The ISME Journal (2017) 11, 1736–1745

www.nature.com/ismej

ORIGINAL ARTICLE

Genomic exploration of individual giant ocean viruses

William H Wilson[1,2], Ilana C Gilg[1], Mohammad Moniruzzaman[3], Erin K Field[1,4], Sergey Koren[5], Gary R LeCleir[3], Joaquín Martínez Martínez[1], Nicole J Poulton[1], Brandon K Swan[1,6], Ramunas Stepanauskas[1] and Steven W Wilhelm[3]

[1]*Bigelow Laboratory for Ocean Sciences, Boothbay, ME, USA;* [2]*School of Marine Science and Engineering, Plymouth University, Plymouth, UK;* [3]*Department of Microbiology, The University of Tennessee, Knoxville, TN, USA;* [4]*Department of Biology, Howell Science Complex, East Carolina University, Greenville, NC, USA;* [5]*Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA and* [6]*National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA*

Viruses are major pathogens in all biological systems. Virus propagation and downstream analysis remains a challenge, particularly in the ocean where the majority of their microbial hosts remain recalcitrant to current culturing techniques. We used a cultivation-independent approach to isolate and sequence individual viruses. The protocol uses high-speed fluorescence-activated virus sorting flow cytometry, multiple displacement amplification (MDA), and downstream genomic sequencing. We focused on 'giant viruses' that are readily distinguishable by flow cytometry. From a single-milliliter sample of seawater collected from off the dock at Boothbay Harbor, ME, USA, we sorted almost 700 single virus particles, and subsequently focused on a detailed genome analysis of 12. A wide diversity of viruses was identified that included *Iridoviridae*, extended *Mimiviridae* and even a taxonomically novel (unresolved) giant virus. We discovered a viral metacaspase homolog in one of our sorted virus particles and discussed its implications in rewiring host metabolism to enhance infection. In addition, we demonstrated that viral metacaspases are widespread in the ocean. We also discovered a virus that contains both a reverse transcriptase and a transposase; although highly speculative, we suggest such a genetic complement would potentially allow this virus to exploit a latency propagation mechanism. Application of single virus genomics provides a powerful opportunity to circumvent cultivation of viruses, moving directly to genomic investigation of naturally occurring viruses, with the assurance that the sequence data is virus-specific, non-chimeric and contains no cellular contamination.

The ISME Journal (2017) **11,** 1736–1745; doi:10.1038/ismej.2017.61; published online 12 May 2017

Introduction

Virus diversity and functional ecology remain understudied due to the necessary constraints of working with single host–virus systems. To date, established methods for in-depth virus characterization have relied largely on cultivation-dependent techniques (Wilson *et al.*, 1993, 2002); essentially requiring growth of a receptive and permissive host organism or cell line to allow investigation of propagation dynamics of the viral pathogen. Downstream physiological characterization in combination with genomic sequence garnered from virus isolates has established the foundation for diagnostic markers that allow functional analysis of virus dynamics in nature (Martínez Martínez *et al.*, 2012; Moniruzzaman *et al.*, 2016); and additionally provides reference genomes to recruit the virus metagenomics data (Ghedin and Claverie, 2005). Although the metagenomics data in viromes has provided incredible insights into the functional ecology and diversity of oceanic viruses (Angly *et al.*, 2006; Brum *et al.*, 2015), one limitation of viromics remains a lack of reference sequences or quantitative biological context within this uncultivated fraction. Less than 0.001% of the predicted global genomic bacteriophage diversity is represented in the current databases; indeed, the number of bacterial genomes has massively surpassed those of virus representatives (Rohwer, 2003; Bibby, 2014). With the exception of viruses whose genomes are RNA (Culley *et al.*, 2006) or ssDNA (Tucker *et al.*, 2011; Labonté and Suttle, 2013), it is technically challenging and largely not possible to confidently assemble larger (primarily

Correspondence: WH Wilson, Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, Plymouth, PL1 2PB, UK.
Email: wilwil@sahfos.ac.uk
Received 6 October 2016; revised 2 March 2017; accepted 8 March 2017; published online 12 May 2017

dsDNA) virus genomes from shotgun metagenomics data (Aguirre de Cárcer *et al*., 2014). Clearly, the intractable nature of bringing the uncultured microbial majority, along with the viruses that infect them, into culture means that alternative culture-independent approaches are required to help establish reference virus genomes. Many environmental sequences remain database orphans without cultured representation or genomic context; it has thus been argued that expanding the cultivated virus genome examples, or development of culture-independent approaches, is of paramount importance to help determine the functional capacity of viruses (Culley, 2013).

To overcome the challenge of reference genome paucity, several culture-independent approaches have been developed and reviewed recently (Brum and Sullivan, 2015). These can be categorized into: (1) characterizing (sequencing) novel virus-host relationships by mining cellular metagenomes, metatranscriptomes or single amplified genomes (SAGs) (Yoon *et al*., 2011; Anantharaman *et al*., 2014; Roux *et al*., 2014; Labonté *et al*., 2015a, 2015b) and single-cell genomics combined with fosmid microarray hybridization (Martínez-García *et al*., 2014; Santos *et al*., 2014); (2) viral tagging, where fluorescently tagged wild viruses are adsorbed to cultured hosts, detected and sorted by flow cytometry then characterized by targeted metagenomics (Deng *et al*., 2014); (3) analysis of large contiguous fragments of DNA in either fosmid libraries (Garcia-Heredia *et al*., 2012; Chow *et al*., 2015), or extracted from pulsed-field gel electrophoresis bands (Ray *et al*., 2012); and finally, (4) virus particle flow cytometry sorting followed by downstream genomic sequencing (Allen *et al*., 2011; Martínez Martínez *et al*., 2014).

Single-cell genomics has provided extraordinary insight into the functional capacity of environmental microorganisms over the last decade, with fluorescence-activated cell sorting (FACS) as the most common method for cell separation (Stepanauskas, 2012). Flow cytometry is also used routinely as a tool to enumerate virus populations in aquatic environments (Marie *et al*., 1999; Brussaard, 2004). Discrete virus groups can be visualized in scatter plots of side-scatter versus fluorescence after staining with a nucleic acid fluorescent dye such as SYBR Green I (Marie *et al*., 1999). It is this property that allowed the initial proof of principle study with the genomics of individual virus particles, where whole-genome amplification was conducted on cultured, unfixed bacteriophages lambda and T4 sorted by flow cytometry (Allen *et al*., 2011). The method was further developed for giant viruses in natural seawater samples collected from the Patagonian Shelf using targeted metagenomics (viromics) (Martínez Martínez *et al*., 2014). These researchers sorted 5000 virus particles from discrete groups detected by flow cytometry, sequenced their viromes following whole-genome amplification and revealed diverse and distinct populations of giant viruses between groups (Martínez Martínez *et al*., 2014). In medical applications, flow cytometry sorting has been used to characterize the infection process of Herpes Simplex viruses by investigating specific viral intermediates (Loret *et al*., 2012); and to assess infectivity characteristics of the Junin virus following antibody binding (Gaudin and Barteneva, 2015) using a process termed flow virometry (Arakelyan *et al*., 2013).

Here we used flow cytometry to sort individual giant virus particles from seawater off the coast of the State of Maine (USA). Genomic analyses revealed a wide range of novel and diverse viruses and provided insights into their functional capacity. This application of single virus genomics to natural samples adds a new culture-independent approach to recover virus-specific genomic information. The data from such a powerful tool set will be instrumental in providing context for the metagenomic data in terms of the genetic content of individual virus particles, a large portion of which is known to be acquired from eukaryotic hosts. In addition, it will aid the investigation of virus genomic microdiversity, which is often not obvious from consensus genomes assembled from the metagenomics data.

Materials and methods

Sampling

Samples for virus sorting were collected from Coastal water from Boothbay Harbor, ME, from 1 m depth at the old Bigelow Laboratory dock (43°50′40″, 69°38′27″W) on 8 April 2011 during high tide. One milliliter samples were either left unfixed and stored at 4 °C or fixed with 0.1% glutaraldehyde (final concentration) (Martínez Martínez *et al*., 2014) for 30 min at 4 °C and snap-frozen in liquid nitrogen. Unfixed and fixed samples were stored at either 4 °C or − 80 °C, respectively until further processing. Unfixed samples were processed as soon as possible, within 4 h. Fixed samples were stored for several days.

Fluorescence-activated sorting

Samples were thawed on ice if frozen or used directly from the fridge, if unfixed, diluted 100-fold with sterile 0.2 μm-filtered 10:1 TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) and stained with SYBR Green I (Molecular Probes Inc., Eugene, OR, USA) as described by Brussaard (2004) prior to sorting. Sorting of virus groups seen by FCM was conducted at the J. J. MacIsaac Facility for Aquatic Cytometry, at Bigelow Laboratory for Ocean Sciences, with a BD Influx flow cytometer (BD Biosciences, San Jose, CA, USA) using a 488 nm argon laser for excitation, a 70 μm nozzle orifice and a CyClone robotic arm (Cytomation, Fort Collins, CO, USA) for droplet deposition. The 'single 1 drop'

mode was used to ensure the absence of non-target particles within the target particle drop and the surrounding drops. Extreme care was taken to prevent sample contamination by any non-target DNA. Sorting instruments and reagents were decontaminated as previously described (Stepanauskas and Sieracki, 2007; Rinke *et al*., 2014). Sorting was performed in a HEPA-filtered environment. The flow cytometer was triggered on side scatter and the sort gates were based on SYBR Green I green fluorescence and side scatter signals. It is worth noting that flow cytometry of viruses is based on the fluorescence of the nucleic acid-specific dye, and not on particle size; however, we made the assumption that particles sorted from distinct groups between the main (presumably small) virus group and bacteria were likely larger viruses than the bulk of small viruses (Supplementary Figure S2). These so-called giant viruses were sorted into 384-well plates for generating environmental giant virus single-amplified genomes (gvSAGs) at the Single Cell Genomics Center (SCGC) Bigelow Laboratory, East Boothbay, ME, and subsequent sequencing as described below. Sorted samples were stored at −80 °C until further processing.

Multiple displacement amplification (MDA) of DNA from sorted particles and sequencing

Bigelow dock water sorted viruses were lysed by KOH (0.4 M final concentration) at 4 °C for 10 min and the genetic material amplified employing the multiple displacement amplification, as previously described for marine bacterioplankton (Stepanauskas and Sieracki, 2007).We used the Nextera DNA sample prep kit (Epicentre, Madison, WI, USA) for library preparation of gvSAGs and barcoding. gvSAGs were sequenced on a 454 FLX genome sequencer using Titanium chemistry. Initially, 36 gvSAGs were selected for genomic sequencing based on their fastest MDA (indicative of good genome recovery (Labonté *et al*., 2015b)) and the absence of PCR-detectable bacterial 16 S rRNA genes. The reads from individual gvSAG libraries were assembled into contigs using the nGen assembler within the Lasergene suite (DNASTAR, Madison, WI, USA).

Contigs for each gvSAG were used to perform a local blastx search against the NR database and also a custom database comprised of nuclear cytoplasmic large DNA virus (NCLDV) core gene sequences, all known virus genomes in the JGI IMG database plus those from the Gordon and Betty Moore Foundation marine phage-sequencing project (www.broadinstitute.org/annotation/viral/Phage/Home.html). gvSAGs were designated as potentially of giant virus origin if at least one contig per gvSAG had significant hits (e-value $< 10^{-3}$) to any giant virus core genes. gvSAG reference numbers indicate the Single Cell Genomics Center's (SCGC) six-character microplate labels (for example, AB-001) followed by microplate well locations. Plate AB-572 received viral particles from a sample that was fixed with 0.1% glutaraldehyde prior to FACS. Plate AB-566 received viral particles from an unfixed sample. On the basis of the annotation of genes from our initial screening, we chose twelve gvSAGs for more in depth sequencing (selected randomly), 10 from the unfixed 4 °C KOH lysis plate (gvSAG reference AB-566) and two from the 0.1% glutaraldehyde-fixed 4 °C KOH lysis plate (gvSAG reference AB-572). Four of the twelve gvSAGs (AB-566-A18, AB-566-M24, AB-566-O14 and AB-572-I12) were sequenced on a single 454 FLX-titanium plate separated into two regions and the remaining eight gvSAGs (AB-566-A22, AB-566-C13, AB-566-F22, AB-566-F23, AB-566-K07, AB-566-L12, AB-566-O17 and AB-572-A11) were sequenced from another region of a 454 plate.

Because of its novel genes and phylogenetic affiliation, gvSAG AB-566-O17 was selected for deeper sequencing using a combination of PacBio RS (Pacific Biosciences, Menlo Park, CA, USA) and Illumina (Illumina Inc., San Diego, CA, USA) reads. PacBio RS libraries with a 3 Kb insert size were prepared and sequenced at the National Center for Genome Resources (NCGR) in Sante Fe, NM. Illumina paired-end libraries were prepared and sequenced at the Oregon State University Center for Genome Research and Biocomputing.

Bioinformatics analyses

The 454 reads were initially processed to trim linkers, sequences shorter than 65 bp and/or that contained any 'Ns' were removed and a low-complexity threshold of 70 (using entropy) was applied using PRINSEQ v0.20.3 (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi). Natural and artificial duplicates were removed from the pyrosequencing runs using the program cdhit-454 (http://weizhong-lab.ucsd.edu/cdhit_454/cgi-bin/index.cgi?cmd=cdhit_454).

The PacBio RS sequences were corrected using pacBioToCA from Celera Assembler v7.0. All the available Illumina data were used for correction with the ovl Mer Threshold set to 10 000 and the repeat separation threshold (coverage cutoff) for correction set to 2000. The correction generated 19 264 sequences with a maximum length of 3,196 bp and a median of 645 bp. The corrected sequences were assembled using Celera Assembler v7.0. The Illumina sequences were assembled using Velvet-SC with the covCutoff parameter set to 2000. The resulting assemblies were merged using PCAP. Finally, the resulting assembly was manually inspected. Assemblies were validated by recruiting Illumina and PacBio RS uncorrected sequences to the contigs. Contigs were mapped to themselves to identify duplications in the assembly. Duplicate contigs having low support from Illumina or both Illumina and PacBio RS uncorrected sequences were removed to generate the final assembly. A secondary

assembly was also generated using Celera Assembler v7.0 to combine corrected PacBio RS sequences with Illumina sequences down sampled to 100× of expected genome size (20.5 Mbp).

Annotation and functional assignment of the PacBio RS/Illumina assembly was performed using MG-RAST and further augmented with open reading frame (ORF) calling and annotation, especially at the ends of contigs, using GeneMarkS (http://exon.gatech.edu/GeneMark/genemarks.cgi) and blastp. The analysis was supplemented by blasting the gvSAG sequences against a custom database comprised of taxonomically diverse complete viral genomes using the blastx algorithm (version 2.2.26+) with an e-value cutoff of $\leqslant 10^{-5}$. Blast results were used to locate appropriate sequences for phylogenetic analysis. Nucleocytoplasmic Virus Orthologous Genes (NCVOG) (Yutin *et al.*, 2009, 2013) were identified within the libraries and selected for phylogenetic analysis. Predicted protein sequences were annotated by blastp against nr and NCVOG databases (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG/).

Discovery of putative viral metacaspases in the GOS data set

A local blast database of the assembled contigs from the Global Ocean Sampling (GOS) project was created. The metacaspase gene from gvSAG AB-566-O17 was queried against this database using tblastx. Contigs having a match (e-value $\leqslant 10^{-5}$) to the query metacaspase gene were selected. Using Prodigal (prodigal.ornl.gov/) we predicted gene boundaries on these contigs. To determine which contigs were of viral origin, we queried these genes against NCBI nr database using blastx. We considered a contig to be of viral origin if at least one of the genes on that contig had a best blast hit to a NCLDV member (Supplementary Data set S3). This approach resulted in 13 contigs we assumed were of NCLDV origin.

Results and Discussion

Giant virus single amplified genomes (gvSAGs)

Real-time monitoring of MDA kinetics (Swan *et al.*, 2011) detected gDNA amplification in 187 out of 690 microplate wells containing individual viral particles. The unfixed 4 °C lysis plate contained the most gvSAGs (a total of 114) with the fixed plates containing between 73 and 79 gvSAGs. PCR analysis of the two plates using bacterial 16 S rDNA primers 27 F and 907 R (Lane *et al.*, 1985) revealed that only four gvSAGs (in the unfixed 4 °C lysis plate) contained potential bacterial contamination, suggesting that the majority of gvSAGs were virally derived. Although more gvSAGs were amplified in the unfixed samples, the quality of the sequence data generated from fixed and unfixed gvSAGs was similar (data not shown).

The sequence data are submitted under NCBI bioproject, accession No. PRJNA369356 (to remain embargoed until manuscript is published). Deep genomic sequencing of 12 gvSAGs resulted in assemblies ranging from 134 kb to 1.12 Mb, comprised of between 16 to over 1000 contigs >300 bp and G+C % ranging from 28% to 36% (Supplementary Table S1). We annotated between 40 to 210 protein coding sequences (CDS) with known functional categories in each gvSAG and 36 to 1,254 CDS with unknown functions (Supplementary Data set S2) from 12

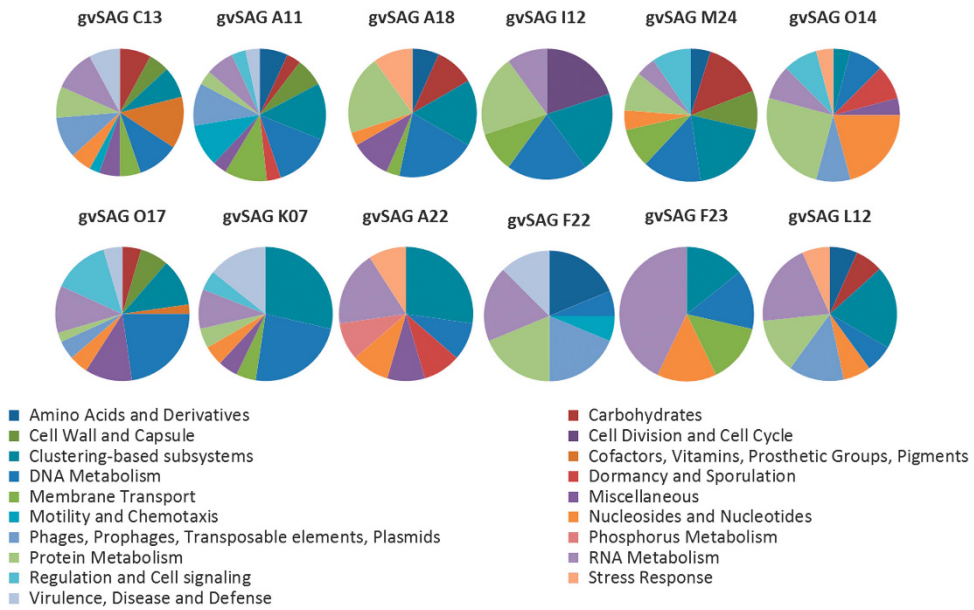

Figure 1 Annotated giant virus single amplified genomes (gvSAGs). Functional groups are colour coded as indicated. gvSAGs C13, A18, M24, O14, O17, K07, A22, F22, F23 and L12 derived from an unfixed seawater sample (plate reference AB-566); and gvSAGs A11 and I12 derived from a glutaraldehyde fixed (0.1% v/v) seawater sample (plate reference AB-572).

Table 1 NCLDV core genes used for phylogenetic analyses of gvSAGs (Supplementary Dataset S1) and their putative phylogenetic affiliations

gvSAG	*Taxonomy*	*NCLDV core genes used in phylogenetic inference*	*Phylogenetic tree*
SCGC AB-566-C13	Iridoviridae	B-family DNA polymerase	Supplementary Data set S1*H*
SCGC AB-572-A11	Mimiviridae (algal)	B-family DNA polymerase	Supplementary Data set S1*A*
SCGC AB-566-A18	Mimiviridae (algal)	D5-primase/helicase	Supplementary Data set S1*G*
SCGC AB-572-I12	Unresolved	A32- Virion packaging ATPase; Major Capsid Protein	Supplementary Data set S1*C, D*
SCGC AB-566-M24	Mimiviridae (algal)	D5-primase/helicase	Supplementary Data set S1*F*
SCGC AB-566-O14	Mimiviridae	B-family DNA polymerase	Supplementary Data set S1*L*
SCGC AB-566-O17	Mimiviridae	A32-virion packaging ATPase	Supplementary Data set S1*M*; Figure 3
SCGC AB-566-K07	Mimiviridae (algal)	A32-virion packaging ATPase	Supplementary Data set S1*E*
SCGC AB-566-A22	Mimiviridae (algal)	A32-virion packaging ATPase	Supplementary Data set S1*B*
SCGC AB-566-F22	Mimiviridae (algal)	B-family DNA polymerase	Supplementary Data set S1*I*
SCGC AB-566-F23	Mimiviridae (algal)	A32-virion packaging ATPase	Supplementary Data set S1*J*
SCGC AB-566-L12	Mimiviridae (algal)	D5-primase/helicase	Supplementary Data set S1*K*

gvSAGs (Figure 1). Phylogenetic analyses, using the conserved nucleocytoplasmic large DNA virus (NCLDV) core genes (Yutin *et al*., 2009, 2013), inferred that our gvSAGs likely derived from the virus families *Iridoviridae,* or candidate *Mimiviridae* (Table 1; Supplementary Data set S1; Supplementary Data set S1 legend). One gvSAG, AB-572-I12, had unresolved taxonomy (Supplementary Data set S1C and S1D), although phylogenetic inference from the A32 virion packaging ATPase phylogeny suggested it may represent an outgroup of the *Phycodnaviridae* (dsDNA viruses that infect algae (Wilson *et al*., 2009)) and *Mimiviridae* families. However, major capsid protein phylogeny of gvSAG AB-572-I12 suggested it might be a highly divergent member of the *Phycodnaviridae* clade, with phylogenetic affinity to *Heterosigma akashiwo* virus 01 (HaV-01) and *Emiliania huxleyi* virus 86 (EhV-86) (Supplementary Data set S1*C*, S1*D*). Five gvSAGs (AB-566-K07, AB-566-A22, AB-566-F23, AB-566-F22 and AB-572-A11) fell within a well-supported clade that also contained the *Phaeocystis globosa* virus 16 T (PgV-16T), a giant algal virus that is not actually classified as *Phycodnaviridae*, and despite being regularly referred to as candidate *Megaviridae* (Santini *et al*., 2013), it appears to be taxonomically affiliated with what is referred to now as the extended *Mimiviridae* (Yutin *et al*., 2013); a term we have adopted in this study. These viruses are part of the 'algal *Mimiviridae'* sister clade, a group of algal viruses showing high diversity and possibly widespread in world's oceans (Moniruzzaman *et al*., 2014; Wilson *et al*., 2014; Moniruzzaman *et al*., 2016).

Since the majority of the gvSAGs showed high phylogenetic affinity to PgV or other algal *Mimiviridae* members (Supplementary Data set S1), we recruited raw 454 reads from these gvSAGs to the PgV genome using SMALT (https://www.sanger.ac.uk/resources/software/smalt/). Total recruitment percentages ranged from 4% to 6.4% for the gvSAGs in the PgV clade; and 1.7% to 4.4% for the other gvSAGs, with relatively even coverage of the PgV--12 T genome in each case (Supplementary Figure S1). It is worth noting that *Phaeocystis globosa* blooms are commonly observed in the Boothbay coastal region. Viruses are thought to be instrumental in the decline of *P. globosa* blooms and can be readily isolated (Brussaard *et al*., 2004; Wilson *et al*., 2006). However, if some of these gvSAGs are infecting *P. globosa* in nature, it is likely their genome similarities would actually be much greater than 6% of the isolated PgV strains. It is likely that the phylogenetic affinity of many of our gvSAGs to PgVs is merely a consequence of the dearth of representative genomes within this clade. A new diagnostic marker has been developed for algal *Mimiviridae* members based on the DNA mismatch repair gene (MutS), which has revealed that algal *Mimiviridae* are diverse and likely abundant in coastal waters off Boothbay (Wilson *et al*., 2014). Hence, it is most likely these gvSAGs are novel algal viruses in the extended *Mimiviridae* clade and new sequence information from them should be considered in this functional context.

Annotation of gvSAGs

The sequence information from each gvSAG revealed a large mix of known Nucleo-Cytoplasmic Virus Orthologous Groups (NCVOGs) (Supplementary Data set S3), helping to confirm their viral origin. But, we also discovered a number of novel virus genes never seen before in NCLDVs (several genes of interest are highlighted in Supplementary Table S4, with a detailed annotation in Supplementary Data set S2). Among the noteworthy examples:

(1) The putative Iridovirus gvSAG AB-566-C13 appears to contain all four eukaryote histone-like proteins, including H4, which is the first for a virus. Lausannevirus and Marseillevirus both encode three histone-like proteins and are missing H4 (Thomas *et al*., 2011). H4 alone has been found in the *Cotesia plutellae* bracovirus (CpBV), a polydnavirus that infects a parasitic wasp (Ibrahim *et al*., 2005) and is thought to play a critical role in suppressing

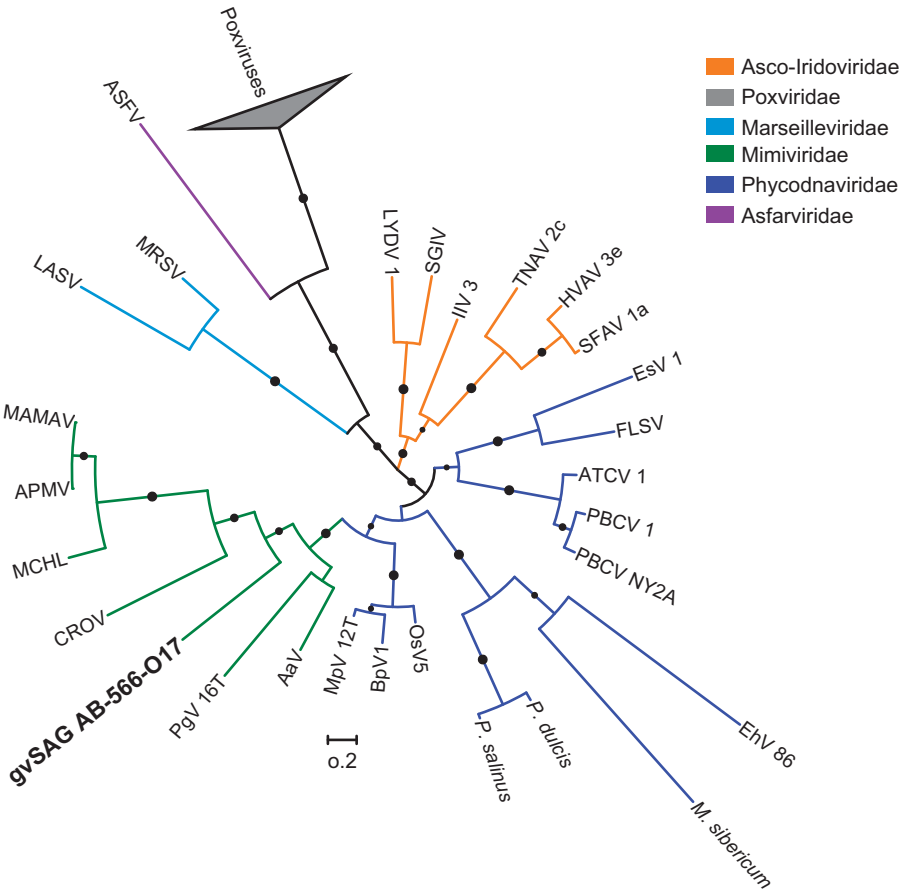

Figure 2 Maximum likelihood phylogenetic tree of gvSAG-566-O17 and other NCLDV members constructed from A32 virion packaging ATPase, a NCLDV core gene. Support values >50% are shown at the nodes as black circles. NCBI gi numbers: MAMAV, Mamavirus (351737587); APMV, *Acanthamoeba polyphaga* Mimivirus (311977822); MCHL, *Megavirus chilensis* (363539780), CROV, *Cafeteria roenbergensis* virus (310831328); PgV, *Phaeocystis globosa* virus (508181864); AaV, *Aureococcus anophagefferens* virus (672551239); MpV, *Micromonas pusilla* virus (472342694); BpV, *Bathycoccus prasinos*. virus (313768078); OsV, *Ostreococcus* sp. virus (163955071); *P. salinus*, *Pandoravirus salinus* (531037319); *P. dulcis, Pandoravirus dulcis* (526119111); *M. sibericum, Mollivirus sibericum* (927594479); EhV, *Emiliania huxleyi* virus (73852542); PBCV, *Paramecium bursaria Chlorella* virus (340025835); ATCV, *Acanthamoeba turfacea Chlorella* virus (155371384); FLSV, *Feldmannia* sp. virus (197322436); EsV, *Ectocarpus siliculosus* virus (13242498); SFAV, *Spodoptera frugiperda* ascovirus (115298612); HVAV, *Heliothis virescens* ascovirus (134287264); TNAV, *Trichoplusia ni* ascovirus (116326781); IIV, Invertebrate iridescent virus (109287966); SGIV, Singapore grouper iridovirus (56692771); LYDV, Lymphocystis disease virus (13358448); ASFV, African swine fever virus (9628186); MRSV, Marseillevirus (284504185); LASV, Lausannevirus (327409704).

host immune response during parasitization (Gad and Kim, 2008).

(2) Putative Mimivirus gvSAG AB-566-O14 was notable since it contained both a putative reverse transcriptase and a transposase. Retrotransposons are known to intersperse giant virus genome segments (Maumus *et al.*, 2014) and reverse transcriptase has been annotated in entomopoxvirus (YP_009001626.1). However, containing both is novel for NCDLVs and, although highly speculative, suggests a unique propagation strategy that exploits a latency mechanism. Latency is certainly known to exist as a propagation mechanism in NCDLVs, for example, many of the viruses that infect seaweeds (phaeoviruses) can exhibit a latent or persistent propagation strategy (Stevens *et al.*, 2014), however, the molecular mechanism is poorly understood.

(3) The same virus also contains six tRNA sythetases, including glutaminyl- and lysyl-tRNA synthetase, neither of which have been previously observed in NCLDVs. It also contains Translation Initiation factor 4E, which in combination with the tRNA synthetases makes this gvSAG more 'cell-like' since it encodes many of its own translation genes, further blurring the boundary between a virus and a cell, an open debate that continues to evolve in the literature (Legendre *et al.*, 2012; Yutin *et al.*, 2014).

(4) Putative Mimivirus gvSAG AB-566-L12 contains least eight tRNAs, including lysine, glutamine, glutamic acid, methionine, arginine, leucine, isoleucine and asparagine.

In-depth analysis of gvSAG AB-566-O17

To test the limits of genomic resolution for one gvSAG, we employed PacBio RS sequencing on

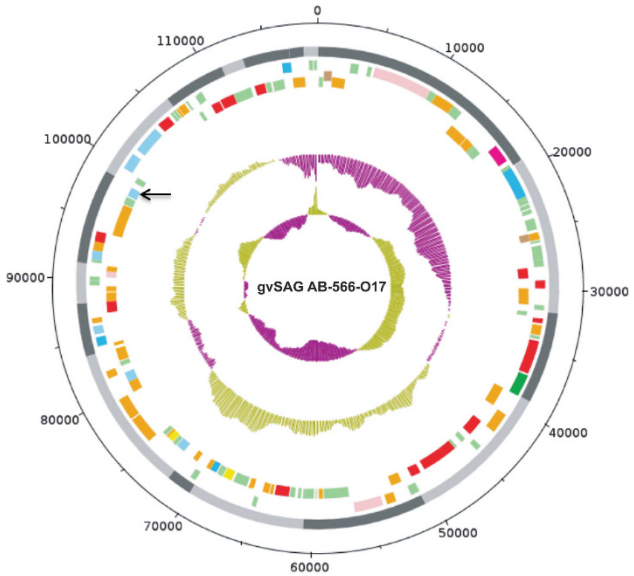

Figure 3 Circular representation of gvSAG AB-566-O17. The outside scale is numbered clockwise in bp. Circle 1 (outermost) represents the 16 contigs in no particular order; circles 2 and 3 (from outside in) are coding sequences (CDSs) on forward and reverse strands, respectively, starting with CDS gvSAG 566-O17_0001 through to gvSAG 566-O17_0122 (listed in Supplementary Table 4). The innermost tracks display GC % and GC skew, respectively. The location of the putative metacaspase is marked with an arrow. CDSs are colour coded by putative functional groups: red, transcription/translation/DNA+RNA modification; dark green, surface or secretory; blue, degradation of large molecules; dark pink, degradation of small molecules; yellow, secondary metabolites/miscellaneous metabolism; light green, novel unknown; light blue, regulators; orange, conserved hypothetical; brown, pseudogenes and partial genes; light pink, virus structural proteins and virion packaging.

gvSAG AB-566-O17, a putative *Mimiviridae* (Figure 2; Supplementary Data set S1M). We assembled a 134 kb concatenated genome comprising 16 contigs (Figure 3). The assembled genome had a G+C content of 36%, with functional genes from 12 different COG functional groups, many of which were consistent with coding sequences (CDSs) previously identified in giant virus genomes; 58 CDSs with unknown function; and 64 CDSs with predicted functions based on blastp analysis (Figure 3; Supplementary Table S2). To date, giant virus genomes within the *Mimiviridae* range from the 371 kb genome of 'little giant', AaV (that infects the brown tide causing microalgae *Aureococcus anophagefferens* (Moniruzzaman *et al.*, 2014)), to the *Acanthamoeba* infecting *Megavirus chilensis* at 1.3 Mb (Arslan *et al.*, 2011). Although the 36% G+C content was similar to other A+T-rich *Mimiviridae* (G+C range 23–32% for PgV, CroV, Megavirus, Mimivirus, AaV (Aherfi *et al.*, 2016)), the concatenated gvSAG AB-566-O17 genome length of 134 kb was considerably smaller than we anticipated. This suggests that we only covered a portion of the gvSAG AB-566-O17genome. For sequenced single bacterial cells, this range is not so unusual, with coverage ranging from a few percent to greater than 90% (Swan *et al.*, 2013; Rinke *et al.*, 2014). However, despite low coverage of individual genomes, we did observe relatively even coverage as demonstrated with recruitment of gvSAGs on the PgV-12T genome (Supplementary Figure S1), in addition to low but even coverage observed on a test EhV-86 SAG (data not shown). The advantage of such even distribution is that we can annotate a wide range of functional and novel CDSs with the knowledge that they are likely virus-specific and they are derived from the same virus.

One of the more exciting virus-specific CDSs from gvSAG AB-566-O17 was an 813 bp CDS (gene number gvSAG-O17_0097, see arrow in Figure 3) that appears to encode the metacaspase *casA*, with the closest blast hit to a metacaspase from a symbiotic fungus *Leucoagaricus sp.*, known to be cultivated by the fungus-growing ant *Cyphomyrmex costatus*. Metacaspases are cysteine-dependent proteases found in taxonomically diverse organisms, including plants, fungi and unicellular protists (Tsiatsiani *et al.*, 2011); in the marine environment they are widespread among prokaryotic and phytoplankton genomes (Bidle, 2015). We can only speculate at the function of this virally derived metacaspase, particularly since we cannot verify which host is infected by gvSAG AB-566-O17. Metacaspases are synonymous with autocatalytic cell death in single-celled organisms (akin to programmed cell death (PCD) in multicellular organisms), and involves a highly coordinated molecular and biochemical regulation of cellular death machinery (Vardi *et al.*, 1999). If viruses could exploit this mechanism during infection by encoding their own metacaspase, it may serve to enhance the efficiency of infection by rewiring host metabolism (Bidle and Vardi, 2011). Different studies have examined the interplay between host-derived metacaspases (Bidle *et al.*, 2007) and viral glycosphingolipids (Vardi *et al.*, 2009) in regulating cell's fate in this system. The discovery of a sphingolipid biosynthesis pathway in the EhV-86 genome, a pathway known to mediate PCD-like processes (Wilson *et al.*, 2005), illustrates the potential utility of this type of strategy.

Phylogenetic analysis of metacaspases from the global ocean survey (GOS) samples (Venter *et al.*, 2004) revealed that they form a separate cluster with the gvSAG AB-566-O17 metacaspase (Figure 4) and that virus-encoded metacaspases may be widespread. Interestingly, the existence of a single host-derived sequence (from *E. huxleyi*) within the virus clade highlights the potential for gene transfer between viruses and hosts.

Our observations suggest that metacaspase-mediated cellular rewiring by viruses may be widespread in ocean systems (though we acknowledge the GOS contigs used in Supplementary Table S3 only cover part of the North West Atlantic Ocean and Sargasso Sea). Indeed, further analysis of the 13 GOS contigs that contain putative viral metacaspases reveal a range of other CDSs on these contigs that may now

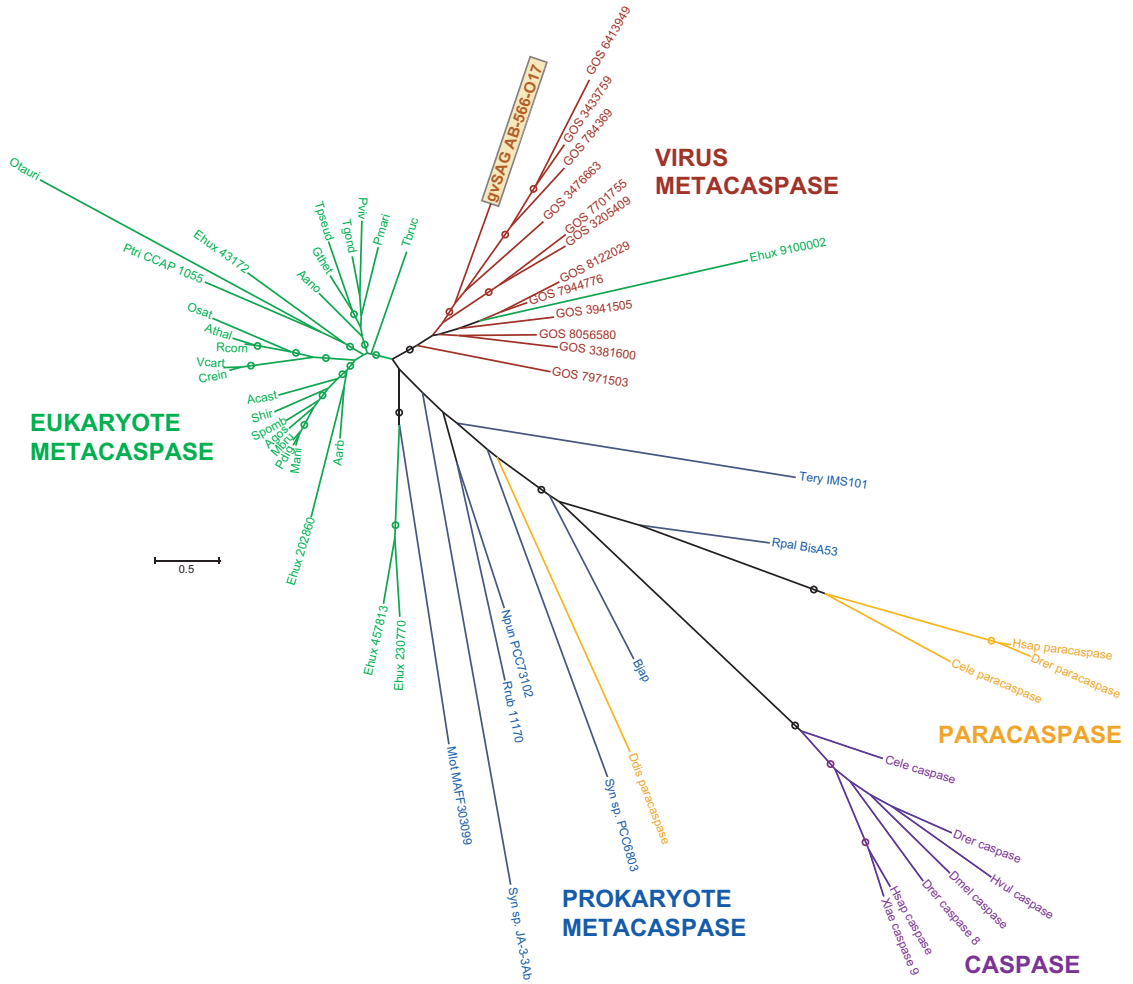

Figure 4 Metacaspase phylogenetic tree. Maximum likelihood inference of the p10 and p20 catalytic cores of metacaspases using the LG substitution model with an estimated proportion of invariable sites and four substitution rate categories. Caspases and paracaspases are included as outgroups (purple and yellow branches, respectively). Expected Likelihood Weights (1000 replications) greater than 75% are indicated as circles at the nodes. Green branches denote eukaryotic metacaspase sequences, blue prokaryotic metacaspase sequences and red putative and known viral metacaspase sequences (see Supplementary Data set S4 for abbreviation definitions). Scale bar represents the number of inferred amino acid substitutions per site.

be considered of viral origin (Supplementary Table S3). As a first step, this type of sequence analysis will help populate databases providing parallel reference to intact virus genomes. However, it should be noted that without some level of validation, these relationships remain only conjectural.

Conclusions

Expansion and evolution of modern molecular biology into the realms of ecology and marine biology has brought with it new discoveries and observations that have reshaped our basic understanding of ocean systems. This study has outlined our ability to investigate viruses on a particle–by-particle basis and has created yet another opportunity to resolve how these obligate pathogens constrain processes in the marine environment. Development of such a tool is clearly important for virus ecology since it opens the door to a culture independent and contamination-free approach to assess functional capacity of viruses. As with all the sequence data, it can only provide inference to the potential function of the biological entity it was derived from, yet it provides a tantalizing glimpse into novel biological strategies in nature. Here we identified a wide range of potentially novel functions for viruses such as a latent propagation strategy as speculated for gvSAG AB-566-O14. Such discoveries provide fodder for new hypothesis-driven ecological research in the future, for example, is latency or persistent infections more prevalent than previously thought? Importantly, single virus genomics will allow genomic databases to be populated with the novel virus data to help provide functional context for metagenomic sequence. Identification of the first known virus metacaspase in gvSAG-566-O17 allowed us to mine similar metacaspases from the GOS metagenomics data set; these would previously have been called as eukaryotic in origin. Finally, the broad spectrum of observations arising from the techniques we describe will, in the future, shape our

understanding of processes ranging from gene flow to global carbon cycling.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We acknowledge the support from NSF grants EF-0949162 (to WHW, RS, and SWW) and OCE-0821374 (to RS and WHW). Support for this project was also received from the Kenneth & Blaire Mossman endowment to the University of Tennessee. SK was supported by the Intramural Research Program of the National Human Genome Institute, National Institutes of Health. JMM was supported by the Gordon and Betty Moore Foundation through Grant GBMF5334.

References

Aguirre de Cárcer D, Angly FE, Alcamí A. (2014). Evaluation of viral genome assembly and diversity estimation in deep metagenomes. *BMC Genomics* **15**: 1–11.

Aherfi S, Colson P, La Scola B, Raoult D. (2016). Giant viruses of amoebas: an update. *Front Microbiol* **7**: 349.

Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ. (2011). Single virus genomics: a new tool for virus discovery. *PLoS ONE* **6**: e17722.

Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**: 757–760.

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

Arakelyan A, Fitzgerald W, Margolis L, Grivel J-C. (2013). Nanoparticle-based flow virometry for the analysis of individual virions. *J Clin Invest* **123**: 3716–3727.

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci* **108**: 17486–17491.

Bibby K. (2014). Improved bacteriophage genome data is necessary for integrating viral and bacterial ecology. *Microb Ecol* **67**: 242–244.

Bidle KD, Haramaty L, Barcelos e Ramos J, Falkowski P. (2007). Viral activation and recruitment of metacaspases in the unicellular coccolithophore, Emiliania huxleyi. *Proc Nail Acad Sci* **104**: 6049–6054.

Bidle KD, Vardi A. (2011). A chemical arms race at sea mediates algal host-virus interactions. *Curr Opin Microbiol* **14**: 449–457.

Bidle KD. (2015). The molecular ecophysiology of programmed cell death in marine phytoplankton. *Annu Rev Mar Sci* **7**: 341–375.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A *et al.* (2015). Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498.

Brum JR, Sullivan MB. (2015). Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Micro* **13**: 147–159.

Brussaard CPD. (2004). Optimization of procedures for counting viruses by flow cytometry. *Appl Environ Microbiol* **70**: 1506–1513.

Brussaard CPD, Short SM, Frederickson CM, Suttle CA. (2004). Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton Phaeocystis globosa (Prymnesiophyceae). *Appl Environ Microbiol* **70**: 3700–3705.

Chow C-ET, Winget DM, White RA, Hallam SJ, Suttle CA. (2015). Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front Microbiol* **6**: 265.

Culley AI, Lang AS, Suttle CA. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**: 1795–1798.

Culley AI. (2013). Insight into the unknown marine virus majority. *Proc Natl Acad Sci* **110**: 12166–12167.

Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P *et al.* (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513**: 242–245.

Gad W, Kim Y. (2008). A viral histone H4 encoded by Cotesia plutellae bracovirus inhibits haemocyte-spreading behaviour of the diamondback moth, Plutella xylostella. *J Gen Virol* **89**: 931–938.

Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Antón J *et al.* (2012). Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. *PLoS ONE* **7**: e33802.

Gaudin R, Barteneva NS. (2015). Sorting of small infectious virus particles by flow virometry reveals distinct infectivity profiles. *Nat Commun* **6**: 6022.

Ghedin E, Claverie JM. (2005). Mimivirus relatives in the Sargasso Sea. *Virol J* **2**: 62.

Ibrahim AMA, Choi JY, Je YH, Kim Y. (2005). Structure and expression profiles of two putative cotesia plutellae bracovirus genes (CpBV-H4 and CpBV-E94a) in Parasitized Plutella xylostella. *J Asia-Pac Entomol* **8**: 359–366.

Labonté JM, Suttle CA. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J* **7**: 2169–2177.

Labonté JM, Field EK, Lau M, Chivian D, Van Heerden E, Wommack KE *et al.* (2015a). Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front Microbiol* **6**: 349.

Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ *et al.* (2015b). Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J* **9**: 2386–2399.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985). Rapid determination of 16 S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci* **82**: 6955–6959.

Legendre M, Arslan D, Abergel C, Claverie J-M. (2012). Genomics of Megavirus and the elusive fourth domain of Life. *Commun Integr Biol* **5**: 102–106.

Loret S, El Bilali N, Lippé R. (2012). Analysis of herpes simplex virus type I nuclear particles by flow cytometry. *Cytometry A* **81A**: 950–959.

Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D. (1999). Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl Environ Microbiol* **65**: 45–52.

Martínez-García M, Santos F, Moreno-Paz M, Parro V, Antón J. (2014). Unveiling viral–host interactions within the 'microbial dark matter'. *Nat Commun* **5**: 4542.

Martínez Martínez J, Schroeder DC, Wilson WH. (2012). Dynamics and genotypic composition of *Emiliania huxleyi* and their co-occurring viruses during a coccolithophore bloom in the North Sea. *FEMS Microbiol Ecol* **81**: 315–323.

Martínez Martínez J, Swan BK, Wilson WH. (2014). Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J* **8**: 1079–1088.

Maumus F, Epert A, Nogué F, Blanc G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. *Nat Commun* **5**: 4268.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH *et al.* (2014). Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. *Virology* **466–467**: 60–70.

Moniruzzaman M, Gann ER, LeCleir GR, Kang Y, Gobler CJ, Wilhelm SW. (2016). Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, Aureococcus anophagefferens. *FEMS Microbiol Ecology* **92**: fiw058.

Ray J, Dondrup M, Modha S, Steen IH, Sandaa R-A, Clokie M. (2012). Finding a needle in the virus metagenome haystack - micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS ONE* **7**: e34238.

Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N *et al.* (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nat Protoc* **9**: 1038–1048.

Rohwer F. (2003). Global phage diversity. *Cell* **113**: 141.

Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R *et al.* (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**: e03125.

Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C *et al.* (2013). Genome of Phaeocystis globosa virus PgV-16 T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci* **110**: 10800–10805.

Santos F, Martinez Garcia M, Parro V, Antón J. (2014). Microarray tools to unveil viral-microbe interactions in nature. *Front Ecol Evol* **2**: 90–94.

Stepanauskas R, Sieracki ME. (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci* **104**: 9052–9057.

Stepanauskas R. (2012). Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**: 613–620.

Stevens K, Weynberg K, Bellas C, Brown S, Brownlee C, Brown MT *et al.* (2014). A novel evolutionary strategy revealed in the phaeoviruses. *PLoS ONE* **9**: e86040.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**: 1296–1300.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci* **110**: 11463–11468.

Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A *et al.* (2011). Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ Microbiol* **13**: 1454–1466.

Tsiatsiani L, Van Breusegem F, Gallois P, Zavialov A, Lam E, Bozhkov PV. (2011). Metacaspases. *Cell Death Differ* **18**: 1279–1288.

Tucker KP, Parsons R, Symonds EM, Breitbart M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**: 822–830.

Vardi A, Berman-Frank I, Rozenberg T, Hadas O, Kaplan A, Levine A. (1999). Programmed cell death of the dinoflagellate Peridinium gatunense is mediated by CO2 limitation and oxidative stress. *Curr Biol* **9**: 1061–1064.

Vardi A, Van Mooy BAS, Fredricks HF, Popendorf KJ, Ossolinski JE, Haramaty L *et al.* (2009). Viral Glycosphingolipids Induce lytic infection and cell death in marine phytoplankton. *Science* **326**: 861–865.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wilson WH, Joint IR, Carr NG, Mann NH. (1993). Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp strain WH7803. *Appl Environ Microbiol* **59**: 3736–3743.

Wilson WH, Tarran GA, Schroeder D, Cox M, Oke J, Malin G. (2002). Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. *J Mar Biol Assoc UK* **82**: 369–377.

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.

Wilson WH, Schroeder DC, Ho J, Canty M. (2006). Phylogenetic analysis of PgV-102 P, a new virus from the english channel that infects Phaeocystis globosa. *J Mar Biol Assoc UK* **86**: 485–490.

Wilson WH, van Etten JL, Allen MJ. (2009). The *Phycodnaviridae* the story of how tiny giants rule the world. *Curr Top Microbiol Immunol* **328**: 1–42.

Wilson WH, Gilg IC, Duarte A, Ogata H. (2014). Development of DNA mismatch repair gene, MutS, as a diagnostic marker for detection and phylogenetic analysis of algal Megaviruses. *Virology* **466–467**: 123–128.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH *et al.* (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**: 714–717.

Yutin N, Wolf YI, Raoult D, Koonin EV. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**: 223–223.

Yutin N, Colson P, Raoult D, Koonin EV. (2013). Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Viroly J* **10**: 106.

Yutin N, Wolf YI, Koonin EV. (2014). Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466–467**: 38–52.

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