Genome sequence of *Ostreococcus tauri* virus OtV-2 enlightens the role of picoeukaryote niche separation in the ocean

Running title: Genome sequence of a low light *Ostreococcus tauri* virus

Karen D. Weynberg[1,2], Michael J. Allen[1], Ilana C. Gilg[3], David J. Scanlan[2] and William H. Wilson[1,3]*

[1]Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, PL1 3DH, UK

[2]School of Life Sciences, University of Warwick, Coventry, CV4 7AL.

[3]Bigelow Laboratory for Ocean Sciences, 180 McKown Point, West Boothbay Harbor, Maine 04575-0475, USA

E-mail addresses:

kwey@pml.ac.uk

mija@pml.ac.uk

igilg@bigelow.org

D.J.Scanlan@warwick.ac.uk

wwilson@bigelow.org*

*Mailing address: Bigelow Laboratory for Ocean Sciences, 180 McKown Point, West Boothbay Harbor, Maine 04575-0475, USA
*Phone: (207)633-9666
*Fax: (207) 633-9641

***Corresponding author**

Abstract Word Count = 240

Text Word count = 5173

Abstract

Ostreococcus tauri, a unicellular marine green alga, is the smallest known free-living eukaryote and is ubiquitous in the surface oceans. The ecological success of this organism has been attributed to distinct low- and high-light adapted ecotypes existing in different niches at a range of depths in the ocean. Viruses have already been characterised that infect the high-light adapted strains. *Ostreococcus tauri* virus isolate OtV-2 is a large double stranded DNA algal virus that infects a low-light adapted strain of *O. tauri* and was assigned to the algal virus family *Phycodnaviridae*, genus *Prasinovirus*. Our working hypothesis for this study was that different viruses infecting high-light vs. low-light adapted *O. tauri* strains would provide clues to propagation strategies that would give them selective advantages within their particular light niche. Sequence analysis of the 184,409 base pair linear OtV-2 genome revealed a range of core functional genes exclusive to this low-light genotype and included a variety of unexpected genes, such as those encoding a RNA polymerase sigma factor, at least four DNA methyltransferases, a cytochrome b_5 and a high affinity phosphate transporter. It is clear that OtV-2 has acquired a range of potentially functional genes from its host, other eukaryotes and even bacteria over evolutionary time. Such piecemeal accretion of genes is a trademark of large double-stranded DNA viruses that has allowed them to adapt their propagation strategies to keep up with host niche separation in the sunlit layers of the oceanic environment.

Introduction

Ostreococcus tauri is the smallest free-living eukaryote described to date, with a size of less than 1 µm (11). The cellular organisation of *O. tauri* is very simple with only a single chloroplast, a single mitochondrion, a single Golgi body and a very reduced cytoplasmic compartment (22). *O. tauri* also lacks flagella and there is no cell wall surrounding the cell membrane.

The *Ostreococcus* genus includes distinct genotypes physiologically adapted to high- or low-light environments, providing evidence of niche adaptation in eukaryotic picophytoplankton (39). Such adaptation has been well-characterised in recent studies on the diversity and ecophysiology of the cyanobacterium, *Prochlorococcus*. The global success of this abundant prokaryotic primary producer has been in part attributed to distinct low- and high-light adapted ecotypes existing in different niches and utilising different resources (38). Strains of *O. tauri* have been isolated in geographically different locations and depths and were shown to be genetically (based on 18S rRNA and internal transcribed spacer (ITS) sequencing) and physiologically (light-limited growth rates) different from one another (39). Growth rates of strains isolated from deep in the euphotic zone were reported to display severe photoinhibition at high light intensities (and are thus commonly referred to as low-light-adapted strains), while strains isolated from surface waters have very slow growth rates at the lowest light intensities (and are thus commonly referred to as high-light-adapted strains). Genetic distances between isolates appear to result from a contrast in both light and nutrient conditions experienced between surface and deep isolates, which drive their genetic divergence (7, 39, 44).

Another factor that has not been considered in determining niche separation in *Ostreococcus* spp. is the role that viruses play. There are two primary mechanisms

that viruses use to shape the diversity and magnitude of microbial populations. The first is simply killing cells leading to host-specific lysis. Here, viruses exert an important influence on the biogeochemistry of the oceans as nutrients are shunted between the particulate and dissolved phases (20, 51). A second and arguably more important function that viruses play is their role in horizontal gene transfer (HGT). Viruses can simply be seen as vectors that facilitate gene shuttling, a role that has been poorly described in marine systems. However, genes transferred between hosts and viruses can give selective advantages in growth (for the host) or propagation (for the virus) in particular environmental niches.

Information on virus propagation strategies and HGT events can be inferred and deduced, respectively, from genome sequence information. *Ostreococcus* spp. is an excellent model system since there are two host genomes, both of which are high-light adapted species (15, 32) and two virus genomes (14, 50) already sequenced. All grow or propagate in high-light adapted systems. Our working hypothesis for this study was that different viruses infecting high-light vs low-light adapted *O. tauri* strains would provide clues to propagation strategies that would give them selective advantages within their particular light niche. Here, we report the genomic sequence of a virus (OtV-2) that infects a low-light adapted strain of *O. tauri* and we compare inferred functionality of coding sequences with its high-light counterparts.

Materials and methods

Virus isolation

The virus OtV-2 was isolated from surface seawater collected on 07/05/2007 at the L4 sampling station in the Western English Channel (co-ordinates are 50º15'N, 04º13'W). The OtV-2 host, *Ostreococcus tauri* strain RCC 393 was grown in Keller (K) medium (25) at 20°C, under a 16:8 light:dark cycle, at irradiance of 30 µmol m^{-2} s^{-1} in a Sanyo MLR-350 incubator. In order to obtain clonal virus stocks, OtV-2 was purified by serial dilution to extinction, as the host strain failed to grow successfully on agarose solid bottom plates preventing the use of plaque purification techniques Briefly, virus lysate was obtained by adding 100 µl of concentrated seawater from station L4 to exponentially growing *O. tauri.* Once clearing of the host culture was observed, the lysate was passed through a PVDF 0.2 µm filter (Durapore, Millipore).

Filtered lysate was used to infect exponentially growing phytoplankton cells in an 8-step series of 10-fold dilution steps from 1 to 10^{-7}. Aliquots (100 µl) of each dilution were added to 3 wells of a 24-well assay plate, each well containing 200 µl of exponentially growing *O. tauri* RCC 393 culture. Cell lysis was recorded as the appearance of a virus group and a decline in cell numbers on a FACScan analytical flow cytometer (Becton Dickinson, Oxford, UK) equipped with a 15 mW laser exciting at 488 nm and with a standard filter set up. Phytoplankton abundance estimates were analysed at high flow rate (~70 µL min^{-1}) and were discriminated by differences in their forward or right angle light scatter (FALS, RALS) and chlorophyll fluorescence. Samples for viral abundance analysis were fixed with glutaraldehyde (0.5 % final concentration) for 30 min at 4°C, snap frozen in liquid nitrogen and stored at –80°C. Samples were subsequently defrosted at room temperature and diluted 500 fold with TE buffer (10 mmol L^{-1} Tris-HCL pH8, 1 mmol L^{-1} EDTA),

stained with SYBR Green 1 (Molecular Probes; Marie et al. 1999) at a final dilution of 5 x 10^{-5} the commercial stock, incubated at 80°C for 10 min in the dark, then allowed to cool for 5 min before flow cytometric analysis. Samples were analysed for 2 min at a flow rate of ~ 35 µL min^{-1} and virus groups were discriminated on the basis of their RALS versus green fluorescence. Data files were analysed using WinMDI 2.8 software (Joseph Trotter, [http://facs.scripps.edu]). Algal lysate from the most dilute step was filtered through a 0.2 µm PVDF filter and the procedure repeated a further two times. The clonal virus sample obtained was filtered and stored at 4°C in the dark.

DNA preparation and sequencing

For preparation of large quantities of viruses for genome sequencing, ten litre volumes of exponentially growing *O. tauri* culture were inoculated with 10 ml of OtV-2 at a multiplicity of infection (MOI) of one. Lysed cultures were passed sequentially through 0.8 µm and 0.2 µm filters to remove large cellular debris. Virus filtrates were concentrated by ultrafiltration to ~ 50 ml using a Quixstand benchtop system and hollow fibre cartridges with a 30,000 molecular weight cut-off (MWC) (GE Healthcare Amersham Biosciences). The 50 ml concentrate was then further concentrated to ~10 ml using the Mid-Gee benchtop system and hollow fibre cartridge with a 30,000 MWC (GE Healthcare Amersham Biosciences). Aliquots (3 ml) of the concentrated OtV-2 lysate were adjusted with CsCl to densities of 1.1, 1.2, 1.3 and 1.4, and gradients from 1.1 to 1.4 were formed by ultracentrifugation at 100,000 × *g* at 22°C for 2 hours in a SW40-Ti Beckman rotor. Virus bands were removed with a syringe and dialysed against 4 × 1 litre volumes of filtered sterile seawater.

Sequence assembly and finishing

Complete sequencing of the OtV-2 genome was performed at the NERC Biomolecular Analysis Facility based at the University of Liverpool, UK, using a GS-FLX 454 genome sequencer. The resulting reads were then *de novo* assembled with the 454's own Newbler assembler software, version 1.1.03.24. Resulting contigs were next screened according to coverage depth to filter out the low coverage algal host contamination and bacterial contamination from the viral DNA. The remaining contigs were subsequently ordered and oriented using *Ostreococcus virus* OsV5 complete genome (NC_010191) as a reference sequence, using MUMmer 3.2. The OtV-2 genome was sequenced to an average of 10-fold coverage. Primers were designed to fill gaps between contigs, and the resultant Sanger sequence data was merged with the 454-generated contigs to form the completed genome sequence. Putative open reading frames (ORFs) were then identified, *de novo*, using glimmer3 using the provided g3-iterated.csh script. These ORFs were subject to BLAST analysis (Altschul et al. 1997) against the reference sequence to identify those ORFs, which may have been split due to frameshift errors caused by the 454 sequencer. Where found, relevant putative coding sequence (CDS) regions were joined and annotated accordingly. Coding regions of less than 65 amino acids were excluded from subsequent analysis. All remaining coding regions were then preliminarily annotated after BLAST search against the reference to identify similarity. Putative tRNA genes were identified with the aid of tRNAscan-SE version 1.23. Annotation was then checked and supplemented manually within the Artemis software tool (release 11) (41).

Sequence annotation

Whole genome sequences of OtV-2 were analysed and annotated using the software program Artemis (release 11) (41) with CDSs being generated based on predicted ORFs, correlation scores for each potential reading frame, GC content and codon usage indices. A CDS was defined as a continuous stretch of DNA that translates into a polypeptide that is initiated by an ATG translation start codon and extends for 65 or more additional codons. Similarities of putative CDSs were detected using BLAST and any homologous sequences were recorded. CDSs were assigned putative functions and were colour coded based on their function. Each CDS was used in a search for homologues using the protein-protein BLAST (BLASTP) program (2). Comparative genomic analysis was conducted between the genomic data obtained in this study and similar virus genomes in the database using the software program Artemis Comparison Tool (ACT) (release 8) (9).

Nucleotide sequence accession number

The OtV-2 sequence has been deposited in the GenBank database (Accession No. FN600414).

Phylogenetic analysis

Phylogenetic analyses were performed, as described previously (50), with the amino acid sequences of the DNA polymerase gene and high affinity phosphate transporter genes obtained from the GenBank database. DNA polymerase amino acid sequences from a range of representative viruses were multiply aligned with the ClustalW program on the region spanning the highly conserved regions I and IV of the DNA *pol* genes. Entire genes of high affinity phosphate transporter genes were aligned.

Results and discussion

Virus isolation.

Ostreococcus tauri-specific virus isolate OtV-2 was isolated from surface seawater collected on 07/05/2007 at the L4 sampling station in the Western English Channel (co-ordinates 50º15'N, 04º13'W). Despite being isolated from surface seawater, host range analysis revealed it did not infect the high-light adapted *Ostereococcus* strains OTH 95, RCC 501, and RCC 356 (results not shown). The OtV-2 host, *O. tauri* strain RCC 393 belongs to clade B, based on analysis of the ITS region of the SSU ribosomal RNA operon (21, 39). RCC 393 is currently one of four characterised Clade B *Ostreococcus* strains, and was isolated at a depth of 90 m from the Mediterranean Sea (39). *O. tauri* strain RCC 393 displays a different pigment composition compared to high-light strains, with a higher chlorophyll *b*: chlorophyll *a* ratio and the additional presence of the chlorophyll pigment chl c_{cs170}, which is absent in high-light adapted strains of the species (39). This latter chlorophyll pigment most likely plays a light-harvesting function in the blue-green region, which dominates at low ambient light levels.

Description of the OtV-2 genome

Sequence analysis of the OtV-2 genome (accession number FN600414) revealed a linear genome of 184,409 bp. Features of the OtV-2 genome sequence include i) a nucleotide composition of 42.15% G+C ii) a total of 237 predicted coding sequences (CDSs) that were defined as having a start codon followed by at least 65 additional codons prior to a stop codon iii) CDSs equally distributed on both strands (53% on the positive strand and 47% on the negative strand) iv) an average gene length of 725 bp and v) a coding density of 1.285 genes per kbp (Table 1). The OtV-2 genome was

oriented with the genomes of both OtV-1 and OtV-5 and a high degree of colinearity was observed among all three genomes (data not shown).

Of the 237 CDSs identified in the OtV-2 genome, 165 (69.6%) encode putative proteins where no function can be attributed (Table 1). The remaining 72 CDSs (30.4%) with homology to known proteins were split into 9 functional groups (Table 2), based on their predicted metabolic functions. At least ten CDSs can be assigned putative functions involved in DNA replication, recombination and repair; 7 in nucleotide metabolism and transport; 8 in transcription; 15 in protein and lipid synthesis, modification and degradation; one in signalling; 5 in DNA methylation; 9 in sugar metabolism; 8 encoding capsid proteins; and a further 8 with other miscellaneous functions.

OtV-2 phylogeny

BLASTP searches identified a putative DNA polymerase (OtV-2_207). Phylogenetic analysis of highly conserved regions of the DNA polymerase of OtV-2 and other viral DNA polymerases showed close similarities with the DNA polymerases of other members of the *Phycodnaviridae* (Fig. 1). As the three currently described OtV viruses cluster with strong bootstrap support with the *M. pusilla* viruses and other phycodnaviruses, OtV-2 can be putatively assigned to the *Phycodnaviridae* family.

The OtV-2 low-light genotype

CDSs exclusive to OtV-2 are highlighted in Tables 2 and 3. Of particular note, OtV-2 is the only one of the three OtV viruses that encodes a cytochrome b_5 (OtV-2_201), RNA polymerase sigma factor (OtV-2_202) and a high affinity phosphate transporter (OtV-2_222) (Table 2) (discussed in detail below). These three CDSs share

homology with proteins encoded in the host genus, *Ostreococcus* (Table 3) (15, 32). A notable feature of most of these exclusive OtV-2 host-like genes is their position in the OtV-2 genome. They occur predominantly in the final quarter of the genome between 147,332 bp (start of OtV-2_183) to 180,613 bp (end of OtV-2_229). Most intriguingly is the clustering of the genes encoding a putative cytochrome b_5 (OtV-2_201) and a putative RNA polymerase sigma factor (OtV-2_202), which are adjacent to one another. Moreover, in this same region of the OtV-2 genome four genes encoding predicted proteins with no known function were identified as having high similarity to proteins encoded within the *Ostreococcus* host. This clustering of islands of host-like genes raises the possibility of a 'hotspot' region within the OtV-2 genome, with a propensity to acquire genes from the host. This may be driven by environmental selection pressure such as growth/propagation in low-light environments. Several host-like genes have been detected localised to a genomic region of a number of cyanomyovirus genomes, indicating this hyperplastic region may be site-specifically associated with the acquisition of host-like genes (29). Accretion of host genes is believed to play a role in the evolution of viruses (24). The evolution of bacteriophages is believed to be driven in part by the acquisition of more and more foreign DNA, the so-called 'moron' hypothesis (23). The existence of several host-like genes in the OtV-2 genome provides strong evidence of HGT events and functional analysis of the genes involved should enable a greater insight into this close interaction.

Novel functionality of the OtV-2 genome.

Arguably the most environmentally relevant acquisition of a host gene by OtV-2 is the presence of a putative high affinity phosphate transporter gene (OtV-2_222). This

is in addition to a gene encoding a putative phosphate starvation inducible protein of the PhoH protein family found in all 3 *Ostreococcus tauri* viruses (OtV-2_024). A database BLASTP search for other viral proteins with homology to the high affinity phosphate transporter (OtV-2_222) in OtV-2 gave only one other match to a putative phosphate-repressible phosphate permease (PPRPP) in the coccolithovirus, EhV-86. The high affinity phosphate transporter gene reported in OtV-2 encodes a PHO4 family protein. With an amino acid identity of 57% to the host protein, it is likely the virus has acquired this gene from its *Ostreococcus* host. This gene has been identified and characterised in other phytoplankton species, such as *Tetraselmis chui* (12). Phylogenetic analysis of the OtV-2 high affinity phosphate transporter (Fig. 2) shows the viral protein forms a distinct cluster with the *Ostreococcus* host proteins (bootstrap value of 100% for 1,000 trials). These results suggest the gene encoding this putative high affinity phosphate transporter has been acquired by the virus through horizontal transfer from the host. Indeed, a ClustalW alignment of the putative high affinity phosphate transporter in the virus and the equivalent proteins in the host species' *O. tauri* and *O. lucimarinus* shows a relatively high degree of conservation between the three proteins (Fig. 3). However, the virus version is missing 54 amino acid residues in the centre of the protein (a result of a 162 bp in-frame deletion in the middle area of the gene) as well as several residues at the N-terminus found in both host versions.

A putative cytochrome b_5 gene (OtV-2_201) was identified in the OtV-2 genome, which shares homology with a gene encoded by the host *O. tauri.* Cytochrome b_5 is a small hemoprotein and a ubiquitous electron transport carrier found in animals, yeasts and plants. The role cytochrome b_5 plays in the synthesis of unsaturated fatty acids

has been well-characterised in animals and plants (45). Cytochrome b_5 consists of two domains, namely a hydrophobic tail that anchors the protein to the membrane and a hydrophilic portion, the heme-binding domain, which is active in redox reactions (48).

BLASTP analysis of the OtV-2 protein against public databases resulted in a best match (an amino acid identity of 60% and an E-value of $2e^{-29}$) to a cytochrome b_5 in *O. tauri.* An alignment of the putative cytochrome b_5 in OtV-2 and the *O. tauri* homologue shows the virus protein aligns with the C-terminal region of the host protein. Across most of this alignment, there is a high level of conservation between the two proteins. However, much of the remainder of the host protein is missing in the viral version (Fig. 4). The core of the heme-binding domain is formed by the characteristic cytochrome b_5 motif His-Pro-Gly-Gly (HPGG) (45). This diagnostic cytochrome b_5 motif 2 of the heme-binding domain was identified both in the OtV-2 cytochrome b_5, at residues 36-39 of the 91 aa polypeptide, and the *O. tauri* host cytochrome b_5 at residues 545-548 of the 586 aa polypeptide. From the evidence presented, there is a possibility the gene encoding this cytochrome b_5 in OtV-2 was acquired from the algal host, particularly as the adjacent gene encodes a host-like RNA polymerase sigma factor. A putative cytochrome b_5 gene has also been identified in *Acanthamoeba polyphaga* mimivirus (36). OtV-2 is, therefore, only the second virus reported to encode a putative cytochrome b_5.

Nucleotide transport and metabolism

The OtV-2 genome encodes a putative deoxycytidylate deaminase (dCD) (OtV2_190), not found in OtV-1 or OtV5. This enzyme converts deoxycytidine 5'-monophosphate (dCMP) to deoxyuridine 5'-monophosphate (dUMP) (28) and is a

major supplier of the substrate for thymidylate synthase, an important enzyme in DNA synthesis (8). This is of significance as the OtV-2 virus encodes a thymidylate synthase, ThyX (OtV2_051), also found in OtV-1 and OtV5. Both enzymes have been shown to be simultaneously elevated in rapidly dividing cells and have minimal activity in non-dividing cells (28). The enzyme dCD is present in most eukaryotes and bacteria, with those present in humans and T4-bacteriophage being the most extensively studied. Viruses known to encode a dCD include certain bacteriophage, e.g. T4, the chloroviruses and the Mimivirus (18, 19, 36). Members of the cytidine deaminase superfamily of enzymes have been investigated extensively in eukaryotes, as they play a role in antibody production in the immune system. As dCD is a major provider of dUMP, and thymidylate synthase is the only *de novo* source of dTMP in most biological systems, these enzymes have also become potential targets for anticancer therapy (40).

Transcription

The OtV-2 genome encodes eight CDSs involved in transcription (Table 2). A putative RNA polymerase sigma factor (OtV-2_202) has been identified in the OtV-2 genome, which is not found in OtV-1 or OtV5. Genes encoding RNA polymerase sigma factors have previously only been reported in bacteriophages and this is believed to be the first finding of such a gene in a virus infecting eukaryotes. These factors assist the polymerase in binding to the promoter selectively and initiating transcription (6). The initiation of transcription from promoter elements is triggered by the reversible association of sigma factors to the complex to form a holoenzyme. The sigma-70 (σ^{70}) factors are known as the primary or major sigma factors and can initiate the transcription of a wide variety of genes. Members of the σ^{70} family are

components of the RNA polymerase holoenzyme that direct bacterial or plastid core RNA polymerase to specific promoter elements that are situated 10 and 35 base-pairs upstream of transcription-initiation points (31). The primary family σ factor, which is essential for general transcription in exponentially growing cells, is reversibly associated with RNA polymerase. The σ^{70} family members have four conserved regions, the highest conservation being found in regions 2 and 4, which are involved in binding to RNA polymerase recognising promoters and separating DNA strands (DNA 'melting'). Bacteriophages, such as T4, encode a σ^{70} factor (47).

BLASTP analysis revealed the protein product of CDS OtV-2_202 shared homology (amino acid identity of 42% and an E-value of $1e^{-09}$) with a RNA polymerase σ^{70} factor encoded in the host *O. tauri* genome. All members of the σ^{70} factor superfamily contain region 2, the most conserved domain of this protein. Region 2 of this protein is highly conserved between organisms as it contains both the -10 promoter recognition helix and the primary core RNA *pol*-binding determinant. Analysis of the putative OtV-1_202 protein using the Pfam database indicated this protein contains a recognisable region 2 (Fig. 5) as does the host protein, thus confirming this as a putative σ^{70} factor. An alignment of this conserved domain of host and virus proteins and those of closest homologues was performed (Fig. 5). The virus gene has deletions at both ends in comparison to the host gene, indicating the virus has undergone significant changes in its sequence, and possibly function, over time. Moreover, the virus gene has a GC content of 37.95% but the host gene has a GC content of 60%. Therefore, if this gene was acquired by the virus from the host it is unlikely to have been a recent event. An alternative source of this gene is from a bacterium. Aside from the closest BLASTP hit to a small region of the host protein,

all closest BLASTP hits are with equivalent proteins from bacterial species, e.g. *Rickettsia* and *Synechococcus.* The acquisition of genes by phycodnaviruses from a bacterium has been reported previously (16, 17). A more convoluted possibility for the transfer of this gene to the virus may have been from a bacterial source to the host, followed by a second HGT event from host to virus. Both BLASTP and Pfam searches of the host protein gave closest hits to similar proteins in cyanobacterial species, with approximate E-values of $3e^{-31}$ and amino acid identities of 33%. The only close hit to a similar protein in a eukaryote species is to one in *Micromonas pusilla,* which is also a prasinophyte species. This finding suggests an alternative hypothesis to all those previously outlined. The Prasinophyceae diverged early at the base of the Chlorophyta (5). As members of the Prasinophyceae, both *Micromonas* and particularly *Ostreococcus,* with its small cell size (less than 1 μm), lack of flagella and simple cellular organisation hold key phylogenetic positions in the eukaryotic tree of life. As a primitive species, *Ostreococcus* may have acquired this gene from a bacterial evolutionary ancestor. The absence of this gene from the OtV-1 and OtV5 genomes indicates only OtV-2 has undergone a further HGT event that has not occurred in the two viruses infecting the high-light host strain. The functionality of the gene encoding a putative RNA polymerase sigma factor is unconfirmed at present but such features of the OtV-2 genome may indicate past HGT events between the virus and bacterial/phage sources.

Protein and lipid synthesis/modification

The OtV-2 genome encodes a putative procollagen-lysine 2-oxoglutarate 5-dioxygenase (PLOD) (OtV-2_158), an enzyme which exists in eukaryotes but has not been reported in prokaryotes or viruses, with the exception of the Mimivirus (36).

This enzyme has been extensively characterised in animals, where its role is to hydroxylate lysine residues in collagens. The resulting hydroxylysines serve as attachment sites for carbohydrate units and are essential for the stability of intermolecular collagen crosslinks (26). The putative PLOD protein in OtV-2 shares approximately 30% amino acid identity with a similar protein in several eukaryotic organisms. The functional presence of this gene in the OtV-2 genome may be to ensure the stability of carbohydrate units in structures such as the capsid protein. Without functional analysis, this hypothesis remains speculative.

A putative 3-methyl-2-oxobutanoate hydroxymethyltransferase is encoded by OtV-2 (OtV-2_029). This enzyme is the first in the pantothenate biosynthesis pathway and catalyses the committant step in synthesis of pantothenate, or vitamin B5 (46). Pantothenate is a necessary precursor to coenzyme A and phosphopantetheine, the prosthetic group of the acyl carrier protein, both of which are vital to a multitude of metabolic processes. Coenzyme A assists in transferring fatty acids from the cytoplasm to mitochondria (30). Pantenthoate is synthesized by microorganisms, i.e. bacteria and fungi, and plants but not animals, which require it as part of their diet. This biosynthesis pathway has been well studied in bacteria such as *E. coli* (27).

The putative 3-methyl-2-oxobutanoate hydroxymethyltransferase CDS in the OtV-2 genome has closest similarity to a gene encoded by several bacterial species (amino acid identity of approximately 45% and E-value of e^{-60}). Although it is not known if this gene is functional in OtV-2, its protein product is predicted to exhibit a Rossmann fold (35), which is highly conserved in these enzymes, and the active site domains are also conserved.

The OtV-2 genome encodes three putative 2OG-Fe(II) oxygenases (OtV-2_165, OtV-2_198, OtV-2_200) and a putative prolyl 4-hydroxylase (OtV-2_106). These enzymes are related, as their family contains members of the 2-oxoglutarate and Fe-dependent oxygenase superfamily and includes the C-terminus of the prolyl 4-hydroxylase alpha subunit. The enzyme 2OG-Fe(II) oxygenase belongs to a class of enzymes that are widespread in eukaryotes and bacteria and catalyze a variety of reactions typically involving the oxidation of an organic substrate using a dioxygen molecule (34). An extensively characterised reaction involving this enzyme is the hydroxylation of proline and lysine side chains in collagen (3). The presence of a putative prolyl-4-hydroxylase, three putative 2OG-Fe(II) oxygenases and a procollagen-lysine,2-oxoglutarate 5-dioxygenase in the OtV-2 genome indicates this virus may encode for the stabilisation of complex structures, such as carbohydrate units, during its replication. Both the OtV-1 and OtV5 genomes encode only one putative 2OG-Fe(II) oxygenase each. Therefore, this begs the question why a virus infecting a low-light strain would require three of these genes? Perhaps the assembly and stabilisation requirements of the virus capsid necessitate the involvement of these enzymes.

Of note, OtV-2 does not encode two enzymes involved in lipid metabolism, namely a patatin-like phospholipase and an oxo-acyl carrier dehydrogenase or a 3-dehydroquinate synthase, which are found encoded on the viruses infecting the high-light host strain. This may be due to the OtV-2 virus utilising host genes involved in these metabolic processes, thus negating the requirement for these genes in the virus genome. Alternatively, the infection process of OtV-2 may differ somewhat from the other OtV viruses and thus exclude the need for genes involved in lipid synthesis.

Sugar manipulation enzymes

Several proteins encoded within the OtV-2 genome have close identity to enzymes involved in sugar manipulation, polysaccharide synthesis and the transfer of sugars to proteins. Some viruses can affect expression of host glycosyltransferases and a few express their own glycosyltransferases. The OtV-2 genome encodes at least four glycosyltransferases: two from family 1, and one each from family 2 and 25 (Table 2). This is two fewer than those encoded by the OtV-1 genome, which also encodes an alpha galactosyltransferase not encoded by OtV5 or OtV-2. All three genomes encode a dTDP-d-glucose 4,6-dehydratase. The OtV-2 genome also contains the *gnd* gene encoding a putative 6-phosphogluconate dehydrogenase (OtV-2_167), not found in viruses infecting the high-light *O. tauri* strain. This enzyme catalyses the decarboxylating reduction of 6-phosphogluconate into ribulose 5-phosphate, in the presence of NADP (52). This reaction is part of the hexose mono-phosphate shunt and pentose phosphate pathways. Virus DNA can be synthesised from the intermediates of the reductive pentose phosphate pathway during photosynthesis, from the intermediates of the oxidative pentose phosphate pathway, and also from nucleotide precursors degraded from host DNA (43). The oxidative pentose phosphate pathway metabolises glucose-6-phosphate to ribose-5-phosphate, which is necessary for the *de novo* biosynthesis of purine and pyrimidine nucleotides of viral DNA (42). Prokaryotic and eukaryotic 6-phosphogluconate dehydrogenases are proteins of approximately 470 amino acids with highly conserved sequences. The OtV-2 protein has approximately 30% amino acid identity and E-values of approximately $2e^{-50}$ with several bacterial species. It is possible the virus acquired the *gnd* gene from a bacterial source. The only other virus reported to encode this

enzyme is cyanophage syn9 (49) and this is therefore only the second report of this enzyme in a virus. Presumably, the majority of viruses, which do not contain the *gnd* gene, hijack the host's biosynthetic pathway for nucleotide synthesis during infection. The OtV-2 genome may possess this gene due to the niche its host occupies thus making a viral-encoded version a beneficial acquisition.

DNA methylation

The OtV-2 genome encodes a number of putative methyltransferases (Table 2). DNA methyltransferases are rare among viruses but are reported most commonly in bacteriophages, and also in the chloroviruses (1) and phaeoviruses (13, 33). Of particular note is the presence of a putative 6-adenine methyltransferase encoded on the OtV-2 genome. DNA adenine methyltransferases (DAM) genes are found in the genomes of many fungi, bacteria, protists and archaea (10). However, the methylation of this base is a rare modification in eukaryotes (37). Restriction/modification systems in prokaryotes help protect cells against invading phages and plasmids (4). The presence of genes encoding methyltransferases in the OtV-2 genome may indicate coevolution enabling protection of viral DNA from the host restriction/modification system during infection.

Miscellaneous

OtV-2 has several genes encoding other putative proteins, including a rhodanese domain-containing protein (OtV-2_203) and an ABC domain protein (OtV-2_052). The product of a large CDS (OtV-2_109) resembles a virus-like inclusion body protein encoded by the protist *Trichomonas vaginalis.* The OtV-2 CDS is 2037 amino acids in length with a G+C content of 41.75%. This gene in OtV-2 aligns with three

consecutive CDSs in the reference genome, OtV-5 and a single large CDS (3398 amino acids long and G+C content of 43.76%) in the OtV-1 genome. Re-sequencing of the large gene in OtV-1 and OtV-2 confirmed these genes are single CDSs. No conserved domains were identified within these CDSs and further functional work is required.

A putative tail assembly fibre protein (OtV-2_065) is encoded by OtV-2 but not seen in the other OtV genomes. A PSI-BLAST of this protein against the public databases resulted in close hits to bacteriophage tail fibre assembly proteins with an average amino acid identity of 30% and an E-value of $2e^{-32}$.

Ostreococcus tauri virus isolate OtV-2 is a virus that specifically infects a low-light adapted *O. tauri* strain (RCC 393). Our original working hypothesis was that different viruses infecting high-light vs low-light adapted *O. tauri* strains would provide clues to propagation strategies that would give them selective advantages within their particular light niche. It is clear that OtV-2 has acquired a range of metabolically diverse genes from its host, other eukaryotes and even bacteria, over evolutionary time. Such piecemeal accretion of genes is a trademark of large double-stranded DNA viruses that has allowed them to adapt their propagation strategies in order to respond to selection pressures, including host niche separation in the surface layers of the oceanic environment. As more viruses are isolated and their genomes sequenced, we will start to see a clearer picture of what core genes are necessary to function in a particular oceanic niche. This study essentially represents a starting point in defining this low-light core virus genotype. In addition, it may help researchers establish a function for many of the unknown genes exclusive to a low-light core

genotype (or indeed any other niche genotype). Crucially, collection of environmental metadata will become a necessary component to help establish functionality of genes of organisms isolated from a particular niche.

Acknowledgements

This research was supported by a standard PhD studentship (ref. NER/S/A/2005/13204) awarded to WHW and DJS, small projects grant (MGF196) awarded to MJA from the Natural Environment Research Council (NERC) and National Science Foundation grant (EF0949162) awarded to WHW. We would like to acknowledge technical help from Margaret Hughes, Kevin Ashelford and Neil Hall at the NERC Biomolecular Analysis Facility at the University of Liverpool.

References

1. **Agarkova, I. V., D. D. Dunigan, and J. L. Van Etten.** 2006. Virion-Associated Restriction Endonucleases of Chloroviruses, p. 8114-8123, vol. 80.
2. **Altschul, S. F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.** 1990. Basic local alignment search tool. Journal of Molecular Biology **215:**403-410.
3. **Aravind, L., and E. V. Koonin.** 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol **2:**RESEARCH0007.
4. **Bertani, G., and J. J. Weigle.** 1953. Host Controlled Variation in Bacterial Viruses, p. 113-121, vol. 65.
5. **Bhattacharya, D., and L. Medlin.** 1998. Algal phylogeny and the origin of land plants. Plant Physiology **116:**9-15.
6. **Burgess, R., A. Travers, J. Dunn, and E. Bautz.** 1969. Factor stimulating transcription by RNA polymerase. Nature **221:**43 - 46.
7. **Cardol, P., B. Bailleul, F. Rappaport, E. Derelle, D. Beal, C. Breyton, S. Bailey, F. A. Wollman, A. Grossman, H. Moreau, and G. Finazzi.** 2008. An original adaptation of photosynthesis in the marine green alga *Ostreococcus*. Proceedings of the National Academy of Sciences of the United States of America **105:**7881-7886.
8. **Carreras, C. W., and D. V. Santi.** 1995. The catalytic mechanism and structure of thymidylate synthase. Annual Review of Biochemistry **64:**721-762.
9. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:**3422-3423.

10. **Cheng, X. D.** 1995. Structure and Function of DNA Methyltransferases. Annual Review of Biophysics and Biomolecular Structure **24:**293-318.

11. **Chretiennotdinet, M. J., C. Courties, A. Vaquer, J. Neveux, H. Claustre, J. Lautier, and M. C. Machado.** 1995. A new marine picoeukaryote - *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). Phycologia **34:**285-292.

12. **Chung, C.-C., S.-P. L. Hwang, and J. Chang.** 2003. Identification of a high affinity phosphate transporter gene in a Prasinophyte alga, *Tetraselmis chui*, and its expression under nutrient limitation. Applied Environmental Microbiology **69:**754-759.

13. **Delaroque, N., D. G. Muller, G. Bothe, T. Pohl, R. Knippers, and W. Boland.** 2001. The complete DNA sequence of the Ectocarpus siliculosus virus EsV-1 genome. Virology **287:**112-132.

14. **Derelle, E., C. Ferraz, M.-L. Escande, S. EycheniÃ©, R. Cooke, G. l. Piganeau, Y. Desdevises, L. Bellec, H. Moreau, and N. Grimsley.** 2008. Life-Cycle and Genome of OtV5, a Large DNA Virus of the Pelagic Marine Unicellular Green Alga *Ostreococcus tauri*. PLoS ONE **3:**e2250.

15. **Derelle, E., C. Ferraz, S. Rombauts, P. Rouze, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynie, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piegu, S. G. Ball, J. P. Ral, F. Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, and H. Moreau.** 2006. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proceedings of the National Academy of Sciences of the United States of America **103:**11647-11652.

16. **Filee, J., N. Pouget, and M. Chandler.** 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evolutionary Biology **8:**320.

17. **Filee, J., P. Siguier, and M. Chandler.** 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends in Genetics **23:**10-15.

18. **Fitzgerald, L. A., M. V. Graves, X. Li, T. Feldblyum, J. Hartigan, and J. L. Van Etten.** 2007. Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect *Chlorella* Pbi. Virology **358:**459-471.

19. **Fitzgerald, L. A., M. V. Graves, X. Li, J. Hartigan, A. J. P. Pfitzner, E. Hoffart, and J. L. Van Etten.** 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic chlorella strain of the heliozoon *Acanthocystis turfacea*. Virology **362:**350-361.

20. **Fuhrman, J. A.** 1999. Marine viruses and their biogeochemical and ecological effects. Nature **399:**541-548.

21. **Guillou, L., W. Eikrem, M. J. Chretiennot-Dinet, F. Le Gall, R. Massana, K. Romari, C. Pedros-Alio, and D. Vaulot.** 2004. Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. Protist **155:**193-214.

22. **Henderson, G. P., L. Gan, and G. J. Jensen.** 2007. 3-D ultrastructure of *O. tauri*: electron cryotomography of an entire eukaryotic cell. PLoS ONE **2:**e749.

23. **Hendrix, R. W., J. G. Lawrence, G. F. Hatfull, and S. Casjens.** 2000. The origins and ongoing evolution of viruses. Trends in Microbiology **8:**504-508.

24. **Iyer, L. A., S. Balaji, E. V. Koonin, and L. Aravind.** 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Research **117:**156-184.

25. **Keller, M. D., R. C. Selvin, W. Claus, and R. R. L. Guillard.** 1987. Media for the culture of marine ultraphytoplankton. Journal of Phycology **23:**633-638.

26. **Kivirikko, K. I., R. Myllyla, and T. Pihlajaniemi.** 1992. Hydroxylation of proline and lysine residues in collagens and other animal and plant proteins. , p. 1-51. *In* J. J. Harding and M. J. C. Crabbe (ed.), Focus on post-translational modifications of proteins. CRC Press, Boca Raton, FL.

27. **Lobley, C. M. C., F. Schmitzberger, M. L. Kilkenny, H. Whitney, H. H. Ottenhof, E. Chakauya, M. E. Webb, L. M. Birch, K. L. Tuck, C. Abell, A. G. Smith, and T. L. Blundell.** 2003. Structural insights into the evolution of the pantothenate-biosynthesis pathway. Biochemical Society Transactions **31:**563-571.

28. **Maley, F., and G. F. Maley.** 1999. Structure, function analysis of T4-phage dexoycytidylate deaminase and its role in the phage metabolic pathway. Paths Pyrimidines **7:**1-7.

29. **Millard, A. D., Zwirglmaier, K., Downey, M., Mann, N.H., Scanlan, D.J.** 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanism of cyanophage evolution. Environmental Microbiology **11:**2370-2387.

30. **Ottenhof, H. H., J. L. Ashurst, H. M. Whitney, S. A. Saldanha, F. Schmitzberger, H. S. Gweon, T. L. Blundell, C. Abell, and A. G. Smith.**

2004. Organisation of the pantothenate (vitamin B-5) biosynthesis pathway in higher plants. Plant Journal **37:**61-72.

31. **Paget, M., and J. Helmann.** 2003. The sigma70 family of sigma factors. Genome Biology **4:**203.

32. **Palenik, B., J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S. S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G. J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, and I. V. Grigoriev.** 2007. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. PNAS **104:**7705-7710.

33. **Park, Y., G. D. Kim, and T. J. Choi.** 2007. Molecular cloning and characterization of the DNA adenine methyltransferase gene in Feldmannia sp virus. Virus Genes **34:**177-183.

34. **Prescott, A. G.** 1993. A dilemma of dioxygenases: or where molecular biology and biochemistry fail to meet. Journal of Experimental Botany **44:**849-861.

35. **Rao, S., and M. G. Rossmann.** 1973. Comparison of super-secondary structures in proteins. Journal of Molecular Biology **76:**241-256.

36. **Raoult, D., S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, and J. M. Claverie.** 2004. The 1.2-megabase genome sequence of mimivirus. Science **306:**1344-1350.

37. **Ratel, D., J. L. Ravanat, F. Berger, and D. Wion.** 2006. N6-methyladenine: the other methylated base of DNA. Bioessays **28:**309-315.

38. **Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm.** 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature **424:**1042-1047.

39. **Rodriguez, F., E. Derelle, L. Guillou, F. Le Gall, D. Vaulot, and H. Moreau.** 2005. Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). Environmental Microbiology **7:**853-859.

40. **Rose, M. G., M. P. Farrell, and J. C. Schmitz.** 2002. Thymidylate synthase: a critical target for cancer chemotherapy. Clinical Colorectal Cancer **I:**220-229.

41. **Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944-945.

42. **Sindelar, L.** 1986. The content of ATP, ADP, AMP, PI, the activity of enzymes involved in the glycolytic pathway and some problems of its regulation and energy balance in tobacco plants infected with potato virus Y. Biologia Plantarum **28:**449-459.

43. **Sindelar, L., M. Sindelarova, and L. Burketova.** 1999. Changes in activity of glucose-6-phosphate and 6-phosphogluconate dehydrogenase isozymes upon potato virus Y infection in tobacco leaf tissues and protoplasts. Plant Physiology and Biochemistry **37:**195-201.

44. **Six, C., Z. V. Finkel, F. Rodriguez, D. Marie, F. Partensky, and D. A. Campbell.** 2008. Contrasting photoacclimation costs in ecotypes of the marine eukaryotic picoplankter *Ostreococcus*. Limnology and Oceanography **53:**255-265.

45. **Smith, M. A., Stobart, A.K., Shewry, P.R., Napier, J.A.** 1998. Cytochrome b_5 and polyunsaturated fatty acid biosynthesis. , p. 181-188. *In* J. A. N. P.R. Shewry, P.J. Davis (ed.), Engineering Crop Plants for Industrial End Uses. Portland Press, London.

46. **Teller, J. H., Powers, S.G., Snell, E.E.** 1976. Ketopantoate hydroxymethyltransferase. I. Purification and role in pantothenate biosynthesis. Journal of Biological Chemistry **251:**3780-3785.

47. **Travers, A. A.** 1969. Bacteriophage sigma factor for RNA polymerase. Nature **223:**1107-1110.

48. **Vergères, G., and L. Waskell.** 1995. Cytochrome b5, its functions, structure and membrane topology. Biochimie **77:**604-620.

49. **Weigele, P. R., W. H. Pope, M. L. Pedulla, J. M. Houtz, A. L. Smith, J. F. Conway, J. King, G. F. Hatfull, J. G. Lawrence, and R. W. Hendrix.** 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. Environmental Microbiology **9:**1675-1695.

50. **Weynberg, K. D., M. J. Allen, K. Ashelford, D. J. Scanlan, and W. H. Wilson.** 2009. From small hosts come big viruses: the complete genome of a second *Ostreococcus tauri* virus, OtV-1. Environmental Microbiology **11:**2821-2839.

51. **Wilhelm, S. W., and C. A. Suttle.** 1999. Viruses and Nutrient Cycles in the Sea - Viruses play critical roles in the structure and function of aquatic food webs. Bioscience **49:**781-788.

52. **Wood, T.** 1986. Distribution of the pentose phosphate pathway in living organisms. Cell Biochemical Functioning **4:**235-240.

Figure Legends

Figure 1. Neighbour-joining tree of DNA *pol* fragments. The designations and GenBank accession numbers of DNA polymerase sequences used for phylogenetic analysis were as follows (scientific name, with abbreviation in parentheses, followed by the database accession number [referring to the National Center for Biotechnology Information database]): Alcephaline herpesvirus, AAK00812; *Autographa californica* nuclear polyhedrosisbaculovirus (AcNPV), P18131; *Bombyx mori* nucleopolyhedral virus (BmNPV), NP047469; Chilo iridescent virus (CIV), AF303741; *Chrysochromulina brevifilum* virus PW1 (CbV- PW1), AAB49739; *Chrysochromulina ericina* virus (CeV01), ABU23716; *Ectocarpus siliculosus* virus (EsV-1), NP_077578; *Emiliania huxleyi* virus-86 (EhV-86), AJ890364; Equid herpesvirus, YP_053075.1; *Feldmania* species virus (FsV), AAB67116; *Feldmannia irregularis* virus-1 (FirrV-1), AY225133.1; Fowlpox virus (FowPV), NP_039057; *Heterosigma akashiwo* virus (HaV), AB194136.1; Human herpesvirus, CAA28464; *Lymantria dispar* virus (LdNPV), AAC70269; *Micromonas pusilla* virus SP1 (MpV-SP1), AAB66713; *Micromonas pusilla* virus SG1 (MpV-SG1), AAB49746; *Micromonas pusilla* virus PL1 (MpV-PL1), AAB49747; *Acanthamoeba polyphaga* mimivirus (Mimivirus), YP_142676; *Molluscum contagiosum* virus (MOCV), NP043990; *Ostreococcus tauri* virus 5 (OtV5), EU304328; *Ostreococcus tauri* virus 1 (OtV-1), FN386611; *Paramecium bursaria Chlorella* virus (PBCV-1), AAC00532; *Paramecium bursaria Chlorella* virus NY-2A (PBCV NY-2A), AAA88827; *Phaeocystis globosa* virus 1 (PgV), ABD65727; *Pyramimonas orientalis* virus (PoV01), ABU23717; *Sulfurisphaera ohwakuensis,* BAA 93703; *Vaccinia* virus (VACV), A24878; and *Ostreococcus tauri* virus 2, (OtV-2), FN600414. The scale bar indicates a distance of 0.2 fixed mutations per amino acid position.

Figure 2. Neighbour-joining tree of high affinity phosphate transporter sequences. The designations and GenBank accession numbers of high affinity phosphate transporter

sequences used for phylogenetic analysis were as follows (scientific name with abbreviation in parentheses, followed by database accession number [referring to the National Center for Biotechnology Information database]): *Alteromonas macleodii* (NC_011138.1); *Chlamydomonas reinhardtii* (XP_001695776.1); *Emiliania huxleyi* (AF334403.1); *Emiliania huxleyi* virus-86 (YP_293871.1); *Neurospora crassa* (XP_959489.1); *Ostreococcus lucimarinus* (XP_001416412.1); *Ostreococcus tauri* (CAL52329); *Saccharomyces cerevisiae* (CBK39373.1); *Tetraselmis chui* (AAO47330.1); *Thalassiosira pseudonana* (XP_002281698.1). The scale bar indicates a distance of 0.1 fixed mutations per amino acid position.

Figure 3. Amino acid alignment of high affinity phosphate transporter sequences of OtV-2, *O.tauri* and *O. lucimarinus.* Dots (.) represent consensus sequences. Dashes (-) represent gaps.

Figure 4. Alignment of the cytochrome b_5 proteins encoded by *Ostreococcus tauri* and the virus OtV-2. Conserved amino acids are identified by an asterisk underneath the corresponding residue. The conserved heme-binding domain is the shaded region.

Figure 5. Amino acid alignment of the highly conserved region 2 of putative RNA polymerase sigma factor gene in OtV-2 and closest matching homologues.

Table 1. Comparison of the general characteristics of the genomes of OtV-2 (this study), OtV-1 (Weynberg *et al*., 2009) and OtV-5 (Derelle *et al*., 2008).

	OtV-1	OtV-2	OtV-5
Genome size (bp)	191,761	184,409	185,373
% G + C	45.36	42.15	45.29
tRNA	4	5	5
Predicted CDSs (65 aa)	232	237	268
Mean gene size (bp/gene)	756	725	692
Coding density (genes/kb)	1.303	1.285	1.424
% known genes	31	30.4	22
% hypothetical genes	16.5	18.5	20
% genes with no known function	69	69.6	77

Table 2. Putative proteins encoded within the OtV-2 genome grouped by function. Shaded proteins indicate those exclusive to OtV-2 (i.e. not present in either OtV-1 or OtV5).

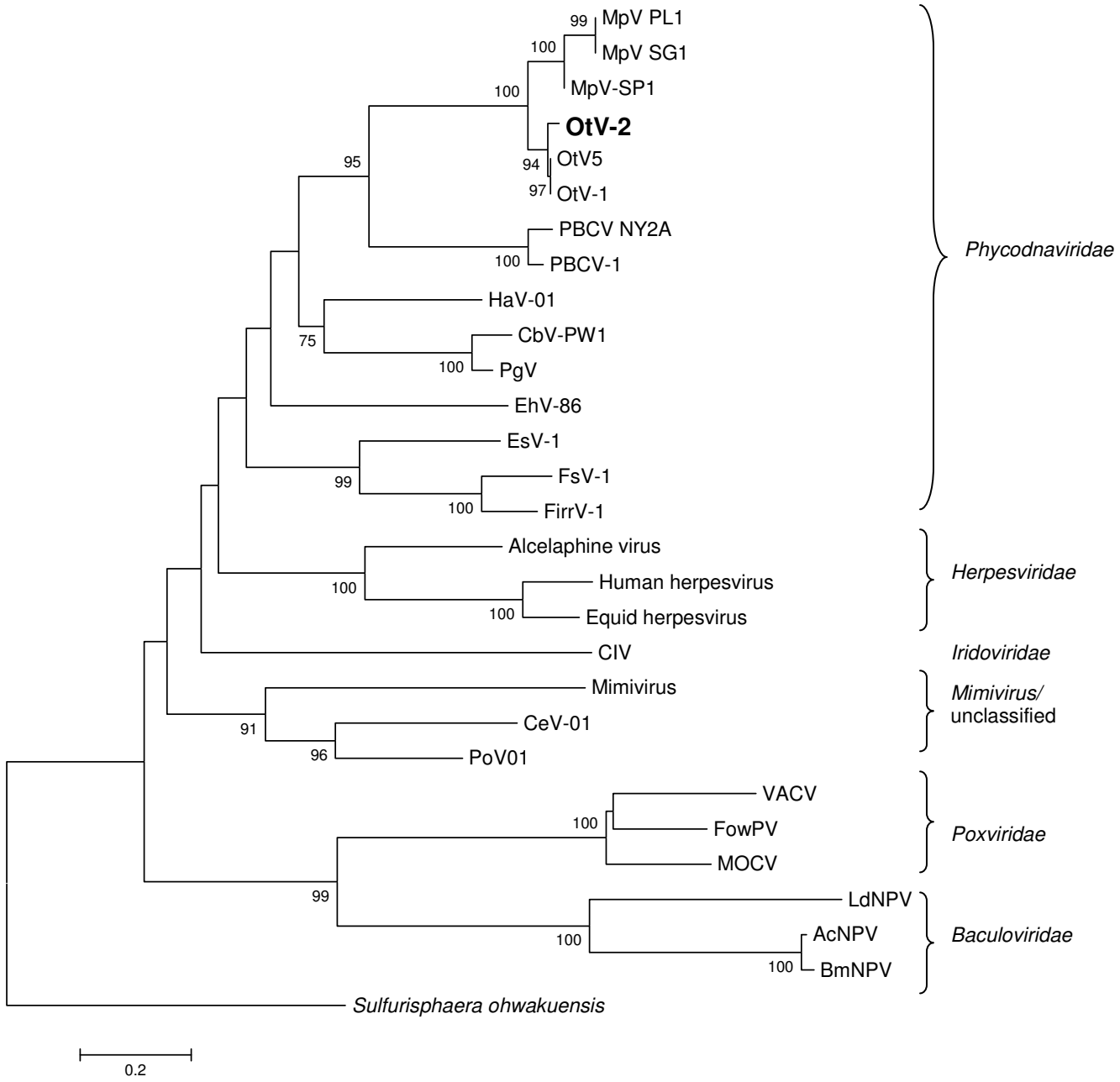
DNA, RNA, replication, recombination and repair	CDS#	**DNA restriction/methylation**	CDS#
DNA polymerase	OtV2_207	Adenine-specific methyltransferase	OtV2_083
ATP dependent DNA ligase	OtV2_169	DNA-cytosine methyltransferase [a]	OtV2_050
DNA topoisomerase I	OtV2_161	DNA-cytosine methyltransferase [a]	OtV2_003
DNA topoisomerase II	OtV2_211	Methyltransferase FkbM	OtV2_010
PCNA	OtV2_100	DNA methylase	OtV2_078
Exonuclease	OtV2_002	**Protein and lipid synthesis/modification**	
Exonuclease	OtV2_118	Prolyl 4-hydroxylase alpha-subunit	OtV2_106
RNase H	OtV2_156	2OG-Fe(II) oxygenase [a]	OtV2_165
VV A18 Helicase	OtV2_056	2OG-Fe(II) oxygenase [a]	OtV2_198
NTPase/helicase	OtV2_107	2OG-Fe(II) oxygenase [a]	OtV2_200
Nucleotide transport and metabolism		Procollagen-lysine 2-oxoglutarate 5-dioxygenase	OtV2_158
Ribonucleotide reductase, large subunit	OtV2_123	Ubiquitin C-terminal hydrolase	OtV2_146
Ribonucleotide reductase, small subunit	OtV2_140	ATP-dependent protease proteolytic subunit	OtV2_171
dUTP pyrophosphatase	OtV2_194	FtsH2 metalloprotease	OtV2_009
ATPase (VV A32 virion packaging)	OtV2_084	Aminotransferase family protein	OtV2_032
Thymidine kinase	OtV2_188	Acetolactate synthase	OtV2_030
CMP/dCMP deaminase, Zinc-binding	OtV2_190	N-myristoyltransferase	OtV2_076
FAD dependent thymidylate synthase	OtV2_051	TPR domain-containing protein	OtV2_027
Transcription		Asparagine synthetase	OtV2_022
RNA transcription factor TFIIB	OtV2_142	Aspartyl/Asparaginyl beta-hydroxylase	OtV2_166
RNA transcription factor TFIIS	OtV2_019	3-methyl-2-oxobutanoate hydroxymethyltransferase	OtV2_029
RNase III	OtV2_117	**Capsid protein**	
RNA polymerase sigma factor	OtV2_202	Capsid protein 1	OtV2_057
SW1/SNF helicase domain protein	OtV2_119	Capsid protein 2	OtV2_062
TATA-box binding protein	OtV2_130	Capsid protein 3	OtV2_066
mRNA capping enzyme [b]	OtV2_079	Capsid protein 4	OtV2_085
mRNA capping enzyme [b]	OtV2_145	Capsid protein 5	OtV2_086
Sugar manipulation enzymes		Capsid protein 6	OtV2_154
GDP-D-mannose dehydratase	OtV2_007	Capsid protein 7	OtV2_206
Glycosyltransferase family 1	OtV2_008	Capsid protein 8	OtV2_210
Glycosyltransferase family 2	OtV2_028	**Miscellaneous**	
Glycosyltransferase family 25	OtV2_163	Virus inclusion body protein	OtV2_109
Glycosyltransferase family 1	OtV2_216	Rhodanese domain-containing protein	OtV2_203
6-phosphofructokinase	OtV2_159	ABC1 domain protein	OtV2_052
NAD-dependent epimerase/dehydratase	OtV2_031	Fibronectin-binding protein	OtV2_199
dTDP-D-glucose 4,6-dehydratase	OtV2_033	Tail fibre assembly protein	OtV2_065
6-phosphogluconate dehydrogenase	OtV2_167	Cytochrome b_5	OtV2_201
Signalling		Phosphate starvation inducible protein	OtV2_024
Serine/Threonine protein kinase	OtV2_135	High affinity phosphate transporter	OtV2_222

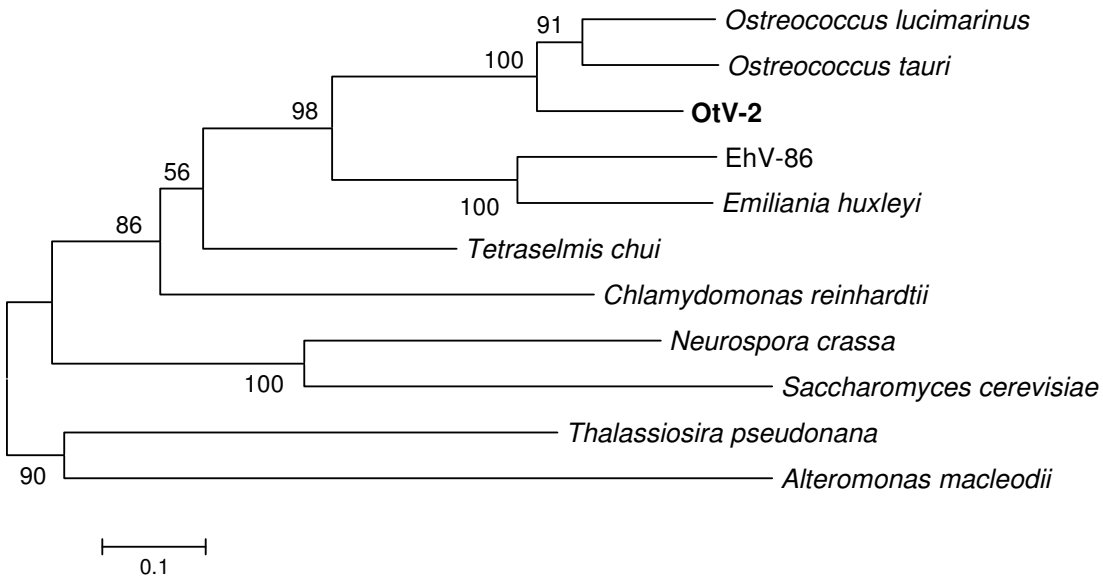
[a] Paralogous genes
[b] Related genes but with different functions

Table 3. Features of low-light (OtV-2) versus high-light (OtV-1 & OtV5) genotypes.

	Genome				
CDS Putative function	**OtV-1**	**OtV-2**	**OtV5**	**Host***	
Cytochrome b_5		✓		✓	
RNA polymerase sigma factor		✓		✓	
High affinity phosphate transporter		✓		✓	
DNA methylase		✓			**Low**
Procollagen-lysine2-oxoglutarate5					**Light**
dioxygenase		✓			
3-methyl-2-oxobutanoate					
hydroxymethyltransferase		✓			
Tail fibre assembly protein		✓			
CMP/dCMP deaminase, Zinc-binding		✓			
6-phosphogluconate dehydrogenase		✓			
ATP-dependent protease proteolytic subunit	✓				
33 kDa *in vitro* translation peptide	✓		✓		
3-dehydroquinate synthase	✓		✓		**High**
Echinonectin	✓				**Light**
Proline dehydrogenase	✓		✓		
Patatin phospholipase	✓		✓		
Oxo-acyl carrier protein dehydrogenase	✓		✓		

*Host *Ostreococcus tauri* RCC 745 (Derelle et al. 2006)


MpV PL1
MpV SG1
MpV-SP1
OtV-2
OtV5
OtV-1
PBCV NY2A
PBCV-1
HaV-01
CbV-PW1
PgV
EhV-86
EsV-1
FsV-1
FirrV-1
Alcelaphine virus
Human herpesvirus
Equid herpesvirus
CIV
Mimivirus
CeV-01
PoV01
VACV
FowPV
MOCV
LdNPV
AcNPV
BmNPV
Sulfurisphaera ohwakuensis
99
100
100
95
94
97
100
75
100
99
100
100
100
91
96
100
99
100
100
Phycodnaviridae
Herpesviridae
Iridoviridae
Mimivirus/
unclassified
Poxviridae
Baculoviridae
0.2


91
100
98
56
100
86
100
90
Ostreococcus lucimarinus
Ostreococcus tauri
OtV-2
EhV-86
Emiliania huxleyi
Tetraselmis chui
Chlamydomonas reinhardtii
Neurospora crassa
Saccharomyces cerevisiae
Thalassiosira pseudonana
Alteromonas macleodii
0.1

```
OtV-2           ----------------------------------------------------------------------------MLQQFEWIV
O.lucimarinus   MNFLPSLDSGLDWNQSRLLSRFTNREFVPQGIFSESCLFKFPRGNMRGRRQNTSRAVSSRTTMSSSDTLAQASRHLLAGATETKR.....
O.tauri         ------------MTDESRTAKYFSRKSRKKEYSLIFSEYFTFLRPGDADGVTSRRALPVSMSGSEALAHAGRALLMAGGEATKN.....
                                                                                                :.*****

OtV-2           VTGAIFAFLASFGIGANDVANAFATSVGSGALKIKQAVPLAAVCEFSGSLFMGSHVVKTIRKGISDQDCFVDD-----PGLLMYGCLCVI
O.lucimarinus   .C.SFL..F.A.....................T..N..V...IF..C.AM...G.........AN.K..AGTGGAND............
O.tauri         .I.GFL..F.A.....................S..S..V..G.F..C.AM...G...E.M...VT..K..NGKSGSHD.V..........
                * *.::**:*:**********************.**.** **.: **.*::***.***:*:***:::*.** .      * **********

OtV-2           LSVAMWLIVASYLEMPVSTTHSCIGGMIGMTMVARGSSCVTWIAKSDQFPYVKGVVAIIISWLLSPIISGLFASALFFVVRSLILRSEDS
O.lucimarinus   FA..I..VI..AF..A.......V.......L.....E..I.TK.A.E.........V.........V...A..FVF.VTL.T.VM...H.
O.tauri         FA.GV..V...A...........V..L....L....DD..V.SK.A.E.......A..VV..V...V...C.SF.M.LIL.T.VM...N.
                ::*.:**::** :**.*******:**:****:****..** *  *:*:*******.*:::**:***:*** *: .:*. :*:*::***.*

OtV-2           YKRIQYGFPVLVAGTFIINTFFIVYKGAKFLKLDDTPLWKACAISFGIGGGTGIFSYFFINPIIFN-----------------------
O.lucimarinus   .S.TAVA....L.C.L...I...........E.....VGT....A......C..IA...VT.Y.LKTTDELFEKQQLEKAERGSGKKAEE
O.tauri         FS.VS.V....FGA.L...V.......SPH.G.K...VGT...AA..V...A..L...L.S.Y.VRRTEELYAQEQAEQAERGSGKKVEE
                                      :.*    ****.. *:*** *******: .* *.***: .*** :**:*** **::**::.* *..

OtV-2           ----------------------------DEERRIEEFCHTDDSVQSIHDNAEVFDTKTEHSMRYLQILTACCDSFAHGANDVANSIGP
O.lucimarinus   KVVRQPREYPVGVFGAPRRMWYALQDHLESSLAHKA.DILDE.MA.LA..E...K..E...LC..........................
O.tauri         TVVRQPREYPRGVFGAPTRVWYAIQDHLETSLAVKASDLIEE.MA.NA......K..E...L.............A.............
                                        .   : .:: . * :* :**:*** ** *** .************:**************

OtV-2           FAAIYAIYKSGTVS--KNAEMGNDAYWILSLGAGGIVVGLATYGYKILEALGTKMAKLTPSRGICIELGAAAVIILGSRLGWPLSTTHCQ
O.lucimarinus   ..SMVVVF...K..--.E....D.S....G...A...C...L......H.....I................C...M.............
O.tauri         ..T.VLV..A.K..TK.KDP..D..........L.......L...R..H.....L................C.......M..........
                **::  ::*:*.**  *:  **:*:****.*** *** *** ***:**.*****:****************.***:***:**********

OtV-2           VGATVGVALFEGTGGVNWKLLYKTIAGWLLTLVVVGSTTAFLFAQGAYAPMVKYPSYVINA------------
O.lucimarinus   .........L..RK.I..FIIG..VF..II...I..FS..AF.........KS..C.ITGSC------------
O.tauri         ....M...ML..RK.I..FI.A..V...II......FS.SAF.....F....H..C.IAATAKDSISEALCTI
                          ****:***::**  *:** :: **: **::***:** :*: :*****:***  **.*:  :
```

```
O._tauri          MPDARESSTNDRVVTRRRRSRSIAMAEQVIVIGGGLAGLSAAHTVLEHGARVVLLDKCSFLGGNSTKATSGINGALTRTQARLNIPDSAD
OtV-2             ----------------------------------------------------------------------------------------

O._tauri          KFEDDIIKGAAGVGHTEAPAHTIPLAKVLAQGSGSSVDWLCEKFKLDLSLVAQLGGHSYPRTHRGKERFPGFTITYALMEGLEKVMEDSN
OtV-2             ----------------------------------------------------------------------------------------

O._tauri          GETARIITKAEAKRLLTDGSGTVIGVEYEKDGVLNQEYGPVVIATGGFGADYKPDSLLKKYRPDLQALPTTNGDHCTGDGIKMAMAVGAD
OtV-2             ----------------------------------------------------------------------------------------

O._tauri          TIDMTSVQVHPTGLVNPAEPDSKVKFLAAEALRGVGGILLDANGNRFADELGRRDYVSGEMNRNKGPFRLILNGKASTEIEWHCKHYVGR
OtV-2             ----------------------------------------------------------------------------------------

O._tauri          GIMKRHDSGAEVAKELGISPQKLADTFAKYNEAARTKNCPFGKKFFTNAPFEMNDFFHSAIVCTVVHYTMGGLAINTDSQIVGPRGPIPG
OtV-2             ----------------------------------------------------------------------------------------

O._tauri          LFGAGEVVGGIHGRNRLGGNSLLDCVVFGRVAGSAVSRHLMSTAIRALRSGQTTAMNRVANLNDKINPPAMSAAPAAASAASGGGSRALT
OtV-2             ------------------------------------------------------MNRIKTINDHINP------------------RDLS
                                                                          ***: .:**:***                  * *:

O._tauri          MDEINKHNTEGDLWVIIEGNVYDLTKFLPDHPGGKKAIMLFAGKDATEEFNMLHPPNVLKKYLSPDAKIGTVLG
OtV-2             LTEIAKHNTEEDCWVIIKDIVYDLTKFLPDHPGGKKAIILFAGKDATEEFDMLHPPNVLKKYLTPEVVLGPVKK
                  : ** ***** * ****:. ******************:***********:************:*:. :*.*
```

```
Micromonas pusilla              RGLLSVSDLAQEGCAGLAQAADKFDPSRGYKFSTYAYNW
Micromonas sp.                  RGLLSVSDLMQEGCAGLATAADKFDPQRGYRFSTYAFFW
Thalassiosira_pseudonana        ASSVQFQDACQQGIIGLTRATEKFDPELGFRFSTYAIWW
Synechococcus sp.               LGQLEFGDLVQEGSIGLNRAVEKFDPELGYKFSTYAYWW
Vibrio harveyi                  NRGLALLDLIEEGNLGLIRAVEKFDPERGFRFSTYATWW
Ostreococcus_tauri              RRSISRQDMIHEGVTGLVRASEKFDPALGCAFSTYAYNW
OtV-2                           RGIHTRKDINSVGLHGLVRAAQKFNPELGFKFSTYAYPW
                                       *    *  **  * :**:*  *  *****  *
```