The ISME Journal (2011) 5, 107–121

www.nature.com/ismej

ORIGINAL ARTICLE

Ecogenomics and genome landscapes of marine Pseudoalteromonas phage H105/1

Melissa Beth Duhaime[1,2], Antje Wichels[3], Jost Waldmann[1], Hanno Teeling[1] and Frank Oliver Glöckner[1,2]

[1]*Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, Bremen, Germany;* [2]*School of Engineering and Sciences, Jacobs University, Bremen, Germany and* [3]*Biosciences Division, Alfred Wegener Institute for Polar and Marine Research, BAH, Helgoland, Germany*

Marine phages have an astounding global abundance and ecological impact. However, little knowledge is derived from phage genomes, as most of the open reading frames in their small genomes are unknown, novel proteins. To infer potential functional and ecological relevance of sequenced marine Pseudoalteromonas phage H105/1, two strategies were used. First, similarity searches were extended to include six viral and bacterial metagenomes paired with their respective environmental contextual data. This approach revealed 'ecogenomic' patterns of Pseudoalteromonas phage H105/1, such as its estuarine origin. Second, intrinsic genome signatures (phylogenetic, codon adaptation and tetranucleotide (tetra) frequencies) were evaluated on a resolved intragenomic level to shed light on the evolution of phage functional modules. On the basis of differential codon adaptation of Phage H105/1 proteins to the sequenced *Pseudoalteromonas* spp., regions of the phage genome with the most 'host'-adapted proteins also have the strongest bacterial tetra signature, whereas the least 'host'-adapted proteins have the strongest phage tetra signature. Such a pattern may reflect the evolutionary history of the respective phage proteins and functional modules. Finally, analysis of the structural proteome identified seven proteins that make up the mature virion, four of which were previously unknown. This integrated approach combines both novel and classical strategies and serves as a model to elucidate ecological inferences and evolutionary relationships from phage genomes that typically abound with unknown gene content.
The ISME Journal (2011) **5,** 107–121; doi:10.1038/ismej.2010.94; published online 8 July 2010
Subject Category: evolutionary genetics
Keywords: ecogenomics; genome signatures; genomics; marine; phage; *Pseudoalteromonas*

Introduction

Viruses are the most abundant biological entity and the largest source of genetic material on the planet (Suttle, 2007), and are likely the major vehicle for gene transfer in the ocean. Considering the global volume of seawater, the worldwide abundance of marine phages and bacteria and the frequency of gene transfers per infection, virus-mediated transfers occur up to 10^{15} times per second in the ocean (Bushman, 2002), with an extrapolated 10^{28} bp of DNA transduced by phages per year (Paul *et al.*, 2002). Evidence shows that these transfers include host-derived metabolic genes central to the metabolism of the world's oceans (Lindell *et al.*, 2005), carried by the virus and expressed during infection (Lindell *et al.*, 2004).

Correspondence: MB Duhaime, Microbial Genomics, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, Bremen 28359, Germany.
E-mail: mduhaime@mpi-bremen.de
Received 14 January 2010; revised 4 May 2010; accepted 19 May 2010; published online 8 July 2010

Pseudoalteromonads are ubiquitous heterotrophic members of marine bacterial communities, which, as with most microbial life, are ecologically and evolutionarily influenced by phages (Moebus, 1992; Wichels *et al.*, 1998, 2002; Männistö *et al.*, 1999; Médigue *et al.*, 2005; Thomas *et al.*, 2008). All three sequenced Pseudoalteromonads contain integrated prophages, two of which are dominated by P2-like myovirus proteins (Prophinder) (Lima-Mendez *et al.*, 2008). Pseudoalteromonas phage H105/1, the focus of this study, is a member of the *Siphoviridae* family isolated from the North Sea, on *Pseudoalteromonas* sp. H105 (Figure 1). Phage H105/1 also infects *Pseudoalteromonas* spp. H103 and H108, which were isolated along with H105 from the same water sample (Wichels *et al.*, 1998). The host, *Pseudoalteromonas* sp. H105, is susceptible to lysis or growth inhibition by other Helgoland and North Sea phages, including members of both the *Myovridae* and *Siphoviridae* families (Wichels *et al.*, 1998, 2002).

Phage genomes are small (3–300 kb) compared with the *Bacteria* and *Archaea* they infect (1000–13 000 kb), and typically abound with unknown gene content. The majority of the open reading

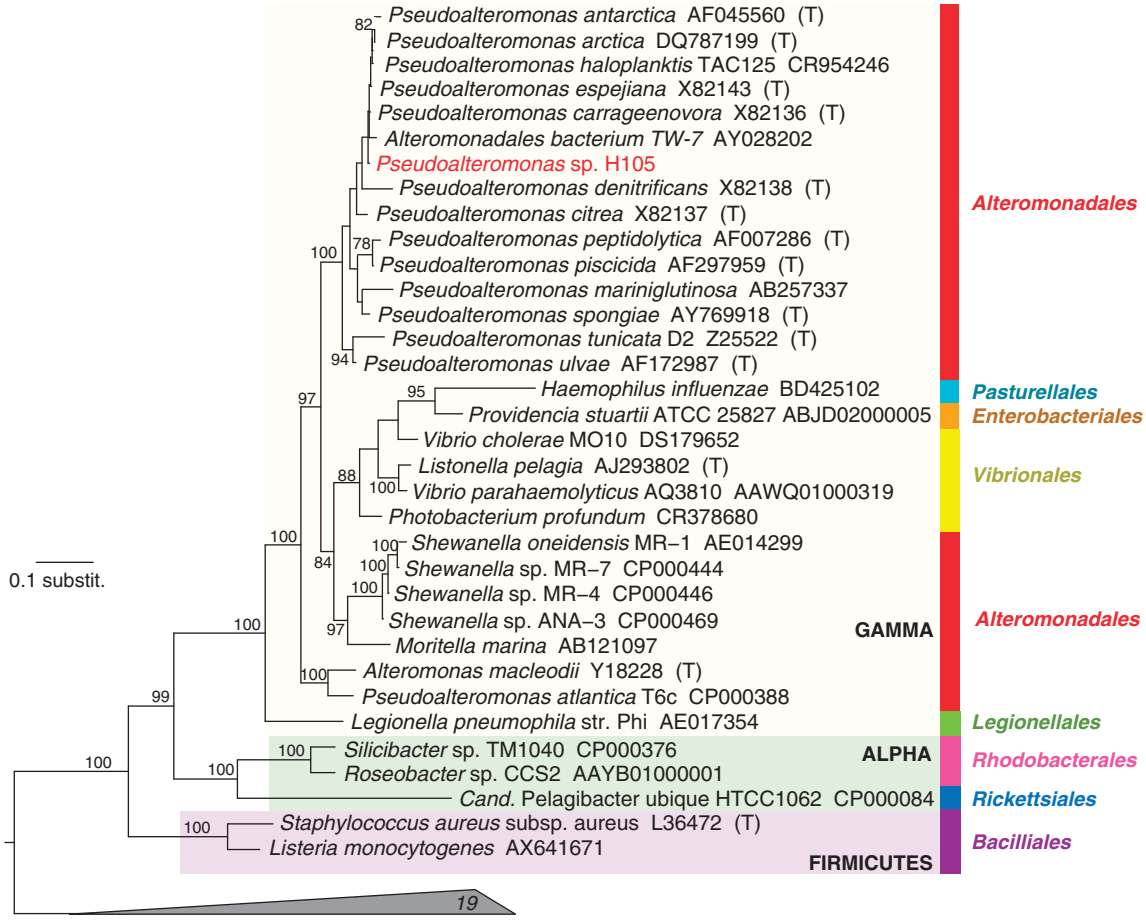


Figure 1 Phylogenetic characterization of host *Pseudoalteromonas* sp. H105 16S rRNA gene. Maximum likelihood tree calculated with 1000 bootstraps using RAxML. RefSeq accession numbers follow the organism name; (T) indicates a type strain. Background shading highlights the taxonomic classification of the organism. Bootstrap values >75 are shown on the branches; 19 sequences were used as an outgroup. Bar represents 10% estimated sequence change. ALPHA and GAMMA denote the class of Proteobacteria per cluster.

frames (ORFs) (over 60%) of sequenced marine phage genomes are hypothetical proteins (unique in public sequence databases) or conserved hypothetical proteins (similar only to other unknown proteins). As the public sequence databases are insufficient to grasp phage protein diversity, traditional approaches to genome analysis, which rely on similarity searches (blast (Altschul *et al.*, 1990) against NCBI-nr), or protein family classification (Pfam (Finn *et al.*, 2010)), reveal little about phage evolution. We expanded the similarity searches of Phage H105/1 to include the Global Ocean Sampling data set (GOS) (Rusch *et al.*, 2007) and five globally distributed marine viral metagenomes missing from NCBI-nr and -env (Angly *et al.*, 2006; McDaniel *et al.*, 2008). Such an approach lends itself to 'ecogenomic' interpretations, whereby ecological inferences about Phage H105/1 can be made on the basis of genomic patterns and their respective environmental contextual data (Kottmann *et al.*, 2010).

Further complicating phage genomics, their evolution is driven by the rampant exchange of functional genome modules (Botstein, 1980; Hendrix *et al.*, 1999; Pedulla *et al.*, 2003) that are frequently swapped between phages infecting diverse hosts (Lucchini *et al.*, 1999; Filée *et al.*, 2006). As such, phage genomes can be thought of as veritable 'concatenated metagenomes', in that consecutive fragments can have very dissimilar origins and evolutionary pasts. Tetranucleotide (tetra) usage frequencies, a feature increasingly used to cluster sequence fragments originating from discrete organisms of a community (Woyke *et al.*, 2006; Andersson and Banfield, 2008; Dick *et al.*, 2009), were considered in this study as a tool to differentiate and shed light on the evolutionary history of Phage H105/1 'functional modules'.

We set out to test (a) whether the isolation habitat of Phage H105/1 (Helgoland, North Sea), in light of its respective environmental parameters, influences the distribution of H105/1 protein sequences in the currently available 'global virome', and (b) whether a resolved intra-genomic (tetra) frequency signature exists, and how this may be related to host codon adaptation. These novel approaches, integrated with experimental characterization of the phage's infection dynamics and structural proteome, offer strategies to elucidate ecological and evolutionary

patterns and understand genomic features of Pseudoalteromonas phage H105/1.

Materials and methods

Phage harvesting, DNA isolation and sequencing
Pseudoalteromonas sp. strain H105 and Pseudoalteromonas phage H105/1 were isolated at 1 m, in September 1990 (Moebus, 1992), off the coast of Helgoland (54°11′3 N, 7°54′0 W) in the North Sea. The host was stored in liquid nitrogen, and the phage at 4 °C in SM buffer (100 mM NaCl, 81.2 mM $MgSO_4$. 7H2O, 50 mM Tris-HCl (pH 7.5), 0.01% gelatin). The host was reconstituted in marine media and infected with H105/1 (October 2006) using the agar overlay method (Wichels *et al.*, 1998). Phages were harvested from plates with SM, precipitated (polyethylene glycol/NaCl method) (Sambrook and Russell, 2001) and then recovered in SM. Purified lysates were incubated (1 h, 65 °C) with proteinase K (100 μg ml^{-1} final) and sodium dodecyl sulphate (0.5% final). DNA was phenol:chloroform extracted, ethanol precipitated and resuspended in 1X TE Buffer (tris ethylenediaminetetraacetic acid). The genome was sequenced by Agowa GmbH (Berlin, Germany) using a linear *Escherichia coli* vector, pJAZZ-KA (BigEasy-pTEL, Lucigen; Middleton, WI, USA). The Pseudoalteromonas phage H105/1 genome sequence has been deposited in GenBank under accession number: HM588722.

Virion structural proteome analysis
Lysates were purified by CsCl centrifugation (Sambrook and Russell, 2001). Briefly, debris was extracted from polyethylene glycol-purified lysates with chloroform (1:1), vortexed and centrifuged (3000 × g, 15 min, 4 °C), and the aqueous phase was used in CsCl purification. The gradient tube (Ultra-Clear, Beckman, Fullerton, CA, USA) was layered with 1.125 ml each of (1) 1.7 g CsCl ml^{-1}, (2) 1.5 g CsCl ml^{-1}, (3) 1.45 CsCl ml^{-1} and (4) topped with 1.15 g CsCl ml^{-1}, and centrifuged (87 000 × g, 2 h, 4 °C). A blue–white band containing the phage was removed (2 ml total volume) and dialysed (Pierce Slide-A-Lyzer 10 K MWCO, Rockford, IL, USA) twice in 1 l buffer (10 mM NaCl, 50 mM Tris-HCl (pH 8), 10 mM $MgCl_2$) to remove CsCl. Phages were concentrated 10 × (Microcon 30 kD; Millipore, Billerica, MA, USA) and proteins denatured by five freeze–thaw (96 °C) cycles and 12% sodium dodecyl sulphate, then separated by sodium dodecyl sulphate-polyacrylamide gel electrophoresis, as described by Paul *et al.* (2005). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry peptide mass fingerprint spectra were generated from trypsin-digested bands excised from polyacrylamide gel electrophoresis gel (TOPLAB GmbH; Martinsried, Germany). Each peptide fingerprint was best matched to its original peptide sequence in Phage H105/1 genome using a probability-based Mowse score (−10*Log(P), where P is the probability that the observed match is random). Protein scores >29 were considered significant ($P<0.05$).

Genome annotation
Genes were predicted on the basis of (i) GeneMark.hmm (prokaryotic version using bacterial or archaeal genetic code, precomputed *Pseudoalteromonas haloplanktis* chromosome 1 model and default settings) (Besemer *et al.*, 2001) and (ii) FGENESB (generic bacterial model, default settings; Softberry, Mount Kisco, NY, USA), also used to predict operons. Rho-independent bacterial transcriptional terminators were predicted using FindTerm (energy threshold −11, default settings; Softberry, Mount Kisco, NY, USA). Promoters were predicted searching regions 150 bp upstream of predicted starts (BPROM, threshold 0.2, default settings; Softberry) with custom Perl wrappers. Annotation and comparative genomics of Phage H105/1 used JCoast (Richter *et al.*, 2008), streamlining annotation protocols and results of Blastp (low-complexity filter) against NCBI-nr, Pfam (Finn *et al.*, 2010), SignalP (Emanuelsson *et al.*, 2007) and TMHMM (Krogh *et al.*, 2001). Predicted ORFs were searched against the ACLAME MeGO (Mobile Element Gene Ontology) database, which provides functional annotations based on a manually curated database of viruses and mobile genetic elements (Toussaint *et al.*, 2007). When Phage H105/1 proteins were most similar to proteins of a bacterial or archaeal genome, hits were classified as 'prophage' if (a) they have a 'phage-like neighbourhood' (i.e., phage-like proteins 10 genes up or downstream of a 100-kb range) or (b) they lie in a profinder-predicted prophage (Lima-Mendez *et al.*, 2008).

Ecogenomic analysis
Reads from GOS (0.1–0.8 μm fraction, plus two 0.8–3.0 μm fraction samples) (Rusch *et al.*, 2007) were retrieved from the CAMERA database (Seshadri *et al.*, 2007) (Supplementary Material S1). Five marine virus metagenomes were retrieved from NCBI, representing pooled viromes from the Arctic, British Columbia (Strait of Georgia estuary), Gulf of Mexico, Sargasso Sea and a coastal Tampa Bay community, the integrated prophages of which were induced to undergo lysis (gpids: 18 225 and 28 619; Supplementary Material S2) (Angly *et al.*, 2006; McDaniel *et al.*, 2008). Specialized tblastn (BLOSUM62 substitution matrix) of predicted Phage H105/1 proteins and of nine other marine phages was carried out against all six reading frames of virome nucleotide reads using Decypher hardware (TimeLogic, Inc., Carlsbad, CA, USA). Hits with *e*-values $<10^{-4}$ and 20% query coverage (GOS) or 10% query coverage (viromes) were accepted to minimize false positives as determined by behaviour over a range of thresholds (Supplementary Material S3). Raw hit counts were normalized by gene size,

number of reads or site and (for GOS) number of sites or habitat to reduce the effects of uneven sampling of different habitats; values were multiplied by a constant (10^8 for GOS, 10^7 for virome hits) to bring final counts near whole numbers. To determine the extent to which habitat influences the distribution of Phage H105/1 proteins, environmental physiochemical parameters (temperature, salinity, nitrate, phosphate, silicate, dissolved oxygen, oxygen saturation, oxygen utilization) for the marine viromes were interpolated using megx.net GIS tools (Kottmann *et al.*, 2010) on the basis of published location, depth and time parameters (Supplementary Material S2), except for the Arctic sites, where interpolation is not possible. Unfortunately, many of the viromes were pooled samples collected over a range of locations and depths. When a depth range was reported, data for minimum and maximum (and for Bay of British Columbia, intermediate) depths were collected. Sample sites were clustered (average linkage) on the basis of a distance matrix (Euclidean distance) of *z*-score-transformed environmental data. The R project for statistical computing (v. 2.10) was used to calculate correlation matrices and perform principal component analyses to project variation between (i) virus metagenomes, with respect to the relative abundance of proteins from 10 marine phages (on the basis of total blast hits normalized by genome and metagenome size, and double-centred to identify biases among all phages and all viromes) and (ii) virus metagenome and phage isolation sites (on the basis of *z*-score-transformed environmental parameters, interpolated as described above).

Host phylogenetic analysis and tree reconstruction

16S rRNA gene tree. An overnight culture of *Pseudoalteromonas* sp. H105 culture was used in a polymerase chain reaction under standard conditions with GM3F and GM4R primers (Muyzer *et al.*, 1995). Products were gel extracted, purified, used to construct clone libraries with the pGEM-T Easy Vector System I (Promega; Madison, WI, USA), and inserts were sequenced. The host 16S rRNA gene sequence was imported into ARB (Ludwig *et al.*, 2004) with the SILVA 98 SSU Ref data set (Pruesse *et al.*, 2007), from which additional sequences for the tree were selected and exported using a 5% similarity filter to remove highly variable positions. A maximum likelihood tree was calculated using a randomly seeded rapid bootstrap analysis ($n = 1000$) and search for best-scoring tree was performed using RAxML (Stamatakis, 2006), version 7.0.4 (MPI master process), with the Generalized Time Reversible γ-model, which optimizes substitution rates and uses a GTR model of nucleotide substitution and a γ-model of rate heterogeneity.

Single-stranded binding (ssb) and terminase large subunit (terL) protein trees. Sequences were retrieved from GenBank and aligned (ClustalW, default parameters) (Larkin *et al.*, 2007). Maximum likelihood tree (described above) was calculated using the Jones-Taylor-Thornton matrix model for protein distance and empirical base frequencies, and bootstrapped ($n = 1000$).

Genome signature analysis

Codon adaptation index (CAI). As its specific host is not sequenced, the relative 'host' codon adaptation of the Pseudoalteromonas Phage H105/1 proteins was modelled using *Pseudoalteromonas atlantica* and *Pseudoalteromonas haloplanktis* genomes as a reference. The CAI of the phage to these bacteria was calculated using Jcat (Grote *et al.*, 2005), which implements the algorithm proposed by Carbone *et al.* to distinguish highly expressed genes based on internal codon biases (Carbone *et al.*, 2003).

Tetra frequencies. To examine tetra correlations, all large sequence fragments (>25 kb, $n = 347\ 886$) were retrieved from GenBank (July 2008). The H105/1 genome was split into 30 fragments (1 kb steps, 10 kb window). Fragments were extended by their reverse complement to account for strand biases. Observed and expected frequencies for the 256 possible tetrads were computed by a maximal-order Markov model; differences between observed and expected frequencies were transformed into *z*-scores (Teeling *et al.*, 2004). To determine the GenBank sequences that are most similar to Phage H105/1 genome fragments, squared *z*-scores were correlated and Pearson's coefficient of all pairwise correlations was calculated (Waldmann, 2010). The coefficient cutoff was determined as the minimum value resulting in a score in each of the 30 windows. A balance between the highest possible correlation coefficients (large window size) and the most granular resolution of tetra signal along the genome (small window size) was achieved using a 10 kb window and 0.61 Pearson's coefficient cutoff (Supplementary Material S4). Correlation scores for each 10 kb genome fragment were normalized, such that the sum of all coefficients of each GenBank fragment type recruited ('bacteria', 'phage' or 'unassigned') was divided by the total sum, and cumulatively mapped onto the Phage H105/1 genome for each window. Considering the window and step sizes, each 1 kb portion of the genome is represented by 10 overlapping 10 kb fragments normalized to one. Thus, the sum of bacteria, phage and unassigned scores will equal 10 at all points along the genome.

Results and discussion

Pseudoalteromonas phage H105/1 biology

Phage H105/1 has a long, noncontractile tail (characteristic of siphoviruses), with unique knobs (Supplementary Material S5A) (Wichels *et al.*,

1998). Infection with Phage H105/1 led to rapid lysis, as evidenced by plaque formation in 10–12 h. Intriguingly, the plaques had clear centres surrounded by fuzzy haloes (Supplementary Material S5B). Haloed plaques are thought to indicate either (i) the presence of a phage-encoded polysaccharide depolymerase (Erskine, 1973; Vandenbergh and Cole, 1986) or (ii) 'pseudolysogeny', a poorly understood condition used to describe the sustained coexistence of a large number of both virus and host. Phage H105/1 does not seem to carry a polysaccharide depolymerase (Figure 2a, Table 1). The latter case, pseudolysogeny, is thought to be caused by stalled or incomplete lysis of the host population, as the phage on infection passively resides in its host, neither integrating nor lysing nor replicating as a plasmid in host progeny (Miller and Day, 2008). Haloed plaques have been observed in other marine phages, in which pseudolysogeny has been implicated: Pseudoalteromonas phages H24/1 and H24/2 (isolated from Helgoland on *Pseudoalteromonas* sp. H24 (Moebus, 1997)) and Listonella phage HSIC (Williamson *et al*., 2001). Its plaque behaviour and λ-like genome content and architecture (Figure 2a) suggest that Pseudoalteromonas phage H105/1 is a temperate phage, able to (or with the past ability to) integrate into its host genome.

Genome features and annotations

Pseudoalteromonas phage H105/1 is 30.7 kb with 52 predicted ORFs. The total G + C% content is 40.85% and the genome-wide coding density is 91%, which is comparable to the average coding density of all marine phages: 89% (data not shown). Over 60% of Phage H105/1 ORFs are unknown (Figure 2b), although the genome organization shows remarkable functional synteny with other λ-like siphoviruses (Figure 2a), which is likely preserved by the temporal control under which phage genes are transcribed (Calendar, 1970). Phage H105/1 has two distinct functional supermodules, whereby the proteins that require direct interaction with the host genome, replicative machinery, metabolic or stress response processes and cell lysis ('host interaction module') are physically separated from those involved in structure and assembly ('phage structural module'; Figure 2a). Intriguingly, of the six most similarly sequenced phages (those sharing the greatest number of proteins), three are marine (Figure 2a), suggesting an overall 'marine' character (Figure 2c).

Phylogenetic signature. Although 33 proteins have no homologues in GenBank (hypothetical proteins), based on their best-blast hits, bacterial homologues show a distinct trend towards the host class (*Gammaproteobacteria*), and the phage homologues are dominated by either *Siphoviridae* (Phage H105/1 class) or prophages (Figure 2b), providing further bioinformatic support that Phage H105/1 integrates as well. Of the 11 phage hits, eight have Gammaproteobacteria hosts. The phylogenetic signature of Phage H105/1 suggests that a majority of its proteins come from a common pool of *Gammaproteobacteria* or phages that infect *Gammaproteobacteria*. Such a host phylogenetic trend has been seen previously in phage genomes (Sullivan *et al*., 2005) and supports the view that phages are mobile genomic extensions of the hosts they infect (Siefert, 2009).

Host interaction: recombination and replication. Containing a MazG pyrophosphohydrolase domain, it is likely that the *ORF 1* gene product is involved in transcriptional repression (Table 1). In *E. coli*, MazG is known to interfere with (or reverse) starvation-induced programmed cell death by decreasing the cellular pool of effector nucleotide, guanosine 3′,5′-bispyrophosphate (ppGpp) (Gross *et al*., 2006). When cyanobacteria are subjected to nitrate starvation, their pool of ppGpp increases and amino acid levels drop (Friga *et al*., 1981), but this process can be impeded by phage infection (Borbély *et al*., 1980). Thus, if functional, a phage MazG protein may help maintain the metabolism of a starving host (Clokie and Mann, 2006; Bryan *et al*., 2008) long enough for the phage to propagate. Of the 12 phage proteins in this domain family, 6 are marine (Figure 2c): Phage H105/1, Roseobacter phage SIO1 and Cyanophages P-SSM2, P-SSM4, S-PM2 and Syn9, suggesting a unique marine signature to this protein family not seen in any other Phage H105/1 protein, and implicating an important role for MazG in marine phage systems.

ORF 6, encoding an ssb protein, is often found in an operon with essential recombination function proteins (ORF 8; Figure 2a). They are known to interact, as essential recombination function specifically binds single-stranded DNA to facilitate phage genome circularization (Poteete *et al*., 1983; Iyer *et al*., 2002). Of all similar ssb proteins in GenBank, the H105/1 ssb clusters most closely with host-like *Alteromonadales* homologues (Figure 3), none of which are from integrated prophages, suggesting that ssb is of host origin. Considering that such a strong host phylogenetic association is not seen in homologues of any other Phage H105/1 ORF (Table 1), and that other phage ssb proteins cluster most closely with those of their host, or host affiliation (Figure 3), single-stranded binding proteins may serve as an informative diagnostic of phage–host associations, especially for temperate phages that could benefit from host-like recombination proteins.

Triggered by (host) stress-inducing environmental conditions, temperate phages rely on a 'genetic switch' to initiate the lytic replication cycle (Ptashne, 2004). The lysogenic state of integrated prophages is maintained by the binding of a repressor protein, which prevents the expression of phage genes needed for lytic replication. Containing a helix-turn-helix domain found in phage and

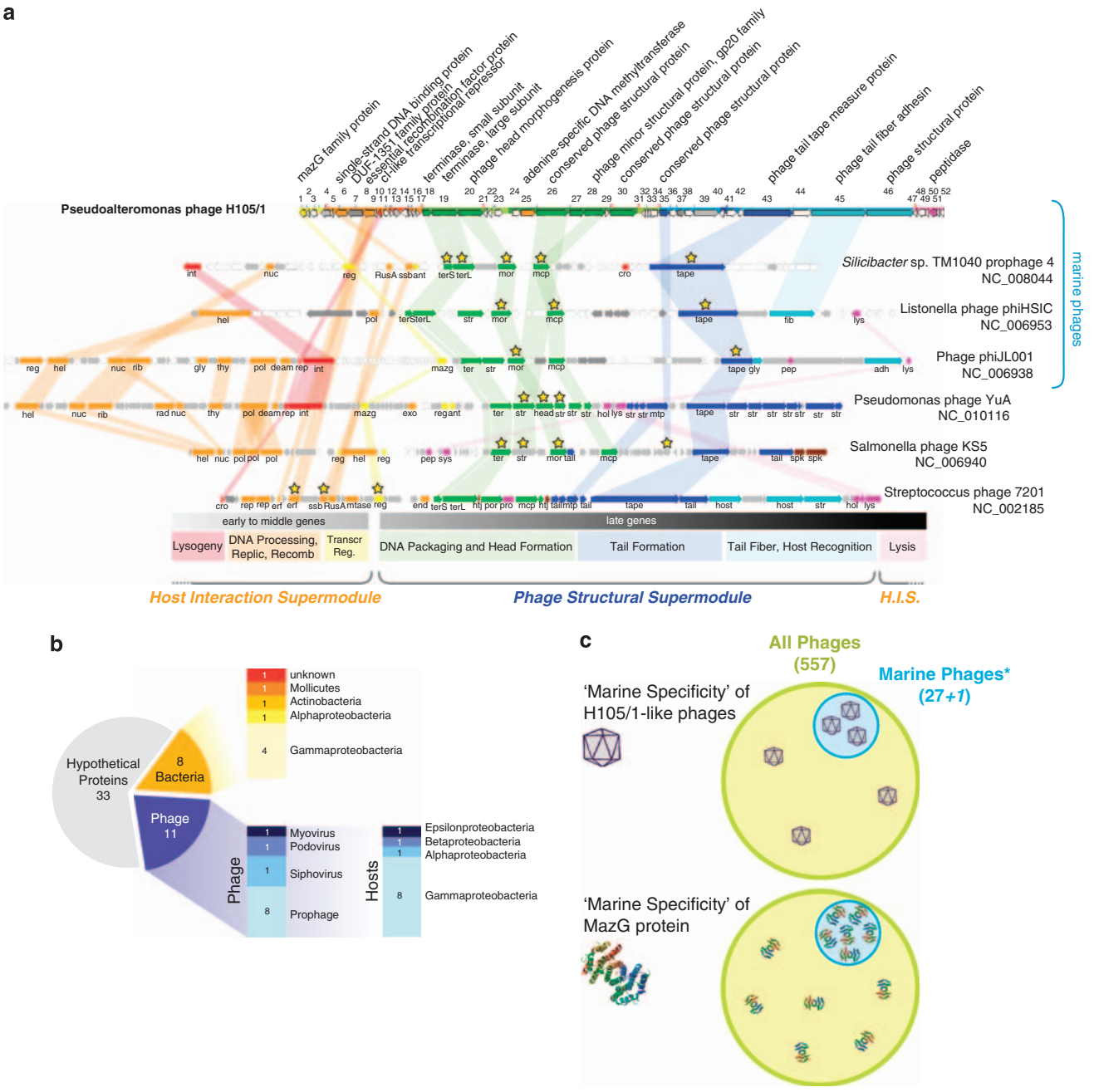


Figure 2 Pseudoalteromonas phage H105/1 genome and synteny with other siphoviruses. Conserved ORFs with homologues in GenBank or those in the phage proteome (Figure 4) are colour coded based on their functional module (labelled in the bottom row); hypothetical proteins (white); conserved hypothetical bacterial proteins (light grey); conserved hypothetical phage proteins (dark grey); promoters (blue lines), transcription terminators (red squiggles); operons are delineated by colour blocks behind the Phage H105/1 genome. Also shown are the six phages sharing the greatest number of homologues with Phage H105/1. Background vertical colour blocks connect phage proteins of similar function, whereas stars indicate explicit Blastp-based sequence similarity to Phage H105/1 ($e < 10^{-5}$). The three 'marine phages' are indicated. Early, middle and late genes describe the temporal transcription of the different modules typical of λ-like phages. The 'Host Interaction Supermodule' and 'Phage Structural Module Supermodule' are indicated as defined in this study. Int, integrase; nuc, exodeoxyribonuclease; reg, transcriptional regulator; RusA; ssb, single-stranded binding protein; ant, antirepressor; terS, terminase, small subunit; terL, terminase, large subunit; ter, terminase; mor, morphogenesis protein; mcp, major capsid protein; cro, repressor; tape, tape measure protein; hel, helicase; pol, DNA polymerase; str, structural protein; fib, tail fibre; lys, lysin; rib, ribonucleotide reductase; gly, glycosyl transferase; thy, thymidylate synthase; deam, deaminase; rep, repressor; mazg; pep, peptidase; adh, adhesin; rad, repair activity DNA; exo, exonuclease; hol, holin; mtp, minor tail protein; spk, tail spike protein; erf, essential recombination function protein; end, endonuclease; mtase, methyltransferase; htj, head-tail joining protein; por, portal; tail, tail protein; host, host recognition protein. **(b)** Best Blastp hits to Pseudoalteromonas phage H105/1 are characterized as hypothetical proteins (no similarity to proteins in NCBI-nr), phage (including manually determined prophages) or bacteria. Phage hits are classified by virus family and host class; bacterial hits are classified by class. **(c)** Overrepresentation of marine phages ($n = 27$) among the six most similar to Phage H105/1, relative to all available phage genomes ($n = 557$), and overrepresentation of marine phages among all phages containing the MazG protein domain. *One marine prophage is considered among the marine phages.

Table 1 Pseudoalteromonas phage H105/1 ORF and annotation summary based on homology to NCBI-nr, Pfam and MeGO databases

ORF	*AA*	*Annotation*	*e-Value (% id/% orf coverage) BlastP (taxonomy); accn*	*Pfam; MeGO*
1/−	103	Transcriptional repressor, MazG family protein	4.0E-12 (46/83) *Nitratiruptor* sp. SB155-2 prophage (Epsilonproteobacteria); YP_001355920	PF03819: MazG, 3.7E-2; MeGO: transcriptional repressor activity (phi:0000127), maintenance of prophage immunity (phi:0000057)
2/−	64	Hypothetical protein	Non-significant.	
3/−	109	Hypothetical protein	Non-significant	
4/−	156	Conserved hypothetical protein	3.0E-19 (38/96) *Vibrio cholerae* V51 (Gammaproteobacteria; Vibrionales); ZP_01487122	
5/−	59	Hypothetical protein		
6/−	184	single-stranded DNA binding protein	9.0E-43 (71/61) Alteromonadales bacterium TW-7 (Gammaproteobacteria; Alteromonadales); ZP_01613173	PF00436: SSB, 1.6E-40; MeGO: single-stranded DNA binding (GO:0003697)
7/−	239	Conserved hypothetical phage protein, DUF1351 family	9e-05 (47) Iodobacteriophage phiPLPE (Myovirus)	PF07083: DUF1351 protein family of unknown function, 1E-1; MeGO: phage function unknown (phi:0000326)
8/−	191	Essential recombination function protein	2.0E-40 (63/72) Enterobacteria phage P22 (Podovirus), host: *Salmonella enterica* serovar typhimurium; NP_059596	PF04404: ERF superfamily, 2.2E-28; MeGO: DNA strand annealing activity (GO:0000739), DNA recombination (phi:0000130)
9/−	34	Hypothetical protein	Non-significant	
10/−	62	Transcriptional repressor	7.7E-2 (46) Bacteriophage APSE-2, host: Candidatus Hamiltonella defensa; ACJ10163	PF01381: HTH_3, 9.7E-2; MeGO: transcriptional repressor activity (phi:0000127), maintenance of prophage immunity (phi:0000057)
11/+	91	Hypothetical protein	Non-significant	
12/+	65	Hypothetical protein	Non-significant	
13/+	101	Conserved hypothetical protein	5.0E-4 (24/77) Pseudomonas fluorescens Pf0-1; YP_349064	
14/−	89	Hypothetical protein	Non-significant	
15/+	55	Hypothetical protein	Non-significant	
16/+	63	Hypothetical protein	Non-significant	
17/+	56	Hypothetical protein	Non-significant	
18/+	153	Phage terminase, small subunit	8.0E-17 (52/62) *Yersinia enterocolitica* subsp. Enterocolitica 8081 prophage (Gammaproteobacteria; Enterobacteriales); YP_001006550	PF03592: terminase small subunit; 1.9E-13; MeGO: phage terminase small subunit (phi:0000074), DNA binding activity (phi:0000109), phage DNA maturation (phi:0000019)
19/+	415	Phage terminase, large subunit	9.0E-135 (57/97) *Silicibacter* sp. TM1040 prophage (Alphaproteobacteria; Rhodobacterales); YP_612796	PF03237: terminase-like family, 1.9E-31; MeGO: phage terminase large subunit (phi:0000073); phage DNA maturation (phi:0000019)
20/+	388	Phage head morphogenesis protein	Pseudomonas phage YuA (Siphovirus); YP_001595877	PF04233: phage Mu protein F like, 1.2E-2; MeGO: phage head or capsid minor protein (phi:0000185)
21/−	59	Hypothetical protein	Non-significant	
22/−	81	Hypothetical protein	Non-significant	
23/+	114	Hypothetical protein	Non-significant	
24/+	127	Hypothetical protein	Non-significant	
25/+	229	Adenine-specific DNA methyltransferase	9.0E-31 (40/97) *Spiroplasma citri* poss. degenerate prophage (Tenericutes; Mollicutes; Entomoplasmatales); CAK98777	PF01555: N6_N4_Mtase; 1.3E-30; MeGO: DNA methyltransferase activity (phi:0000117)
26/+	489	Conserved phage structural protein	1.0E-33 (27/94) *Pseudomonas fluorescens* Pf-5 prophage (Gammaproteobacteria; Pseudomonadales); YP_260866	MeGO: phage function unknown (phi:0000326)
27/+	237	Phage minor structural protein GP20 family	Non-significant	PF06810: phage minor structural protein GP20 family; 3.2E-3
28/+	320	Conserved phage structural protein	**?** *Delftia acidovorans* SPH-1 prophage (Betaproteobacteria; Burkholderiales); YP_00156426	
29/+	54	Hypothetical protein	No sig.	
30/+	411	Conserved hypothetical phage protein	1.0E-4 (21/89) Vibrio phage KVP40 (Myovirus); NP_899611	
31/−	77	Hypothetical protein	Non-significant	
32/−	61	Hypothetical protein	Non-significant	
33/−	61	Hypothetical protein	Non-significant	

Table 1 (Continued)

ORF	*AA*	*Annotation*	*e-Value (% id/% orf coverage) BlastP (taxonomy); accn*	*Pfam; MeGO*
34/−	134	Hypothetical protein	Non-significant	
35/+	161	Conserved phage structural protein	5.1E-2 (31/78) Pseudomonas phage M6 (Siphovirus); YP_001294532	MeGO: phage function unknown (phi:0000326)
36/+	119	Hypothetical protein	No sig.	
37/+	127	Conserved hypothetical phage protein	4.0E-6 (33/83) Salmonella phage KS7 (Siphovirus)	MeGO: phage function unknown (phi:0000326)
38/+	140	Hypothetical protein	NA	
39/+	391	Conserved hypothetical protein	3.0E-4 (25/74) Alpha proteobacterium BAL199 (Proteobacteria; Alphaproteobacteria); ZP_02186593	
40/−	80	Hypothetical protein	Non-significant	
41/+	154	Hypothetical protein	Non-significant	
42/+	91	Hypothetical protein	Non-significant	
43/+	767	Phage tail tape measure protein	9.0E-19 (33/31) *Verminephrobacter eiseniae* EF01-2 poss. degenerate prophage (Burkholderiales); YP_999425	MeGO: phage tail tape measure protein (phi:0000086)
44/+	292	Hypothetical protein	Non-significant	
45/+	867	Phage tail fibre adhesin Gp38 family protein	Non-significant	PF05268: phage tail fibre adhesin Gp38, 3.E-2
46/+	747	Phage structural protein	Non-significant	
47/−	53	Hypothetical protein	Non-significant	
48/−	126	Hypothetical protein	Non-significant	
49/−	61	Hypothetical protein	Non-significant	
50/−	114	Carboxypeptidase, peptidase M15 family protein	3.0E-15 (36/99) *Magnetococcus* sp. MC-1 poss. degenerate prophage (Proteobacteria); YP_865602	PF08291: Peptidase M15, 3.4E-20; MeGO: carboxypeptidase activity (GO:0004180)
51/−	51	Hypothetical protein	Non-significant	
52/−	83	Hypothetical protein	Non-significant	

Abbreviations: MeGO, Mobile Element Gene Ontology; NA, non-available; ORF, open reading frame.

plasmid transcription control proteins, and with homology to putative phage cI proteins, the ORF 10 gene product may be involved in cI repressor-like activity (Table 1).

Phage structure and DNA packaging. The structural proteins of the mature Phage H105/1 virion were analysed. Seven proteins of the 'phage structural module' were identified in the phage structural proteome (Figure 4, Supplementary Material S6): a phage head morphogenesis protein (ORF 20), a phage tail tape measure protein (ORF 43), a phage tail fibre adhesin (ORF 45) and four novel proteins (ORFs 26, 28, 35 and 46) that are now experimentally verified as structural proteins.

The 'phage structural supermodule' of Phage H105/1, responsible for phage assembly and structure, is syntenous with the morphogenetic operon of other temperate phages and prophages (Figure 2a) (Botstein and Matz, 1970; Canchaya *et al*., 2003). Typical of a λ-like morphogenetic operon (Casjens, 2003), ORF 20, encoding a putative head morphogenesis protein, is found in the 'DNA packaging and head formation' module with genes encoding the large and small terminases (ORFs 18 and 19), ATP-binding proteins that cut the concatenated phage DNA to prepare it for packaging (Black, 1989) (Figure 2a). The large terminase protein sequence of H105/1 falls outside the known function-based phage terminase clusters (Figure 3b), and is most related to the terminase of a marine Silibacter prophage, with similarity to other known lambdoid phages (SO-1 and KS5). ORF 27 shares a domain with the Staphylococcus phage-dominated minor structural protein Gp20 family (PF06810), not to be confused with the synonymous T4-like capsid assembly protein Gp20 (PF07230), a common cyanophage marker gene (Zhong *et al*, 2002). Among the '*tail formation'* genes, ORF 43, encoding a phage tail length tape measure protein, is involved in the regulation of the phage tail length (Abuladze *et al*., 1994). In the 'tail fibre, host recognition' module, ORF 45 contains a domain of the Phage tail fibre adhesin Gp38 Pfam family (Table 1). In T2-like phages, gp38 is responsible for recognition of host cell receptors (Haggard-Ljungquist *et al*., 1992); it is thus considered to be a critical factor for change and is one of the most rapidly evolving components of a phage–host system.

The presence of ORF 25 (Table 1), a methyltransferase gene, among the Phage H105/1 structual 'late genes' (rather than in a DNA modification module of an 'early' operon (Figure 2a) (Mobberley *et al*., 2008)), suggests that the enzyme does not methylate *incoming* phage DNA at the time of infection or insertion in an attempt to mask itself from host restriction enzymes. An alternative strategy, also proposed in Bacteriophage N15 (Ravin *et al*., 2000), may exist: as new virions are assembled during the lytic phase, the replicated DNA is methylated before packaging.

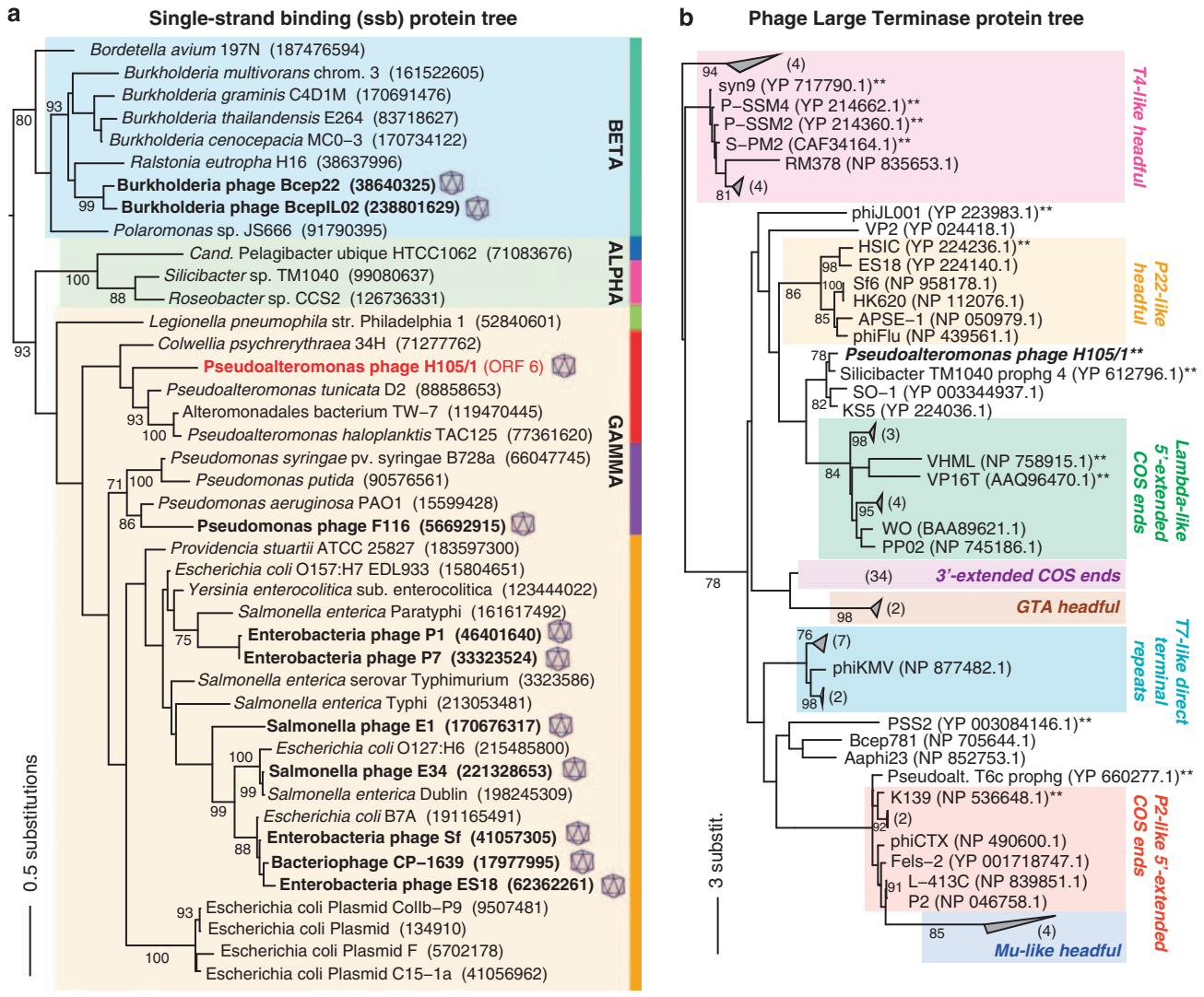


Figure 3 Consensus maximum likelihood trees for two Phage H105/1 genes, generated from 1000 bootstrapped resampled versions of the original data set using the JTT matrix model for protein distance measures. Bootstrap values >75 shown on the branches. (**a**) ORF 8, single-stranded binding (ssb) protein sequence tree. A phage capsid symbol and bold print denote phage sequences. Note that *ssb ORF 6* of Pseudoalteromonas phage H105/1 groups with nonprophage (manually determined by examining gene neighbourhood in the bacterial genome) ssb proteins of *Pseudoalteromonas tunicata* D2, *Pseudoalteromonas haloplanktis* TAC125 and Alteromonadales bacterium TW-7. All other phage ssb proteins group with their host, or with a closely related organism, whereas plasmids cluster independently. NCBI genInfo identifiers (GI numbers) follow each protein in parentheses. ALPHA, BETA, GAMMA denotes the class of Proteobacteria per cluster. Horizontal colour bars indicate order (turquoise: *Burkholderiales*, blue: *Rickettsiales*, magenta: *Rhodobacteriales*, green: *Legionellales*, red: *Alteromonadales*, purple: *Bacilliales*, orange: *Enterobacteriales*). (**b**) ORF 18, phage large terminase (terL) protein sequence tree. Large terminase genes are characterized on the basis of phage DNA packaging mechanisms (Casjens *et al.*, 2005). Double asterisks (**) indicate marine phages or prophages. Parentheses hold either GenBank accession numbers or number of proteins in collapsed nodes.

Host lysis. Host lysis requires both a phage lysin and holin to dissolve the membrane potential and permeabilize the cell wall, respectively (Wang *et al.*, 2000). ORF 50 contains a conserved domain of the Peptidase M15 Pfam family of metallopeptidases (Table 1), lysins likely involved in host cell lysis.

Ecogenomics: H105/1 in GOS and five marine virus metagenomes

Of the 52 ORFs, 14 have homologues in samples from the GOS data set (Figure 5a). These genes, many of which are found in the 'host interaction supermodule' (ORFs 1, 6, 8, 50), are seen proportionately more in the GOS 'Estuary' sites, with the most hits (63) to Delaware Bay (NJ, USA). Among the marine virus metagenomes, there are proportionally more hits to the British Columbia samples ('BBC', Figures 5b and d), a trend again strongest in the 'host interaction supermodule'. The British Columbia surface site (lower than average salinity) clusters with Helgoland, also a lower saline, turbid region of the North Sea, influenced by the Elbe river plume (Becker *et al.*, 1992) (Figures 5c and e; Supplementary Material S2). The BBC surface environmental parameters are most diagnostic of the BBC virome,

as nearly all of the 86 pooled samples were from the upper water column (C. Suttle, personal communication). Furthermore, when nine other marine phage genomes, including 'H105/1-like' marine phages (Figure 2), were subjected to the same analysis, Phage H105/1 proteins were found to be among the most abundant in the BBC sample (Figure 5d). When the virome sites and all 10 phage isolation sites ('*x:hab*', Figure 5e) were analysed on the basis of variation in their environmental parameters, the H105/1 habitat falls most similar to surface BBC and GOM sites, which are all negatively associated with salinity (Figure 5e).

This ecogenomic trend may reflect the original habitat of Phage H105/1 and further intimates the importance of temperate phages in offering genome plasticity to lysogens (hosts with integrated prophages) in unstable habitats, a concept also suggested by a high prevalence of lysogens among microbial populations in a low saline, highly turbid Mississippi River plume (Long *et al*., 2008). In light of the assumption that there are site-specific differences in the relative abundances of virus sequences between viromes (Angly *et al*., 2006), these biases may be influenced by environmental parameters. The estuary-enriched hits tend to be found in the 'host interaction module', which may represent the mechanism through which 'phage organismal ecology' can exist. The adaptation of a phage to its environment happens through close association with its host. Through, for example, phage-mediated transcriptional regulation, a phage can respond to the metabolic state of its host as the host directly responds to its environment.

Painting a genome landscape: codon adaptation and tetra usage

As a large portion of Phage H105/1 genes are new, or have only very distant homology, few evolutionary relationships based on sequence similarity alone can be established to describe the history of phage

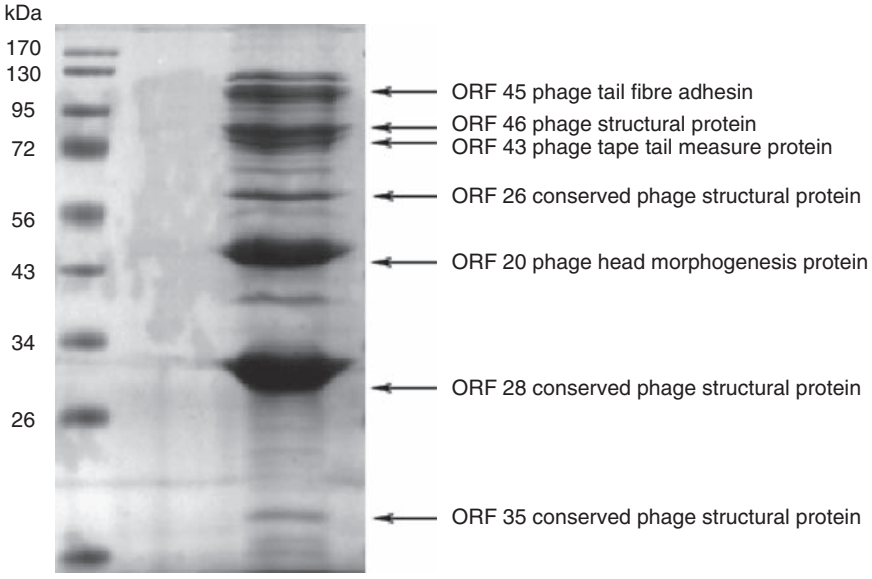
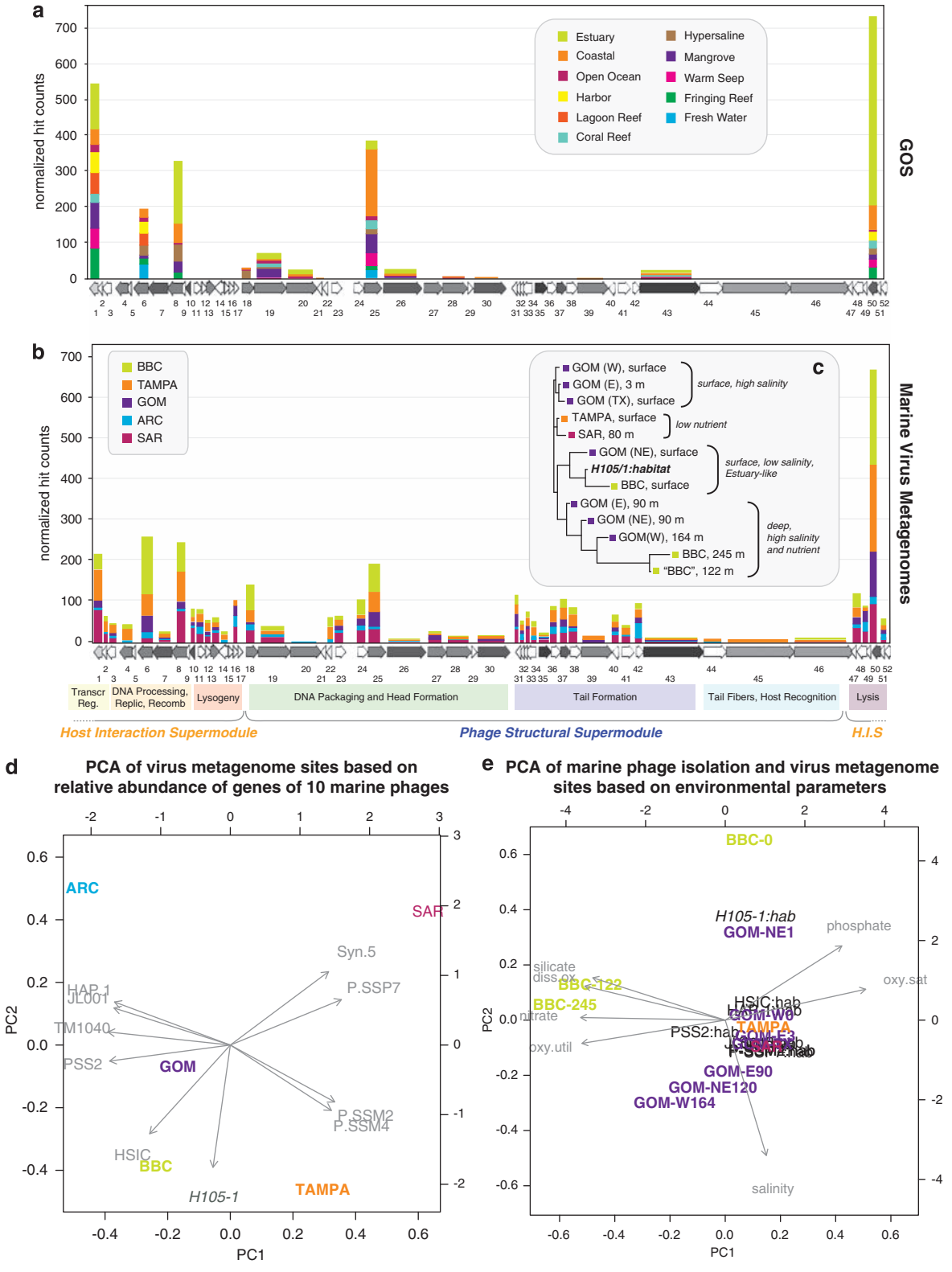


Figure 4 Pseudoalteromonas phage H105/1 structural proteome. Sodium dodecyl sulphate-polyacrylamide gel electrophoresis gel image of Pseudoalteromonas phage H105/1 lysate containing mature phage particles. ORFS 26, 28, 35, 43, 45 and 46, which were previously unknown proteins (hypothetical or conserved hypothetical), are annotated as phage structural proteins on the basis of this proteomic confirmation.

Figure 5 Presence of Pseudoalteromonas phage H105/1 proteins in GOS microbial metagenomes and marine virus metagenomes (viromes). (**a**) Normalized tblastn hit counts to the GOS metagenomes grouped by habitat type, as defined in the original data set. (**b**) Normalized tblastn hit counts to five marine virus metagenomes. (**c**) Neighbour-joining tree clustering virome sample sites based on their environmental parameters as interpolated by megx.net GIS tools. Pooled samples that represent a range of depths were treated as independent sites; depths used for data interpolation are indicated, following the sample name. Overall, sites cluster most strongly by salinity and depth; as such, Helgoland clusters with BBC and northeastern GOM sites, all of which have lower than average salinity and are influenced by major river outflows. Note that data interpolation for Arctic sites is not possible, thus ARC is missing from this depiction. See Supplementary Material S2 for precise coordinates and depths used to retrieve interpolated data. (**d**) Differences between the virus metagenomes (names in colour) on the basis of the total hits of 10 marine phage genomes (grey vectors) are depicted by principal component analysis (PCA; 95.3% of variation explained in the first two components). Note the high relative abundance of H105/1 proteins in the BBC and TAMPA sites, whereas other H105/1-like phages (namely, JL001 and TM1040) fall elsewhere. (**e**) Virus metagenome sites and isolation sites of 10 marine phages ('*x:hab*') are projected on the basis of PCA (95.2% of variation explained in the first two components) of their interpolated environmental parameters. Note the close association between the H105/1 habitat and the surface BBC and GOM samples, as well as their negative association with salinity. diss.ox, dissolved oxygen; oxy.sat, oxygen saturation; oxy.util, oxygen utilization. The marine phages and environmental parameters can be found at www.megx.net/genomes/phages.

proteins and supermodules. Thus, taking an alignment-free approach, we investigate patterns of codon adaptation and tetra frequencies across the genome to study the evolutionary history of Phage H105/1.

Host-indexed CAI. Phage genomes are under codon-selective pressure imposed by the translational biases of their microbial hosts (Carbone, 2008; Lucks *et al.*, 2008; Bahir *et al.*, 2009). The λ-phage genome 'landscape' is subdivided into peaks and



valleys on the basis of its CAI, which reflects the adaptation of each gene to the codon bias of its host (Lucks *et al*., 2008). In the absence of a host genome sequence, Phage H105/1 CAI was calculated on the basis of preferred codon usage of two sequenced *Pseudoalteromonas* spp. (Figure 6). Previous studies have found that genes with the greatest host-indexed CAI (genes most resembling host codon bias) encode phage structural proteins, that is, capsid (Carbone, 2008) and tail genes (Lucks *et al*., 2008). They presume that proteins that are produced rapidly *en mass* during lytic growth most resemble codon usage of their host because of the selection for translational efficiency, the fundamental force thought to drive codon bias in single-cell organisms (Sharp and Li, 1987). The relatively recent observation of Lucks *et al* (2008) is supported by a similar pattern found in select structural proteins of the Phage H105/1 head (ORFs 26–28) and tail (ORFs 34–36 and 39) formation modules (Figure 6). Carbone also found genes responsible for host interaction, inhibition of host functions, single-stranded DNA binding and transcriptional regulation to be strongly host biased (Carbone, 2008), which is also seen in the respective proteins of Phage H105/1's 'host interaction supermodule' (Figure 6).

Tetra frequency correlation. We ask: 'given a portion of the H105/1 genome, which sequences (of the roughly 350 000 large fragments in GenBank) are most correlated based on their respective tetranucleotide frequencies?' Whole-genome tetra frequencies have been shown to correlate with whole genomes of their hosts (Pride *et al*., 2006). However, when examined on a finer scale, only certain portions of a phage genome were found to retain tetra frequencies of their host. In this study, tetra frequency correlations of fragments of a phage genome allow for a more resolved look at the influences on nucleotide adaptation along a genome, thus shedding light on the history of the parts of the concatenated whole.

We found a predominately phage tetra signature across the entire genome (Figure 6). Intriguingly, tetra peaks and valleys coincide with the codon adaptation genome landscape: regions with greatest 'host' codon adaptation have the greatest bacterial tetra signal, whereas regions of low adaptation peak in phage tetra signature (Figure 6; see Supplementary Information 1 for a description of the unassigned fragments). These patterns likely reflect alternative, mutually inclusive selective forces that have an effect on different signatures. When codon usage is biased, codon adaptation reflects selection based on mechanistic properties of efficient translation, whereas tetra frequencies, although poorly understood, are likely to be influenced (a) by stochastic processes that accumulate through time (Pride *et al*., 2006), and (b) by restriction-modification-related processes through the avoidance of restriction sites (Pride *et al*., 2003). However, codon usage and tetra frequency are inevitably intertwined through their coupled reliance on the same nucleotides. A convincing correlation between tetra

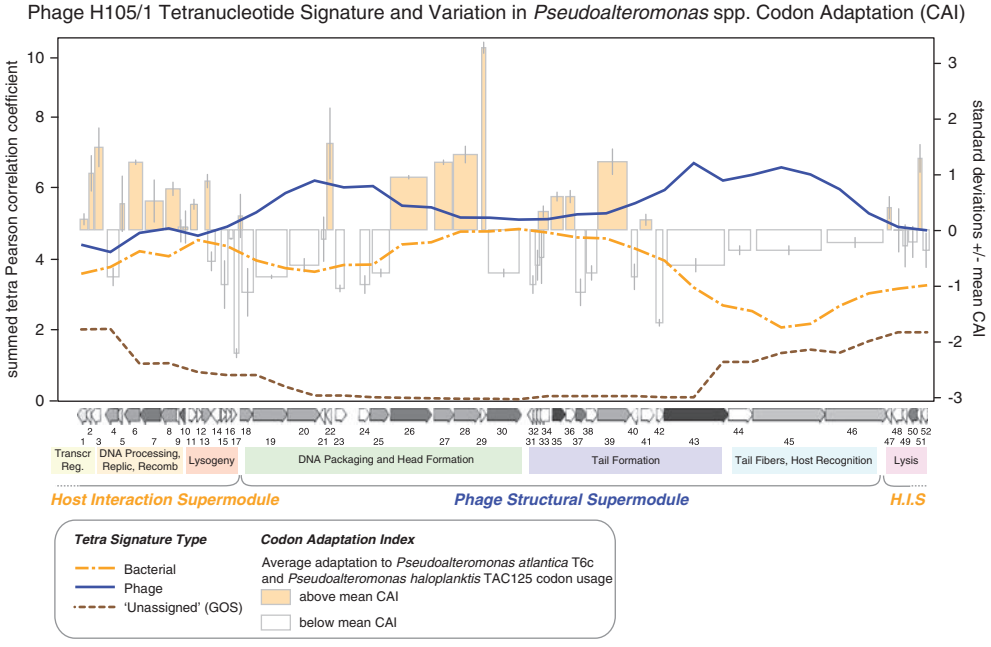


Figure 6 Pseudoalteromonas H105/1 genome signatures. The left axis denotes cumulative tetra correlations between 10 kb genome fragments of Phage H105/1 and large GenBank fragments, including the origin (bacterial, phage, or unassigned) of the GenBank fragment. The right axis quantifies the degree of codon adaptation around the mean for Phage H105/1 genes indexed to the sequenced *Pseudoalteromonas* spp. Proteins with codon usage more adapted to *Pseudoalteromonas* spp. bias have positive values and are labelled in orange. Error bars indicate the standard deviation of the two averaged *Pseudoalteromonas* spp. CAI. The relevance of unassigned (GOS) is described in Supplementary Information 1.

frequencies and preferential codon usage has been observed in genomes assembled from environmental communities as well (Dick *et al.*, 2009).

Such a resolved impression of tetra correlations within a phage genome has led to the observation that host codon adaptation and nucleotide composition along the genome are associated, but why? Phage proteins will remain 'associated with' a certain host by persisting among the new combinations of genes that make up phages infecting it, as gene flow predominately occurs between phages that infect the same host (Figure 2b) (Duffy and Turner, 2008). Guided by selection, some proteins of the combination will differ; some will remain the same. Although phage fitness is influenced by several factors on many levels (Duffy and Turner, 2008), we take advantage of the fact that codon adaptation is a selective force observed at the sequence level. As such, the Phage H105/1 proteins of greatest codon adaptation may be the proteins that are selected to remain 'associated' with its *Pseudoalteromonas* sp. host. As such, they have longer residence time with their host, and thus have the time to ameliorate a host or bacteria-like tetra signature (Pride *et al.*, 2006). Contrarily, nonadapted proteins are under less selective pressure to remain associated with a specific host, and, as many are structural proteins highly conserved in other phages (Table 1), may be more mobile components of a greater phage protein pool. In the absence of host amelioration, these proteins retain a phage tetra signature common to the phage pool. Thus, the bacterial or phage tetra pattern could reflect different stages of amelioration. However, little is known of how mutation rates differ in different portions of phage genomes (Duffy and Turner, 2008), nor how rates of swapping may differ between phage functional modules, which remains among the most compelling questions pertaining to mechanisms of phage evolution.

Phage diversity and evolution in light of marine phage H105/1

The ecogenomic and evolutionary influences on the Phage H105/1 genome content are highlighted by the phylogenetic signature of its functional annotations, the global distribution of its protein-coding genes, as well as by codon adaptation and tetra frequency correlations. These approaches (i) extend beyond the commonly searched public databases; (ii) take advantage of the invaluable environmental context of the sequenced organisms, a frequently neglected asset (Field, 2008) that, through integration with marine ecology, will shed light on the hidden pool of phage functional diversity; and (iii) are not restricted by limitations of sequence similarity. When integrated with experimental approaches (i.e., proteomic validation), such analyses enrich our ecological and evolutionary understanding of phage genomics, especially valid considering the nearly 10-fold increase in marine phage genome sequences that will soon be available (Broad Institute, 2010).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

MBD is grateful to Amelia Rotaru for SDS-PAGE assistance, Marianne Jacob for contributions to tetra analysis, Matthew B Sullivan and Vincent Denef for critically reading the paper, and Matthias Ullrich, Renzo Kottmann and Ivaylo Kostadinov for fruitful discussions and support. A Marie Curie Early Stage Training Fellowship to MBD supports funding for this project (MEST-CT-2004-007776) through the MarMic program of the MPI for Marine Microbiology. Further acknowledgement goes to two anonymous reviewers who offered useful suggestions, strengthening this analysis and paper.

References

Abuladze NK, Gingery M, Tsai J, Eiserling FA. (1994). Tail length determination in bacteriophage T4. *Virology* **199**: 301–310.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andersson AF, Banfield JF. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**: 1047–1050.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Bahir I, Fromer M, Prat Y, Linial M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* **5**: 311.

Becker GAB, Dick SD, Dippner JWD. (1992). Hydrography of the German Bight. *Mar Ecol Prog Ser* **91**: 9–19.

Besemer J, Lomsadze A, Borodovsky M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607–2618.

Black LW. (1989). DNA Packaging in dsDNA Bacteriophages. *Annu Rev Microbiol* **43**: 267–292.

Borbély G, Kaki C, Gulyás A, Farkas GL. (1980). Bacteriophage infection interferes with guanosine 3′-diphosphate-5′-diphosphate accumulation induced by energy and nitrogen starvation in the cyanobacterium *Anacystis nidulans*. *J Bacteriol* **144**: 859–864.

Botstein D. (1980). A theory of modular evolution for bacteriophages. *Ann NY Acad Sci* **354**: 484–491.

Botstein D, Matz MJ. (1970). A recombination function essential to the growth of bacteriophage P22. *J Mol Biol* **54**: 417–440.

Broad Institute (2010). Marine Phage Sequencing Project##http://www.broadinstitute.org/annotation/viral/Phage.

Bryan MJ, Burroughs NJ, Spence EM, Clokie MR, Mann NH, Bryan SJ. (2008). Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. *PLoS One* **3**: e2048.

Bushman F. (2002). *Lateral DNA Transfer: Mechanisms and Consequences*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Calendar R. (1970). The regulation of phage development. *Annu Rev Microbiol* **24**: 241–296.

Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. (2003). Prophage genomics. *Microbiol Mol Biol R* **67**: 238.

Carbone A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* **66**: 210–223.

Carbone A, Zinovyev A, Képès F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19**: 2005–2015.

Casjens S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* **49**: 277–300.

Casjens SR, Gilcrease EB, Winn-Stapley DA, Schicklmaier P, Schmieger H, Pedulla ML. *et al.* (2005). The generalized transducing Salmonella bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J Bacteriol* **187**: 1091–1104.

Clokie MR, Mann NH. (2006). Marine cyanophages and light. *Environ Microbiol* **8**: 2074–2082.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Duffy S, Turner PE. (2008). Phage evolutionary biology. In: Abedon, S. (ed). *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Cambridge University Press: Cambridge, UK, pp 147–176.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* **2**: 953–971.

Erskine JM. (1973). Characteristics of Erwinia amylovora bacteriophage and its possible role in the epidemology of fire blight. *Can J Microbiol* **19**: 837–845.

Field D. (2008). Working together to put molecules on the map. *Nature* **453**: 978.

Filée J, Bapteste E, Susko E, Krisch HM. (2006). A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol* **23**: 1688–1696.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE *et al.* (2010). The Pfam protein families database. *Nucleic Acids Res* **38**: D211–D222.

Friga GM, Borbély G, Farkas GL. (1981). Accumulation of guanosine tetraphosphate (ppGpp) under nitrogen starvation in *Anacystis nidulans*, a cyanobacterium. *Arch Microbiol* **129**: 341–343.

Gross M, Marianovsky I, Glaser G. (2006). MazG—a regulator of programmed cell death in *Escherichia coli*. *Mol Microbiol* **59**: 590–601.

Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC *et al.* (2005). JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res* **33**: W526–W531.

Haggard-Ljungquist E, Halling C, Calendar R. (1992). DNA sequences of the tail fiber genes of bacteriophage P2: evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. *J Bacteriol* **174**: 1462.

Hendrix RW, Smith M, Burns RN, Ford ME, Hatfull GF. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world'sa phage. *PNAS* **96**: 2192.

Iyer LM, Koonin EV, Aravind L. (2002). Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. *BMC Genomics* **3**: 8.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W *et al.* (2010). Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391–D395.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**: 863–865.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–89.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *PNAS* **101**: 11013.

Long A, McDaniel LD, Mobberley J, Paul JH. (2008). Comparison of lysogeny (prophage induction) in heterotrophic bacterial and Synechococcus populations in the Gulf of Mexico and Mississippi River plume. *ISME J* **2**: 132–144.

Lucchini S, Desiere F, Brüssow H. (1999). Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. *J Virol* **73**: 8647–8656.

Lucks JB, Nelson DR, Kudla GR, Plotkin JB. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* **4**: e1000001.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

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**: 355–363.

McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F *et al.* (2008). Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS One* **3**: e3263.

Médigue C, Krin E, Pascal G, Barbe V, Bernsel A, Bertin PN *et al.* (2005). Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res* **15**: 1325–1335.

Miller RV, Day MJ. (2008). Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology. In: Abedon, S. (ed). *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Cambridge University Press: Cambridge, UK, pp 114–143.

Mobberley JM, Authement RN, Segall AM, Paul JH. (2008). The temperate marine phage PhiHAP-1 of *Halomonas aquamarina* possesses a linear plasmid-like prophage genome. *J Virol* **82**: 6618–6630.

Moebus K. (1992). Further investigations on the concentration of marine bacteriophages in the water around Helgoland, with reference to the phage-host systems encountered. *Helgoland Mar Res* **46**: 275–292.

Moebus K. (1997). Investigations of the marine lysogenic bacterium H24. 2. Development of pseudolysogeny in nutrient rich broth. *Mar Ecol Prog Ser* **148**: 229–240.

Muyzer G, Teske A, Wirsen CO, Jannasch HW. (1995). Phylogenetic relationships of Thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Arch Microbiol* **164**: 165–172.

Paul JH, Sullivan MB, Segal AM, Rohwer F. (2002). Marine phage genomics. *Comp Biochem Phys* **133**: 463–476.

Paul JH, Williamson SJ, Long A, Authement RN, John D, Segall AM *et al.* (2005). Complete genome sequence of phiHSIC, a pseudotemperate marine phage of Listonella pelagia. *Appl Environ Microbiol* **71**: 3311–3320.

Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**: 171–182.

Poteete AR, Sauer RT, Hendrix RW. (1983). Domain structure and quaternary organization of the bacteriophage P22 Erf protein. *J Mol Biol* **171**: 401–418.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**: 145–158.

Pride DT, Wassenaar TM, Ghose C, Blaser MJ. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* **7**: 8.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Ptashne M. (2004). *A Genetic Switch*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Ravin V, Ravin N, Casjens S, Ford ME, Hatfull GF, Hendrix RW. (2000). Genomic sequence and analysis of the atypical temperate bacteriophage N15. *J Mol Biol* **299**: 53–73.

Richter M, Lombardot T, Kostadinov I, Kottmann R, Duhaime MB, Peplies J *et al.* (2008). Jcoast—a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics* **9**: 177.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Sambrook J, Russell DW. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.

Sharp PM, Li WH. (1987). The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281–1295.

Siefert JL. (2009). Defining the mobilome. *Methods Mol Biol* **532**: 13–27.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**: e144.

Suttle CA. (2007). Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**: 938–947.

Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, Penesyan A *et al.* (2008). Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. *PLoS One* **3**: e3252.

Toussaint A, Lima-Mendez G, Leplae R. (2007). PhiGO, a phage ontology associated with the ACLAME database. *Res Microbiol* **158**: 567–571.

Vandenbergh PA, Cole RL. (1986). Cloning and Expression in Escherichia coli of the Polysaccharide Depolymerase Associated with Bacteriophage-Infected *Erwinia amylovora*. *Appl Environ Microbiol* **51**: 862–864.

Waldmann J. (2010). ocount2: a library for the calculation and comparison of oligonucleotide patterns##http://www.promedici.de/ocount2.

Wang IN, Smith DL, Young R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu Rev Microbiol* **54**: 799–825.

Wichels A, Biel SS, Gelderblom HR, Brinkhoff T, Muyzer G, Schütt C. (1998). Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* **64**: 4128–4133.

Wichels A, Gerdts G, Schütt C. (2002). *Pseudoalteromonas* spp. phages, a significant group of marine bacteriophages in the North Sea. *Aquat Microb Ecol* **27**: 233–239.

Williamson SJ, McLaughlin MR, Paul JH. (2001). Interaction of the PhiHSIC virus with its host: lysogeny or pseudolysogeny? *Appl Environ Microbiol* **67**: 1682–1688.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO. *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Zhong Y, Chen F, Wilhelm SW, Poorvin L, Hodson RE. (2002). Phylogenetic Diversity of Marine Cyanophage Isolates and Natural Virus Communities as Revealed by Sequences of Viral Capsid Assembly Protein Gene gp20. *AEM* **68**: 4.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)