBI, January 2012).

For genome size comparison, all *Caudovirales* and *Microviridae* genomes were extracted from NCBI, including draft genome sequences, November 27, 2012. For this, all phage nucleotide sequences were extracted from NCBI and all sequences <1,000 nt were removed because they are unlikely to represent a full genome. The origin of the remaining sequences was manually examined to evaluate if they were complete or nearly complete genomes, and 186 (*Podoviridae*), 284 (*Myoviridae*), 625 (*Siphoviridae*), and 54 (*Microviridae*) genomes were used for the analyses and are referred to in the text. To search for specific genes within NCBI, the following search phrases were used within the Protein search module accessed in January 2013: ((GroEl OR Chaperonin OR Cpn10) AND "viruses"[porgn:__txid10239]), (thymidylate synthase AND "viruses"[porgn:__txid10239]), and ((nrdA OR nrdB OR ribonucleoside triphosphate reductase) AND "viruses"[porgn:__txid10239]). The results were manually verified to be the protein in question (e.g., searched in Pfam if annotation was unclear) and replicate sequences were removed.

Metagenomic reads were recruited to all ORFs in one representative of each of the 12 *Cellulophaga* phage genera (both representatives for genus Cba101) by using the TBLASTN workflow (e-value cutoff $<1 \times 10^{-3}$; percent amino acid identity cutoff >20%; alignment length >45 nt; alignments per query, 300) in the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (https://portal.camera.calit2.net/) against Broad Phage Metagenomes, a database consisting of 137 viral metagenomes sequenced through the Moore Foundation (www.broad.mit.edu/annotation/viral/Phage/Home.html). The same analysis was also performed on *Prochlorococcus* phage P-SSM4 (NC_006884), *Vibrio* phage KVP40 (NC_005083), and *Enterobacteria* phage T4 (NC_000866). To calculate the proportion of novel data contributed by the

Cellulophaga phages, all reads that recruited to the *Cellulophaga* phages were BLASTed against the NCBI nonredundant database (as of November 2012) by using BLASTX. The number of reads lacking significant hits (e-value cutoff $<1 \times 10^{-3}$) to NCBI was divided by the total number of reads in the metagenome to give the proportion of novel data in a metagenome to which the *Cellulophaga* phages contributed. The fraction of queuosine (Que) genes of phage origin was calculated by dividing the total number of reads of phage origin per gene by the total number of reads recruiting to the same *Cellulophaga* Que gene. For a read to be of phage origin, it (*i*) lacked significant hits (e-value cutoff $<1 \times 10^{-3}$) to NCBI, (*ii*) had higher bitscore to the *Cellulophaga* phages than any other sequences in NCBI, or (*iii*) had highest bitscore to a phage in NCBI.

EM. Transmission EM grids were prepared by placing 10 μL of CsCl-purified lysate (as described earlier) onto 200 mesh Formvar-coated copper grids (Ted Pella) for 5 min. The solution was subsequently removed with filter paper, and grids were negatively stained with 2% (wt/vol) uranyl acetate solution by rinsing the grids with two drops of the solution and staining for 45 s with a third drop. The grids were examined by using a Philips CM12 microscope with an accelerating voltage of 80 kV.

MS-Based Proteomics Analyses. Phages were harvested with MSM buffer from fully lysed plates and CsCl-purified as described earlier. The purified phage particles were prepared before 2D LC-MS/MS analyses by using an optimization of the Filter-Aided Sample Preparation kit (Protein Discovery) (5). All reagents were provided for in the kit. Briefly, purified phage were resuspended in 8 M urea/10 mM DTT, denatured, and passed over the 30-kDa filter, then washed with 8 M urea and treated with iodoacetamide to label cysteine residues. Iodoacetamide was washed away with 8 M urea and then 50 mM ammonium bicarbonate. Sequencing-grade trypsin was then added, and digestion processed overnight. The next day, peptides were eluted from the 30-kDa filter via ammonium bicarbonate buffer, NaCL buffer and water/0.1% formic acid. Three aliquots were prepared per sample and frozen at −80 °C until 2D LC-MS/MS analyses. The Filter-Aided Sample Preparation–prepared peptides (>500 ng) were loaded onto the back column of a split-phase 2D column (~3–5 cm SCX and 3–5 cm C-18; all packing materials purchased from Phenomenex). The column was loaded to the HPLC and washed with 100% aqueous solution for 5 min, followed by a ramp from 100% aqueous to 100% organic solution for 10 min. The column was connected to a front column (RP C-18; 15 cm) with a nanospray source on Velos Orbitrap (dsDNA phages) or LTQ or LTQVelos (ssDNA phages) and run for 5 to 12 h 2D separation of increasing salt pulses (ammonium acetate), followed by water to organic gradients (6). All instruments were run in a data-dependent manner as previously described (6, 7). To recruit peptides to the phage genomes, the resultant MS/MS spectra were searched against a database consisting of annotated phage proteins, all phage ORFs >30 aa (to identify ORFs possibly missed through the annotation), and proteins from sequenced *Bacteroidetes* bacteria (*Dokdonia donghaensis* MED134, *Leeuwenhoekiella blandensis* MED217, *Robiginitalea biformata* HTCC2501), and eukaryotic organisms (human and mouse) to use as indicator for false positives. Data analyses were performed by using SEQUEST and filtered with DTASelect with conservative filters (6).

1. Besemer J, Borodovsky M (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* 27(19):3911–3920.
2. Marchler-Bauer A, et al. (2011) CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39(database issue):D225–D229.
3. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(database issue):D211–D222.
4. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955–964.
5. Wiśniewski JR, Zougman A, Mann M (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J Proteome Res* 8(12):5674–5678.
6. Verberkmoes NC, et al. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 3(2):179–189.
7. Erickson AR, et al. (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* 7(11):e49138.

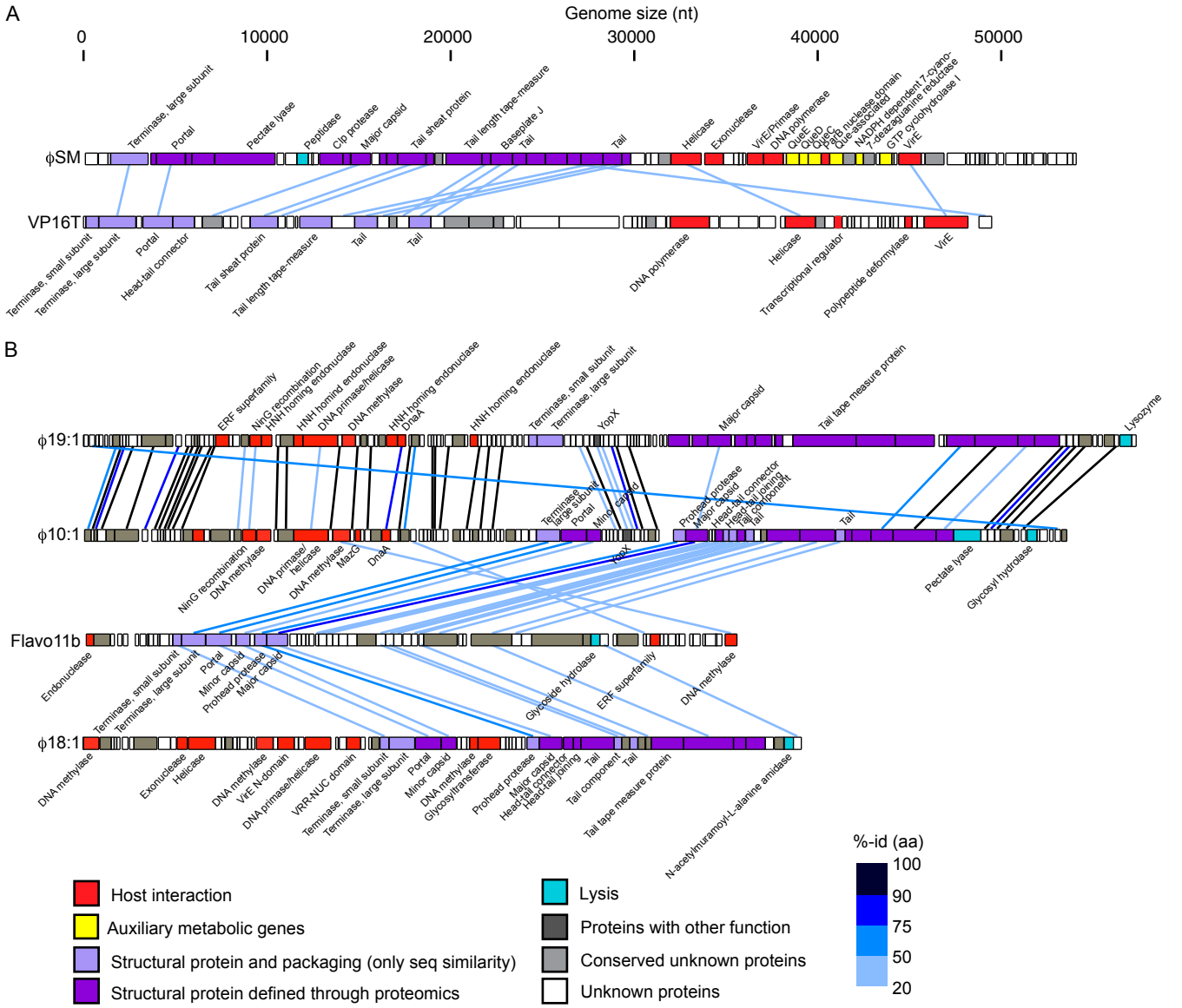

Fig. S1. Genome comparison of (*A*) *Cellulophaga* genus CbaSM phage (ϕSM) and *Vibrio* phage VP16T (AY328852), (*B*) *Cellulophaga* genera Cba181 (ϕ18:1) and Cba101 (ϕ10:1 and ϕ19:1), and *Flavobacterium* phage 11b (NC_006356). *Cellulophaga* genus Cba101b shares only five genes with *Flavobacterium* phage 11b. The annotation for VP16T and 11b is directly retrieved from GenBank annotations.

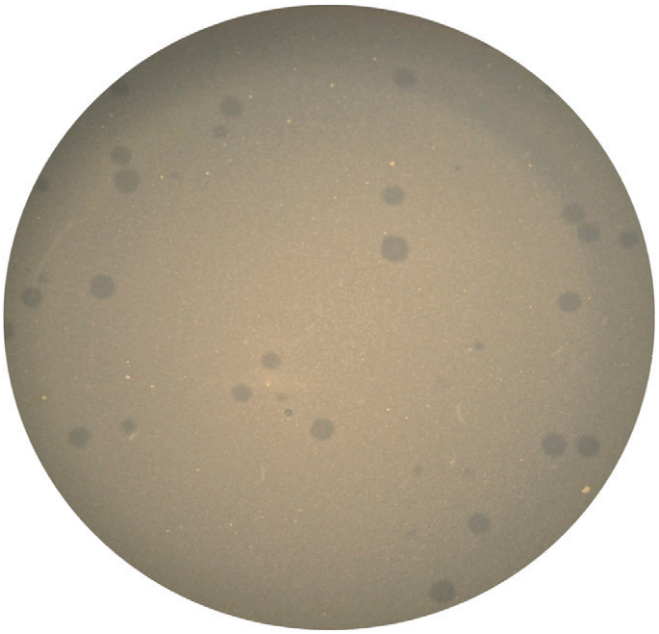

Fig. S2. Plaque morphology of phage ϕ48:2 when growing on its original host *C. baltica* strain NN016048.

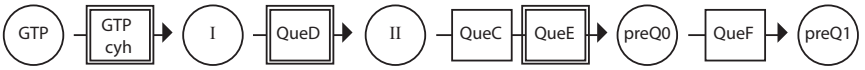

Fig. S3. Genes coding for enzymes involved in bacterial de novo Que biosynthesis. Boxes with a single line represent genes occurring only in the genus CbaSM phages, whereas boxes with double lines occur in genera CbaSM and Cba131 phages. Functional abbreviation: cyh, cyclohydrolase I. Intermediate abbreviations: I, 7,8-dihydroneopterin triphosphate; II, 6-pyruvoyltetrahydropterin; preQ0, 7-cyano-7-deazaguanine; preQ1, 7-aminomethyl-7-dezaguanine.

Table S1. General features of *Cellulophaga* phages

Phage isolate	Family	Putative genus	DNA type	Genome size, nt	G+C, %	ORFs	tRNA	ORFs with hits to databases, %	ORFs with function, %
ϕ38:1	*Podoviridae*	Cba401likevirus	dsDNA	72,534	38.06	101	16	34.7	14.9
ϕ40:1	*Podoviridae*	Cba401likevirus	dsDNA	72,529	38.06	101	16	34.7	14.9
ϕ13:2	*Podoviridae*	Cba183likevirus	dsDNA	72,369	32.87	127	1	36.2	14.2
ϕ18:3	*Podoviridae*	Cba183likevirus	dsDNA	71,442	32.88	123	—	44.7	15.4
ϕ19:3	*Podoviridae*	Cba183likevirus	dsDNA	75,991	32	130	1	41.5	13.1
ϕ46:3	*Podoviridae*	Cba183likevirus	dsDNA	72,960	32.7	121	—	40.5	13.2
ϕ14:2	*Podoviridae*	Cba142likevirus	dsDNA	100,356	29.6	133	—	25.6	15.0
ϕ4:1	*Podoviridae*	Cba41likevirus	dsDNA	146,105	32.62	197	24	31.5	14.2
ϕ17:2	*Podoviridae*	Cba41likevirus	dsDNA	145,340	32.65	198	23	31.8	14.6
ϕS_M	*Myoviridae*	CbaSMlikevirus	dsDNA	54,566	33.46	81	—	40.7	28.4
ϕ3:1	*Myoviridae*	CbaSMlikevirus	dsDNA	54,427	33.45	81	—	39.5	28.4
$\phi 3S_T$:2	*Myoviridae*	CbaSMlikevirus	dsDNA	54,014	33.47	80	—	41.3	28.8
ϕ38:2	*Myoviridae*	CbaSMlikevirus	dsDNA	54,012	33.48	80	—	41.3	28.8
ϕ47:1	*Myoviridae*	CbaSMlikevirus	dsDNA	54,016	33.47	80	—	41.3	28.8
ϕ39:1	*Siphoviridae*	Cba391likevirus	dsDNA	28,756	31.33	49	—	42.9	22.4
ϕ46:1	*Siphoviridae*	Cba461likevirus	dsDNA	34,844	38.26	54	—	51.9	27.8
ϕ12:1	*Siphoviridae*	Cba181likevirus	dsDNA	39,148	36.59	64	—	53.1	39.1
ϕ12:3	*Siphoviridae*	Cba181likevirus	dsDNA	39,151	36.58	64	—	53.1	39.1
ϕ17:1	*Siphoviridae*	Cba181likevirus	dsDNA	38,776	36.48	65	—	50.8	33.8
ϕ18:1	*Siphoviridae*	Cba181likevirus	dsDNA	39,189	36.52	65	—	50.8	35.4
ϕ18:2	*Siphoviridae*	Cba181likevirus	dsDNA	38,476	36.58	63	—	52.4	36.5
ϕ10:1	*Siphoviridae*	Cba101likevirus	dsDNA	53,664	31.54	107	2	37.4	18.7
ϕ19:1	*Siphoviridae*	Cba101likevirus	dsDNA	57,447	31.24	117	4	29.1	12.8
ϕS_T	*Siphoviridae*	Cba131likevirus	dsDNA	78,267	30.18	109	—	39.4	23.9
ϕ19:2	*Siphoviridae*	Cba131likevirus	dsDNA	78,276	30.18	109	—	39.4	23.9
ϕ13:1	*Siphoviridae*	Cba131likevirus	dsDNA	76,665	30.22	104	—	42.3	25.0
ϕ12:2	*Microviridae*	Cba184likevirus	ssDNA	6,453	34.81	13	—	23.1	15.4
ϕ12a:1	*Microviridae*	Cba184likevirus	ssDNA	6,478	34.02	13	—	23.1	15.4
ϕ18:4	*Microviridae*	Cba184likevirus	ssDNA	6,478	34.27	13	—	23.1	15.4
ϕ48:1	*Microviridae*	Cba184likevirus	ssDNA	6,478	34.21	13	—	23.1	15.4
ϕ48:2	Novel	Cba482likevirus	ssDNA	11,480	28.8	30	—	26.7	3.3

Table S2. Distribution of thymidylate synthase or thymidylate synthase complementing protein (Thy) and ribonucleoside-diphosphate reductase (RNR) among *Cellulophaga* phages and 625 sequenced siphoviruses

Phage	Accession no.	Genome size, bp	Thy	RNR
Cellulophaga phage ϕST	This study	78,267	X	I
Cellulophaga phage ϕ19:2	This study	78,276	X	I
Cellulophaga phage ϕ13:1	This study	76,665	X	I
Bacillus phage PBC1	NC_017976	41,164	X	—
Bacillus phage phi3T	PH3THYP3	NA	X	—
Bacteroides phage B124-14	NC_016770	47,159	X	—
Bacteroides phage B40-8	NC_011222	44,929	X	—
Caulobacter phage CcrColossus	NC_019406	279,967	—	I
Caulobacter phage CcrKarma	NC_019410	221,828	—	I
Caulobacter phage CcrMagneto	NC_019407	218,929	—	I
Caulobacter phage CcrSwift	NC_019411	219,216	—	I
Caulobacter phage phiCbK	NC_019405	215,710	—	I
Clavibacter phage CMP1	NC_013698	58,652	X	—
Clostridium phage phi8074-B1	NC_019924	47,595	—	III
Clostridium phage phiCP13O	NC_019506	38,329	X	—
Clostridium phage phiCP26F	NC_019496	39,188	X	—
Clostridium phage phiCP34O	NC_019508	38,309	X	—
Clostridium phage phiCP39-O	NC_011318	38,753	X	—
Clostridium phage phiCP9O	JF767210	39,594	X	—
Clostridium phage phiCTP1	NC_014457	59,199	X	III
Colwellia phage 9A	NC_018088	104,936	X	—
Cyanophage PSS2	NC_013021	107,530	X	II
EBPR siphovirus 2	JF412297	48,954	X	—
Enterobacteria phage EPS7	NC_010583	111,382	X	III
Enterobacteria phage H8	AC171169	104,373	X	I, III
Enterobacteria phage SPC35	NC_015269	118,351	X	I, III
Enterobacteria phage T5	NC_005859	121,750	X	I, III
Erwinia phage phiEaH2	NC_019929	243,050	X	—
Escherichia phage bV_EcoS_AKFV33	NC_017969	108,853	X	I, III
Gordonia phage GTE2	NC_015720	45,530	X	—
Lactococcus phage 949	NC_015263	114,768	—	III
Mycobacterium phage Adjutor	NC_010763	64,511	X	—
Mycobacterium phage Airmid	JN083853	51,241	X	II
Mycobacterium phage Alma	JN699005	53,177	X	II
Mycobacterium phage Arturo	JX307702	51,500	—	II
Mycobacterium phage Astro	JX015524	52,494	—	II
Mycobacterium phage Backyardigan	JF704093	51,308	X	II
Mycobacterium phage Benedict	JN083852	51,083	X	II
Mycobacterium phage Blue7	JN698999	52,288	X	II
Mycobacterium phage Butterscotch	NC_011286	64,562	X	—
Mycobacterium phage Bxz2	NC_004682	50,913	X	II
Mycobacterium phage Che12	NC_008203	52,047	X	II
Mycobacterium phage D29	NC_001900	49,136	X	II
Mycobacterium phage DaVinci	JF937092	51,547	X	II
Mycobacterium phage Eagle	HM152766	51,436	X	II
Mycobacterium phage EricB	JN049605	51,702	X	II
Mycobacterium phage Flux	JQ809701	51,370	X	II
Mycobacterium phage George	JF704107	51,578	X	—
Mycobacterium phage Gladiator	JF704097	52,213	X	II
Mycobacterium phage Goose	JX307704	50,645	—	II
Mycobacterium phage Gumball	NC_011290	64,807	X	—
Mycobacterium phage Hammer	JF937094	51,889	X	II
Mycobacterium phage HelDan	JF957058	50,364	X	II
Mycobacterium phage ICleared	JQ896627	51,440	X	II
Mycobacterium phage Jeffabunny	JN699019	48,963	X	II
Mycobacterium phage JHC117	JF704098	50,877	X	II
Mycobacterium phage L5	NC_001335	52,297	X	II
Mycobacterium phage LHTSCC	JN699015	51,813	X	II
Mycobacterium phage MeeZee	JN243856	51,368	X	II
Mycobacterium phage Microwolf	JF704101	50,864	X	II
Mycobacterium phage Nova	JN699014	65,108	X	—

Table S2. Cont.

Phage	Accession no.	Genome size, bp	Thy	RNR
Mycobacterium phage PackMan	JF704110	51,339	X	II
Mycobacterium phage PBI1	NC_008198	64,494	X	—
Mycobacterium phage Peaches	NC_013694	51,376	X	II
Mycobacterium phage PLot	NC_008200	64,787	X	—
Mycobacterium phage Pukovnik	NC_011023	52,892	X	II
Mycobacterium phage Rebeuca	JX411619	51,235	—	II
Mycobacterium phage RedRock	GU339467	53,332	X	II
Mycobacterium phage Rockstar	JF704111	47,780	X	II
Mycobacterium phage Sabertooth	JX307703	51,377	—	II
Mycobacterium phage Saintus	JN831654	49,228	—	II
Mycobacterium phage Send513	JF704112	71,547	X	—
Mycobacterium phage Shaka	JF792674	51,369	X	II
Mycobacterium phage SirHarley	JF937107	64,791	X	—
Mycobacterium phage SWU1.	NC_017973	52,474	—	II
Mycobacterium phage TA17A	JN400277	52,868	X	II
Mycobacterium phage Tiger	JQ684677	50,332	X	II
Mycobacterium phage Timshel	JF957060	53,278	—	II
Mycobacterium phage TiroTheta9	JN561150	51,367	X	II
Mycobacterium phage Trixie	JN408461	53,526	X	II
Mycobacterium phage Troll4	NC_011285	64,618	X	—
Mycobacterium phage Turbido	JN408460	53,169	X	II
Mycobacterium phage Twister	JQ512844	51,094	X	II
Mycobacterium phage Vix	JF704114	50,963	X	II
Mycobacterium phage Wile	JN243857	51,308	X	II
Mycobacterium phage Wonder	HM755814	48,491	X	II
Pectobacterium phage My1	NC_018837	122,024	X	I, III
Phage phiJL001	NC_006938	63,649	X	—
Pseudomonas phage M6	NC_007809	59,446	X	—
Pseudomonas phage MP1412	NC_018282	61,167	X	—
Pseudomonas phage YuA	NC_010116	58,663	X	—
Rhodococcus phage ReqiPepy6	GU580941	76,797	X	—
Rhodococcus phage ReqiPoco6	GU580942	78,064	X	—
Rhodococcus phage RER2	NC_016653	46,586	X	II
Rhodococcus phage RGL3	NC_016650	48,072	X	II
Roseobacter phage RDJL Phi 1	NC_015466	62,668	X	—
Salmonella phage SSU5	NC_018843	103,299	X	—
Streptomyces phage phiC31	NC_001978	41,491	X	—
Streptomyces phage phiELB20	JX262376	51,160	—	II
Streptomyces phage phiHau3	NC_018836	50,255	—	II
Streptomyces phage R4	NC_019414	51,071	—	II
Synechococcus phage S-CBS2	NC_015463	72,332	X	II
Synechococcus phage S-CBS4	NC_016766	69,420	X	—
Thermus phage P23-45	NC_009803	84,201	X	II
Thermus phage P74-26	NC_009804	83,319	X	II
Vibrio phage pVp-1	NC_019529	111,506	—	II
Vibrio phage SSP002	JQ692107	76,350	X	—
Yersinia phage phiR201	NC_019919	112,795	X	II

Table displays complete and draft genomes as of November 2012 retrieved from the NCBI. NA, not available; RNR, ribonucleoside-diphosphate reductase; Thy, thymidylate synthase or thymidylate synthase complementing protein.

Table S3. Distribution of thymidylate synthase or thymidylate synthase complementing protein (Thy) and ribonucleoside-diphosphate reductase (RNR) among *Cellulophaga* podoviruses and 186 podoviruses

Phage	Accession no.	Genome size, nt	Thy	RNR	Taxonomic affiliation	Aquatic	Host	Accession no.	RNR
Cellulophaga phage ϕ14:2	This study	100,356	X	I	Cba142likevirus*	Yes	*C. baltica*	KC859630	I
Cellulophaga phage ϕ4:1	This study	146,105	X	I	Cba41likevirus*	Yes	*C. baltica*	KC859630	I
Cellulophaga phage ϕ17:2	This study	145,340	X	I	Cba172likevirus*	Yes	*C. baltica*	KC859630	I
Synechococcus phage Syn5	NC_009531	46,214	X	—	*Autographivirinae*	Yes	*Synechococcus sp.* WH8109	WP_006851477	
Synechococcus phage P60	NC_003390	47,872	X	II	*Autographivirinae*	Yes	*Synechococcus* sp. WH 7803	NC_009481	II
Cyanophage NATL2A-133	NC_016659	47,536	—	II	*Autographivirinae*	Yes	*Prochlorococcus marinus* str. NATL2A	NC_007335	II
Cyanophage NATL1A-7	NC_016658	47,741	—	II	*Autographivirinae*	Yes	*Prochlorococcus marinus* str. NATL1A	NC_008819	II
Cyanophage 9515–10a	NC_016657	47,055	—	II	*Autographivirinae*	Yes	*Prochlorococcus marinus* MIT 9515	YP_001011050	II
Prochlorococcus phage P-SSP7	NC_006882	45,176	—	II	*Autographivirinae*	Yes	*Prochlorococcus marinus* MED 4	CAE19120	II
Roseobacter phage SIO1	NC_002519	39,898	X	II	*Autographivirinae*	Yes	Host RNR classification not determined	—	—
Roseovarius Plymouth podovirus 1	FR719956	74,704	—	II	N4likevirus	Yes	Host RNR classification not determined	—	—
Roseovarius sp 217 phage 1	FR682616	74,583	X	II	N4likevirus	Yes	*Roseovarius* sp. 217	PRJNA54245	II
Sulfitobacter phage EE36phi1	NC_012696	73,325	X	II	N4likevirus	Yes	*Sulfitobacter* sp. EE-36	PRJNA54191	II
Silicibacter phage DSS3phi2	NC_012697	74,611	X	II	N4likevirus	Yes	*Ruegeria pomeroyi* DSS-3	NC_003911	II
Erwinia phage vB_EamP-S6	NC_019514	74,669	X	—	N4likevirus	No	No phage RNR	—	—
Enterobacter phage IME11	NC_019423	72,570	X	—	N4likevirus	No	No phage RNR	—	—
Escherichia phage vB_EcoP_G7C	NC_015933	72,917	X	—	N4likevirus	No	No phage RNR	—	—
Escherichia phage N4	NC_008720	70,153	X	—	N4likevirus	No	No phage RNR	—	—
Enterobacter phage EcP1	NC_019485	59,080	X	—	N4likevirus	No	No phage RNR	—	—
Enterobacter phage Eco32	NC_010324	77,554	X	—	Phieco32likevirus	No	No phage RNR	—	—
Cronobacter phage vB_CsaP_GAP52	NC_019402	76,631	—	III	Phieco32likevirus	No	Host RNR classification not determined	—	—

Table displays complete and draft genomes as of November 2012 retrieved from the NCBI, including taxonomic affiliation beyond family (if available) and whether the phages are isolated from aquatic environments. For phages containing RNR, the host's RNR class is given if information is available. NA, not available; RNR, ribonucleoside-diphosphate reductase; Thy, thymidylate synthase or thymidylate synthase complementing protein.

Table S4. Summary of recruitment data for the T4- phages hierarchical core analyses

Phage	Phage type	Reads recruit to genes, %		Redundantly recruited reads with higher bitscore to P-SSM4, %
		Within core	Redundantly to other T4	
P-SSM4	Marine cyano	27	21	—
Kvp40	Marine	44	51	86
T4	Nonmarine	64	68	89

Dataset S1. Phage genome annotation details based on similarity to NCBI nonredundant, CDD, and Pfam databases

Dataset S1

If applicable, structural proteomic data and metagenomic recruitment summary are provided.