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.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOI</td>
<td>10.1111/j.1462-2920.2011.02502.x</td>
</tr>
<tr>
<td>Publisher</td>
<td>John Wiley &amp; Sons Ltd.</td>
</tr>
<tr>
<td>Version</td>
<td>Accepted Manuscript</td>
</tr>
<tr>
<td>Terms of Use</td>
<td><a href="http://cdss.library.oregonstate.edu/sa-termsofuse">http://cdss.library.oregonstate.edu/sa-termsofuse</a></td>
</tr>
</tbody>
</table>
A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages.

Theo W. Dreher\textsuperscript{1,2*}, Nathan Brown\textsuperscript{1,†}, Connie S. Bozarth\textsuperscript{1}, Andrew D. Schwartz\textsuperscript{1}, Erin Riscoe\textsuperscript{1}, J. Cameron Thrash\textsuperscript{1}, Samuel E. Bennett\textsuperscript{3}, Shin-Cheng Tzeng\textsuperscript{4} and Claudia S. Maier\textsuperscript{4}

\textsuperscript{1}Department of Microbiology, \textsuperscript{2}Center for Genome Research and Bioinformatics, \textsuperscript{3}Department of Environmental & Molecular Toxicology, \textsuperscript{4}Department of Chemistry, Oregon State University, Corvallis, Oregon 97331.

Keywords: \textit{Myoviridae}, Klamath River, cyanomyophage, freshwater phage, \textit{nusG}, \textit{spoT}, phage tRNA, T4-like phage

\textbullet Corresponding author: Mailing address: Theo W. Dreher, Department of Microbiology, Oregon State University, Corvallis, Oregon 97331-3804. Phone: (541) 737-1795. Fax: (541) 737-0496. E-mail: theo.dreher@oregonstate.edu

\textdagger Present address: Pittsburgh Bacteriophage Institute, University of Pittsburgh, Pittsburgh, PA 15260
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*. 


INTRODUCTION

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

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

MATERIALS AND METHODS
Sample collection and cyanophage enrichment. Water was collected from the top 0.5 m of Copco Reservoir (mid-channel near the dam wall at latitude 41.979°N and longitude 122.333°W) during a *Microcystis* bloom on the Klamath River in Northern California on 10 September, 2008, and transferred to the laboratory in the dark on ice. In order to enrich for cyanophages present in the sample, a 200 mL aliquot of 0.2 μm filtered water was supplemented with 4 mL of 50x BG-11 medium (Sigma-Aldrich, St. Louis, MO) and 20 mL of *Microcystis*-dominated cultures derived from the Klamath River system (August 2007). The enriched water sample was incubated at 24°C under fluorescent lamps at 10 μE/m²/s for 10 days. Water quality data relevant to the collected sample are available at http://www.pacificorp.com/es/hydro/hl/kr.html#.

Cyanophage isolation and amplification. The enriched culture was treated with chloroform, and cellular material was removed by centrifugation. The 0.2 μm-filtered supernatant (Supor-200; Pall Life Sciences) was ultracentrifuged using a Ti60 rotor (Beckman) at 4°C and 177,520 g, for 90 min. Pellets were resuspended in SM (50 mM Tris-HCl, 100 mM NaCl, 10 mM MgSO₄·7H₂O, pH 7.5) and stored at 4°C. Plaque assays were conducted using BG-11 top agar layered onto BG-11 agar plates that were incubated under the growth conditions described above until plaques were visible (~ one week). Three serial plaque isolations were performed using the Klamath River system culture, which was also used for phage amplification in liquid culture. Phage particles were collected by ultracentrifugation. During the course of phage isolation, it was noticed that the culture characteristics had changed. The identity of the resultant culture (LC16) was examined by PCR amplification of genomic DNA using the cyanobacteria-wide primers CS1F and ULR directed at the internal transcribed spacer of the rRNA operon (ITS), followed by DNA sequencing of the PCR product as described (Bozarth et al., 2010). A single derived sequence indicated culture purity. Comparison of the sequence with the GenBank database identified LC16 as a member of the *Cyanobium gracile* cluster that includes freshwater *Synechococcus* isolates (Ernst et al., 2003; Chen et al., 2006) (Fig. S1). The closest known relative has been isolated from Lake Balaton (Hungary), with other related isolates from freshwater sources in Germany and Wisconsin (USA) and brackish or saline sources in California, Baltic Sea (Denmark) and White Sea (Russia).
**Genome structure analysis.** Genome size was estimated using pulsed field gel electrophoresis (PFGE) after genome preparation as described (Lingohr et al., 2008), using a 1.4% agarose gel in a CHEF II PFGE unit (Bio-Rad) set to run at 6 V cm\(^{-1}\) for 18 h with a 0.1 sec switch time.

To determine whether the genome was linear, circular or circularly permuted, BAL-31 nuclease digestion of the phage genome was performed as described (Yoshida et al., 2008). Purified phage DNA (200 ng) was incubated with 0.1 U/µl BAL-31 nuclease (NEBiolabs) at 30°C for 0, 10, 20, 40 and 60 minutes. The DNA was then extracted with phenol/chloroform, ethanol precipitated and digested overnight at 37°C with BamH1 endonuclease. The restriction products were separated on 0.8% agarose gel and visualized with ethidium bromide. 1kB DNA ladder (GeneRuler, Fermentas) and a plasmid (pGEM, Promega) were included as linear and supercoiled circular controls.

**Genome sequencing.** The genome (500 ng) was sequenced by Roche, Inc. (Branford, CT) using GS FLX Titanium Sequencing, accumulating 232 kbp at an average coverage of 350X. A first draft of the genome sequence was assembled using Newbler (gsAssembler) software (Roche). Five contigs larger than 5,000 bp in length were assembled, amounting to a total of 173.5 kbp. To determine the order of contigs and to fill the gaps, a multiplex PCR approach was used (Tettelin et al., 1999). Briefly, pairs of outward-oriented primers positioned about 100 bp from the two ends of each contig were used in all combinations to direct PCR amplifications using genomic DNA as template. Products visible on gels after 30 cycles were confirmed by PCR using individual primer pairs, and were extracted from agarose gels and submitted for direct Sanger sequencing (Genewiz, Inc.). Each gap was sequenced in both directions, with additional primers designed when needed. Contigs and junction sequences were arranged and assembled in Geneious (Biomatters, Ltd., http://www.geneious.com). The gap sequences could be confirmed using smaller contigs and unassembled reads from the 454 data. The average sequencing coverage for contigs ranged between 360 and 500 with an average Phred Equivalent (Roche) quality score of 64.0.

**Genome annotation and analysis.** Regions of coding sequence were predicted using Glimmer 3 (Delcher et al., 1999) and Genemark S (Besemer et al., 2001); tRNAs were predicted using tRNAscan-SE (Lowe and Eddy, 1997). Annotated protein coding sequences were determined using a BLASTx search against the NCBI nr database. Annotations were made based on an E-Value 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 NH₄HCO₃ 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 CapTrapp
a C18 column (Agilent Zorbax 300SB-C18, 250 x 0.3 mm, 5 μm). A binary gradient system was used consisting of solvent A, 0.1% aqueous formic acid and solvent B, acetonitrile containing 0.1% formic acid. Peptides were trapped and washed with 1% solvent B for 3 min. Peptide separation was achieved using a linear gradient from 10% B to 30% B at a flow rate of 4 μL/min over 35 minutes.

For the LC-MS/MS analysis, the LTQ-FT mass spectrometer was operated in a data-dependent mode. A full FT-MS scan (m/z 350-2000) was alternated with collision-induced dissociation (CID) MS/MS scans of the 5 most abundant doubly or triply charged precursor ions. As the survey scan was acquired in the ICR cell, the CID experiments were performed in the linear ion trap where precursor ions were isolated and subjected to CID in parallel with the completion of the full FT-MS scan. CID was performed with helium gas at a normalized collision energy 35% and activation time of 30 ms. Automated gain control (AGC) was used to accumulate sufficient precursor ions (target value, 5 x 10⁴/micro scan; maximum fill time 0.2 s). Dynamic exclusion was used with a repeat count of 1 and exclusion duration of 60 s. Data acquisition was controlled by Xcalibur (version 2.0.5) software (Thermo).

For the sequence database search, Thermo RAW data files were processed with Proteome Discoverer v1.0 using default parameters except for a S/N threshold setting of 10. A Mascot (v2.2.04) search of a phage-encoded protein database (299 sequences; 54263 residues) was launched from Proteome Discoverer with the following parameters: the digestion enzyme was set to Trypsin/P and two missed cleavage sites were allowed. The precursor ion mass tolerance was set to 10 ppm, while fragment ion tolerance of 0.8 Da was used. Dynamic modifications included carbamidomethyl (+57.0214 Da) for Cys and oxidation (+15.9994 Da) for Met.

**PCR to detect S-CRM01 in the Klamath system.** Enriched phage fractions were prepared (see above) from samples collected from the Klamath River system during August-October, 2009. Multiple plaques isolated on LC16 top agar plates from each sample were tested for the presence of S-CRM01 by PCR analysis using two primer pairs directed at g44 (gp23) and g34 (Ma-LMM01-like hypothetical protein). In all cases, reactions were either positive or negative for both primer pairs. The primer pair for detecting g34 was CRM01-g34(F) 5’ GTCAAATAGAATCCAGGATGAATTA and CRM01-g34(R) 5’ TACCATAGTCTCCACCGTTTC. The primer pair detecting g44 was CRM01-g44(F) 5’ GACGTATGTGGCGTTCAGCCAATGA and CRM01-g44(R) 5’ CGGTTGATTCTGCAAGGATTTTC. PCR reactions used High Fidelity Platinum Taq polymerase.
(Invitrogen) in the provided buffer with 0.2 µM of each primer, 0.2 mM dNTPs, and 2.5 mM 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 were run. These primers were designed to be specific for S-CRM01, avoiding amplification from known related cyanomyophage genomes; detection of S-CRM01 was scored as positive only when products of the expected size were amplified with both primer pairs.

RESULTS AND DISCUSSION

Isolation and physical characteristics. Phage S-CRM01 was isolated from a surface sample taken in September 2008 from a *Microcystis*-dominated bloom in Copco Reservoir on the Klamath River in Northern California. At 319.8 river km from the coast, this site is far removed from salt water habitats. S-CRM01 infects and lyses LC16, a culture belonging to the *Cyanobium gracile* cluster of mostly freshwater *Synechococcus* (see Materials and Methods). We have observed lytic infection of no other hosts, including the freshwater cyanobacteria *Microcystis aeruginosa*, *Synechococcus elongatus* PCC 7942 and *Synechocystis sp.* PCC6803. During 2009, a survey was conducted to assess the distribution of S-CRM01 between Upper Klamath Lake and the Klamath River estuary. On the basis of PCR identification of plaques on LC16 plates, S-CRM01 was present across a length of about 250 km along the Klamath River valley, from the Williamson River Delta at the northern end of Upper Klamath Lake as far downstream as Seiad Valley (Fig. 1).

Negative staining electron microscopy revealed a T4-like morphology, with isometric heads about 85-100 nm in diameter and rigid contractile tails 15-20 nm in diameter and 140-170 nm long (Fig. 2). Contracted tails reveal a sheath and core 20-30 nm and about 10 nm in diameter, respectively. A double-ringed baseplate and narrow neck are visible.

The genome migrated as a 180 kbp band in PFGE. Nuclease BAL-31 digestion followed by cleavage with *Bam*HI resulted in simultaneous loss of material from all bands, indicative of circularly permuted linear DNA (not shown). This is consistent with the presence of a gene homologous to the *gp17* large terminase subunit of phage T4, which determines a circularly permuted packaging strategy (Casjens and Gilcrease, 2008).

S-CRM01 is a “photosynthetic” cyanomyophage most closely related to marine phages of *Synechococcus* and *Prochlorococcus*. Consistent with a circularly permuted organization and its migration in PFGE, the S-CRM01 genome assembled into a circular map 178,563 bp long (Fig. 3).
The G+C content of the genome is 39.7%, and 297 protein-coding genes and 33 tRNA genes are predicted. Protein-coding genes have been annotated with the following conventions (Table 1): S-CRM01 genes are designated as $g1$ etc.; in addition, genes with homology to numbered phage T4 genes are designated as $gp1$, etc.; genes lacking direct homology with T4 but with similarity to other cyanomyophage genes that themselves have significant homology to T4 are designated as $gp$-like; other genes are designated with conventional gene names, e.g., $nrdA$.

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

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

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

The S-CRM01 genome possesses an unusually large number (182) of ORFs with no significant homologs in the GenBank database (as of 30 August, 2010), representing 61% of predicted ORFs (genes shown in black in Fig. 3). This is far higher than the number of unique genes in the marine 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
genes during viral infection, indicating that these proteins (4 of which are unique to S-CRM01) are functionally relevant.

As expected, abundant mass spectrometry signals were registered for gp23 major capsid protein \( (g44) \), gp18 contractile tail sheath protein \((g36)\) and gp19 tail tube protein \((g37)\), which are present in multiple copy number in the T4 capsid (Miller et al., 2003a). Abundant signals were also observed for peptides from g9, at 271 kDa the largest protein encoded by the S-CRM01 genome. This huge protein has a predicted strong \( \beta \)-strand character and no identifiable sequence motifs, but has BLAST matches to a wide range of phage proteins (including putative tail protein) and to glycosyl hydrolase bacterial neuraminidase repeat (BNR) proteins. An internal dot plot analysis shows numerous different internal repeat elements 10 to 20 residues long, mostly repeated only once. These properties suggest that g9 is a tail fiber gene (Weigele et al., 2007), although we have not observed tail fibers under the electron microscope.

Three proteins with collagen-like triple helix repeat domains were detected by mass spectrometry: g13 with 36 GXY repeats, g20 with 85 GXQ repeats, and g23 with 40 GXQ repeats. Collagen-like proteins have been reported from some phage, participating in spike formation in phage PRD1 (Caldentey et al., 2000) and suggested to be found in tail fibers (Smith et al., 1998; Sullivan et al., 2005). More careful virion characterization will be needed to determine whether S-CRM01 possesses the tail fibers or whiskers that are predicted by the presence of these proteins.

**T4-like nonstructural genes predominantly on the minus strand.** A second major region of synteny with the exoT-even marine cyanomyophage genomes is the 51 kbp segment between genes \( g69 \) and \( g165 \) (Region 2, nts 72,094–123,515) (Figs. 3, 4; Table 1)(c.f. Sullivan et al., 2010). These genes, all of which are on the minus strand, include core T4-like genes with roles in DNA replication, recombination and repair (e.g., \( gp43 \) DNA polymerase, \( gp61 \) primase, \( gp45 \) sliding clamp, \( uvsX \)), nucleotide metabolism (\( nrdA, nrdB, nrdC, td \) thymidylate synthase) and 5 of the 6 marine cyanophage signature genes (\( phoH, hli03, psbA, mazG, hsp20 \)). Other significant genes also present in Region 2 are: a member of the 2OG-FeII oxygenase superfamily that is present in nearly all marine cyanomyophages, a \( nusG \) transcription anti-termination factor homolog (not found in T4-related phages), \( regA \) regulator of early gene translation (common to all marine cyanomyo and T4-like phages), and RNase H (\( rnh \), common to T4-like phages but found in only S-PM2 and Syn19 among the marine cyanomyophages).
This segment of the genome is also predicted to be expressed predominantly via late transcription, though some putative early promoters have been identified, notably for the expression of gp45, gp44, gp62 and gp33, all of which are involved in establishing and maintaining T4 late gene transcription (Miller et al., 2003a). The other key gene needed for late transcription, gp55 alternative sigma factor, is located in Region 1B between the two structural gene segments and on the plus strand, and it also appears to be expressed from an early promoter (Table S4). The recombination genes uvsY, uvsW, gp47 and gp46 are also located in Region 1B, all on the plus strand in another region of synteny with marine cyanomyophage genomes (Figs. 3, 4).

As in the structural gene region, but even more so, the conserved genes in Region 2 are interspersed with additional genes at multiple sites. Forty-eight (49%) of ORFs in Region 2 are unique to S-CRM01.

49 kbp (27%) of the genome contains 132 ORFs, 85% of which are unique genes. The remainder of the genome (Region 3, genes g166-g297) possesses no overall similarity to the genomes of other phages apart from g166 and g167 (Fig. 4) and contains only 20 genes with homology to previously identified ORFs. Fifteen of these have homologs in at least one T4-like or cyanomyophage. This part of the genome also contains all but two of the 33 tRNA genes. The significant genes present in this segment are cobS (g167), DNA ligase (g194), speD (g109), spoT (g215) and HNH endonuclease (g205). cobS, involved in cobalamin biosynthesis, may support the synthesis of deoxynucleotides (Sullivan et al., 2005) and is one of the marine cyanophage signature genes (Millard et al., 2009). DNA ligase is common in T4-like phage genomes, though not found in the marine cyanomyophages; the S-CRM01 DNA ligase gene is most closely related to Chorella virus and PB1-like myophage ligases. speD encodes S-adenosylmethionine decarboxylase, which is involved in polyamine synthesis and found in a few marine cyanomyophage genomes (Mann et al., 2005). spoT, which encodes ppGpp synthetase and hydrolase, has not previously been recognized in myophage genomes but is present in the Aeromonas myophages Aeh1, 44RR, PHG25 and PHG31. HNH endonuclease is possibly a member of a family of homing endonucleases, although no introns have been detected in the S-CRM01 genome.

All but one of the genes (g186) in Region 3, including the tRNA genes, reside on the plus strand. Region 3 genes are atypical in a number of ways. Identifiable promoters are sparse; the G+C content 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
promoter recognition. Interestingly, S-CRM01 gp55 is clearly more closely related to marine cyanomyophage gp55 genes (which recognize ATAAATA promoters) than to RB49 or Aeh1 gp55.

A full set of tRNAs. The 33 tRNA genes in the S-CRM01 genome (Table 2) is more than found in some bacteria and includes all of the 20 amino acid specificities, although there is no initiator tRNA$^{\text{Met}}$ as found in some phage genomes. KVP40 also has all 20 specificities represented, though some are thought to be pseudogenes that may not be functional (Miller et al., 2003b). The S-PM2 genome contains 24 tRNA genes representing all but the cysteine and phenylalanine specificities, but the other marine cyanomyophage genomes have much smaller subsets of these. tRNA$^{\text{His}}$ is the only S-CRM01 tRNA gene with an encoded 3’ CCA terminus, while in S-PM2 five tRNA genes (including tRNA$^{\text{His}}$) include the 3’ CCA. Three tRNA genes with a CAU anticodon are present. Based on features in the anticodon loop (C32 and A38) that are recognized by the TilS enzyme, which modifies C34 in the anticodon to lysidine (Nakanishi et al., 2009), the $t2$ and $t31$ genes are proposed to be tRNA$^{\text{Ile}}$ genes, leaving one tRNA$^{\text{Met}}$ gene ($t22$). Gene $t2$ has multiple additional nucleotides that are proposed to be characteristic of lysidine-containing tRNA$^{\text{Ile}}$ in cyanobacteria (Freyhult et al., 2007)(see Table S5).

It cannot be determined without experimentation whether all predicted tRNAs are functional, although key identity elements and conserved features (Giegé et al., 1998) are generally present (Table S5). Some of the predicted tRNAs do have highly unusual features (see Table S5 and refer to tRNA database; Juhling et al., 2009) that may compromise function: e.g., U73 and long variable loop in tRNA$^{\text{Glu}}$ $t12$, A1 in tRNA$^{\text{Glu}}$ $t13$, U73 in tRNA$^{\text{Ile}}$ $t31$, U10;U25 in tRNA$^{\text{Ser}}$ $t21$ (probable cause for identification as possible pseudogene by tRNAscan-SE). The tRNA$^{\text{His}}$ gene $t23$, like other phage tRNA$^{\text{His}}$ occurrences (e.g., S-PM2), lacks the additional 5’ residue G-1 that is in most systems critical for histidine identity (Giegé et al., 1998). G-1 would be either replaced with U-1 (adjacent nucleotide in the genome) or lacking in the mature S-CRM01 tRNA$^{\text{His}}$, a situation that in E. coli allows only partial translational function (Yan and Francklyn, 1994). Another possibility is that G-1 could be added post-transcriptionally by a Thg1-like activity (Heinemann et al., 2010) or that a variant histidyl-tRNA synthetase that does not rely on G-1 is present in host cells (Wang et al., 2007). No Thg1 or histidyl-tRNA synthetase is evident among the S-CRM01 genes. Establishing the functionality of phage tRNA genes is important in determining their role during infection and the 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 \textit{nrdA} and \textit{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, \textit{spoT} and \textit{nusG} (indicated in pink in Fig. 3) are unusual genes in T4-like phages. SpoT is a pyrophosphohydrolase that in \textit{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
that NusG functions in concert with Rho to decrease doubling time and prevent expression of the
cryptic rac prophage in *E. coli* MG1655. Over-expression of NusG by S-CRM01 may be able to
increase the metabolic activity of the infected host and prevent interference or competition for
transcriptional machinery from prophages endemic to the host. If such benefits do exist, the absence
of identifiable *nusG* from other phage genomes might suggest that unrecognized transcriptional
regulators are encoded by other phages.

**Relationship between freshwater and marine phages.** There has not been enough data to
derive a clear picture of the genetic relationships between phages in freshwater and marine
environments. Some early observations with podophages (Breitbart et al., 2004) and myophages
(Short and Suttle, 2005) emphasized the discovery of very similar sequences in the viral populations
of the two aquatic environments. For myophages, additional phylogenetic studies with the *gp20*,
*psbA* and *psbD* genes have verified this observation, but have also shown that most (though not all)
freshwater sequences map at high resolution to clades that do separate from marine isolates
(Wilhelm et al., 2006; Chénard and Suttle, 2008; Sullivan et al., 2008; Wang et al., 2010). S-
CRM01 seems to represent both of these scenarios, with the closest currently known *gp20* sequence
found in the Atlantic Ocean isolate P-ShM1 (Sullivan et al., 2008)(Fig. S3). On the other hand, the
*psbA* gene locates to a unique branch between freshwater and marine clades (Fig. S3).

The complete S-CRM01 genome sequence allows more meaningful consideration of the
relationship between related marine and freshwater phages. The data in Fig. 5 show the
relationships to be complex. On the basis of synteny (Fig. 4A) and protein-coding (Fig. 5D) and
tRNA gene content, S-CRM01 is most similar to S-PM2, while on the basis of some phylogenomic
comparisons, P-SSM2, S-SM2 and S-SSM7 can be considered more similar (Fig. 5D). It thus seems
likely that there have been multiple gene exchanges between the S-CRM01 and marine
cyanomyophage lineages during evolution. Brackish estuarine waters are likely connections
between freshwater and marine environments, where genetic exchanges could occur between phages
that principally exist in one or the other aquatic habitat, especially for phages of bacteria such as
*Synechococcus*, which have broad distributions. Highly unique phages do seem to exist in
freshwater, however, such as the *Microcystis*-infecting Ma-LMM01 (Yoshida et al., 2008), and it
will be interesting to learn in the future whether this is the pattern for phages infecting hosts that do
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


Table 2. tRNA genes in the S-CRM01 genome

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

\(^1\) 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 \(t3/t4, t20/t21, t22/t23/t24\).
FIGURE LEGENDS

Fig. 1  Distribution of S-CRM01 in the Klamath River valley, 2009. Positive identifications were made in samples from the Williamson River delta area of Upper Klamath Lake (WRD; 15 October), Copco Reservoir (near dam; 13 October), Iron Gate Reservoir (IG, near dam, 18 August & 15 September), and at two sites on the lower Klamath River: Brown Bear sampling site at Horse Creek (BB, 6 August & 15 September) and Seiad Valley (SV, 15 September).

Fig. 2. Electron micrograph of S-CRM01 phage particles negatively stained with phosphotungstic acid. An intact particle is shown at left and a contracted particle at right. Note the icosahedral head, prominent neck, two-ring baseplate and injection tube.

Fig. 3. Circular genetic map showing the gene organization of phage S-CRM01. The genome can be considered as comprised of Regions 1-3 (outer ring) that are dominated by structural genes, replication-related genes, and unique genes, respectively. Subsequent rings represent: positive strand ORFs, negative strand ORFs, tRNAs (maroon), and positive mass spectrometry identification (mauve). The inner rings show GC content around a 50% mid-line (black), and GC skew (G+ = cyan, C+ = blue). Putative ORF function is described by color as listed below the circular map.

Fig. 4. Extensive synteny between S-CRM01 and marine cyanomyophages. A. Dot plots comparing gene order between S-CRM01 and phage T4 and indicated cyanomyophage genomes. The S-CRM01 genome regions 1-3 (see Fig. 3) are indicated by colored shading and are labeled at right. Syntenous segments I-IV are gp13-gp46 (SCRM01 g29-g55), gp5-td (g59-g80), nrdB-gp45 (g116-g156), and g166-167 (cobS), respectively. B. Diagram indicating the location of the inversion between the S-CRM01 and S-PM2 genomes. Note that the P-SSM2 and S-PM2 genomes have insertions between syntenous segments II and III, and that P-SSM2 and Syn9 genome numbering convention places the regions syntenous with Region 1C at the start of the genomes.

Fig. 5. Phylogenetic relationship of S-CRM01 to other myophages. A. Consensus tree for the most highly supported topology across all tested values of included non-universal genes (missing data). Values at nodes indicate the average bootstrap value for all trees with that node/the number of trees for which that node occurred out of a total of 27 trees (see Materials and Methods). The marine
cyanomyoviruses have been grouped into the clades MC1-4 as indicated. **B and C.** Consensus trees showing alternative topologies among clades MC1-3 when including 0-10% and 15-40% included non-universal genes, respectively. **D.** The number of orthologous clusters analyzed for each phage as a function of % included non-universal genes are indicated and color-matched. The genes used for each level of analysis are indicated in Table S2.

**Fig. 6. Late promoter consensus sequence.** Weblogo representation of the sequences of 81 predicted late promoters (Table S3), indicating the prevalence of both CTAAATA and ATAAATA core sequences. Nucleotide 0 marks the predicted start of transcription based on T4 late promoters (Miller et al., 2003a).
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.
Table S3. Mass spectrometry protein identification data.

Table S4. Predicted transcription control signals in the S-CRM01 genome.

Table S5. tRNA gene details