The ISME Journal (2008), 1–10
© 2008 International Society for Microbial Ecology All rights reserved 1751-7362/08 $32.00
www.nature.com/ismej

ORIGINAL ARTICLE

Detection of inteins among diverse DNA polymerase genes of uncultivated members of the *Phycodnaviridae*

Alexander I Culley, Brenda F Asuncion and Grieg F Steward
Department of Oceanography, University of Hawai‘i at Mãnoa, Honolulu, HI, USA

Viruses in the family *Phycodnaviridae* infect autotrophic protists in aquatic environments. Application of a PCR assay targeting the DNA polymerase of viruses in this family has revealed that phycodnaviruses are quite diverse and appear to be widespread, but a limited number of environments have been examined so far. In this study, we examined the sequence diversity among viral DNA *pol* genes amplified by PCR from subtropical coastal waters of O‘ahu, Hawai‘i. A total of 18 novel prasinovirus-like sequences were detected along with two other divergent sequences that differ at the genus-level relative to other sequences in the family. Of the 20 new sequence types reported here, three were serendipitously found to contain protein introns, or inteins. Sequence analysis of the inteins suggested that all three have self-splicing domains and are apparently capable of removing themselves from the translated polymerase protein. Two of the three also appear to be 'active', meaning they encode all the motifs necessary for a complete dodecapeptide homing endonuclease, and are therefore capable of horizontal transfer. A subsequent PCR survey of our samples with intein-specific primers suggested that intein-containing phycodnaviruses are common in this environment. A search for similar sequences in metagenomic data sets from other oceans indicated that viral inteins are also widespread, but how these genetic parasites might be influencing the ecology and evolution of phycodnaviruses remains unclear.

The ISME Journal advance online publication, 11 December 2008; doi:10.1038/ismej.2008.120
Subject Category: microbial population and community ecology
Keywords: intein; phycodnavirus; marine virus ecology

Introduction

Marine phytoplankton have a global influence on many biogeochemical cycles and provide nearly all of the fixed carbon supporting the ocean's food web. Viral lysis of phytoplankton can have a significant influence on how carbon is partitioned in the food web (Fuhrman, 1999; Suttle, 2005). Models (Murray and Eldridge, 1994) and experimental data (Bratbak *et al.*, 1998) suggest that lysis of phytoplankton by viruses results in an efficient transfer of carbon from primary producers to heterotrophic bacteria at the expense of higher trophic levels. The most extensively studied viruses infecting eukaryotic phytoplankton of marine or freshwater origin are the large, double-stranded-DNA-containing viruses in the family *Phycodnaviridae* (Van Etten *et al.*, 1991; Reisser, 1993). The first phycodnavirus isolated, MpV, infects the universally distributed prymnesiophyte *Micromonas pusilla* (Mayer and Taylor, 1979). Other phycodnaviruses have since been isolated that infect a number of important members of the phytoplankton including toxic bloom formers such as *Heterosigma akashiwo* (Nagasaki and Yamaguchi, 1997) and *Aureococcus anophagefferens* (Milligan and Cosper, 1994), and the coccolithophorid *Emiliania huxleyi* (Castberg *et al.*, 2002).

The phycodnavirus genomes that have been sequenced thus far have several interesting features. Viruses within this family are some of the largest known, with genomes up to 560 kbp in length and particles up to 220 nm in size (Dunigan *et al.*, 2006). Comparative genomic studies suggest that phycodnaviruses are ancient, predating the diversification of the eukaryote lineage (Iyer *et al.*, 2001, 2006). The large genomes of phycodnaviruses harbor genes that are unusual for viruses. *Paramecium bursaria* chlorella virus 1 (PBCV-1), for example, is capable of constructing complex oligosaccharides independent of its host (Markine-Goriaynoff *et al.*, 2004). EhV-86, a coccolithovirus, encodes genes involved in the synthesis of sphingolipids (Wilson *et al.*, 2005).

Correspondence: GF Steward, Department of Oceanography, University of Hawai‘i at Mãnoa, 1000 Pope Road, Marine Sciences Building, Honolulu, HI 96822, USA.
E-mail: grieg@hawaii.edu
Received 4 August 2008; revised 29 October 2008; accepted 4 November 2008

A few phycodnavirus genomes have been found to harbor inteins, which are parasitic genetic elements that typically insert themselves within the conserved motifs of the essential genes of their hosts (Gogarten *et al.*, 2002). Inteins are translated and transcribed with their host protein and then extract themselves autocatalytically before the final conformation of the mature protein (Gogarten and Hilario, 2006). Although all inteins are capable of self-splicing, those that are considered to be 'active' also encode a homing endonuclease that enables them to infect intein-free alleles (Pietrokovski, 1998). Those with missing or degraded homing endonuclease domains are considered 'fixed'. Inteins appear to be most common in *Archaea*, but have been identified in all three domains of life and in viruses (Perler, 2002). Among the phycodnaviruses, inteins have been identified in the DNA polymerase I of *H. akashiwo* virus (HaV; Nagasaki *et al.*, 2005) and *Chrysochromulina ercina* virus (CeV; Larsen *et al.*, 2008) and in the helicase and ribonucleotide reductase of the PBCV (NY-2A; Fitzgerald *et al.*, 2007). Little is known about the prevalence or diversity of inteins among the large number of marine phycodnaviruses that have not been cultivated.

Most culture-independent surveys of phycodnavirus populations have involved PCR amplification and sequence analysis of a fragment of the DNA polymerase gene (DNA *pol*) using degenerate primers (Chen *et al.*, 1996; Short and Suttle, 2002, 2003; Short and Short, 2008), although one recent study has evaluated the utility of using the major capsid protein (Larsen *et al.*, 2008). The diverse sequences detected by these PCR assays in various marine and freshwater environments suggest that a wide range of protists is prone to infection by members of the family *Phycodnaviridae* and that much of the diversity remains undescribed. Of the more than 150 distinct phycodnavirus DNA *pol* sequences recovered by PCR from aquatic environments, none has been reported to contain an intein.

Although phycodnaviruses are likely to be globally distributed, their distribution and diversity are still poorly characterized in subtropical waters. Our objective in this study was to investigate the sequence diversity among DNA polymerase genes from phycodnaviruses in coastal subtropical waters of Hawai'i and to compare those sequences to those from cultivated isolates and those amplified directly from other marine and freshwater environments.

Materials and methods

Station description

Water samples were collected in Kāne'ohe Bay, a reef-protected embayment on the windward side of O'ahu, Hawai'i, in June of 2006 and December of 2007. In June 2006, samples were collected from three sites within the bay and pooled whereas in December 2007, samples were collected from five sites in Kāne'ohe Bay, including two of the three sites from 2006 (Table 1).

Collection

Samples of whole surface seawater (<0.5 m depth) were collected by submersion of polyethylene carboys from the side of a small boat. Viruses and other plankton in the water were collected by direct filtration using a peristaltic pump. The three samples from June 2006 (250–550 ml each; Table 1) were prefiltered through 0.22 μm pore-size polyethersulfone membrane filter cartridges (Sterivex; Millipore, Billerica, MA, USA). Samples were then pooled and filtered through a 0.02 μm aluminum oxide filter (Anotop; Whatman, Middlesex, UK). The five samples from December 2007 (500–1100 ml each; Table 1) were filtered directly onto 0.02 μm filters. The filter inlets and outlets were sealed and the filters were stored at −80 °C.

Extraction

Total nucleic acids were extracted from the 0.02 μm aluminum oxide filters with a commercial kit (MasterPure; Epicentre, Madison, WI, USA). The protocol used in this study was based on the one described in Culley and Steward (2007). Briefly, 400 μl of 2 × T + C lysis buffer containing 50 μg μl^{-1} proteinase K was added to a 3 cc syringe, which was then locked to the filter inlet. After gently pushing

Table 1 Station abbreviations and locations, collection dates and sample volumes

Name	°*Latitude*[a] *(N)*	°*Longitude*[a] *(W)*	*Collection date (dd/mm/yyyy)*	*Sample volume (ml)*
NR2	21.515000	157.809167	28/06/2006	550[b]
AR	21.467639	157.836528	28/06/2006	250[b]
SB	21.436536	157.777361	28/06/2006	350[b]
NB	21.490483	157.833283	11/12/2007	1100
CB	21.457050	157.811350	11/12/2007	750
AR	21.467639	157.836528	11/12/2007	800
SB	21.436536	157.777361	11/12/2007	650
SBE	21.419017	157.780783	11/12/2007	500

[a] Latitude and longitude provided in decimal coordinates.
[b] 2006 samples were pooled after 0.22 μm pore-size filtration.

the buffer into the filter, the filter outlet was flame-sealed. The whole assembly was incubated at 65 °C in air for 10 min and then placed on ice for 5 min. Afterward, the sealed tip was cut and the extracted material was gently pushed from the filter into a sterile microcentrifuge tube. Subsequent steps were performed according to the manufacturer’s protocol.

PCR

AVS primers. The degenerate primers AVS1 and AVS2, designed to target viruses from the family *Phycodnaviridae* (Chen and Suttle, 1995), were used in PCR amplification reactions using conditions described previously (Short and Suttle, 2002). The PCR reaction mixture consisted of 5 μl of template DNA, 5 μl 10 × DNA polymerase assay buffer, 1.5 μl of 1.5 mM $MgCl_2$, 1 μl of 0.2 mM deoxyribonucleoside triphosphates (dNTPs), 1 μl of 10 μM AVS1 primer, 3 μl of 10 μM AVS2 primer and 0.2 μl of 1 U Platinum *Taq* DNA polymerase (Invitrogen Corporation, Carlsbad, CA, USA). Blanks were prepared by substituting ultrapure water (Barnstead Nanopure Life Science TOC; Thermo Fisher Scientific, Dubuque, IA, USA) for the template DNA. The PCR thermal cycling protocol consisted of denaturation at 94 °C for 75 s, followed by 40 cycles of denaturation at 94 °C for 45 s, annealing at 45 °C for 45 s and extension at 72 °C for 1 min. PCR product sizes and yields were estimated relative to standards by agarose gel electrophoresis. Gels were stained with a fluorescent DNA stain (SYBR Safe; Invitrogen Corporation) and DNA bands were visualized with UV transillumination. Images were captured on a digital gel documentation system (VersaDoc; Bio-Rad Laboratories, Hercules, CA, USA). Bands were excised from the gel and purified using a gel extraction kit (MinElute; Qiagen, Valencia, CA, USA).

Degenerate intein primers. A viral intein identified in this study, KBvi-1, was aligned with the homologous intein in the *Acanthamoeba polyphaga* mimivirus (APMV) DNA polymerase using DIALIGN-TX (Subramanian *et al.*, 2008). Conserved regions were used to create a degenerate viral DNA polymerase intein primer pair (VDPI-F and -R; Table 2). Each reaction consisted of 1 μl of extracted DNA (or ultrapure water for blanks), 5 μl of 10 × DNA polymerase assay buffer, 3 μl of 1.5 mM $MgCl_2$, 1 μl of a 0.2 mM stock of dNTPs, 5 μl of each 10 μM VDPI primer stock and 0.2 μl of a 1 U μl^{-1} DNA polymerase stock (Platinum *Taq*; Invitrogen) in a final volume of 50 μl. Thermal cycling consisted of denaturation at 94 °C for 2 min and 15 s, followed by 40 cycles of denaturation at 94 °C for 45 s, annealing at 45 °C for 45 s, extension at 72 °C for 1 min and a final extension at 72 °C for 9 min.

Table 2 PCR primer sequences and annealing temperatures

Name	*Sequence (5′ - 3′)*	$^{\circ}C_{annealing}$
AVS1	GARGGIGCIACIGTIYTIGAYGC	45
AVS2	GCIGCRTAICKKTTKTTISWRTA	
SL1	CAGTCCAGTTACGCTGGAGTC	51
SR2	GGTCAGGTATGATTTAAATGGTCAGT	
VDPI-F	GWGTACTYACTCATACWGGA	45
VDPI-R	TTGATKGAWACRTTRTAYCC	

Cloning and sequencing. Before cloning, the ends of the amplified product were rendered blunt using an end-repair kit (PCRTerminator; Lucigen, Middleton, WI, USA) according to the manufacturer’s protocols. Products were ligated into pSMART-HCKan vectors and transformed into *E. cloni* 10G Chemically Competent Cells using the manufacturer’s protocols (CloneSmart Blunt Cloning Kit; Lucigen). Clones were screened for inserts by PCR amplification with primers SL1 and SR2 provided in the cloning kit (Table 2). Sequencing was performed using dye-terminator cycle-sequencing reactions followed by analysis by capillary electrophoresis (Sequencer model 3730XL; Applied Biosystems, Foster City, CA, USA).

Analysis

Sequences sharing ⩽97% identity on the nucleotide level (Short and Short, 2008) were considered to be different phylotypes. Sequences were labeled with a code indicating the location from which they were derived (KB for Kāne‘ohe Bay) and the nature of the sequence (vp for viral polymerase; vi for viral intein). DNA *pol* and intein sequences were analyzed independently. For those DNA *pol* genes containing an intein, the intein was removed before analysis. Representatives from each phylotype were compared to the NCBI database with BLASTp (Altschul *et al.*, 1997). Translated phylotype representatives were aligned with DIALIGN-TX (Subramanian *et al.*, 2008) using the following parameters: length of a low-scoring region = 4, maximum fragment length that is allowed to contain regions of low quality = 40 and sensitivity (−1) = 0. In cases where a region was designated as unaligned in at least one sequence, the corresponding region was removed from all sequences (Supplementary Figures S1 and S2). For the DNA *pol* alignment, this resulted in the removal of 13 regions ranging in size from 1 to 23 positions (including gaps) for 86 positions removed out of 350. For the intein alignment, 18 regions ranging in size from 1 to 25 positions (including gaps) were removed for 127 positions removed out of 582. The remaining aligned regions were concatenated and phylogenetic trees based on Bayesian inference were constructed with MrBayes version 3.1.2 (Altekar *et al.*, 2004) and 1 000 000 generations. The details of the sequences used for phylogenetic analyses are listed in Supplementary Table S1 and S2 (DNA *pol* and inteins, respectively). Estimates of phylotype richness were computed using EstimateS (version 7.5, RK Colwell, http://purl.oclc.org/estimates). The CAMERA website

(Seshadri *et al*., 2007) was used to search environmental metagenomic libraries for sequences similar to the DNA *pol* and intein sequences recovered in this study.

Nucleotide sequence accession numbers
Sequences have been deposited in the GenBank database and sequences KBvp-1 through KBvp-20 were assigned accession numbers EU889354–EU889373.

Results

Phycodnavirus-like DNA polymerase sequences were successfully amplified from samples collected throughout Kāne‘ohe Bay in different years and in different seasons. We identified 20 phylotypes from Kāne‘ohe Bay in this study, all of which were novel ($\leqslant$97% identity with any previously described sequence). Two phylotypes were observed in both summer 2006 and winter 2007 (KBvp-8 and KBvp-1) whereas 14 phylotypes were unique to the summer sample and 4 were unique to the winter samples. Phylotype KBvp-8 (Figure 1; Supplementary Table S3) dominated the clone libraries in both sampling periods, accounting for 58% of the sequences recovered from the summer sample and 94% of the sequences recovered from the winter samples. Of the 20 phylotypes, 3 were found by sequencing a second, larger amplification product that was observed in two of the samples after amplification with the AVS primers. These sequences contained an intein inserted in the pol-C insertion site, that is, between YGD and TDS amino-acid motifs of the B family DNA polymerase (Figure 2). KBvi-1 (present in the KBvp-11 phylotype) was found in both the summer 2006, pooled sample, and in the winter 2007 sample from Station SBE. KBvi-2 (in KBvp-16) and KBvi-3 (in KBvp-2) were detected at lower frequency and only at Station SBE from the winter 2007 sampling (Figure 1; Supplementary Table S4). The larger, intein-containing PCR product was not detectable in the other four winter 2007 stations. However, when we assayed with an intein-specific primer set, amplification occurred in the samples from the remaining four winter 2007 stations. Sequencing of 4–5 clones from each of these samples revealed that all of the directly amplified intein fragments were identical to KBvi-1 (Supplementary Table S4).

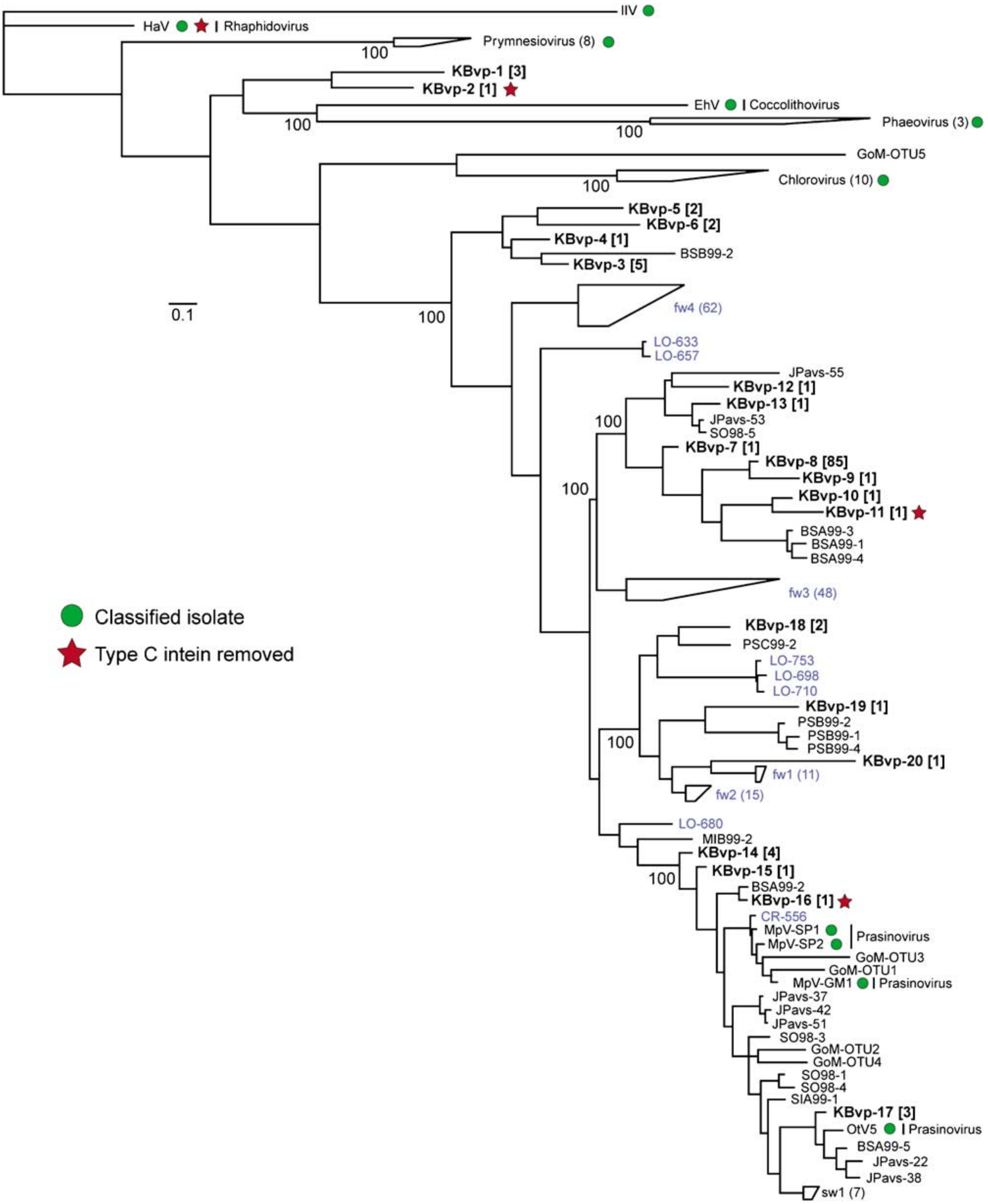
According to BLASTp comparisons (Altschul *et al*., 1997) with the NCBI database, all phylotypes were most similar to phycodnavirus polymerase sequences. Translated polymerase sequences contained the amino-acid motif YGDTDS, a region conserved among all classified members of the family Phycodnaviridae (Chen and Suttle, 1995). The results of BLASTp searches of the NCBI and NEB intein databases showed that all three inteins identified in this study were most similar to the intein region of the APMV DNA polymerase (Raoult *et al*., 2004). All three inteins exhibited the expected features of a functional autocatalytic self-splicing domain (Perler, 2002; Figure 2), including the residues that provide the nucleophilic groups in the self-splicing reactions (Pietrokovski, 1998; Perler, 2002). KBvi-1 and KBvi-2 appeared also to have all four signature motifs of a dodecapeptide (DOD) homing endonuclease, whereas KBvi-3 lacked at least one of the four motifs.

Searches of the CAMERA ‘all metagenomic ORF peptides database’ with the intein sequences from this study resulted in significant similarity (*e*-value $<1\times10^{-3}$) to metagenomic sequences from stations ranging from the North Atlantic Ocean (Bedford Basin, Nova Scotia) to the South Pacific (Rangirora Atoll) Ocean. A total of 82 significant matches were found to each of KBvi-1 and KBvi-2, and 78 matches were found to KBvi-3. The highest identity of a BLAST hit (77%) was between KBvi-2 and a sequence from Chesapeake Bay. Five of the identified intein-like sequences appeared to be full length, as determined by the presence of conserved intein-splicing domains at the N- and C termini. Two of the five (JCVI-PEP-1105114456875 and JCVI_PEP_1105132855101) appeared to encode a DOD endonuclease domain. A BLASTp search of the NCBI database with only the extein regions of these sequences resulted in significant similarity only to phycodnavirus DNA polymerase sequences. We found no significant similarity between the inteins from this study and metagenomic sequences from a depth profile at station ALOHA (DeLong *et al*., 2006), a location approximately 100 km North of O‘ahu, Hawai‘i.

Rank abundance curves for both sampling periods suggest a similar community composition, characterized by a few abundant members and numerous,

Figure 1 Phylogenetic tree of phycodnavirus DNA polymerase sequences based on Bayesian inference. The designations for sequences retrieved in this study are in bold type and begin with ‘KB’ for Kāne‘ohe Bay. The number of sequences that comprise each KB phylotype is shown in brackets. Clades and sequences followed by a green circle are from phycodnavirus isolates with classifications recognized by the International Committee on Taxonomy of Viruses (ICTV). All others shown are environmental sequences from previously published investigations of seawater (black type; Chen *et al*., 1996; Short and Suttle, 2002, 2003) or freshwater (blue type; Short and Short, 2008). The number of sequences that comprise each clade is in parentheses adjacent to the clade. A red star adjacent to a sequence indicates that the DNA polymerase contains a type C intein that has been removed for this analysis. Invertebrate iridescent virus (IIV) is the outgroup. The name, accession number, description and source of all the sequences used in this analysis are listed in Supplementary Table S1. Bayesian inference of phylogeny was calculated with MrBayes version 3.1.2 (Altekar *et al*., 2004) using 1 000 000 generations. Bayesian clade credibility values are shown for relevant nodes. The Bayesian scale bar indicates a distance of 0.1.

rare phylotypes (data not shown). An estimate of the mean population size identifiable with these methods using EstimateS (version 7.5, RK Colwell, http://purl.oclc.org/estimates) resulted in mean population size estimates of 45 ± 19 phylotypes for the summer and 35 ± 13 for the pooled winter data. With the intein removed, phylotype KBvp-16 polymerase is 97% identical to BSA 99-2, a sequence from a



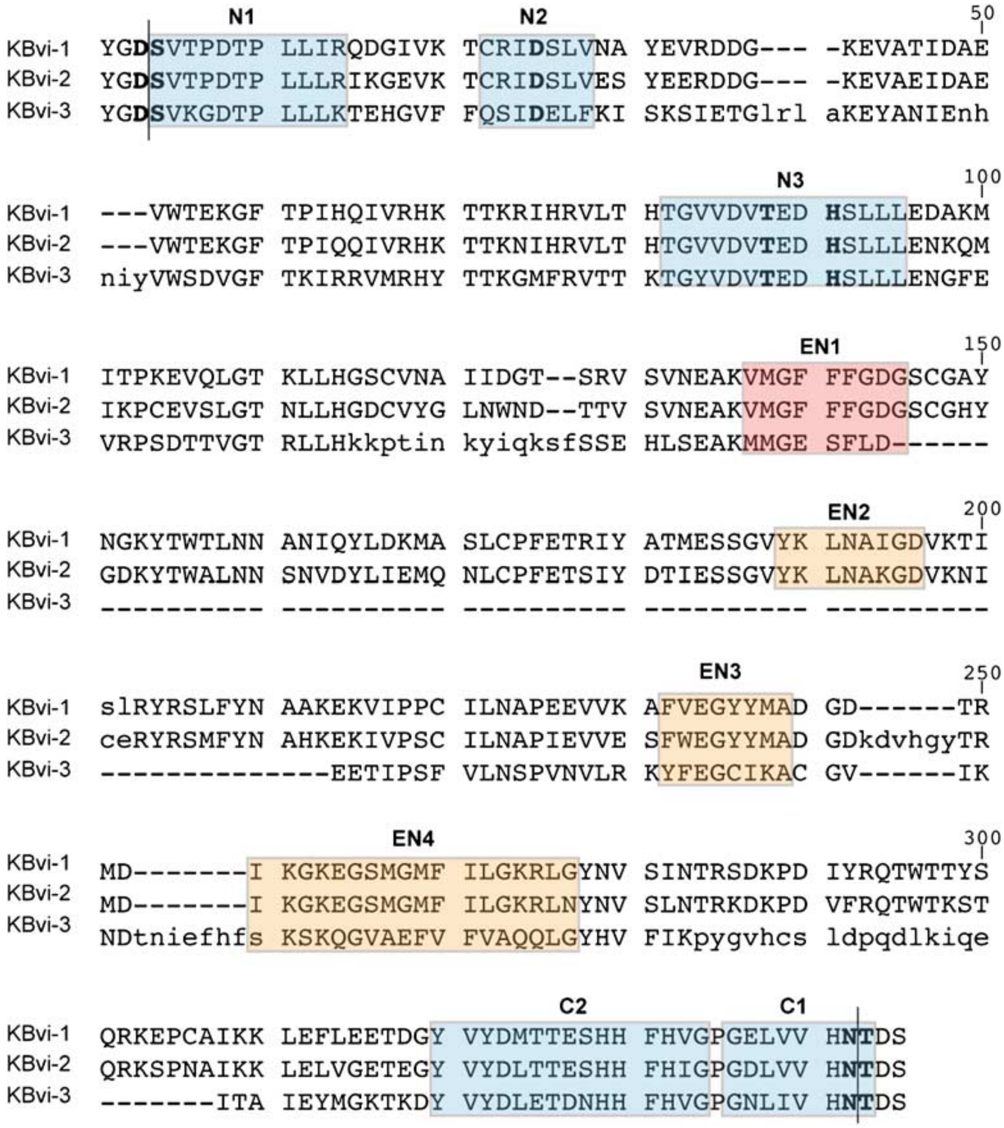

N1
N2
50
KBvi-1 YGDSVTPDTP LLIRQDGIVK TCRIDSLVNA YEVRDDG--- -KEVATIDAE
KBvi-2 YGDSVTPDTP LLLRIKGEVK TCRIDSLVES YEERDDG--- -KEVAEIDAE
KBvi-3 YGDSVKGDTP LLLKTEHGVF FQSIDELFKI SKSIETGlrl aKEYANIEnh
N3
100
KBvi-1 ---VWTEKGF TPIHQIVRHK TTKRIHRVLT HTGVVDVTED HSLLLEDAKM
KBvi-2 ---VWTEKGF TPIQQIVRHK TTKNIHRVLT HTGVVDVTED HSLLLENKQM
KBvi-3 niyVWSDVGF TKIRRVMRHY TTKGMFRVTT KTGYVDVTED HSLLLENGFE
EN1
150
KBvi-1 ITPKEVQLGT KLLHGSCVNA IIDGT--SRV SVNEAKVMGF FFGDGSCGAY
KBvi-2 IKPCEVSLGT NLLHGDCVYG LNWND--TTV SVNEAKVMGF FFGDGSCGHY
KBvi-3 VRPSDTTVGT RLLHkkptin kyiqksfSSE HLSEAKMMGE SFLD------
EN2
200
KBvi-1 NGKYTWTLNN ANIQYLDKMA SLCPFETRIY ATMESSGVYK LNAIGDVKTI
KBvi-2 GDKYTWALNN SNVDYLIEMQ NLCPFETSIY DTIESSGVYK LNAKGDVKNI
KBvi-3 ---------- ---------- ---------- ---------- ----------
EN3
250
KBvi-1 slRYRSLFYN AAKEKVIPPC ILNAPEEVVK AFVEGYYMAD GD------TR
KBvi-2 ceRYRSMFYN AHKEKIVPSC ILNAPIEVVE SFWEGYYMAD GDkdvhgyTR
KBvi-3 --------------EETIPSF VLNSPVNVLR KYFEGCIKAC GV------IK
EN4
300
KBvi-1 MD-------I KGKEGSMGMF ILGKRLGYNV SINTRSDKPD IYRQTWTTYS
KBvi-2 MD-------I KGKEGSMGMF ILGKRLNYNV SLNTRKDKPD VFRQTWTKST
KBvi-3 NDtniefhfs KSKQGVAEFV FVAQQLGYHV FIKpygvhcs ldpqdlkiqe
C2
C1
KBvi-1 QRKEPCAIKK LEFLEETDGY VYDMTTESHH FHVGPGELVV HNTDS
KBvi-2 QRKSPNAIKK LELVGETEGY VYDLTTESHH FHIGPGDLVV HNTDS
KBvi-3 -------ITA IEYMGKTKDY VYDLETDNHH FHVGPGNLIV HNTDS

Figure 2 Alignment of intein sequences from Kāne'ohe Bay. The vertical bars indicate the beginning and end of the intein within the viral DNA polymerase. Amino-acid residues in bold type are essential for self-splicing. Regions of the alignment highlighted in color are conserved intein domains. Blue boxes delineate conserved motifs of the splicing domain (N1, N2, N3, C2 and C1). The region highlighted in red (EN1) is the signature dodecapeptide motif found in endonucleases in the LAGLIDADG family, whereas the regions in orange (EN2, EN3 and EN4) identify additional conserved regions in homing endonucleases capable of horizontal transfer.

sample collected in 1999 from the west coast of Vancouver Island, British Columbia (Short and Suttle, 2002). The greatest similarity between a phylotype identified in this study and a cultivated virus was found between KBvp-17 and the DNA *pol* of the putative prasinovirus OtV5, the sequences of which share 94% amino-acid identity. The levels of amino-acid identity between phylotypes from within this study ranged from the predetermined upper cutoff of 97% (KBvp-14 and KB-15) to a minimum of 48% (KBvp-1 and KBvp-16 with the intein removed).

Bayesian phylogenetic analysis based on alignments of partial DNA polymerase sequences reflected established phycodnavirus taxonomy in which viruses cluster according to host type (Figure 1). All environmental phylotypes from this study appeared to fall within the family *Phycodnaviridae* and 20% of these (4 of 20) formed a cluster with known prasinoviruses (Bayesian support value = 100). This cluster was embedded in a larger well-supported cluster (Bayesian support value = 100) that contains sequences from freshwater and marine environments and included 90% (18 of 20) of the sequences in this study. The remaining two sequences formed a deeply branching cluster with their next nearest neighbor sharing <46% amino-acid identity.

Construction of a Bayesian phylogenetic tree based on an alignment of the three inteins from this study, putative inteins from the CAMERA metagenomic database, and DNA polymerase I motif C inteins from viral and archaeal isolates resulted in three distinct clusters (Figure 3). Inteins from viral isolates and phycodnavirus-like DNA *pol* genes amplified from the environment formed one cluster (Bayesian support value = 100), whereas inteins from halophilic and thermophilic archaeal isolates each formed monophyletic clusters (Bayesian support values = 100).

Discussion

Investigations of eukaryotic algal virus diversity were facilitated by the introduction of a PCR assay targeting the phycodnavirus DNA *pol* gene over a decade ago (Chen and Suttle, 1995). Although this assay has now been applied in a number of marine and freshwater environments, this study is the first

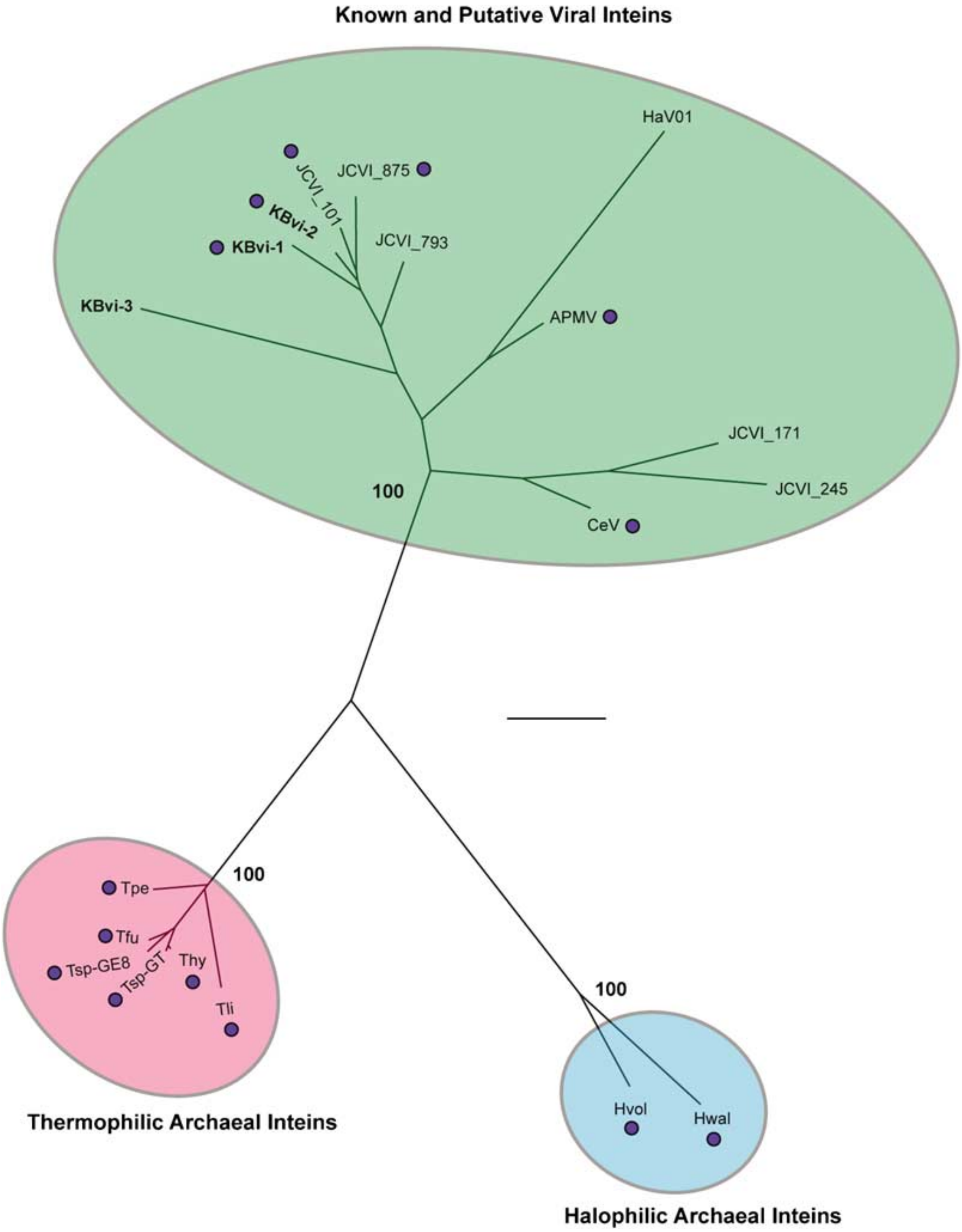


Figure 3 Phylogenetic tree of DNA polymerase I motif C inteins. The designations for sequences retrieved in this study are in bold type and begin with 'KB' for Kāne'ohe Bay. Sequences starting with 'JCVI' are environmental sequences from the Global Ocean Survey (Yooseph *et al.*, 2007). A blue circle adjacent to a sequence indicates the intein encodes a homing endonuclease. Details for the sequences used in this analysis are provided in Supplementary Table S2. Bayesian inference of phylogeny was calculated with MrBayes version 3.1.2 (Altekar *et al.*, 2004) using 1 000 000 generations. Bayesian clade credibility values are shown for relevant nodes. The Bayesian scale bar indicates a distance of 0.3.

characterization of phycodnavirus diversity in the subtropical Pacific Ocean and the first report of the direct detection of inteins in uncultivated viruses using this primer set.

In Kāne'ohe Bay, like all others in which these primers have been applied, most of the sequences recovered formed a distinct and well-supported cluster in which the few cultivated members known so far are viruses that infect prasinophytes. The maximum distance between any two sequences in this clade is similar to the maximum distance between any two DNA *pol* sequences within the established genus *Phaeovirus*. It is therefore reasonable to hypothesize that all of the viruses within this large clade belong to the genus *Prasinovirus*. Sequences from additional cultivated and rigorously classified viruses are needed to determine whether the relationship between DNA *pol* sequences and genus assignment that is apparent in the existing data is robust.

The preponderance of prasinovirus-like sequences in this and other studies may be because of the dominance of these types among the phycodnaviruses in marine and freshwater environments. It is also possible that this distribution is simply a result of bias during sample processing or in the amplification. For example, it is likely that 0.22 μm prefiltration removes some phycodnaviruses. However, we found that prasinovirus-like phylotypes (and KBvp-8 in particular) dominated both the winter 2006 sample, which was prefiltered, and the summer samples, which were not. Some PCR bias does seem likely, as it has been reported that the AVS primers do not amplify the DNA polymerase gene of several representative members of genera within the family (Sandaa *et al.*, 2001; Nagasaki *et al.*, 2005). Different approaches, such as quantitative PCR assays targeting specific viral groups, will be needed to determine the actual *in situ* abundance of prasino-like viruses relative to other genera within the family Phycodnaviridae.

Although the proportions of different sequences obtained from this study cannot be interpreted quantitatively, the data do indicate that the phycodnaviruses in Kāne'ohe Bay are quite diverse and expand our knowledge of phycodnavirus diversity. In addition to the 18 unique prasinovirus-like sequences, 2 other sequences (KBvp-1 and KBvp-2) clustered together, but diverged significantly from any established genera. The distance from this cluster to its nearest neighbor is greater than the distances between other established genera within the Phycodnaviridae. One other sequence (GoM-OTU5) from a separate study (Chen *et al.*, 1996) also diverges significantly from other known genera and from the KBvp-1 and -2 clusters. This suggests the presence of at least two new genera within the family Phycodnaviridae, which have yet to be properly described. Additional genetic and phenotypic information about the viruses from which these DNA *pol* sequences derive is required to establish their taxonomic status properly.

Although this is the first study to report the presence of inteins in viral DNA *pol* sequences amplified from the environment, this is unlikely to be a result of inteins having a restricted biogeography. Rather, intein-containing sequences have probably been overlooked in the past, as they are about 900 bp larger than the fragment size expected for *pol* genes having no intein. When examining our PCR products by gel electrophoresis, the larger intein-containing fragment, when detectable at all, was fainter than the major band representing intein-free amplicons. This faint band was assumed to be a nonspecific amplification product until cloning and sequencing proved otherwise. Amplification with intein-specific primers revealed that the intein-containing *pol* genes were more common than indicated by the AVS primer results. These results demonstrate that amplification of viral inteins with AVS primers may not occur even if an intein is present. Therefore, any evaluation of the presence or absence of viral inteins based on AVS amplification must be considered a minimum estimate.

The detection of inteins from all of our samples in this study, the presence of a variety of putative inteins from every station from the Global Ocean Survey data set (Yooseph *et al.*, 2007) and the presence of DNA polymerase I motif C inteins in the phycodnavirus isolates HaV (Nagasaki *et al.*, 2005) and CeV (Monier *et al.*, 2008) all suggest that inteins are common among marine viruses. This raises the question of how DNA polymerase I motif C inteins are acquired by phycodnaviruses. Several modes of transmission are possible. One possibility is that an intein is inserted from an alternative intein insertion site in the same genome. Sequence analysis of inteins found in type B DNA polymerase I, however, shows that these inteins are specific for their insertion site (Ogata *et al.*, 2005), suggesting that exchange between insertion sites is infrequent. Another scenario is the exchange of inteins between viral and host genomes with similar insertion sites. If this exchange were frequent, we would expect to see similarity between DNA polymerase I motif C inteins of viruses and their hosts, but this does not appear to be common. Within the intein database, we found no examples of identical inteins shared between host and virus. BLAST analyses of the extein sequences associated with putative GOS inteins returned significant hits only to phycodnavirus DNA polymerase sequences and not to polymerases of protists, their presumptive hosts (data not shown). Finally, inteins may exchange among viruses. This mechanism is consistent with the strong support for a distinct viral clade of DNA polymerase I motif C inteins (Figure 3), regardless of host. Exchange of inteins among viruses presumably requires simultaneous infections of the same host by distinct viruses. The observation that multiple phycodnavirus strains are capable of infecting the same host (Nagasaki *et al.*, 1999) offers some support to this scenario.

We did not directly determine the rate of propagation of inteins within phycodnavirus populations. However, we found no examples of a DNA polymerase phylotype occurring with and without an intein among the sequences reported so far. This suggests that the rate at which new inteins are introduced into a population is slow relative to the rate at which inteins spread throughout that population. The absence of homing endonuclease motifs from KBvi-3 indicates that this intein is now fixed (incapable of replicating itself into a homologous intein-free site). In contrast, the presence of homing endonuclease motifs in the KBvi-1 and KBvi-2 inteins suggests that these may still be capable of horizontal transfer. However, the very similar percent amino-acid identities between the inteins (76.6%) and exteins (77.4%) of these two polymerases suggest that, at least in this instance, they have diverged from a common intein-containing ancestor.

In summary, we successfully recovered novel phycodnavirus-like DNA polymerase sequences from subtropical waters. Moreover, several of these polymerase sequences harbored inteins, two of which appear to be capable of self-propagation by horizontal transfer. These results extend the known geographical range of phycodnaviruses in seawater and contribute to known phycodnavirus diversity, including the first identification of phycodnavirus inteins *in situ* by direct PCR amplification. These data provide new information about inteins, a poorly understood class of genetic parasites, and raise interesting questions about their distribution, replication cycle and effect on the evolution and ecology of viruses in the sea.

Acknowledgements

We thank Sara Yeo for her help with sample collection in Kāne‘ohe Bay, Olivia Nigro for advice and assistance, Matthias Fischer for enlightening conversations on intein biology, and Amarendran Subramanian for assistance with DIALIGN-TX. We thank Barbara Bruno, Education Director of the Center for Microbial Oceanography: Research and Education (C-MORE) and Sharon Ziegler-Chong of the Pacific Internship Programs for Exploring Science for supporting BFA through the University of Hawai‘i Hawaiian Internship Program. This work was supported by grants from the National Science Foundation (EF 04-2599) and The University of Hawai‘i Sea Grant College Program (R/EL-37). This is contribution no. 91 to the C-MORE publication series.

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**: 407–415.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bratbak G, Jacobsen A, Heldal M. (1998). Viral lysis of *Phaeocystis pouchetii* and bacterial secondary production. *Aquat Microb Ecol* **16**: 11–16.

Castberg T, Thyrhaug R, Larsen A, Sandaa RA, Heldal M, Van Etten JL *et al.* (2002). Isolation and characterization of a virus that infects *Emiliania huxleyi* (Haptophyta). *J Phycol* **38**: 767–774.

Chen F, Suttle CA. (1995). Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Appl Environ Microbiol* **61**: 1274–1278.

Chen F, Suttle CA, Short SM. (1996). Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl Environ Microbiol* **62**: 2869–2874.

Culley AI, Steward GF. (2007). New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl Environ Microbiol* **73**: 5937–5944.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Dunigan DD, Fitzgerald LA, Van Etten JL. (2006). Phycodnaviruses: a peek at genetic diversity. *Virus Res* **117**: 119–132.

Fitzgerald LA, Graves MV, Li X, Feldblyum T, Nierman WC, Van Etten JL. (2007). Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect Chlorella NC64A. *Virology* **358**: 472–484.

Fuhrman JA. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Gogarten JP, Hilario E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol Biol* **6**: 94.

Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. (2002). Inteins: structure, function, and evolution. *Annu Rev Microbiol* **56**: 263–287.

Iyer LA, Balaji S, Koonin EV, Aravind L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* **117**: 156–184.

Iyer LM, Aravind L, Koonin EV. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* **75**: 11720–11734.

Larsen JB, Larsen A, Bratbak G, Sandaa RA. (2008). Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* **74**: 3048–3057.

Markine-Goriaynoff N, Gillet L, Van Etten JL, Korres H, Verma N, Vanderplasschen A. (2004). Glycosyltransferases encoded by viruses. *J Gen Virol* **85**: 2741–2754.

Mayer JA, Taylor FJR. (1979). A virus which lyses the marine nanoflagellate *Micromonas pusilla*. *Nature* **281**: 299–301.

Milligan KLD, Cosper EM. (1994). Isolation of virus capable of lysing the brown tide microalga, *Aureococcus anophagefferens*. *Science* **266**: 805–807.

Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. (2008). Marine mimivirus relatives are probably large algal viruses. *Virol J* **5**: 12.

Murray AG, Eldridge PM. (1994). Marine viral ecology: incorporation of bacteriophage into the microbial planktonic food web paradigm. *J Plankton Res* **16**: 627–641.

Nagasaki K, Shirai Y, Tomaru Y, Nishida K, Pietrokovski S. (2005). Algal viruses with distinct intraspecies host specificities include identical intein elements. *Appl Environ Microbiol* **71**: 3599–3607.

Nagasaki K, Tarutani K, Yamaguchi M. (1999). Cluster analysis on algicidal activity of HaV clones and virus sensitivity of *Heterosigma akashiwo* (Raphidophyceae). *J Plankton Res* **21**: 2219–2226.

Nagasaki K, Yamaguchi M. (1997). Isolation of a virus infectious to the harmful bloom causing microalga *Heterosigma akashiwo* (Raphidophyceae). *Aquat Microb Ecol* **13**: 135–140.

Ogata H, Raoult D, Claverie JM. (2005). A new example of viral intein in Mimivirus. *Virol J* **2**: 8.

Perler FB. (2002). InBase: the intein database. *Nucleic Acids Res* **30**: 383–384.

Pietrokovski S. (1998). Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci* **7**: 64–71.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H *et al.* (2004). The 1.2-megabase genome sequence of mimivirus. *Science* **306**: 1344–1350.

Reisser W. (1993). Viruses and virus-like particles of freshwater and marine eukaryotic algae—a review. *Arch Protistenkd* **143**: 257–265.

Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G. (2001). Isolation and characterization of two viruses with large genome size infecting *Chrysochromulina ericina* (Prymnesiophyceae) and *Pyramimonas orientalis* (Prasinophyceae). *Virology* **290**: 272–280.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: 394–397.

Short SM, Short CM. (2008). Diversity of algal viruses in various North American freshwater environments. *Aquat Microb Ecol* **51**: 13–21.

Short SM, Suttle CA. (2002). Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl Environ Microbiol* **68**: 1290–1296.

Short SM, Suttle CA. (2003). Temporal dynamics of natural communities of marine algal viruses and eukaryotes. *Aquat Microb Ecol* **32**: 107–119.

Subramanian AR, Kaufmann M, Morgenstern B. (2008). DIALIGN_TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol Biol* **3**: 6.

Suttle CA. (2005). Viruses in the sea. *Nature* **437**: 356–361.

Van Etten JL, Lane LC, Meints RH. (1991). Viruses and virus-like particles of eukaryotic algae. *Microbiol Rev* **55**: 586–620.

Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG *et al.* (2005). Complete genome sequence and lytic phase transcription profile of a coccolithovirus. *Science* **309**: 1090–1092.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: 432–466.

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