Supporting Information

Yau et al. 10.1073/pnas.1018221108

SI Materials and Methods

Samples and DNA Sequencing. Water samples collected from Organic Lake were (*i*) surface water from the eastern side of the ice-free lake (68° 27' 25.48" S, 78° 11' 28.06" E), December 24, 2006; (*ii*) a depth profile collected through a 30-cm hold drilled through the surface ice above the deepest point in the lake (68° 27' 22.15" S, 78° 11' 23.95" E), November 10, 2008; and (*iii*) surface water from the northeast side of the partially ice-covered lake (68° 27' 21.02" S, 78° 11' 42.42" E), December 12, 2008. Samples were sequentially filtered through a 20-μm prefilter and biomass captured onto 3.0-, 0.8-, and 0.1-μm membrane filters as described previously (1, 2). The samples from 2008 also included 50% (vol/vol) RNAlater. DNA extraction, sequencing, and quality validation was performed as previously described (1, 2). DNA sequencing was performed at the J. Craig Venter Institute in Rockville, MD.

Transmission Electron Microscopy. Unfiltered Organic Lake surface water from December 24, 2006 (fixed on site in 1% vol/vol formalin) was concentrated and a solvent exchange performed with sterile filtered ammonium acetate solution 1% (wt/vol) using a 50-kDa cutoff Microcon centrifugal filter device (Millipore) according to the manufacturer's instructions. Formvar-coated 200-mesh copper grids were floated on a droplet of sample for 30 min, excess liquid wicked off, and the grid negatively stained for 30 s with uranyl acetate 2% (wt/vol). The sample was visualized using a JEOL1400 transmission electron microscope (JEOL) at 100 kV at 150,000–250,000× magnification.

Metagenomic Assembly and Annotation. Mosaic metagenomic assemblies were generated as previously described (1, 2). For the 0.1-μm Organic Lake 2006 sample, assembly was a hybrid of Sanger and 454 read data (Table S1). For all other sample size fractions, runtime parameters used were standard for 454 sequencing data. Low-GC (≥51%) scaffolds >10 kb from the 0.1-μm 2006 assembly had high coverage (>45×), indicating that these were from the dominant taxa. One of these scaffolds was binned as virophage and the rest as phycodnavirus (PV). To further separate the Organic Lake phycodnavirus (OLPV) types and assess the completeness of their genomic content, highly conserved single-copy PV orthologs were identified as follows. An all-against-all BLASTp search was conducted with protein sequences from the 10 available PV genomes (*Acanthocystis turfacea* Chlorella virus 1, PbCV-1, PbCV AR158, PbCV FR483, PbCV NY2A, *Emiliania huxleyi* virus 86, *Ectocarpus siliculosus* virus 1, *Feldmannia* sp. virus, *Ostreococcus* virus 5, and *Ostreococcus tauri* virus 1) and *Acanthamoeba polyphaga* mimivirus (APMV) (which was included as a close PV relative). BLASTp results were parsed and clustered using orthoMCL V1.4 (3, 4). Pairs of each ortholog were located on eight of the PV scaffolds. The location of each ortholog pair had a complementary distribution, so the eight scaffolds were able to be sorted unambiguously into two strains, OLPV-1 and OLPV-2. OLPV-1 ribonucleotide reductase α-subunit appeared as duplicated on different scaffold ends, likely as an artifact of its proximity to an assembly break point. The remaining high-coverage scaffolds were searched for predicted proteins present in one OLPV strain but not in the other and assigned to the strain in which it was absent. Comparison of OLPV-1 and OLPV-2 scaffolds was performed using tBLASTn of concatenated scaffolds from each strain and visualized using the Artemis Comparison Tool (ACT) (5). DNA sequence data are available in GenBank and CAMERA (http://web.camera.calit2.net).

Organic Lake Virophage Genome Completion and Annotation. The high coverage (77×), large number of Sputnik homologs that encode essential functions, and length of the putative Organic Lake virophage (OLV) scaffold from the 0.1-μm 2006 hybrid assembly indicated that it was a near-complete genome. Reads from this scaffold were reassembled at high stringency and visualized using Phred/Phrap/Consed (6) to complete the sequence. Mate-pair data indicated a circular molecule, and primers were designed to span the ends of the scaffold and sequence across the gap. Touch-down PCR was performed with eDNA from the 0.1-μm 2006 sample, the product used for nested PCR, and the final product was cloned and sequenced. The complete genome was manually annotated and visualized using Artemis (7). Translated ORFs (minimal size 120 aa) were compared (BLASTp) with GenBank, with the all-metagenomic ORF peptide database on CAMERA (http://web.camera.calit2.net), and with predicted proteins from OLPV-1 and OLPV-2 scaffolds. Comparisons between the OLV genome and OLPV-1 /OLPV-2 were performed with tBLASTn and visualized using ACT (5). Primers used to close the OLV genome were as follows (listed as primer function, name, orientation, sequence 5′-3′): Outer gap spanning primer, SY11, forward, TTG TCT TAT GTA TTA CAA ATC ATT GAA; Outer gap spanning primer, SY12, reverse, CGA CAT TAA TCG GTT GTT TT; Nested gap spanning primer, SY13, forward, GCA TTA CGA ATG TGT TCC AG; Nested gap spanning primer, SY14, reverse, TTC TCC GTG ATT GAT ATC GT; Sequencing primer, SY23, forward, TCC CTA TTG ATG TCA AAA CC; and Sequencing primer, SY24, forward, GAT TCT GGT TGG AGC ATA TAT TT.

Phylogenetic Analysis. Translated amino acid sequences from viral marker genes of interest were retrieved from the 0.1-μm 2006 metagenomic assemblies from this study, GenBank, and CAMERA all-metagenomic reads ORF peptide database. Homologous sequences were aligned using MUSCLE v3.6 (8). Neighbor-joining analysis, test for clade support (bootstrap analysis, 2,000 replicates), and tree drawing was performed with Molecular Evolutionary Genetics Analysis (MEGA) software v4 (9). Maximum likelihood analysis (JTT substitution model) and test for clade support (aLRT analysis) was performed with PhyML (10) and the tree visualized using MEGA. 18S rRNA gene sequences were retrieved from reads of all filter sizes, compared (BLASTn, e-value <1.0e-5) with the SILVA100 SSURef database, aligned, and phylogeny performed using ARB software as previously described (1, 2). The abundance and similarity of virophages in all lake samples and filter sizes was estimated using BLASTp (e-value <1.0e-5) to search using the OLV major capsid protein (MCP) sequence against a database of proteins predicted from sequencing reads. The database was generated as previously described (1), and the percentage identity of the BLAST hit was used as a proxy for species similarity.

Metaproteomic Analysis. Metaproteomics of proteins from the 0.1-μm filter from 2006 was performed as previously described (1, 2), with minor modifications. The protein sequence database was generated by combining ORFs from the 3.0-, 0.8-, and 0.1-μm mosaic assemblies with 130,581 sequences in the database. Scaffold 3.0 (version 3_00_05; Proteome Software) was used to

validate MS/MS-based peptide and protein identifications. Protein identification data are available in Table S2.

Model of Algal Host–Virus and Virophage Dynamics. To model the effect a virophage would have on algal *Pyramimonas* algal host populations in Organic Lake, modified Lotka-Volterra equations were used describing the OLV as a predator of predator OLPV. The original equations are given by:

$$\frac{dA}{dt} = \alpha A - \varepsilon PA \qquad \textbf{[S1]}$$

$$\frac{dP}{dt} = \theta PA - \mu P, \qquad \textbf{[S2]}$$

where A is the number of *Pyramimonas* (prey), P is the number of OLPV (predator), α and θ are the specific growth and production rates of the prey and predator, respectively, ε is the rate of predator-mediated death of prey and μ is the specific decay rate of the predator. Eq. **S1** describes the change in *Pyramimonas* abundance and Eq. **S2** the change in OLPV abundance in the absence of OLV. In the presence of OLV, *Pyramimonas,* OLPV and OLV dynamics are described by the following equations:

$$\frac{dP}{dt} = \theta PA - \mu P - \omega PV \qquad \textbf{[S3]}$$

$$\frac{dV}{dt} = \beta PV - \gamma V, \qquad \textbf{[S4]}$$

where V is the number of OLV (predator of predator), ω the rate of OLV mediated reduction in OLPV infective particles, and β and γ the production and decay rates of OLV, respectively. Eq. **S3** is a modified version of Eq. **S2**, which includes the effect of OLV on the change in abundance of OLPV. Eq. **S4** describes the growth properties of OLV as a predator of OLPV. Values for the variables for the solution shown (Fig. 4, main text) were as follows: initial prey (10), predator (1), and predator of predator (10) numbers, $\alpha = 0.1$, $\theta = 0.015$, $\varepsilon = 0.01$, $\mu = 0.05$, $\omega = 0.01$, $\beta = 0.015$, and $\gamma = 0.15$). COmplex PAthway Simulator (COPASI) software (11) was used to simulate prey, predator, and predator of predator dynamics using the deterministic (LSODA) method.

1. Lauro FM, et al. (2010) An integrative study of a meromictic lake ecosystem in Antarctica. *ISME J*, 10.1038/ismej.2010.185.
2. Ng C, et al. (2010) Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *ISME J* 4:1002–1019.
3. Li L, Jr., Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.
4. Chen F, Mackey AJ, Jr., Stoeckert CJ, Jr., Roos DS (2006) OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34 (Database issue):D363–D368.
5. Carver TJ, et al. (2005) ACT: The Artemis comparison tool. *Bioinformatics* 21:3422–3423.
6. Gordon D (2004) Viewing and editing assembled sequences using Consed. *Current Protocols in Bioinformatics*, eds Baxevanis AD, Davison DB (John Wiley & Sons, New York), Sections 11.2.1–11.2.43.
7. Rutherford K, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16:944–945.
8. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
9. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* 9:299–306.
10. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
11. Hoops S, et al. (2006) COPASI—a COmplex PAthway SImulator. *Bioinformatics* 22: 3067–3074.

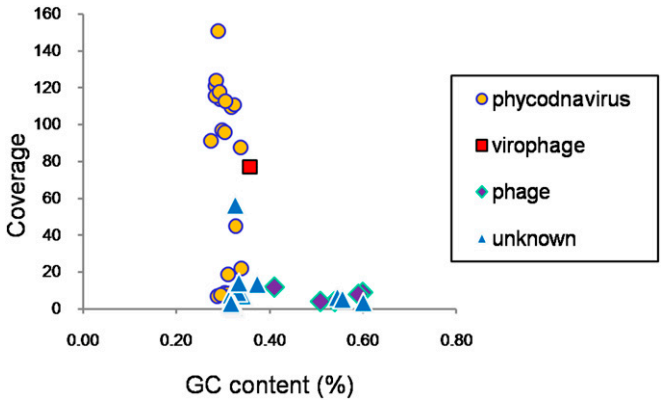

Fig. S1. Plot of GC% vs. coverage for the 2006 Organic Lake 0.1-μm hybrid assembly scaffolds >10 kb in size.

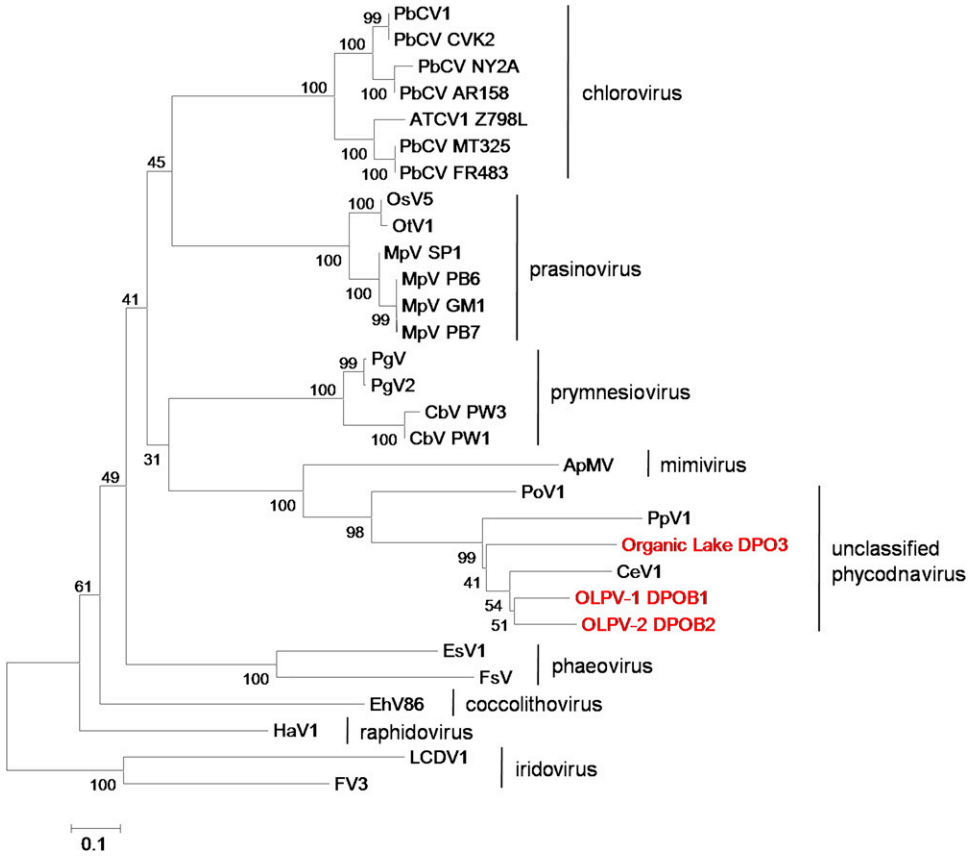

Fig. S2. Phylogeny of OLPV using B family DNA polymerase sequences. Neighbor-joining tree of B family DNA polymerase amino acid sequences of OLPV, and nucleo-cytoplasmic large DNA virus (NCLDV) sequences from GenBank. Abbreviations and accession numbers from bottom to top: PbCV1, *Paramecium bursaria* chlorella virus 1 (AAC00532.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35142.1); PbCV NY2A, *P. bursaria* chlorella virus NY2A (ABT14648.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU43776.1); AtCV1, *Acathocystis turfacea* chlorella virus (ABT16932.1); PbCV MT325, *P. bursaria* chlorella virus MT325 (ABT13573.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15308.1); OsV5, *Ostreococcus* virus 5 (ABY28020.1); OtV1, *Ostreococcus tauri* virus 1(YP_003495047.1); MpV SP1, *Micromonas pusilla* virus SP1(AAB66713.1); MpV PB6, *M. pusilla* virus PB6 (AAB49743.1); MpV GM1, *M. pusilla* virus GM1 (AAB49742.1); MpV PB7, *M. pusilla* virus PB7 (AAB49744.1); CbV PW1, *Chrysochromulina brevifilum* virus PW1 (AAB49739.1); CbV PW3, *C. brevifilum* virus PW3 (AAB49740.1); ApMV, *Acathamoeba polyphaga* mimivirus (AAV50591.1); PoV, *Pyramimonas orientalis* virus (ABU23717.1); PpV, *Phaeocystis pouchetii* virus (ABU23718.1); CeV1, *Chrysochromulina ericinia* virus 1(ABU23716.1); EsV1, *Ectocarpus silicolosus* virus (AAK14511.1); FsV, *Feldmannia* sp. virus(AAB67116.1); EhV86, *Emiliania huxleyii* virus 86 (CAI65453.1); HaV1, *Heterosigma akashiwo* virus 1(BAE06251.1); FV3, Frog virus 3(AAT09720.1); and LCDV1, Lymphocystis disease virus 1 (NP_078724.1).

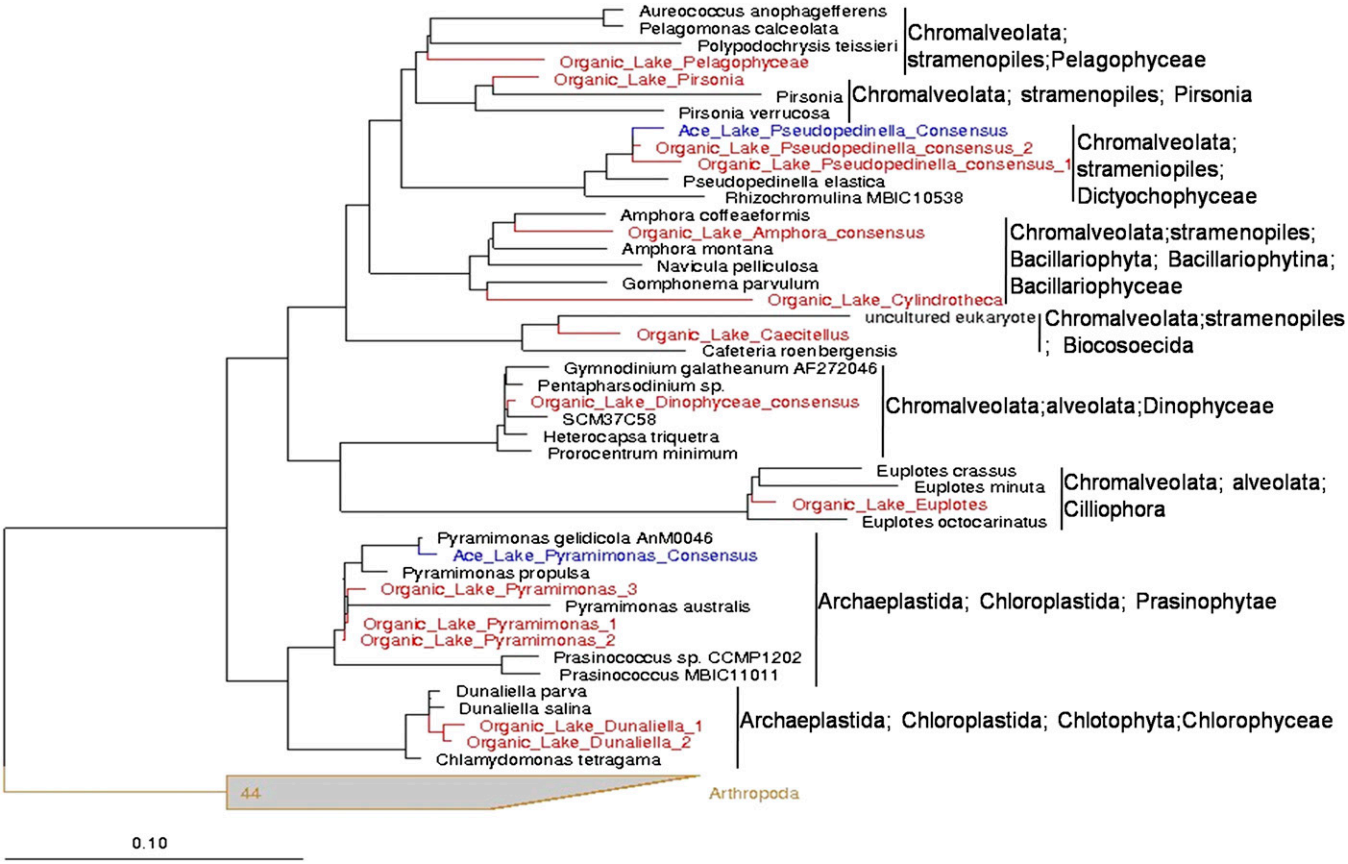

Fig. S3. Phylogeny of the 18S rRNA genes from the 2006 Organic Lake data.

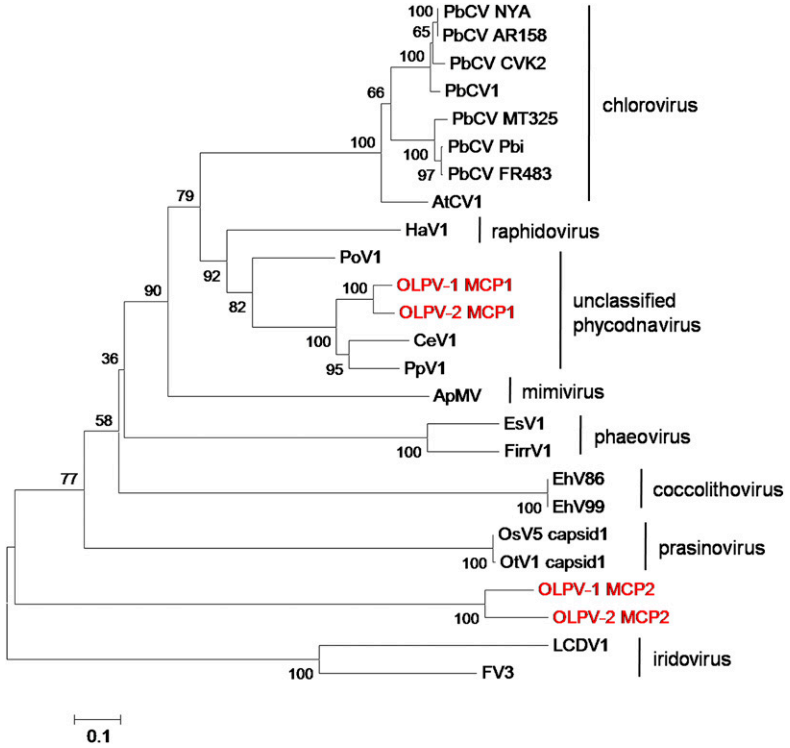

Fig. S4. Phylogeny of OLPV using MCP sequences. Neighbor-joining tree of MCP sequences of OLPV, and NCLDV sequences from GenBank. Abbreviations and accession numbers from top to bottom: PbCV NYA, *Paramecium bursaria* chlorella virus NY2A(ABT14984.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU44077.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35143.1); PbCV1, *P. bursaria* chlorella virus 1 (AAA88828.1); PbCV MT325, *P. bursaria* chlorella virus MT325 (ABT14017.1); PbCV Pbi, *P. bursaria* chlorella virus Pbi (AAC27492.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15755.1); AtCV1, *Acathocystis turfacea* chlorella virus 1 (ABT16414.1); HaV1, *Heterosigma akashiwo* virus 1 (BAE06835.1); PoV, *Pyramimonas orientalis* virus (ABU23714.1); CeV1, *Chrysochromulina ericinia* virus 1 (ABU23712.1); PpV1, *Phaeocystis pouchetii* virus (ABU23715.1); ApMV, *Acathamoeba polyphaga* virus (Q5UQL7.2); EsV1, *Ectocarpus siliculosus* virus (AAK14534.1); FirrV1, *Feldmannia irregularis* virus 1 (AAR26925.1); EhV86, *Emiliania huxleyi* virus 86(CAI65508.2); EhV99, *E. huxleyi* virus 99 (ABU23713.1); OsV5, *Ostreococcus* virus 5 (ABY27849.1); OtV1, *Ostreococcus tauri* virus 1 (CAY39653.1); LCDV1, Lymphocystis diseases virus 1 (AAC24486.2); and FV3, Frog virus 3 (AAT09750.1).

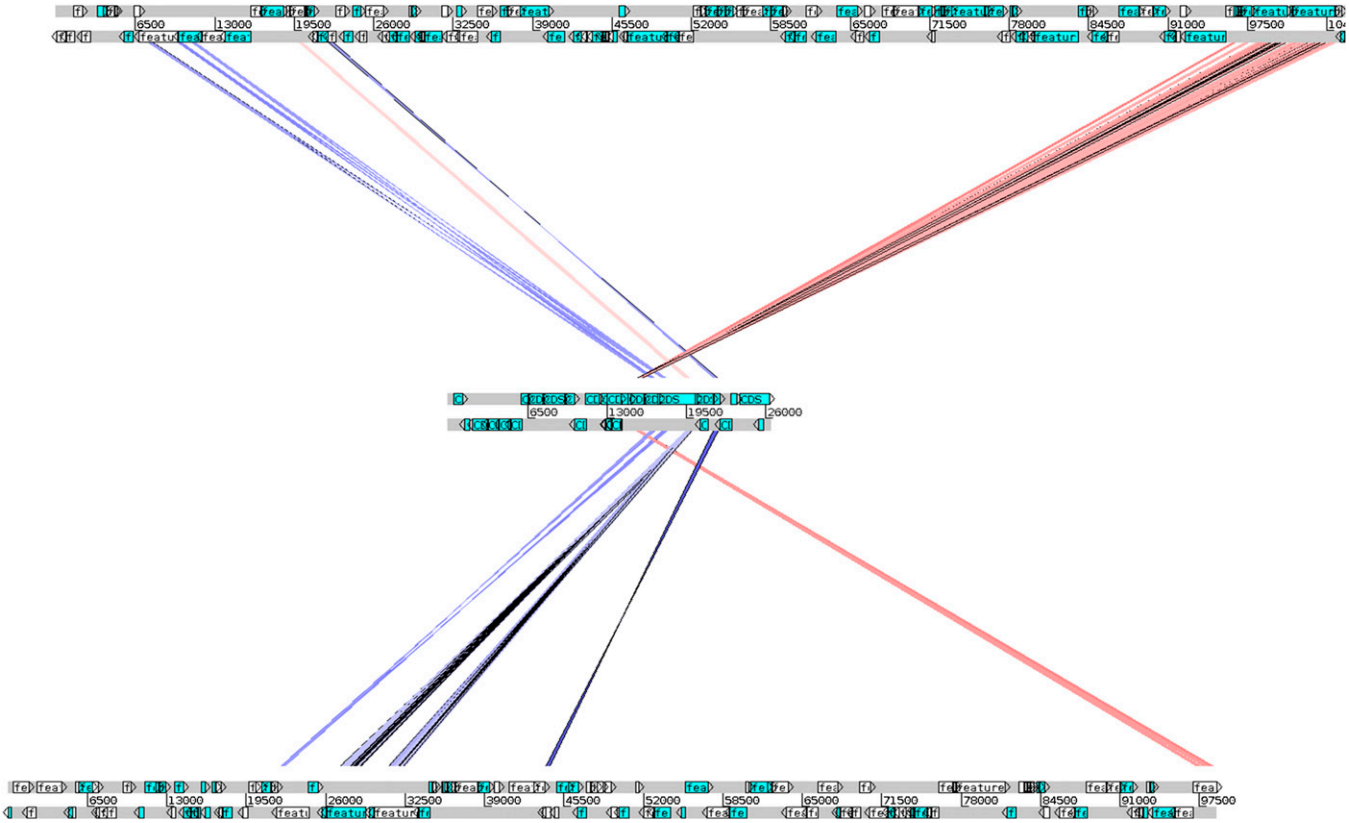

Fig. S5. Comparison of the location of genes in OLV compared with OLPVs. OLV (*Middle*), OLPV-1 (*Top*), and OLPV-2 (*Bottom*).

Table S1. Summary of metagenome data for Organic Lake samples

Sample	Trimmed reads	Contigs over 10 kb (reads in contigs)	Annotated predicted ORFs from contigs (total)
December 2006 Sanger/454, 0.1 μm	418,265 (28,481, Sanger + 389,784, 454)	45 (221,573)	7,318 (21,961)
November 2008 454, 0.1 μm	494,573	5 (771)	33,262 (83,684)
December 2008 454, 0.1 μm	446,200	2 (40,314)	23,012 (64,779)

Table S2. Peptide data for Organic Lake metaproteomics: OLPV and OLV proteins from the 2006 Organic Lake 0.1-μm filter sample

Gene ID	Source	Normalized spectral abundance	GenBank accession no.	Description	Coverage (%)	No. peptides (unique)	Peptide sequences
162322530*	OLPV-1	0.000661	A7U6F0	Major capsid protein [*Phaeocystis pouchetii* virus]	33.0	15 (4)	R.AIDECLWAVSSLSPSSSADVK.V K.ALGAQPFNYTDAVDALPNSIK.A R.EGTYFDQVQPFQHHTR.Y R.HSNFAMESIEQTFNGQADFGR.R K.HYGDWMQIWCQLTLDK.N R.IDNATLQLVLSNATVEGTNTAK.V R.IMSGMGGLAYSN K.INLDLR.A R.LNFNHPCK.E K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N
162322348	OLPV-1	0.000120	—	—	11.3	2 (2)	K.GNVDVYQENK.L K.IESDAEPSWVR.G
162322406†	OLPV-1	0.000177	—	—	29.4	4 (4)	R.QNQSCGGVNQVNGTHVNR.T R.TAFHLGDGLSR.Q K.TNDGTLVGK.S K.YVSESSTYTR.F
162313481	OLPV-1	0.000010	YP_002714448	Leucine rich repeat-containing Miro-like protein [*Synechococcus* sp. PCC 7335]	3.96	2 (2)	K.ITTIPENIGQLVK.I R.SNLQGVTEEQLMSNK.I
162276060	OLPV-2	0.000897	—	—	28.9	2 (2)	K.TPTGLEFSLTGR.A R.VNHTDACSTGNK.E
162300260	OLPV-2	0.000226	—	—	34.6	2 (2)	R.VDIEGGTPFFLK.E K.YTFQPSELSNTYFSK.E
162276024†	OLPV-2	0.000127	—	—	16.0	3 (3)	K.LGGGISSR.S R.SEVGFQSTMVGSDVAMQR.K R.TSLHMGDVLSR.K
162275992	OLPV-2	0.000098	NP_048709	Hypothetical protein PBCV1_A352L [*Paramecium bursaria* Chlorella virus 1]	16.6	2 (2)	K.NINLLSAGANYGINTVGSSLR.N R.NPNLQIR.S
162300108	OLPV-2	0.000046	ZP_01471812	BNR containing hypothetical protein RS9916_28494 [*Synechococcus* sp. RS9916]	7.66	5 (5)	K.NDNIITLLDTK.Q K.NVVINSEGTIISAVNNK.G K.QDVITDQTNLNVGR.L K.YENGSWNTLGQLIR.G R.LTVNNSIISK.E
162319393*	OLPV-2	0.000016	A7U6F0	MCP [*Phaeocystis pouchetii* virus]	26.3	13 (2)	R.AIDECLWAVNTLSPDSSSDVK.V K.ALGAQPFNYTDAIDALPNSVK.A R.EGTYFDQVQPFQHHTR.S R.IDNATLQLVLSNATVEGTNTAK.V R.IMSGMGGLAYSN K.INLDLR.A R.LNFNHPCK.E K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N
162300134‡	OLPV-1/2	0.000100	AAR21578	Heat shock protein 70 [*Phytophthora nicotianae*]	6.97	3 (3)	K.ATAGDTHLGGEDFDNR.M R.IINEPTAAAIAYGLDK.K R.VEIIANDQGNR.T
162286324‡	OLPV	0.000176	NP_048575	Hypothetical protein PBCV1_A227L [*Paramecium bursaria* Chlorella virus 1]	14.7	2 (2)	K.DVPLVANFSAK.F K.MKLENTVEK.M

Table S2. Cont.

Gene ID	Source	Normalized spectral abundance	GenBank accession no.	Description	Coverage (%)	No. peptides (unique)	Peptide sequences
OLV9	OLV	0.001681	YP_002122381	Capsid protein V20 [Sputnik virophage]	31.1	15 (15)	K.AGLLSEMDAYSLYQMSR.R K.ELVLSFSSGVK.F K.FGTQASTLFLK.D R.GSSATMSGLLTK.S K.GVASAVESAIGGAK.T K.HIQTTPSMVDK.Y R.ITSDVQVAVK.D K.IYLVVRPQYR.S R.NGSQQTWNEFR.G K.NILPYDEFVAYK.T F.NVNVPSENTLVDR.N K.SEVLEAK.E K.VSVQSADILNVITK.Q R.YISLHPSQYAK.L K.YTSLGSIIVIDPVR.D
OLV8	OLV	0.000334	YP_002122379/ YP_002122380	*N*-term:hypothetical protein V18 [Sputnik virophage] C-term: minor capsid protein V19 [Sputnik virophage]	19.1	8 (8)	K.AGTPIPGVIYEPSYPR.W K.DIGTDMPYFIFDK.D K.GGYADYR.S K.TLLEFGQSK.D K.TLPVFIPTIK.Y K.TNGTTPPR.F K.YSEDDTNESIR.N K.QAFIGLQK.T

*Proteins that have some shared peptides.
†162322406 and 162276024 are protein homologs (Fig. 2, main text).
‡A group of proteins containing similar peptides that could not be differentiated by the mass spectral analysis. Only one gene number of that group is displayed.

Table S3. Top BLASTp matches of predicted coding sequences from the OLV genome compared with the OLPV, GenBank nonredundant (NR) protein database, and CAMERA metagenomic reads ORF peptide database

Gene ID	Start	End	GenBank NR match (accession no./% aa identity/e-value)	OLPV match (gene ID/% aa identity/e-value)	Metagenomic ORF peptide match sample, location (accession no./% aa identity/e-value)
OLV1	460	1077	—	—	GS003, 0.1, North American East Coast, 1m (JCVI_PEP_1105157870626/41%/1e-29)
OLV2	1701	1333	hypothetical protein [*Tetrahymena thermophila* SB210] (XP_001029204.1/38% /3e-04)	—	GS012, 0.1, North American East Coast, 13.2m (JCVI_PEP_1105080106223/42%/1.7e-11)
OLV3	3187	2030	—	—	—
OLV4	3991	3224	V3 [Sputnik virophage] (YP_002122364.1/ 39%/4e-24)	—	GS001b, 0.8, Sargasso Sea, 5m (JCVI_PEP_1105131296011/43%/5e-38)
OLV5	5029	4160	V21 [Sputnik virophage] (YP_002122382.1/42%/4.3e-02)	—	GS020, 0.1, Panama Canal, 2m (JCVI_PEP_1105125932065/25%/6e-12)
OLV6	5940	5044	—	—	GS000c, 0.22, Sargasso Sea, 5m (JCVI_PEP_1105136847382/24%/6e-3)
OLV7	5978	6547	V9 [Sputnik virophage] (YP_002122370.1/35%/3e-14)	—	GS020, 0.1, Panama Canal, 2m (JCVI_PEP_1105140820785/26%/1e-12)
OLV8	6574	7740	*N*-term:V18 [Sputnik virophage] (YP_002122379.1/27%/9e-05) C-term: V19 [Sputnik virophage] (YP_002122380.1/26%/0.16)	—	GS008, 0.1, North American East Coast, 1m (JCVI_PEP_1105124194533/32%/6e-8)
OLV9	7791	9518	V20 [Sputnik virophage] (YP_002122381.1/28%/9e-10)	—	GS033, 0.1, Galapagos Islands, 0.2m (JCVI_PEP_1105120114513/28%/2e-14)
OLV10	9563	10273	—	—	GS001b, 0.8, Sargasso Sea, 5m (JCVI_PEP_1105163928413/61%/5e-4)
OLV11	11210	10317	—	—	GS013, 0.1, North America East Coast, 2.1m (JCVI_PEP_1105123792445/ 39%/9e-37)
OLV12	11284	12324	N terminus:hypothetical protein ATCV_Z547R [*Acanthocystis turfacea* Chlorella virus 1] (YP_001427028.1/36%/7e-09), C terminus: Lipase class 3 [*Bacillus thuringiensis* IBL 200] (gb|EEM96541.1|/27%/1.7e-02)	—	GS018, Carribean Sea,1.7m (JCVI_PEP_1105087988121/34%/5.6e-23)
OLV13	12539	14884	collagen-like protein [*Bacillus megaterium*] (YP_001569009.1/ 66.67%/4e-03)	—	GS027, 0.1, Galapogos Islands, 2.2m (JCVI_PEP_1105075498120/43%/6.7e-11)
OLV14	14023	12905	—	—	—
OLV15	13041	14078	—	—	—
OLV16	14151	13372	—	—	GS020, 0.1, Panama Canal, 2m (JCVI_PEP_1105127133835/36%/8.5e-11)
OLV17	15094	16023	putative transmembrane protein [Flavobacteria bacterium BAL38] (ZP_01734433.1/51%/8e-34)	OLPV Lipoprotein Q-like protein (162322444/40%/1e-24)	GS009, 0.1, North American East Coast, 1m (JCVI_PEP_1105137954859/50%/4e-37)
OLV18	16054	17211	hypothetical protein BH13620 [*Bartonella henselae* str. Houston-1] (YP_034083.1/15%/4e-04)	OLPV Cyanothece sp. cce_0037-like protein (162322244/65%/2e-33)	GS000c, 0.1, Carribean Sea, 2m (JCVI_PEP_1105149563549/39%/2e-26)
OLV19	17168	20278	Phage tail fiber repeat family protein [*Trichomonas vaginalis* G3] (XP_001296018.1/ 42%/4e-11)	OLPV Lipoprotein Q-like protein (162322444/65%/9e-33)	GS016, 0.1, Carribean Sea, 2m (JCVI_PEP_1105149563549/29%/1e-27)
OLV20	20266	21570	collagen triple helix containing protein A1Q_3499 [*Vibrio harveyi* HY01] (ZP_01986098.1/ 69%/6e-04)	OLPV hypothetical protein (162322252/32%/1e-07)	GS033, 0.1, Galapagos Islands, 0.2m (JCVI_PEP_1105153074955/69%/1e-5)
OLV21	21089	20622	—	—	—
OLV22	21747	22157	—	OLPV hypothetical protein (162322266/56%/5e-31)	GS017, 0.1, Carribean Sea, 2m (JCVI_PEP_1105100448171/43%/4e-24)

Table S3. Cont.

Gene ID	Start	End	GenBank NR match (accession no./% aa identity/e-value)	OLPV match (gene ID/% aa identity/e-value)	Metagenomic ORF peptide match sample, location (accession no./% aa identity/e-value)
OLV23	23089	22256	D12 class N6 adenine-specific DNA methyltransferase [Candidatus *Koribacter versatilis* Ellin345] (YP_592471.1/28%/1e-24)	—	GS002, 0.1, North America East Coast, 1m (JCVI_PEP_1105085453201/33%/8e-18)
OLV24	23174	23560		—	GS013, 0.1, North American East Coast 2.1m (JCVI_PEP_1105132174179/32%/1e-03)
OLV25	23889	26219	N terminus: V13 [Sputnik virophage] (YP_002122374.1/ 34%/5e-31) C terminus: Primase-polymerase domain containing hypothetical protein [*Ostreococcus lucimarinus* CCE9901] (XP_001421479.1/29%/9e-32)	—	N terminus: GS030, 0.1, Galapagos Islands, 19m (JCVI_PEP_1105105378071/40%/8e-38) C terminus: GS013, 0.1, North America East Coast, 2.1m (JCVI_PEP_1105129419397/51%/8e-71)
OLV26	25666	25376	—	—	GS013, 0.1, North American East Coast,2.1m (JCVI_PEP_1105129419399/54%/8e-6)

Table S4. BLASTp matches for OLV MCP in predicted ORFs of Organic Lake and Ace Lake contigs and CAMERA metagenomic reads ORF peptide database

Sample	Filter size (μm)	Gene ID	Contig ID	Identity (%)	Alignment length (aa)	e-value	Contig coverage (×)
Organic Lake	0.1	OLV9	—	—	—	—	77.12
December 2006	0.8	176157210	scf7180000034275	98.61	575	0.0	16.03
	3.0	181703798	deg7180000108904	98.96	575	0.0	48.65
Organic Lake	0.1	192841413	deg7180000116398	99.64	555	0.0	13.71
November 2008	0.8	193037024	scf7180000086663	93.98	133	2e-61	2.50
	3.0	192638971	deg7180000028400	99.36	156	1e-76	1.86
		192955191	deg7180000024244	93.98	133	1e-61	3.10
Organic Lake	0.1	192709908	scf7180000109753	99.01	304	9e-173	4.38
December 2008		192709920	scf7180000109753	99.59	244	1e-120	4.38
		192712009	deg7180000067104	54.70	117	3e-30	1.58
		192890551	deg7180000061276	36.89	122	2e-13	3.15
	0.8	193060302	deg7180000053149	53.75	160	6e-43	2.30
	3.0	—	—	—	—	—	—
Ace Lake 2006	0.1	167813925	scf7180000126822	28.86	246	3e-14	2.36
		167858124	scf7180000129064	21.78	381	5e-10	1.94
		167891594	scf7180000136823	24.85	326	2e-04	8.15
		167875536	deg7180000086604	22.95	244	6e-04	2.21
	0.8	176091445	deg7180000053588	91.61	143	8e-78	3.35
		175769103	deg7180000078701	88.24	153	1e-74	1.77
		176042318	deg7180000058177	81.77	181	1e-74	2.48
		176000635	deg7180000087166	53.39	221	8e-58	2.50
		176042707	deg7180000058207	50.78	193	5e-46	2.34
		175886340	deg7180000074162	58.90	146	4e-45	2.73
		176249679	deg7180000049481	61.94	155	2e-44	2.75
		175748439	deg7180000058552	76.79	112	2e-35	2.39
		175637390	deg7180000058712	50.91	165	2e-35	2.03
		176100822	deg7180000055966	53.38	133	6e-35	1.48
		176018109	deg7180000086684	59.68	124	4e-27	1.66
		176000624	deg7180000087165	53.85	104	6e-27	1.73
		175805608	deg7180000054222	48.60	107	7e-21	1.93
		175908895	deg7180000061971	51.91	131	4e-20	3.27
		175821062	deg7180000080443	46.46	127	6e-20	1.43
		176026419	deg7180000054364	52.59	116	8e-19	1.47
		176133336	scf7180000089989	38.36	146	3e-12	1.51
		176018257	deg7180000086719	31.21	173	4e-12	1.23
		176137412	deg7180000052688	29.37	126	1e-06	2.47
		175686880	scf7180000092161	24.00	125	2e-06	2.98
	3.0	175741076	deg7180000030508	85.78	232	8e-109	1.25
		175748837	deg7180000027929	55.29	170	4e-44	1.32
		175751996	deg7180000037324	51.27	158	5e-41	1.63
		175859792	scf7180000045944	30.21	288	8e-26	2.69
Hypersaline lagoon,	0.1	JCVI_PEP_1105120114513	—	27.84	273	2e-14	—
Floreana Island,		JCVI_PEP_1105100621559	—	24.76	307	9e-10	—
Ecuador (GS033)		JCVI_PEP_1105161421335	—	25.61	289	2e-6	—
Delaware Bay, NJ,	0.1	JCVI_PEP_1105106741177	—	24.62	264	6e-14	—
United States (GS011)		JCVI_PEP_1105089715877	—	27.16	313	1e-17	
Upwelling near Fernandina Island, Ecuador (GS031)	0.1	JCVI_PEP_1105079267881	—	28.23	170	8e-11	—
Lake Gatun, Panama (GS020)	0.1	JCVI_PEP_1105119255775	—	26.71	149	5e-9	—

e-value cutoff 1e-5, alignment length >100 aa.