OSU Oregon State UNIVERSITY

Open Access Articles

# A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages

The Faculty of Oregon State University has made this article openly available. Please share how this access benefits you. Your story matters.



| | |
|---|---|
| Citation | Dreher, T. W., Brown, N., Bozarth, C. S., Schwartz, A. D., Riscoe, E., Thrash, C., Bennett, S. E., Tzeng, S.-C. and Maier, C. S. (2011), A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. Environmental Microbiology, 13: 1858–1874. doi:10.1111/j.1462-2920.2011.02502.x |
| DOI | 10.1111/j.1462-2920.2011.02502.x |
| Publisher | John Wiley & Sons Ltd. |
| Version | Accepted Manuscript |
| Terms of Use | http://cdss.library.oregonstate.edu/sa-termsofuse |



Oregon State UNIVERSITY OSU

1 **A freshwater cyanophage whose genome indicates close relationships to photosynthetic**
2 **marine cyanomyophages.**

3 Theo W. Dreher[1,2*],Nathan Brown[1,†], Connie S. Bozarth[1], Andrew D. Schwartz[1], Erin Riscoe[1], J.
4 Cameron Thrash[1], Samuel E. Bennett[3], Shin-Cheng Tzeng[4] and Claudia S. Maier[4]
5 [1]Department of Microbiology, [2]Center for Genome Research and Bioinformatics, [3]Department of
6 Environmental & Molecular Toxicology, [4]Department of Chemistry, Oregon State University,
7 Corvallis, Oregon 97331.
8
9
10
11
12
13
14
15 Keywords: *Myoviridae*, Klamath River, cyanomyophage, freshwater phage, *nusG, spoT*, phage
16 tRNA, T4-like phage
17
18
19
20
21 ☐Corresponding author: Mailing address: Theo W. Dreher, Department of Microbiology, Oregon
22 State University, Corvallis, Oregon 97331-3804. Phone: (541) 737-1795. Fax: (541) 737-0496. E-
23 mail: theo.dreher@oregonstate.edu
24
25 † Present address: Pittsburgh Bacteriophage Institute, University of Pittsburgh, Pittsburgh, PA 15260
26
27
28
29
30

**ABSTRACT**

Bacteriophage S-CRM01 has been isolated from a freshwater strain of *Synechococcus* and shown to be present in the upper Klamath River valley in northern California and Oregon. The genome of this lytic T4-like phage has a 178,563 bp circular genetic map with 297 predicted protein-coding genes and 33 tRNA genes that represent all 20 amino acid specificities. Analyses based on gene sequence and gene content indicate a close phylogenetic relationship to the "photosynthetic" marine cyanomyophages infecting *Synechococcus* and *Prochlorococcus*. Such relatedness suggests that freshwater and marine phages can draw on a common gene pool. The genome can be considered as being comprised of three regions. Region 1 is populated predominantly with structural genes, recognized as such by homology to other T4-like phages and by identification in a proteomic analysis of purified virions. Region 2 contains most of the genes with roles in replication, recombination, nucleotide metabolism and regulation of gene expression, as well as 5 of the 6 signature genes of the photosynthetic cyanomyophages (*hli03*, *hsp20*, *mazG*, *phoH* and *psbA; cobS* is present in Region 3). Much of Regions 1 and 2 are syntenous with marine cyanomyophage genomes, except that a segment encompassing Region 2 is inverted. Region 3 contains a high proportion (85%) of genes that are unique to S-CRM01, as well as most of the tRNA genes. Regions 1 and 2 contain many predicted late promoters, with a combination of CTAAATA and ATAAATA core sequences. Two predicted genes that are unusual in phage genomes are homologs of cellular *spoT* and *nusG*.

1 **INTRODUCTION**

2 Research over recent years has brought into focus the important contribution of bacteriophages to the
3 ecology of microbial populations and to biochemical and geochemical cycles in the environment;
4 bacteriophages are both enormously abundant and diverse (Rohwer, 2003; Weinbauer, 2004;
5 Breitbart and Rohwer, 2005; Rohwer et al., 2009). Marine environments have attracted the most
6 attention (Fuhrman, 1999; Wommack and Colwell, 2000; Breitbart et al., 2002; Mann, 2003; Suttle,
7 2005, 2007), with relatively few studies dedicated to investigating bacteriophages present in
8 freshwater environments (Middelboe et al., 2008; Wilhelm and Matteson, 2008). We have been
9 interested in studying the microbial diversity and population dynamics of toxic cyanobacterial
10 blooms (Bozarth et al., 2010), which are an increasingly common ecological dysfunction seen in a
11 wide variety of lakes and reservoirs (Paerl and Huisman, 2009).

12

13 The only currently available genome sequence from a phage infecting freshwater cyanobacteria is
14 that of the myophage Ma-LMM01 infecting *Microcystis aeruginosa* from a Japanese lake (Yoshida
15 et al., 2008). We describe here the genome sequence and other properties of a cyanomyophage
16 isolated from Copco Reservoir on the Klamath River in Northern California in September 2008. The
17 phage was associated with a toxic *Microcystis aeruginosa* bloom, but an endemic *Synechococcus*
18 lineage is the host for this phage. The genome sequence revealed close relationships to a well-
19 studied group of "photosynthetic" exoT4-even cyanomyophages infectious to marine *Synechococcus*
20 and *Prochlorococcus* (Mann et al., 2005; Sullivan et al., 2005; Weigele et al., 2007; Millard et al.,
21 2009; Sullivan et al., 2010). This similarity between freshwater and marine cyanomyophages
22 supports indications from metagenomic (Rodriguez-Brito et al., 2010) and amplicon studies with
23 *psbA* and *gp20* (T4 portal protein) that related phages can be found in freshwater and marine
24 environments (Dorigo et al., 2004; Short and Suttle, 2005; Wilhelm et al., 2006; Chénard and Suttle,
25 2008; Sullivan et al., 2008), although there is also evidence for distinctively freshwater cyanophage
26 lineages (Deng and Hayes, 2008; Yoshida et al., 2008; Wang et al., 2010). Our study is the first
27 involving whole genome characterization to address the relationship between freshwater and marine
28 cyanophages, and supports the possibility that the gene complement of water-borne phages has been
29 shaped by gene pools in both freshwater and marine environments (Sano et al., 2004).

30

31 **MATERIALS AND METHODS**

1 **Sample collection and cyanophage enrichment.** Water was collected from the top 0.5 m of
2 Copco Reservoir (mid-channel near the dam wall at latitude 41.979°N and longitude 122.333°W)
3 during a *Microcystis* bloom on the Klamath River in Northern California on 10 September, 2008,
4 and transferred to the laboratory in the dark on ice. In order to enrich for cyanophages present in the
5 sample, a 200 mL aliquot of 0.2 µm filtered water was supplemented with 4 mL of 50x BG-11
6 medium (Sigma-Aldrich, St. Louis, MO) and 20 mL of *Microcystis*-dominated cultures derived from
7 the Klamath River system (August 2007). The enriched water sample was incubated at 24°C under
8 fluorescent lamps at 10 $\mu E/m^2/s$ for 10 days. Water quality data relevant to the collected sample are
9 available at http://www.pacificorp.com/es/hydro/hl/kr.html#.
10
11 **Cyanophage isolation and amplification.** The enriched culture was treated with chloroform,
12 and cellular material was removed by centrifugation. The 0.2 µm-filtered supernatant (Supor-200;
13 Pall Life Sciences) was ultracentrifuged using a Ti60 rotor (Beckman) at 4°C and 177,520 *g*, for 90
14 min. Pellets were resuspended in SM (50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgSO_4$-$7H_20$, pH
15 7.5) and stored at 4°C. Plaque assays were conducted using BG-11 top agar layered onto BG-11
16 agar plates that were incubated under the growth conditions described above until plaques were
17 visible (~ one week). Three serial plaque isolations were performed using the Klamath River system
18 culture, which was also used for phage amplification in liquid culture. Phage particles were
19 collected by ultracentrifugation. During the course of phage isolation, it was noticed that the culture
20 characteristics had changed. The identity of the resultant culture (LC16) was examined by PCR
21 amplification of genomic DNA using the cyanobacteria-wide primers CS1F and ULR directed at the
22 internal transcribed spacer of the rRNA operon (ITS), followed by DNA sequencing of the PCR
23 product as described (Bozarth et al., 2010). A single derived sequence indicated culture purity.
24 Comparison of the sequence with the GenBank database identified LC16 as a member of the
25 *Cyanobium gracile* cluster that includes freshwater *Synechococcus* isolates (Ernst et al., 2003; Chen
26 et al., 2006) (Fig. S1). The closest known relative has been isolated from Lake Balaton (Hungary),
27 with other related isolates from freshwater sources in Germany and Wisconsin (USA) and brackish
28 or saline sources in California, Baltic Sea (Denmark) and White Sea (Russia).
29

1 **Genome structure analysis.** Genome size was estimated using pulsed field gel electrophoresis
2 (PFGE) after genome preparation as described (Lingohr et al., 2008), using a 1.4% agarose gel in a
3 CHEF II PFGE unit (Bio-Rad) set to run at 6 V $cm^{-1}$ for 18 h with a 0.1 sec switch time.
4 To determine whether the genome was linear, circular or circularly permuted, BAL-31 nuclease
5 digestion of the phage genome was performed as described (Yoshida et al., 2008). Purified phage
6 DNA (200ng) was incubated with 0.1 U/µl BAL-31 nuclease (NEBiolabs) at 30°C for 0, 10, 20, 40
7 and 60 minutes. The DNA was then extracted with phenol/chloroform, ethanol precipitated and
8 digested overnight at 37°C with *Bam*H1 endonuclease. The restriction products were separated on
9 0.8% agarose gel and visualized with ethidium bromide. 1kB DNA ladder (GeneRuler, Fermentas)
10 and a plasmid (pGEM, Promega) were included as linear and supercoiled circular controls.
11
12 **Genome sequencing.** The genome (500 ng) was sequenced by Roche, Inc. (Branford, CT) using
13 GS FLX Titanium Sequencing, accumulating 232 kbp at an average coverage of 350X. A first draft
14 of the genome sequence was assembled using Newbler (gsAssembler) software (Roche). Five
15 contigs larger than 5,000 bp in length were assembled, amounting to a total of 173.5 kbp. To
16 determine the order of contigs and to fill the gaps, a multiplex PCR approach was used (Tettelin et
17 al., 1999). Briefly, pairs of outward-oriented primers positioned about 100 bp from the two ends of
18 each contig were used in all combinations to direct PCR amplifications using genomic DNA as
19 template. Products visible on gels after 30 cycles were confirmed by PCR using individual primer
20 pairs, and were extracted from agarose gels and submitted for direct Sanger sequencing (Genewiz,
21 Inc.). Each gap was sequenced in both directions, with additional primers designed when needed.
22 Contigs and junction sequences were arranged and assembled in Geneious (Biomatters, Ltd.,
23 http://www.geneious.com). The gap sequences could be confirmed using smaller contigs and
24 unassembled reads from the 454 data. The average sequencing coverage for contigs ranged between
25 360 and 500 with an average Phred Equivalent (Roche) quality score of 64.0.
26
27 **Genome annotation and analysis.** Regions of coding sequence were predicted using Glimmer
28 3 (Delcher et al., 1999) and Genemark S (Besemer et al., 2001); tRNAs were predicted using
29 tRNAscan-SE (Lowe and Eddy, 1997). Annotated protein coding sequences were determined using
30 a BLASTx search against the NCBI nr database. Annotations were made based on an E-Value
31 cutoff <1e-5. Phage-associated ORFs were compared to a custom database of T4-like cyanophages

using a BLASTx search. Genome annotation was curated in both Geneious and Artemis (Rutherford et al., 2000). Early promoters were predicted in regions upstream of ORFs by similarity to putative S-PM2 early promoters (Mann et al., 2005) and using BPROM (LDF>5; Softberry, Mount Kisco, NY). Putative terminator sequences were identified by TransTerm (http://uther.otago.ac.nz). The gene-annotated S-CRM01 sequence is available under GenBank accession HQ615693.

Phylogenomic analysis was carried out using the Hal pipeline (http://aftol.org/pages/Halweb3.htm; http://sourceforge.net/projects/bio-hal/), which consists of a set of Perl scripts that automates a series of phylogenomic analyses using existing software and sequence analysis programs (Robbertse et al., 2006); analysis was executed on a 64-bit Linux cluster operating Red Hat Linux 3.2.3, Linux version 2.4.21. The proteins encoded by 21 phage genomes were analyzed at 9 levels of "missing data" (inclusion in the analysis of genes/orthologous clusters not present in all phages), in each case using 3 gap removal methodologies (complete gap removal, and liberal and conservative gap-removal with GBlocks), resulting in a total of 27 phylogenetic trees. Inclusion of genes missing from a few taxa can improve phylogenetic analyses by increasing the number of genes analyzed (Wiens, 2006).

**Transmission electron microscopy.** CsCl gradient-purified phage was applied to a glow-discharged carbon-type B, 300-mesh copper grid (Ted Pella, Redding, CA) and stained with 1% phosphotungstic acid, pH 6.5. Samples were observed on a Philips CM-12 transmission electron microscope at 60 kEV.

**Mass spectrometry-based proteomics.** Phage was purified using a CsCl density gradient. Phage-associated proteins were prepared for SDS-PAGE by boiling for 2 minutes in SDS-PAGE protein sample loading buffer and separated on a 8-16 % gradient gel (Bio-Rad). The sample lane was cut into 11 sections, dehydrated with 50% acetonitrile in 50 mM $NH_4HCO_3$ and dried in a SpeedVac. Reduction and alkylation with DTT and iodoacetamide, and in-gel trypsin digestion were performed as described in the Protease-Max (Promega, Madison, WI) manual. A blank section of the gel was processed and included in all subsequent analyses.

LC-MS/MS analyses of the extracted peptides were performed on a LTQ-FT Ultra mass spectrometer (Thermo) with an IonMax ion source. The mass spectrometer was coupled to a nanoAcquity Ultra performance LC system (Waters) equipped with a Michrom Peptide CapTrap and

1 a $C_{18}$ column (Agilent Zorbax 300SB-C18, 250 x 0.3 mm, 5 μm). A binary gradient system was used
2 consisting of solvent A, 0.1% aqueous formic acid and solvent B, acetonitrile containing 0.1%
3 formic acid. Peptides were trapped and washed with 1% solvent B for 3 min. Peptide separation was
4 achieved using a linear gradient from 10% B to 30% B at a flow rate of 4 μL/min over 35 minutes.
5 For the LC-MS/MS analysis, the LTQ-FT mass spectrometer was operated in a data-dependent
6 mode. A full FT-MS scan (m/z 350-2000) was alternated with collision-induced dissociation (CID)
7 MS/MS scans of the 5 most abundant doubly or triply charged precursor ions. As the survey scan
8 was acquired in the ICR cell, the CID experiments were performed in the linear ion trap where
9 precursor ions were isolated and subjected to CID in parallel with the completion of the full FT-MS
10 scan. CID was performed with helium gas at a normalized collision energy 35% and activation time
11 of 30 ms. Automated gain control (AGC) was used to accumulate sufficient precursor ions (target
12 value, $5 \times 10^4$/micro scan; maximum fill time 0.2 s). Dynamic exclusion was used with a repeat
13 count of 1 and exclusion duration of 60 s. Data acquisition was controlled by Xcalibur (version
14 2.0.5) software (Thermo).
15 For the sequence database search, Thermo RAW data files were processed with Proteome
16 Discoverer v1.0 using default parameters except for a S/N threshold setting of 10. A Mascot
17 (v2.2.04) search of a phage-encoded protein database (299 sequences; 54263 residues) was launched
18 from Proteome Discoverer with the following parameters: the digestion enzyme was set to Trypsin/P
19 and two missed cleavage sites were allowed. The precursor ion mass tolerance was set to 10 ppm,
20 while fragment ion tolerance of 0.8 Da was used. Dynamic modifications included carbamidomethyl
21 (+57.0214 Da) for Cys and oxidation (+15.9994 Da) for Met.
22
23 **PCR to detect S-CRM01 in the Klamath system.** Enriched phage fractions were prepared (see
24 above) from samples collected from the Klamath River system during August-October, 2009.
25 Multiple plaques isolated on LC16 top agar plates from each sample were tested for the presence of
26 S-CRM01 by PCR analysis using two primer pairs directed at *g44* (*gp23*) and *g34* (Ma-LMM01-like
27 hypothetical protein). In all cases, reactions were either positive or negative for both primer pairs.
28 The primer pair for detecting *g34* was CRM01-g34(F) 5´ GTCAAATAGAATCCAGGATGAATTA
29 and CRM01-g34(R) 5´ TACCATAGTCTCCACCGTTTC. The primer pair detecting *g44* was
30 CRM01-g44(F) 5´ GACGTATGTGGCGTTCAGCCAATGA and CRM01-g44(R) 5´
31 CGGTTGATTTCTGCAAGGATTTC. PCR reactions used High Fidelity Platinum Taq polymerase

1 (Invitrogen) in the provided buffer with 0.2 μM of each primer, 0.2 mM dNTPs, and 2.5 mM
2 $MgSO_4$; after initial denaturation, 35 cycles of 0.5 min at 94°C, 0.5 min at 52°C, and 1 min at 68°C
3 were run. These primers were designed to be specific for S-CRM01, avoiding amplification from
4 known related cyanomyophage genomes; detection of S-CRM01 was scored as positive only when
5 products of the expected size were amplified with both primer pairs.
6
7 **RESULTS AND DISCUSSION**
8 **Isolation and physical characteristics.** Phage S-CRM01 was isolated from a surface sample
9 taken in September 2008 from a *Microcystis*-dominated bloom in Copco Reservoir on the Klamath
10 River in Northern California. At 319.8 river km from the coast, this site is far removed from salt
11 water habitats. S-CRM01 infects and lyses LC16, a culture belonging to the *Cyanobium gracile*
12 cluster of mostly freshwater *Synechococcus* (see Materials and Methods). We have observed lytic
13 infection of no other hosts, including the freshwater cyanobacteria *Microcystis aeruginosa,*
14 *Synechococcus elongatus* PCC 7942 and *Synechocystis sp.* PCC6803. During 2009, a survey was
15 conducted to assess the distribution of S-CRM01 between Upper Klamath Lake and the Klamath
16 River estuary. On the basis of PCR identification of plaques on LC16 plates, S-CRM01 was present
17 across a length of about 250 km along the Klamath River valley, from the Williamson River Delta at
18 the northern end of Upper Klamath Lake as far downstream as Seiad Valley (Fig. 1).
19 Negative staining electron microscopy revealed a T4-like morphology, with isometric heads
20 about 85-100 nm in diameter and rigid contractile tails 15-20 nm in diameter and 140-170 nm long
21 (Fig. 2). Contracted tails reveal a sheath and core 20-30 nm and about 10 nm in diameter,
22 respectively. A double-ringed baseplate and narrow neck are visible.
23 The genome migrated as a 180 kbp band in PFGE. Nuclease BAL-31 digestion followed by
24 cleavage with *Bam*HI resulted in simultaneous loss of material from all bands, indicative of
25 circularly permuted linear DNA (not shown). This is consistent with the presence of a gene
26 homologous to the *gp17* large terminase subunit of phage T4, which determines a circularly
27 permuted packaging strategy (Casjens and Gilcrease, 2008).
28
29 **S-CRM01 is a "photosynthetic" cyanomyophage most closely related to marine phages of**
30 ***Synechococcus* and *Prochlorococcus*.** Consistent with a circularly permuted organization and its
31 migration in PFGE, the S-CRM01 genome assembled into a circular map 178,563 bp long (Fig. 3).

1 The G+C content of the genome is 39.7%, and 297 protein-coding genes and 33 tRNA genes are
2 predicted. Protein-coding genes have been annotated with the following conventions (Table 1): S-
3 CRM01 genes are designated as *g1* etc.; in addition, genes with homology to numbered phage T4
4 genes are designated as *gp1,* etc.; genes lacking direct homology with T4 but with similarity to other
5 cyanomyophage genes that themselves have significant homology to T4 are designated as *gp-like*;
6 other genes are designated with conventional gene names, e.g., *nrdA*.

7 The genome contains 34 genes with homology to T4, and a total of 86 ORFs (29% of total) with
8 homology to ORFs found in at least one of a group of 17 closely related myophages that lytically
9 infect marine *Synechococcus* or *Prochloroccus* (Mann et al., 2005; Sullivan et al., 2005; Weigele et
10 al., 2007; Millard et al., 2009; Sullivan et al., 2010)(Tables 1, S1). Phage S-PM2 from this group
11 shares the largest number of genes with S-CRM01 (75), while the others have between 60 and 73
12 genes in common with S-CRM01 (Fig. S2). In contrast, the S-CRM01 genome shares a mere 4
13 genes with the only other known genome from a freshwater cyanomyophage, that of Ma-LMM01,
14 which infects *Microcystis aeruginosa* (Yoshida et al., 2008).

15 The S-CRM01 genome encodes all six genes proposed as (marine) cyanophage signature genes
16 (Millard et al., 2009): *cobS, hli03, hsp20, mazG, phoH* and *psbA*. The *psbA* gene, encoding the D1
17 protein of photosystem II, and *hli03* (high light inducible) gene are characteristic "photosynthetic
18 phage" genes that are thought to function in augmenting the photosynthetic capacity of infected cells
19 or in protecting against oxidative stress resulting from high light intensities (Lindell et al., 2005;
20 Mann et al., 2005; Clokie et al., 2006).

21 The S-CRM01 genome shows extensive synteny with the marine cyanomyophage genomes, but
22 a 64 kbp part of the syntenous region is in an inverted orientation (Fig. 4) with most genes between
23 *g56* and *g165* expressed from the negative strand. Together with the GC-skew profile (Fig. 3, inner-
24 most ring; the leading strand in prokaryotic genomes is enriched in G; Lobry, 1996), this is
25 suggestive of bidirectional replication with an origin near the gene inversion boundary (nt 123K). No
26 such pronounced GC-skew patterns are evident in the marine cyanomyophage genomes (not shown).

27 The S-CRM01 genome possesses an unusually large number (182) of ORFs with no significant
28 homologs in the GenBank database (as of 30 August, 2010), representing 61% of predicted ORFs
29 (genes shown in black in Fig. 3). This is far higher than the number of unique genes in the marine
30 cyanomyophage genomes of similar size (about 60-100) (Millard et al., 2009; Sullivan et al., 2010).

Based on both gene content and sequence relatedness, S-CRM01 is a member of a discrete clade that encompasses all 17 currently sequenced marine cyanomyophages, but it is the most divergent member of this group (Fig. 5). The phylogenetic relationships among the cyanomyophages were explored using the Hal phylogenomics pipeline (Robbertse et al., 2006), which produces whole genome phylogenies using single-copy protein coding genes. The pipeline can be configured for inclusion of different numbers of orthologous clusters to allow the analysis to be expanded by including genes missing from a few genomes (e.g., Fig. 5D; see methods). As shown in the consensus tree in Fig. 5A, the cyanomyophages (including S-CRM01) form a monophyletic group supported by high bootstrap values. However, the group is quite diverse, with S-CRM01 being the most divergent member and the marine phages partitioning into at least 4 clades. To examine the possibility of long branch attraction artifacts, the same analysis was run without the T4, Aeh1 and KVP40 genomes, with the overall topology matching that of the consensus tree in Fig. 5A. Phylogenetic analysis based on gene content (using a subset of the genomes analyzed here) showed similar overall relationships (Millard et al., 2009).

With different numbers of genes (orthologous clusters) included in the analysis, two additional alternative tree topologies with differences in the relative positions of MC1, MC2 and MC3 were observed (Fig. 5B, C). Analysis of individual gene phylogenies produced a similar range of topologies, although the S-CRM01 branch (typically the longest) was at times placed among the MC1-3 clades. No pattern between tree topology and gene type or location in the genome could be discerned. The variable relationships indicated above suggest high levels of horizontal gene exchange among the cyanomyophages, analogous to that proposed among the *Synechococcus* or *Prochlorococcus* hosts of these phage (Zhaxybayeva et al., 2009). Evidence for gene exchange at individual loci among the marine cyanophages has been reported by Zeidner et al. (2005), Sullivan et al. (2006) and Bryan et al. (2008). Note that the analysis of Fig. 5 does not support the classification of phages based on host (*Synechococcus* or *Prochlorococcus*), since both hosts are represented in clades MC2 and MC4 (c.f. Sullivan et al., 2010).

**A structural gene cluster mostly on the plus strand.** The bioinformatically identifiable structural genes are all present in a 72 kbp segment of the genome (Region 1), comprised of two clusters: *g1* through *g45*, covering nts 1-51,380 (genome Region 1A), and *g59* through *g68*, covering nts 62,104-72,062 (Region 1C). Region 1A includes strong synteny with the genomes of marine

exoT-even cyanomyophages and phage T4 (Fig. 4). This region encodes most of the recognizable structural genes, with all but one gene expressed from the plus strand. Expression is predicted to be dominated by the activity of late promoters (Table S4), as appropriate for structural protein genes.

Both transcriptional directions are represented in Region 1C, with transcription again predicted to be dominated by late promoters. This region is also syntenous, although in inverted orientation, to the marine cyanomyovirus genomes (Fig. 4), but is not syntenous to T4.

As has been observed for marine cyanomyophages, the conserved genes that define synteny are variably interspersed with additional genes. Twenty-three ORFs with no homologs in the GenBank database are located within Region 1; other genes are most similar to phage or bacterial proteins (Table 1). Millard et al. (2009) have described a hyperplastic region between *gp15* and *gp18* in S-RSM4 and other cyanomyophages. In S-CRM01 this segment contains only one non-conserved gene, but several genes of varied apparent origins are located upstream of *gp13*.

**Virion proteomics.** Mass spectrometry was used to identify phage proteins present in a preparation of S-CRM01 virions purified through a CsCl density gradient. Forty-three proteins were identified with high certainty (Fig. 3, Tables 1 and S3). All of these, except two proteins encoded in Region 3 of the genome, are encoded by genes that are closely associated with late promoters (Table S4). Most of the identified proteins are encoded in Regions 1A and 1C, emphasizing the clustering of structural protein genes in these parts of the genome. Thirteen of these proteins correspond to structural proteins of T4 (*gp* or *gp-like* genes), 13 correspond to genes with BLAST hits to other phage or bacterial genes, while 9 are encoded by genes unique to S-CRM01. These results indicate that the S-CRM01 virion is composed of proteins with a variety of origins: homologs of T4 proteins, homologs of proteins from the related marine cyanomyophages, proteins with closest BLAST hits in other phage or in bacterial genomes, and proteins that have no currently known homologs.

An additional 8 proteins encoded by genes not clustered with structural genes were identified by our proteomic study. These genes are located in Regions 1B, 2 and 3 (Fig. 3, Table 1). Electron microscopy suggested that host material attached to phage baseplates may have been present in the phage preparation made for proteomic analysis. Consequently, it is uncertain whether these proteins are truly virion-associated (structural) proteins or phage-encoded proteins that have been inserted into host structural components. Their identification does prove the expression of the respective

1 genes during viral infection, indicating tht these proteins (4 of which are unique to S-CRM01) are
2 functionally relevant.
3 As expected, abundant mass spectrometry signals were registered for gp23 major capsid protein
4 (*g44*), gp18 contractile tail sheath protein (*g36*) and gp19 tail tube protein (*g37*), which are present in
5 multiple copy number in the T4 capsid (Miller et al., 2003a). Abundant signals were also observed
6 for peptides from g9, at 271 kDa the largest protein encoded by the S-CRM01 genome. This huge
7 protein has a predicted strong β-strand character and no identifiable sequence motifs, but has
8 BLAST matches to a wide range of phage proteins (including putative tail protein) and to glycosyl
9 hydrolase bacterial neuraminidase repeat (BNR) proteins. An internal dot plot analysis shows
10 numerous different internal repeat elements 10 to 20 residues long, mostly repeated only once.
11 These properties suggest that g9 is a tail fiber gene (Weigele et al., 2007), although we have not
12 observed tail fibers under the electron microscope.
13 Three proteins with collagen-like triple helix repeat domains were detected by mass
14 spectrometry: g13 with 36 GXY repeats, g20 with 85 GXQ repeats, and g23 with 40 GXQ repeats.
15 Collagen-like proteins have been reported from some phage, participating in spike formation in
16 phage PRD1 (Caldentey et al., 2000) and suggested to be found in tail fibers (Smith et al., 1998;
17 Sullivan et al., 2005). More careful virion characterization will be needed to determine whether S-
18 CRM01 possesses the tail fibers or whiskers that are predicted by the presence of these proteins.
19
20 **T4-like nonstructural genes predominantly on the minus strand.** A second major region of
21 synteny with the exoT-even marine cyanomyophage genomes is the 51 kbp segment between genes
22 *g69* and *g165* (Region 2, nts 72,094–123,515) (Figs. 3, 4; Table 1)(c.f. Sullivan et al., 2010). These
23 genes, all of which are on the minus strand, include core T4-like genes with roles in DNA
24 replication, recombination and repair (e.g., *gp43* DNA polymerase, *gp61* primase, *gp45* sliding
25 clamp, *uvsX*), nucleotide metabolism (*nrdA*, *nrdB*, *nrdC*, *td* thymidylate synthase) and 5 of the 6
26 marine cyanophage signature genes (*phoH, hli03, psbA, mazG, hsp20*). Other significant genes also
27 present in Region 2 are: a member of the 2OG-FeII oxygenase superfamily that is present in nearly
28 all marine cyanomyophages, a *nusG* transcription anti-termination factor homolog (not found in T4-
29 related phages), *regA* regulator of early gene translation (common to all marine cyanomyo and T4-
30 like phages), and RNase H (*rnh*, common to T4-like phages but found in only S-PM2 and Syn19
31 among the marine cyanomyophages).

1 This segment of the genome is also predicted to be expressed predominantly via late
2 transcription, though some putative early promoters have been identified, notably for the expression
3 of *gp45, gp44, gp62* and *gp33*, all of which are involved in establishing and maintaining T4 late gene
4 transcription (Miller et al., 2003a). The other key gene needed for late transcription, *gp55*
5 alternative sigma factor, is located in Region 1B between the two structural gene segments and on
6 the plus strand, and it also appears to be expressed from an early promoter (Table S4). The
7 recombination genes *uvsY, uvsW, gp47* and *gp46* are also located in Region 1B, all on the plus strand
8 in another region of synteny with marine cyanomyophage genomes (Figs. 3, 4).
9 As in the structural gene region, but even more so, the conserved genes in Region 2 are
10 interspersed with additional genes at multiple sites. Forty-eight (49%) of ORFs in Region 2 are
11 unique to S-CRM01.
12
13 **49 kbp (27%) of the genome contains 132 ORFs, 85% of which are unique genes.** The
14 remainder of the genome (Region 3, genes *g166-g297*) possesses no overall similarity to the
15 genomes of other phages apart from *g166* and *g167* (Fig. 4) and contains only 20 genes with
16 homology to previously identified ORFs. Fifteen of these have homologs in at least one T4-like or
17 cyanomyo phage. This part of the genome also contains all but two of the 33 tRNA genes. The
18 significant genes present in this segment are *cobS* (*g167*), DNA ligase (*g194*), *speD* (*g109*), *spoT*
19 (*g215*) and HNH endonuclease (*g205*). *cobS*, involved in cobalamin biosynthesis, may support the
20 synthesis of deoxynucleotides (Sullivan et al., 2005) and is one of the marine cyanophage signature
21 genes (Millard et al., 2009). DNA ligase is common in T4-like phage genomes, though not found in
22 the marine cyanomyophages; the S-CRM01 DNA ligase gene is most closely related to Chorella
23 virus and PB1-like myophage ligases. *speD* encodes S-adenosylmethionine decarboxylase, which is
24 involved in polyamine synthesis and found in a few marine cyanomyophage genomes (Mann et al.,
25 2005). *spoT*, which encodes ppGpp synthetase and hydrolase, has not previously been recognized in
26 myophage genomes but is present in the *Aeromonas* myophages Aeh1, 44RR, PHG25 and PHG31.
27 HNH endonuclease is possibly a member of a family of homing endonucleases, although no introns
28 have been detected in the S-CRM01 genome.
29 All but one of the genes (*g186*) in Region 3, including the tRNA genes, reside on the plus strand.
30 Region 3 genes are atypical in a number of ways. Identifiable promoters are sparse; the G+C content
31 is generally high (Fig. 3) and the ORFs are more likely to have a higher G+C% in the third codon

position; the ORFs are shorter and less tightly spaced. A similarly extensive array of unique genes is also present in the Aeh1 genome (Comeau et al., 2007) and to a lesser extent in other myophages, including the marine cyanomyophages (e.g., S-PM2; Mann et al., 2005).

**Genome transcription.** The predicted transcription control signals —start sites of 34 early promoters and 81 late promoters, and coordinates of 17 terminator hairpins — are listed in Table S4. As for T4-like phages such as RB49 (Desplats et al., 2002) and marine cyanomyophages such as S-PM2 (Mann et al., 2005), S-CRM01 gene expression lacks the middle phase of transcription found in T4. Either early or late (in a few instances both) promoters are predicted for genes that are middle-transcribed in T4. Thus, early promoters are predicted upstream of the genes involved in establishing late transcription: *gp55* (sigma factor), *gp44, gp45*, *gp46* and *gp33* (sliding clamp and clamp loader proteins). DNA replication and repair genes that are either middle or late expressed in T4, such as *gp43* (DNA polymerase), *gp32* (ssDNA binding protein), *gp41* (helicase), *uvsX, uvsW*, and *gp46* and *gp47* endonucleases, are associated with putative late promoters. This is also true of nucleotide metabolism genes *td, nrdA* and *nrdC*, as well as each of the marine cyanomyophage signature genes except *cobS.* A few of these genes — *rnh*, *td*, *phoH*, *hsp20* — are associated with both predicted early and late promoters. This is also true of some, though not all, tRNA genes.

All identifiable structural genes are associated with putative late promoters, although in several cases (*gp8*, *gp13* & *gp14*, *gp17*, *gp18* & *gp19*, *gp22*, *gp5*) early promoters have unexpectedly also been predicted with good prediction scores. Transcription patterns will need to be experimentally assessed to verify these predictions.

Late promoters are predicted to fall into two categories, 46 with a CTAAATA core sequence and 35 with an ATAAATA core sequence (Fig. 6, Table S4). ACTAAATA is the most frequent promoter sequence. ATAAATA promoters are most common with T4 (Miller et al., 2003a) and the marine cyanomyophages (Mann et al., 2005; Weigele et al., 2007; Sullivan et al., 2010). The CTAAATA sequence is largely absent from the latter genomes, indicating that these promoters have not merely been overlooked in previous annotations. CTAAATA promoters are unusual among T4-like phages, previously reported for RB49 (Desplats et al., 2002) and Aeh1 (Nolan et al., 2006). Like RB49, S-CRM01 late promoters are a combination of CTAAATA and ATAAATA types, with no evident specialization within gene groups that might suggest separate regulation. We have also only detected a single late transcription *gp55* sigma factor, which may thus have unusually broad

1 promoter recognition. Interestingly, S-CRM01 *gp55* is clearly more closely related to marine
2 cyanomyophage *gp55* genes (which recognize ATAAATA promoters) than to RB49 or Aeh1 *gp55*.
3
4 **A full set of tRNAs.** The 33 tRNA genes in the S-CRM01 genome (Table 2) is more than found
5 in some bacteria and includes all of the 20 amino acid specificities, although there is no initiator
6 $tRNA^{Met}$ as found in some phage genomes. KVP40 also has all 20 specificities represented, though
7 some are thought to be pseudogenes that may not be functional (Miller et al., 2003b). The S-PM2
8 genome contains 24 tRNA genes representing all but the cysteine and phenylalanine specificities, but
9 the other marine cyanomyophage genomes have much smaller subsets of these. $tRNA^{His}$ is the only
10 S-CRM01 tRNA gene with an encoded 3´ CCA terminus, while in S-PM2 five tRNA genes
11 (including $tRNA^{His}$) include the 3´ CCA. Three tRNA genes with a CAU anticodon are present.
12 Based on features in the anticodon loop (C32 and A38) that are recognized by the TilS enzyme,
13 which modifies C34 in the anticodon to lysidine (Nakanishi et al., 2009), the *t2* and *t31* genes are
14 proposed to be $tRNA^{Ile2}$ genes, leaving one $tRNA^{Met}$ gene (*t22*). Gene *t2* has multiple additional
15 nucleotides that are proposed to be characteristic of lysidine-containing $tRNA^{Ile}$ in cyanobacteria
16 (Freyhult et al., 2007)(see Table S5).
17 It cannot be determined without experimentation whether all predicted tRNAs are functional,
18 although key identity elements and conserved features (Giegé et al., 1998) are generally present
19 (Table S5). Some of the predicted tRNAs do have highly unusual features (see Table S5 and refer to
20 tRNA database; Juhling et al., 2009) that may compromise function: e.g., U73 and long variable loop
21 in $tRNA^{Glu}$ *t12*, A1 in $tRNA^{Glu}$ *t13*, U73 in $tRNA^{Ile2}$ *t31*, U10:U25 in $tRNA^{Ser}$ *t21* (probable cause
22 for identification as possible pseudogene by tRNAScan-SE). The $tRNA^{His}$ gene *t23*, like other phage
23 $tRNA^{His}$ occurrences (e.g., S-PM2), lacks the additional 5´ residue G-1 that is in most systems
24 critical for histidine identity (Giegé et al., 1998). G-1 would be either replaced with U-1 (adjacent
25 nucleotide in the genome) or lacking in the mature S-CRM01 $tRNA^{His}$, a situation that in *E. coli*
26 allows only partial translational function (Yan and Francklyn, 1994). Another possibility is that G-1
27 could be added post-transcriptionally by a Thg1-like activity (Heinemann et al., 2010) or that a
28 variant histidyl-tRNA synthetase that does not rely on G-1 is present in host cells (Wang et al.,
29 2007). No Thg1 or histidyl-tRNA synthetase is evident among the S-CRM01 genes. Establishing
30 the functionality of phage tRNA genes is important in determining their role during infection and the
31 selective forces that act on them.

A connection between the suite of S-CRM01 tRNA genes and codon usage in phage ORFs or in those ORFs expected to be expressed late could not be discerned. It is thought that phage tRNAs serve to optimize expression of proteins by allowing more efficient decoding of codons in phage mRNAs that are under-represented in host mRNAs (Bailly-Bechet et al., 2007). In the absence of the host genome sequence, this hypothesis cannot be tested for S-CRM01. However, there must be a different reason for the presence of tRNAs for the two specificities represented by single codons (methionine and tryptophan) or of tRNA genes with unusual features that could be expected to decrease decoding efficacy (see Table S5); perhaps the viral tRNAs provide an advantage by being more flexible in their accommodation to the interactions on the ribosome in a stressed (infected) cell.

In several cyanomyophage and T4-like genomes, some tRNA genes are located in the general region downstream of the *nrdA* and *nrdB* genes. Only two of the 33 tRNA genes in S-CRM01 are found in this location, the remainder being loosely spaced across a 22 kbp segment of Region 3 among mostly genes that are unique to S-CRM01 (Fig. 3). The S-CRM01 genome includes homologs of 22 of the 24 tRNA genes from S-PM2, with nucleotide identity ranging from 42% to 81%, with an average of 68%. tRNAs among the MC1 clade (S-PM2, Syn1 and S-RSM4) share average identities of 76-80%, consistent with their close relationship on the basis of protein-coding genes. This would suggest vertical inheritance of tRNAs, but the lack of extensive synteny, wide range in nucleotide identity, and wide variety of tRNA gene content presents a complex picture of tRNA gene evolution in the cyanomyophages; the MC1 clade members possess 24, 12 and 6 tRNA genes, and the MC2 clade members possess 10, 5 and 1 tRNA genes.

**Noteworthy genes encoded by S-CRM01**. Two predicted genes, *spoT* and *nusG* (indicated in pink in Fig. 3) are unusual genes in T4-like phages. SpoT is a pyrophosphohydrolase that in *E. coli* removes the alarmone (p)ppGpp, which accumulates as a result of RelA action in cells when the stringent response is triggered by amino acid starvation (Srivatsan and Wang, 2008). SpoT is also capable of synthesizing (p)ppGpp under certain conditions, such as fatty acid starvation. Through a network of interactions with other proteins, it seems to sense the physiological status of the cell and modulate (p)ppGpp levels (Potrykus and Cashel, 2008; Srivatsan and Wang, 2008). The effect of elevated (p)ppGpp is a down-regulation of normal macromolecular synthesis and a switch to gene expression governed by alternative sigma factors, turning gene expression to pathways such as amino acid biosynthesis. The introduction into a cell of a replicating phage can be expected to

deplete cellular nutrients, with the risk of inducing elevated (p)ppGpp and shutting off ribosome synthesis, general transcription and replication. Indeed, one may speculate that host cells utilize (p)ppGpp to establish an antiviral state that inhibits viral replication and amplification, perhaps until nucleases can attack the viral nucleic acid. Specific effects on viral gene expression are also possible, as illustrated by the ability of (p)ppGpp to shut down some lambda phage promoters (Potrykus and Cashel, 2008). It could thus be advantageous for an infecting phage to counter this type of innate defense, and a virally expressed SpoT enzyme could do that by hydrolyzing (p)ppGpp or otherwise altering the regulation of (p)ppGpp levels.

The predicted S-CRM01 SpoT possesses most of the key amino acids needed for (p)ppGpp hydrolysis (Hogg et al., 2004), but it is a very small protein, suggesting activity as a (p)ppGpp hydrolase with little or no regulatory control. Phage-encoded *spoT* has not been previously recognized, although homologous genes are present in the *Aeromonas* phages AehI, 44RR, PHG25 and PHG31, which were assigned to T4 gene cluster (T4-GC) 1803 by Sullivan et al. (2010). All of the marine cyanomyophage genomes possess a *mazG* gene (as does S-CRM01), which has been postulated to hydrolyze (p)ppGpp and avoid stationary phase conditions in the cell (Clokie et al., 2010; Bryan et al., 2008). However, MazG proteins have varied nucleotide substrate specificities (Galperin et al., 2006) and it has been cautioned that they may not function in (p)ppGpp hydrolysis (Sullivan et al., 2010). In fact, it may be that the broad substrate specificity of MazG (Lu et al., 2010; Zhang and Inouye, 2002; Zhang et al., 2003) involves this protein both in the removal of mutagenic nucleotides that are produced as a result of oxidative stress (Lu et al., 2010) associated with photosynthesis as well as suppression of (p)ppGpp levels.

S-CRM01 may be the first phage genome to contain a putative *nusG* gene. The NusG protein associates with elongating RNA polymerase to modulate transcription in various ways. Based mainly on *E. coli* studies, NusG is considered a transcription elongation factor because it increases transcription elongation rates (Squires and Zaporojets, 2000; Yakhnin et al., 2008), although in some bacteria the opposite is true (Sevostyanova and Artsimovitch, 2010). In addition, NusG can promote or suppress pausing at different sites (and thereby facilitate attenuation control) (Sevostyanova and Artsimovitch, 2010; Yakhnin et al., 2008), promote transcript release at termination sites (Chalissery et al., 2007), and may exert effects on translation (Squires and Zaporojets, 2000). S-CRM01 may benefit by influencing the elongation and termination phases of transcription of the phage, or even host, genome through the action of viral NusG. Additionally, Cardinale et al. (2008) have shown

1 that NusG functions in concert with Rho to decrease doubling time and prevent expression of the
2 cryptic *rac* prophage in *E. coli* MG1655. Over-expression of NusG by S-CRM01 may be able to
3 increase the metabolic activity of the infected host and prevent interference or competition for
4 transcriptional machinery from prophages endemic to the host. If such benefits do exist, the absence
5 of identifiable *nusG* from other phage genomes might suggest that unrecognized transcriptional
6 regulators are encoded by other phages.
7
8 **Relationship between freshwater and marine phages**. There has not been enough data to
9 derive a clear picture of the genetic relationships between phages in freshwater and marine
10 environments. Some early observations with podophages (Breitbart et al., 2004) and myophages
11 (Short and Suttle, 2005) emphasized the discovery of very similar sequences in the viral populations
12 of the two aquatic environments. For myophages, additional phylogenetic studies with the *gp20*,
13 *psbA* and *psbD* genes have verified this observation, but have also shown that most (though not all)
14 freshwater sequences map at high resolution to clades that do separate from marine isolates
15 (Wilhelm et al., 2006; Chénard and Suttle, 2008; Sullivan et al., 2008; Wang et al., 2010). S-
16 CRM01 seems to represent both of these scenarios, with the closest currently known *gp20* sequence
17 found in the Atlantic Ocean isolate P-ShM1 (Sullivan et al., 2008)(Fig. S3). On the other hand, the
18 *psbA* gene locates to a unique branch between freshwater and marine clades (Fig. S3).
19 The complete S-CRM01 genome sequence allows more meaningful consideration of the
20 relationship between related marine and freshwater phages. The data in Fig. 5 show the
21 relationships to be complex. On the basis of synteny (Fig. 4A) and protein-coding (Fig. 5D) and
22 tRNA gene content, S-CRM01 is most similar to S-PM2, while on the basis of some phylogenomic
23 comparisons, P-SSM2, S-SM2 and S-SSM7 can be considered more similar (Fig. 5D). It thus seems
24 likely that there have been multiple gene exchanges between the S-CRM01 and marine
25 cyanomyophage lineages during evolution. Brackish estuarine waters are likely connections
26 between freshwater and marine environments, where genetic exchanges could occur between phages
27 that principally exist in one or the other aquatic habitat, especially for phages of bacteria such as
28 *Synechococcus*, which have broad distributions. Highly unique phages do seem to exist in
29 freshwater, however, such as the *Microcystis*-infecting Ma-LMM01 (Yoshida et al., 2008), and it
30 will be interesting to learn in the future whether this is the pattern for phages infecting hosts that do
31 not exist in the oceans.

## ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of Susan Corum of the Karuk Tribe of California, Richard Raymond and PacifiCorp Inc., and Mary Lindenberg and Tamara Wood of the US Geological Survey for sample collections from the Klamath valley. We thank Elizabeth Kutter for advice on initial experiments, Josh Powell and Theresa Sawyer for assistance with electron microscopy, Andrea McHugh for assisting in the laboratory, Michael Schwalbach for assistance with PFGE, Joey Spatafora for discussions, Susan Golden and Roger Ely for providing cultures, and Roger Hendrix for comments on the manuscript. The OSU mass spectrometry facility and services core is supported in part by NIH/NIEHS grant P30 ES000210. This work was supported by Oregon Sea Grant award NA060AR4170010 and by Oregon State University Agricultural Experiment Station.

## REFERENCES

Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* **17**: 1486-1495.

Besemer, J., Lomsadze, A., and 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.

Bozarth, C.S., Schwartz, A.D., Shepardson, J.W., Colwell, F.S., and Dreher, T.W. (2010) Population turnover in a *Microcystis* bloom results in predominantly non-toxigenic variants late in the season. *Appl Environ Microbiol* **76**: 5207-5213.

Breitbart, M., Miyake J.H. and Rohwer, F. (2004) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Letts* **236**: 249-256.

Breitbart, M., and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278-284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D. et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Bryan, M.J., Burroughs, N.J., Spence, E.M., Clokie, M.R., Mann, N.H., and Bryan, S.J. (2008) Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. *PLoS One* **3**: e2048.

Caldentey, J., Tuma, R. and Bamford, D.H. (2000) Assembly of bacteriophage PRD1 spike complex: Role of the multidomain protein P5. *Biochemistry* **39**:10566-73.

Cardinale, C.J., Washburn, R.S., Tadigotla, V.R., Brown, L.M., Gottesman, M.E., and Nudler, E. (2008) Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in *E. coli*. *Science* **320**:935-938.

Casjens, S.R., and Gilcrease, E.B. (2008) Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. In *Bacteriophages: Methods and Protocols, Volume 2: Molecular and applied aspects*. Clokie, M.R., and Kropinski, A.M. (eds). New

1 York: Humana Press, pp. 91-111.
2 Chalissery, J., Banerjee, S., Bandey, I., and Sen, R. (2007) Transcription termination defective
3 mutants of Rho: role of different functions of Rho in releasing RNA from the elongation complex. *J*
4 *Mol Biol* **371**: 855-872.
5 Chen, F., Wang, K., Kan, J., Suzuki, M.T., and Wommack, K.E. (2006) Diverse and unique
6 picocyanobacteria in Chesapeake Bay, revealed by 16S-23S rRNA internal transcribed spacer
7 sequences. *Appl Environ Microbiol* **72**: 2239-2243.
8 Chénard, C., and Suttle, C.A. (2008) Phylogenetic diversity of sequences of cyanophage
9 photosynthetic gene psbA in marine and freshwaters. *Appl Environ Microbiol* **74**: 5317-5324.
10 Clokie, M.R., Millard, A.D., and Mann, N.H. (2010) T4 genes in the marine ecosystem: studies of
11 the T4-like cyanophages and their role in marine ecology. *Virol J* **7**: 291.
12 Clokie, M.R., Shan, J., Bailey, S., Jia, Y., Krisch, H.M., West, S., and Mann, N.H. (2006)
13 Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium.
14 *Environ Microbiol* **8**: 827-835.
15 Comeau, A.M., Bertrand, C., Letarov, A., Tetart, F., and Krisch, H.M. (2007) Modular
16 architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology*
17 **362**: 384-396.
18 Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial
19 gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636-4641.
20 Deng, L., and Hayes, P.K. (2008) Evidence for cyanophages acrive against bloom-forming
21 freshwater cyanobacteria. *Freshwater Biology* **53**: 1240-1252.
22 Desplats, C., Dez, C., Tetart, F., Eleaume, H., and Krisch, H.M. (2002) Snapshot of the genome
23 of the pseudo-T-even bacteriophage RB49. *J. Bacteriol.* **184**: 2789-2804.
24 Dorigo, U., Jacquet, S., and Humbert, J.F. (2004) Cyanophage diversity, inferred from g20 gene
25 analyses, in the largest natural lake in France, Lake Bourget. *Appl Environ Microbiol* **70**: 1017-1022.
26 Ernst, A., Becker, S., Wollenzien, U.I., and Postius, C. (2003) Ecosystem-dependent adaptive
27 radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiology*
28 **149**: 217-228.
29 Freyhult, E., Cui, Y., Nilsson, O., and Ardell, D.H. (2007) New computational methods reveal
30 tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie* **89**: 1276-
31 1288.
32 Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature*
33 **399**: 541-548.
34 Galperin, M.Y., Moroz, O.V., Wilson, K.S., and Murzin, A.G. (2006) House cleaning, a part of
35 good housekeeping. *Mol Microbiol* **59**: 5-19.
36 Giegé, R., Sissler, M., and Florentz, C. (1998) Universal rules and idiosyncratic features in tRNA
37 identity. *Nucleic Acids Res* **26**: 5017-5035.
38 Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large
39 phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.
40 Heinemann, I.U., Randau, L., Tomko, R.J., Jr., and Soll, (2010) D. 3'-5' tRNAHis
41 guanylyltransferase in bacteria. *FEBS Lett* **584**: 3567-3572.
42 Hogg, T., Mechold, U., Malke, H., Cashel, M., and Hilgenfeld, R. (2004) Conformational
43 antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates
44 (p)ppGpp metabolism during the stringent response [corrected]. *Cell* **117**: 57-68.
45 Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., and Putz, J. (2009) tRNAdb
46 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* **37**: D159-162.

1 Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis
2 genes in marine viruses yield proteins during host infection. *Nature* **438**: 86-89.
3 Lingohr, E., Frost, S., and Johnson, R. (2008) Determination of bacteriophage genome size by
4 Pulsed-Field Gel Electrophoresis In *Bacteriophages: Methods and Protocols, Volume 2: Molecular*
5 *and applied aspects*. Clokie, M.R., and Kropinski, A.M. (eds). New York: Humana Press, pp. 19-25.
6 Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol*
7 *Evol* **13**: 660-665.
8 Lowe, T.M., and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer
9 RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955-964.
10 Lu, L.D., Sun, Q., Fan, X.Y., Zhong, Y., Yao, Y.F., and Zhao, G.P. (2010) Mycobacterial MazG
11 is a novel NTP pyrophosphohydrolase involved in oxidative stress response. *J Biol Chem* **285**:
12 28076-28085.
13 Mann, N.H. (2003) Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev*
14 **27**: 17-34.
15 Mann, N.H., Clokie, M.R., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J. et al. (2005) The
16 genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus
17 strains. *J Bacteriol* **187**: 3188-3200.
18 Middelboe, M., Jacquet, S., and Weinbauer, M. (2008) Viruses in freshwater ecosystems: an
19 introduction to the exploration of viruses in new aquatic habitats. *Freshwater Biology* **53**: 1069-
20 1075.
21 Millard, A.D., Zwirglmaier, K., Downey, M.J., Mann, N.H., and Scanlan, D.J. (2009)
22 Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of
23 Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of
24 cyanophage evolution. *Environ Microbiol* **11**: 2370-2387.
25 Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003a)
26 Bacteriophage T4 genome. *Microbiol Mol Biol Rev* **67**: 86-156, table of contents.
27 Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A. et al. (2003b)
28 Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a
29 T4-related bacteriophage. *J Bacteriol* **185**: 5220-5233.
30 Nakanishi, K., Bonnefond, L., Kimura, S., Suzuki, T., Ishitani, R., and Nureki, O. (2009)
31 Structural basis for translational fidelity ensured by transfer RNA lysidine synthetase. *Nature* **461**:
32 1144-1148.
33 Nolan, J.M., Petrov, V., Bertrand, C., Krisch, H.M., and Karam, J.D. (2006) Genetic diversity
34 among five T4-like bacteriophages. *Virol J* **3**: 30.
35 Paerl, H.W., and Huisman, J. (2009) Climate change: a catalyst for global expansion of harmful
36 cyanobacterial blooms. *Environ Microbiol Rep* **1**: 27-37.
37 Potrykus, K., and Cashel, M. (2008) (p)ppGpp: still magical? *Annu Rev Microbiol* **62**: 35-51.
38 Robbertse, B., Reeves, J.B., Schoch, C.L., and Spatafora, J.W. (2006) A phylogenomic analysis
39 of the Ascomycota. *Fungal Genet Biol* **43**: 715-725.
40 Rodriguez-Brito, B., Li, L-L., Wegley, L., et al. (2010) Viral and microbial community dynamics
41 in four aquatic environments. *ISME J.* **4:**739-51.
42 Rohwer, F. (2003) Global phage diversity. *Cell* **113**: 141.
43 Rohwer, F., Prangishvili, D., and Lindell, D. (2009) Roles of viruses in the environment. *Environ*
44 *Microbiol* **11**: 2771-2774.
45 Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B.
46 (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944-945.

1 Sano, E., Carlson, S., Wegley, L., and Rohwer, F. (2004) Movement of viruses between biomes.
2 *Appl Environ Microbiol* **70**: 5842-5846.
3 Sevostyanova, A., and Artsimovitch, I. (2010) Functional analysis of Thermus thermophilus
4 transcription factor NusG. *Nucleic Acids Res*.
5 Short, C.M., and Suttle, C.A. (2005) Nearly identical bacteriophage structural gene sequences are
6 widely distributed in both marine and freshwater environments. *Appl Environ Microbiol* **71**: 480-
7 486.
8 Smith, M.C., Burns, N., Sayers, J.R., Sorrell, J.A., Casjens, S.R., and Hendrix, R.W. (1998)
9 Bacteriophage collagen. *Science* **279**: 1834.
10 Squires, C.L., and Zaporojets, D. (2000) Proteins shared by the transcription and translation
11 machines. *Annu Rev Microbiol* **54**: 775-798.
12 Srivatsan, A., and Wang, J.D. (2008) Control of bacterial transcription, translation and replication
13 by (p)ppGpp. *Curr Opin Microbiol* **11**: 100-105.
14 Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three
15 Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*
16 **3**: e144.
17 Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P. and Chisholm, S.W.
18 (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and
19 their hosts. *PLoS Biol* **4**: e234.
20 Sullivan, M.B., Coleman, M.L., Quinlivan, V., Rosenkrantz, J.E., DeFrancesco, A.S., Tan, G., Fu,
21 R., Lee, J.A., Waterbury, J.B., Bielawski, J.P. and Chisholm, S.W. (2008) Portal protein diversity
22 and phage ecology. *Environ. Microbiol.* **10:**2810-23.
23 Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R. et al.
24 (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses
25 from diverse hosts and environments. *Environ Microbiol*.
26 Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356-361.
27 Suttle, C.A. (2007) Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**:
28 801-812.
29 Tettelin, H., Radune, D., Kasif, S., Khouri, H., and Salzberg, S.L. (1999) Optimized multiplex
30 PCR: efficiently closing a whole-genome shotgun sequencing project. *Genomics* **62**: 500-507.
31 Wang, C., Sobral, B.W., and Williams, K.P. (2007) Loss of a universal tRNA feature. *J Bacteriol*
32 **189**: 1954-1962.
33 Wang, G., Murase, J., Asakawa, S., and Kimura, M. (2010) Unique viral capsid assemply protein
34 gene (*g20*) of cyanophages in the floodwater of a Japanese paddy field. *Biol Fertil Soils* **46**: 93-102.
35 Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F. et al. (2007)
36 Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and
37 *Synechococcus*. *Environ Microbiol* **9**: 1675-1695.
38 Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127-181.
39 Wiens, J.J. (2006) Missing data and the design of phylogenetic analyses. *J Biomed Informatics*
40 **39**:34-42.
41 Wilhelm, S.W., and Matteson, A.R. (2008) Freshwater and marine virioplankton: a brief overview
42 of commonalities and differences. *Freshwater Biology* **53**: 1076-1089.
43 Wilhelm, S.W., Carberry, M.J., Eldridge, M.L., Poorvin, L., Saxton, M.A., and Doblin, M.A.
44 (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity
45 assays and molecular analyses of g20 genes. *Appl Environ Microbiol* **72**: 4957-4963.
46 Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: viruses in aquatic ecosystems.

1 *Microbiol Mol Biol Rev* **64**: 69-114.
2 Yakhnin, A.V., Yakhnin, H., and Babitzke, P. (2008) Function of the Bacillus subtilis
3 transcription elongation factor NusG in hairpin-dependent RNA polymerase pausing in the trp
4 leader. *Proc Natl Acad Sci U S A* **105**: 16131-16136.
5 Yan, W., and Francklyn, C. (1994) Cytosine 73 is a discriminator nucleotide in vivo for histidyl-
6 tRNA in Escherichia coli. *J Biol Chem* **269**: 10022-10027.
7 Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y. et al. (2008) Ma-
8 LMM01 infecting toxic Microcystis aeruginosa illuminates diverse cyanophage genome strategies. *J*
9 *Bacteriol* **190**: 1762-1772.
10 Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G. and Béjà, O. (2005) Potential
11 photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral
12 intermediates. *Environ. Microbiol.* **7:**1505-13.
13 Zhang, J., and Inouye, M. (2002) MazG, a nucleoside triphosphate pyrophosphohydrolase,
14 interacts with Era, an essential GTPase in Escherichia coli. *J Bacteriol* **184**: 5323-5329.
15 Zhang, J., Zhang, Y., and Inouye, M. (2003) Thermotoga maritima MazG protein has both
16 nucleoside triphosphate pyrophosphohydrolase and pyrophosphatase activities. *J Biol Chem* **278**:
17 21408-21414.
18 Zhaxybayeva, O., Doolittle, W.F., Papke, R.T. and Gogarten, J.P. (2009) Intertwined evolutionary
19 histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Bio Evol* **1:**325-39.
20
21
22

1

1

**Table 2. tRNA genes in the S-CRM01 genome**

| tRNA gene | Identity | Anticodon | Nt positions | Strand |
|---|---|---|---|---|
| *t1* | Leu | UAA | 88617-544 | - |
| *t2* | Ile2 | CAU (LysAU)[1] | 88643-715 | - |
| *t3* | Gly | UCC | 127547-618 | + |
| *t4* | Arg | UCU | 127623-697 | + |
| *t5* | Lys | UUU | 128953-9025 | + |
| *t6* | Pro | UGG | 129165-236 | + |
| *t7* | Tyr | GUA | 129535-619 | + |
| *t8* | Gly | GCC | 130519-590 | + |
| *t9* | Asp | GUC | 130857-930 | + |
| *t10* | Trp | CCA | 130985-1055 | + |
| *t11* | Asn | GUU | 131716-787 | + |
| *t12* | Glu | UUC | 131866-941 | + |
| *t13* | Glu | CUC | 132043-115 | + |
| *t14* | Pro | UGG | 134674-748 | + |
| *t15* | Ile | GAU | 134944-5014 | + |
| *t16* | Ile | GAU | 135055-128 | + |
| *t17* | Phe | GAA | 135201-274 | + |
| *t18* | Ala | UGC | 135313-385 | + |
| *t19* | Val | GAC | 135762-833 | + |
| *t20* | Ser | GCU | 136579-663 | + |
| *t21* | Ser | GGA | 136668-759 | + |
| *t22* | Met | CAU | 139590-660 | + |
| *t23* | His | GUG | 139662-738[2] | + |
| *t24* | Thr | GGU | 138739-810 | + |
| *t25* | Thr | UGU | 141268-341 | + |
| *t26* | Gln | UUG | 141517-587 | + |
| *t27* | Lys | CUU | 141839-912 | + |
| *t28* | Cys | GCA | 142050-123 | + |
| *t29* | Leu | CAA | 145010-084 | + |
| *t30* | Arg | ACG | 149226-300 | + |
| *t31* | Ile2 | CAU (LysAU)[1] | 149448-521 | + |
| *t32* | Leu | GAG | 149554-627 | + |
| *t33* | Val | UAC | 149712-783 | + |

[1] Probable modification to lysidine-A-U; [2] Includes 3´ CCA

Multiple genes: Arg (2x): *t4, t30*; Glu (2x): *t12, t13*; Gly (2x): *t3, t8*; Ile (2x): *t15, t16*; Ile2 (2x): *t2, t31*; Leu (3x): *t1, t29, t32*; Lys (2x): *t5, t27*; Pro (2x): *t6, t14*; Ser (2x): *t20, t21*; Thr (2x): *t24, t25*; Val (2x): *t19, t33*.

Tightly juxtaposed genes are *t3/t4*, *t20/t21, t22/t23/t24*.

1 **FIGURE LEGENDS**
2
3 **Fig. 1 Distribution of S-CRM01 in the Klamath River valley, 2009.** Positive identifications were
4 made in samples from the Williamson River delta area of Upper Klamath Lake (WRD; 15 October),
5 Copco Reservoir (near dam; 13 October), Iron Gate Reservoir (IG, near dam, 18 August & 15
6 September), and at two sites on the lower Klamath River: Brown Bear sampling site at Horse Creek
7 (BB, 6 August & 15 September) and Seiad Valley (SV, 15 September).
8
9 **Fig. 2. Electron micrograph of S-CRM01 phage particles** negatively stained with
10 phosphotungstic acid. An intact particle is shown at left and a contracted particle at right. Note the
11 icosahedral head, prominent neck, two-ringed baseplate and injection tube.
12
13 **Fig. 3. Circular genetic map showing the gene organization of phage S-CRM01.** The genome
14 can be considered as comprised of Regions 1-3 (outer ring) that are dominated by structural genes,
15 replication-related genes, and unique genes, respectively. Subsequent rings represent: positive
16 strand ORFs, negative strand ORFs, tRNAs (maroon), and positive mass spectrometry identification
17 (mauve). The inner rings show GC content around a 50% mid-line (black), and GC skew (G+ =
18 cyan, C+ = blue). Putative ORF function is described by color as listed below the circular map.
19
20 **Fig. 4. Extensive synteny between S-CRM01 and marine cyanomyophages. A.** Dot plots
21 comparing gene order between S-CRM01 and phage T4 and indicated cyanomyophage genomes.
22 The S-CRM01 genome regions 1-3 (see Fig. 3) are indicated by colored shading and are labeled at
23 right. Syntenous segments I-IV are *gp13-gp46* (SCRM01 *g29-g55*), *gp5-td* (*g59-g80*), *nrdB-gp45*
24 (*g116-g156*), and *g166-167* (*cobS*), respectively. B. Diagram indicating the location of the inversion
25 between the S-CRM01 and S-PM2 genomes. Note that the P-SSM2 and S-PM2 genomes have
26 insertions between syntenous segments II and III, and that P-SSM2 and Syn9 genome numbering
27 convention places the regions syntenous with Region 1C at the start of the genomes.
28
29 **Fig. 5. Phylogenetic relationship of S-CRM01 to other myophages. A.** Consensus tree for the
30 most highly supported topology across all tested values of included non-universal genes (missing
31 data). Values at nodes indicate the average bootstrap value for all trees with that node/the number of
32 trees for which that node occurred out of a total of 27 trees (see Materials and Methods). The marine

1 cyanomyoviruses have been grouped into the clades MC1-4 as indicated. **B and C.** Consensus trees
2 showing alternative topologies among clades MC1-3 when including 0-10% and 15-40% included
3 non-universal genes, respectively. **D**. The number of orthologous clusters analyzed for each phage
4 as a function of % included non-universal genes are indicated and color-matched. The genes used
5 for each level of analysis are indicated in Table S2.
6
7 **Fig. 6. Late promoter consensus sequence.** Weblogo representation of the sequences of 81
8 predicted late promoters (Table S3), indicating the prevalence of both CTAAATA and ATAAATA
9 core sequences. Nucleotide 0 marks the predicted start of transcription based on T4 late promoters
10 (Miller et al., 2003a).
11
12
13
14

## Supplemental Figures and Tables

**Fig. S1. Klamath River watershed isolate LC16 is a member of the *Cyanobium gracile* cluster of *Synechococcus* sp.** Consensus phylogenetic tree depicting relationships between ITS sequences from *Synechococcus* isolates, including selected marine *Synechococcus* and *Prochlorococcus* isolates that serve as hosts for characterized cyanomyophages (indicated next to host names). The indicated *Cyanobium gracile* cluster of mostly freshwater isolates follows the definition of Ernst et al. (2003). The tree was constructed by maximum likelihood analysis with a TN93 (Tamura-Nei) substitution model using the PhyML package (Guindon and Gascuel, 2003). Bootstrap values are indicated.

**Fig. S2. Number of common genes among S-CRM01, marine cyanomyophages and T4-like phages.** Gene similarities are based on BLAST hits with at least a 1e-5 E-value. Colors indicate levels of similarity.

**Fig. S3. Phylogenetic tree showing relationship of S-CRM01 *gp20* gene to other *gp20* sequences from freshwater or marine phage sources.** See Fig. S1 for tree construction details. Cyanophages whose genomes have been sequenced are labeled by name and with a red hexagon. The group designations are taken from Sullivan et al. (2008). EU715812, closest to S-CRM01, represents phage P-ShM1 (Sullivan et al. (2008).

**Fig. S4. Phylogenetic tree showing relationship of S-CRM01 *psbA* gene to other *psbA* sequences from freshwater or marine phage sources.** See Fig. S1 for tree construction details. Cyanophages whose genomes have been sequenced are labeled by name and with a red hexagon. The group designations are taken from Chénard and Suttle (2008). S-CRM01 sits between two a freshwater clade (EU258991-6 and EU404136) and a marine-dominated clade containing phage S-SSM7.

**Table S1. Predicted S-CRM01 ORFs, detailed.** (see Excel file)

**Table S2. Table of genes used in phylogenomic analysis of Fig. 5.**

1 **Table S3. Mass spectrometry protein identification data.**
2
3 **Table S4. Predicted transcription control signals in the S-CRM01 genome.**
4
5 **Table S5. tRNA gene details**
6
7