Complete Genome Sequence and Lytic Phase Transcription Profile of a *Coccolithovirus*

William H. Wilson *et al.*

Science **309**, 1090 (2005);
DOI: 10.1126/science.1113109

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

25. K. J. Chin, T. Lukow, R. Conrad, *Appl. Environ. Microbiol.* **65**, 2341 (1999).
26. C. Erkel *et al.*, *FEMS Microbiol. Ecol.* **53**, 187 (2005).
27. T. Lueders, K. J. Chin, R. Conrad, M. Friedrich, *Environ. Microbiol.* **3**, 194 (2001).
28. S. Lehmann-Richter, R. Grosskopf, W. Liesack, P. Frenzel, R. Conrad, *Environ. Microbiol.* **1**, 159 (1999).
29. H. A. C. Denier van der Gon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12021 (2002).
30. D. A. Wardle *et al.*, *Science* **304**, 1629 (2004).
31. We thank P. Frenzel for help in the preparation of rice microcosms, M. Klose and P. Claus for laboratory technical assistance, T. Lueders and M. Friedrich for introduction of the SIP technique, and S. Kolb for assistance in phylogenetic analysis. Supported by the Deutsche Forschungsgemeinschaft (grant no. SFB 395). All sequences generated from this study have been deposited in the GenBank database under accession nos. AJ878935 to AJ879061.

Supporting Online Material
www.sciencemag.org/cgi/content/full/309/5737/1088/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

11 April 2005; accepted 6 July 2005
10.1126/science.1113435

Complete Genome Sequence and Lytic Phase Transcription Profile of a *Coccolithovirus*

William H. Wilson,[1]* Declan C. Schroeder,[2] Michael J. Allen,[1] Matthew T. G. Holden,[3] Julian Parkhill,[3] Bart G. Barrell,[3] Carol Churcher,[3] Nancy Hamlin,[3] Karen Mungall,[3] Halina Norbertczak,[3] Michael A. Quail,[3] Claire Price,[3] Ester Rabbinowitsch,[3] Danielle Walker,[3] Marie Craigon,[4] Douglas Roy,[4] Peter Ghazal[4]

The genus *Coccolithovirus* is a recently discovered group of viruses that infect the globally important marine calcifying microalga *Emiliania huxleyi*. Among the 472 predicted genes of the 407,339–base pair genome are a variety of unexpected genes, most notably those involved in biosynthesis of ceramide, a sphingolipid known to induce apoptosis. Uniquely for algal viruses, it also contains six RNA polymerase subunits and a novel promoter, suggesting this virus encodes its own transcription machinery. Microarray transcriptomic analysis reveals that 65% of the predicted virus-encoded genes are expressed during lytic infection of *E. huxleyi*.

Large icosahedral viruses from the family *Phycodnaviridae* infect marine or freshwater eukaryotic algae, and all contain dsDNA genomes ranging from 180 to 560 kbase pair (kbp) (*1*, *2*). Phycodnaviruses belong to a group of viruses that replicate, completely or partly, in the cytoplasm of eukaryotic cells and are termed nucleocytoplasmic large DNA viruses (NCLDVs) (*3*). Two other members of the *Phycodnaviridae* have been sequenced before this study: the 335,593–base pair (bp) genome of a virus that infects a marine filamentous brown alga, *Ectocarpus siliculosus* (EsV-1) (*4*), and the 330,744-bp genome of a virus that infects a unicellular chlorella-like green algal symbiont of the freshwater protozoa *Paramecium bursaria* (PBCV-1). These two viruses, which are from the same family, only have 33 genes in common (*1*).

[1]Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, PL1 3DH, UK. [2]Marine Biological Association, Citadel Hill, Plymouth, PL1 2PB, UK. [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]Scottish Centre for Genomic Technology and Informatics, Chancellor's Building, College of Medicine, University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4SB, UK.

*To whom correspondence should be addressed. E-mail: whw@pml.ac.uk

EhV-86, a lytic virus about 170 to 175 nm in diameter, was originally isolated by plaque assay from a seawater sample collected from a dying *Emiliania huxleyi* bloom in the English Channel (*5*). Phylogenetic analysis of its DNA polymerase gene places it in a new genus (*Coccolithovirus*) within the family *Phycodnaviridae* (*6*). The EhV-86 host, *Emiliania huxleyi* (Haptophyta), is a unicellular alga known for its elegant calcium carbonate scales, which it produces intracellularly and sequesters over its cell surface (*7*). It is perhaps best known for its immense coastal and open ocean blooms at temperate latitudes and is a key species for current studies on global biogeochemical cycles and climate modelling (*8*–*10*).

Sequence analysis (*11*) of EhV-86 revealed a circular genome with a length of 407,339 bp (Fig. 1) (*12*), making this the largest *Phycodnaviridae* genome sequenced to date. Other larger algal virus genomes are known to exist, such as the virus that infects *Pyramimonas orientalis*, a marine microalga, which has a genome of about 560 kbp (*13*). Recently, Mimivirus, the largest NCLDV [previously isolated from amoeba growing in the water of a cooling tower (*14*)] was sequenced, revealing a 1,181,404-bp genome (*15*). General features of the EhV-86 genome sequence include a nucleotide composition of 40.2% G+C, a total of 472 predicted genes (coding sequences, CDSs) (Fig. 1) with an average gene length of 786 bp, a coding density of 91%, five tRNA genes [encoding amino acids Leu (containing an intron), Ile, Gln, Asn, and Arg] and two further introns (ehv064 and ehv459). In addition, we identified the location and orientation of three distinct families of repeats (designated A, B, and C) throughout the genome (Fig. 1 and fig. S1). Family A repeats are noncoding but are found immediately upstream of 86 predicted CDSs and are possible promoter elements essential for transcription of associated CDSs. Family B repeats are characterized by GC-rich regions found in CDSs, which encode proline-rich domains. Family C repeats are AT-rich, noncoding, and characteristic of virus genome origins of replication (*16*).

Of the 472 CDSs, only 66 (14%) have been annotated with functional product predictions on the basis of sequence similarity or protein domain matches (Table 1). A large proportion of CDSs exhibited no similarity to proteins in the public databases; only 21% of the CDSs contain protein-protein basic local alignment search tool (BLASTP) results that matched with an *E* value lower than 0.01 (for details of sequence similarity and protein domain database matches, see table S2). Future analysis of marginal similarities may increase this proportion further. Most of the genes with functional predictions are at the "ends" of the genome; surprisingly, there are only 8 such genes within 230 CDSs in the 176-kbp central region between 156 kbp and 332 kbp. The role of this central region is unclear.

Microarray analysis was used to obtain a transcriptome profile during lytic infection and to confirm whether putative CDSs were transcriptionally active. The microarray was constructed from 75-mer oligonucleotides corresponding to the sense strand of each putative CDS on the EhV-86 genome. EhV-86 gene expression (RNA transcription) was determined by hybridization to the microarray with fluorescently labeled cDNAs prepared from an infected *E. huxleyi* culture 33 hours postinfection. This time was chosen because the majority of cells in the culture were infected at a wide range of infection stages (fig. S3); the majority of kinetic classes of transcripts should be expressed at this point, and we detected expression of 308 EhV-86 CDSs

(65%) (Fig. 1 and database S5) (*17*). Analysis of the expression data allowed us to confirm annotation of some overlapping CDSs. Analyses of the transcriptome revealed that many of the unknown CDSs were expressed and are therefore likely to be functional. Up to 35% of EhV-86 CDSs were either not expressed or were beyond the limit of detection at 33 hours postinfection. Interestingly, expression was not observed in a putative phosphate permease (ehv117), a protein predicted to be involved in high affinity phosphate transport (*18*, *19*). Because this study was conducted in nutrient excess, it is possible the permease is only expressed during phosphate-limiting conditions.

Analysis of EhV-86 homologs of known genes revealed several that have never been identified in a virus before. There are at least four genes involved in sphingolipid biosynthesis, encoding sterol desaturase (ehv031), serine palmitoyltransferase (ehv050), transmembrane fatty acid elongation protein (ehv077), lipid phosphate phosphatase (ehv079), and a further two genes encoding desaturases (ehv061 and ehv415). Transcriptomic analysis revealed that all these genes except ehv415 are expressed during infection (Fig. 1). Sphingolipids are membrane lipids present in all eukaryotes and some prokaryotes and also play a key role in several processes, particularly signal transduction (*20*). A sphingolipid biosynthesis pathway has not previously been discovered on a virus genome. Sphingolipid biosynthesis leads to the formation of ceramide (*21*), which is known to suppress cell growth and is an intracellular signal for apoptosis (*22*, *23*). Several viruses are known to induce apoptosis (*24*). Uniquely, EhV-86 appears to encode key components of the ceramide pathway in its genome; furthermore, it is known that there is a connection between protease activation and ceramide-induced apoptosis (*25*), and intriguingly EhV-86 contains eight proteases (including five serine proteases) throughout the genome (ehv021, ehv109, ehv133, ehv151, ehv160, ehv349, ehv361, and ehv447). All proteases except ehv361 are expressed during infection (Fig. 1). One theory is that this algal virus encodes a mechanism for inducing apoptosis as a strategy for killing the host cell and disseminating progeny virions during the infection cycle. Apoptosis, or programmed cell death, has been observed previously in marine phytoplankton (*26*), although usually as a response to nutrient or other physiological stressors (*27*, *28*) rather than virus infection.

Further analysis of the genome reveals that EhV-86 contains 25 of the core set of 40 to 50 conserved virus genes for NCLDVs (*3*) that encode some of the principal features of virion structure and genome replication and expression (Table 1 and table S2). Uniquely for the *Phycodnaviridae*, EhV-86 contains six RNA polymerase subunits (ehv064, ehv105, ehv108, ehv167, ehv399, and ehv434), all of which are expressed. The presence of a putative virus RNA polymerase holoenzyme together with the family A repeats, which we suggest are previously unknown promoters (Fig. 1), would indicate that EhV-86 encodes its own transcription machinery. Hence, expression of some EhV-86 transcripts may occur in the cytoplasm rather than the nucleus (*29*). *Phycodnaviridae* lack RNA polymerases; thus, further phylogenetic analysis may point to coccolithoviruses being reclassified into a distinct subfamily.

EhV-86 is the largest algal virus genome sequenced to date, and it expresses a range of virus genes, revealing functions more commonly observed in animal and plant cells. Viral ceramide production raises questions about cell death in phytoplankton and may have implications for new sources of lipids in the marine food chain. Ongoing definition of the EhV-86 genome, coupled with transcriptomic analysis, will reveal the function of expressed genes that have no homologs in current databases. As more giant viruses are discovered and their genomes sequenced, there is sure to be an explosion of exciting gene discoveries and novel functions.

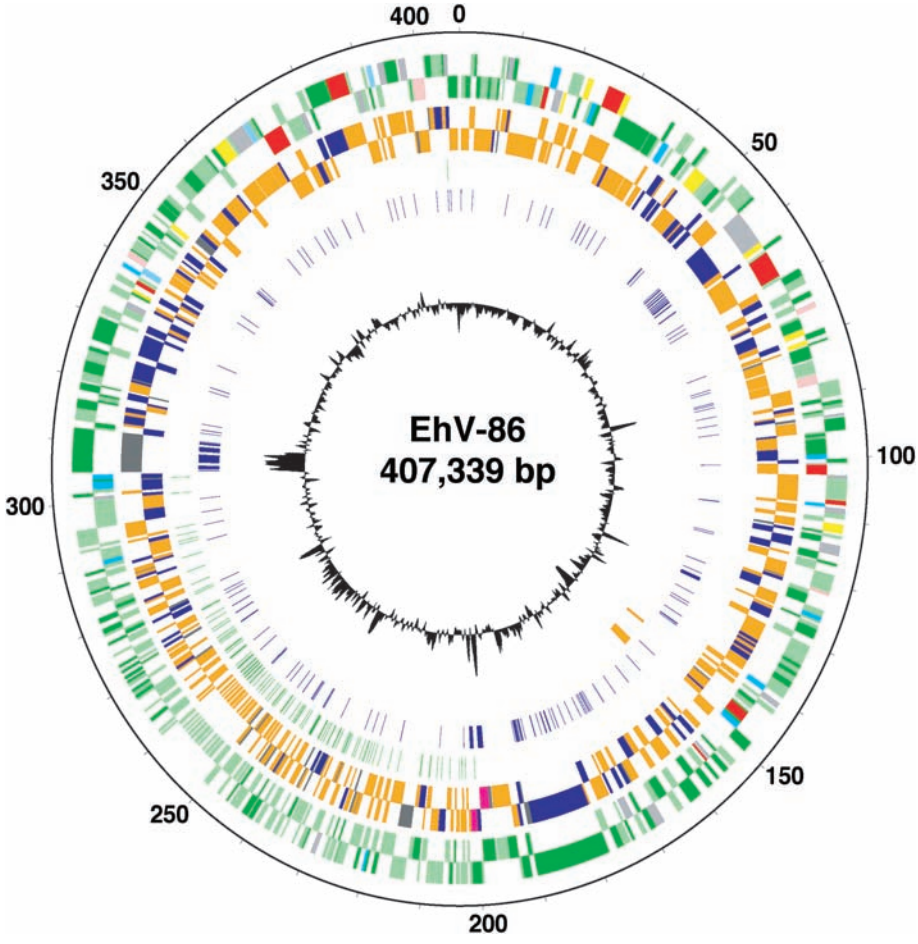

Fig. 1. Circular representation of the 407,339-bp EhV-86 genome. The outside scale is numbered clockwise in kbp. Circles 1 and 2 (from outside in) are CDSs (forward and reverse strands, respectively), starting with CDS ehv001 at position 276 bp (color-coded gray and just below the 0-kb marker on the reverse strand). CDSs are color-coded by putative function: light green, no known function; dark green, no known function but contains transmembrane helices; gray, miscellaneous; sky blue, degradation of large molecules; red, information transfer; yellow, central or intermediary metabolism; pink, virus specific; and light blue, kinases. Circles 3 and 4 are microarray expression results (forward and reverse strands, respectively) color-coded as follows: orange, expression observed; blue, no expression observed; and gray, not determined because oligonucleotides were not designed for that particular CDS. Circles 5 to 7 are the positions of the three families of repeats: green, putative promoter family A; blue, proline-rich family B; and orange, putative origin of replication. Circle 8, G+C content. Initial sequence analysis of EhV-86 suggested a linear genome, because the ends in the sequence alignment always stopped at the same bases. However, polymerase chain reaction amplification across the termini generated a product indicating the genome had a circular stage during virus replication (*30*). Sequence analysis of this product revealed either an A or T (at equal ratio) single base pair overhang at each of the termini, indicating a potential method for circularization of the genome. Consequently, bases T and A were annotated to the "start" of the genome (bases 1 and 2).

Table 1. Genes with functional predictions identified in EhV-86. CDS, systematic gene identifier; CG, CDSs that are known NCLDV core genes (✓) (*3*); Ex, CDSs that are expressed (✓) or not expressed (x) at 33 hours postinfection as determined by microarray transcriptomic analysis.

CDS	Position		Putative function/features	CG	Ex
	Start	End			
ehv014	8433	9299	Longevity-assurance (LAG1) family protein		✓
ehv018	12303	13379	Endonuclease		✓
ehv020	15170	15949	Putative proliferating cell nuclear antigen	✓	✓
ehv021	15998	17212	Serine protease		✓
ehv022	17197	18486	Phosphoglycerate mutase family protein		✓
ehv023	18732	19253	Deoxycytidylate deaminase		✓
ehv026	22027	23004	Ribonucleoside-diphosphate reductase small chain	✓	✓
ehv028	24121	24912	Lipase		✓
ehv030	25675	28713	DNA polymerase delta catalytic subunit	✓	✓
ehv031	28718	29704	Sterol desaturase		✓
ehv041	38172	39305	Endonuclease		✓
ehv050	45226	47838	Serine palmitoyltransferase		✓
ehv060	55942	61926	Lectin protein		x
ehv061	61979	62941	Fatty acid desaturase		✓
ehv064	64018	68501	DNA-dependent RNA polymerase II largest subunit	✓	✓
ehv072	74904	75785	DNA-binding protein	✓	✓
ehv077	78289	79257	Transmembrane fatty acid elongation protein		✓
ehv079	80165	80896	Lipid Phosphate phosphatase		✓
ehv085	86013	87614	Major capsid protein	✓	✓
ehv093	93036	93416	HNH endonuclease family protein		x
ehv101	99368	100174	Hydrolase		x
ehv103	100800	101153	Vesicle-associated membrane protein		✓
ehv104	101158	102741	Putative helicase	✓	✓
ehv105	102837	103331	Transcription factor S-II (TFIIS) family protein	✓	✓
ehv108	106901	107545	DNA-directed RNA polymerase subunit		✓
ehv109	107546	108073	OTU-like cysteine protease		✓
ehv110	108106	108942	RING finger protein		✓
ehv113	110604	112046	Bifunctional dihydrofolate reductase–thymidylate synthase	✓	✓
ehv117	113956	115560	Phosphate permease		x
ehv128	121517	122026	ERV1/ALR family protein	✓	x
ehv133	125567	126283	ATP-dependent protease proteolytic subunit		✓
ehv136	126964	127572	Nucleic acid–binding protein		✓
ehv141	131083	132345	Hypothetical protein	✓	✓
ehv151	139720	140628	Serine protease		✓
ehv152	140747	141610	DNA binding protein		✓
ehv158	145226	147094	DNA ligase	✓	✓
ehv160	147460	148464	Serine protease		✓
ehv166	154991	155719	RING finger protein	✓	x
ehv167	155756	156016	DNA-directed RNA polymerase subunit	✓	✓
ehv179	167551	169176	Major facilitator superfamily protein		✓
ehv184	171775	173091	DNA binding protein		✓
ehv230	219949	220335	Endonuclease		✓
ehv346	228542	229306	Lectin protein		✓
ehv349	289563	290267	Protease		✓
ehv358	300433	300909	Thioredoxin		x
ehv361	302158	303507	Serine protease		x
ehv363	303778	304569	Esterase		x
ehv393	333013	334023	DnaJ domain-containing protein		x
ehv397	335877	336323	Deoxyuridine 5′-triphosphate nucleotidohydrolase	✓	✓
ehv399	337274	337963	DNA-directed RNA polymerase subunit		✓
ehv401	338688	339317	Ribonuclease	✓	✓
ehv402	339307	340335	Protein kinase		x
ehv403	340415	341551	Hypothetical protein	✓	✓
ehv415	348487	349263	Putative fatty acid desaturase		x
ehv428	362962	365202	Ribonucleoside-diphosphate reductase protein	✓	✓
ehv430	365923	368631	Helicase		✓
ehv431	368760	369743	Thymidylate kinase	✓	✓
ehv434	370859	374329	DNA-directed RNA polymerase II subunit	✓	✓
ehv440	378178	379044	Proliferating cell nuclear antigen protein		✓
ehv444	384374	387685	DNA topoisomerase	✓	✓
ehv447	388779	389684	Serine protease		✓
ehv451	391491	392306	Protein kinase	✓	✓
ehv453	393351	394478	MRNA capping enzyme	✓	✓
ehv455	396954	398075	Sialidase		✓
ehv459	398903	401004	Nucleic acid–independent nucleoside triphosphatase	✓	✓
ehv465	404611	405201	Putative thioredoxin protein	✓	x

References and Notes

1. J. L. Van Etten, M. V. Graves, D. G. Muller, W. Boland, N. Delaroque, *Arch. Virol.* **147**, 1479 (2002).
2. W. H. Wilson *et al.*, in *Virus Taxonomy, VIIIth ICTV Report*, C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Dusselberger, L. A. Ball, Eds. (Elsevier/Academic Press, London, 2005), pp. 163–175.
3. L. M. Iyer, L. Aravind, E. V. Koonin, *J. Virol.* **75**, 11720 (2001).
4. N. Delaroque *et al.*, *Virology* **287**, 112 (2001).
5. W. H. Wilson *et al.*, *J. Mar. Biol. Assoc. U.K.* **82**, 369 (2002).
6. D. C. Schroeder, J. Oke, G. Malin, W. H. Wilson, *Arch. Virol.* **147**, 1685 (2002).
7. P. Westbroek, J. R. Young, K. Linschooten, *J. Protozool.* **36**, 368 (1989).
8. R. J. Charlson, J. E. Lovelock, M. O. Andreae, S. G. Warren, *Nature* **326**, 655 (1987).
9. P. Westbroek *et al.*, *Global Planet. Change* **8**, 27 (1993).
10. P. Westbroek *et al.*, in *The Haptophyte Algae*, J. C. Green, B. S. C. Leadbeater, Eds. (Clarendon, Oxford, 1994), vol. 51, pp. 321–334.
11. Materials and methods are available as supporting material on *Science* Online.
12. The complete sequence of EhV-86 is deposited in GenBank/European Molecular Biology Laboratory under accession number AJ890364.
13. R. A. Sandaa, M. Heldal, T. Castberg, R. Thyrhaug, G. Bratbak, *Virology* **290**, 272 (2001).
14. B. La Scola *et al.*, *Science* **299**, 2033 (2003).
15. D. Raoult *et al.*, *Science* **306**, 1344 (2004); published online 14 October 2004 (10.1126/science.1101485).
16. I. Galli, S. M. M. Iguchiariga, H. Ariga, *Nucleic Acids Res.* **20**, 3333 (1992).
17. Microarray data are submitted in the European Bioinformatics Institute ArrayExpress database (www.ebi.ac.uk/arrayexpress) under the accession number e-maxd-2. In addition, this database is also available at the Environmental Genomics Thematic Programme Data Centre (EGTDC) data catalog (http://envgen.nox.ac.uk/) under the accession number egcat:000010.
18. A. Berhe, U. Fristedt, B. L. Persson, *Eur. J. Biochem.* **227**, 566 (1995).
19. A. Berhe, R. Zvyagilskaya, J. O. Lagerstedt, J. R. Pratt, B. L. Persson, *Biochem. Biophys. Res. Commun.* **287**, 837 (2001).
20. A. H. Futerman, Y. A. Hannun, *EMBO Rep.* **5**, 777 (2004).
21. A. H. Merrill Jr., *J. Biol. Chem.* **277**, 25843 (2002).
22. L. M. Obeid, C. M. Linardic, L. A. Karolak, Y. A. Hannun, *Science* **259**, 1769 (1993).
23. Y. A. Hannun, L. M. Obeid, *Trends Biochem. Sci.* **20**, 73 (1995).
24. J. Teodoro, P. Branton, *J. Virol.* **71**, 1739 (1997).
25. S. A. Susin *et al.*, *J. Exp. Med.* **186**, 25 (1997).
26. K. D. Bidle, P. G. Falkowski, *Nat. Rev. Microbiol.* **2**, 643 (2004).
27. C. P. D. Brussaard, A. A. M. Noordeloos, R. Riegman, *J. Phycol.* **33**, 980 (1997).
28. J. A. Berges, P. G. Falkowski, *Limnol. Oceanogr.* **43**, 129 (1998).
29. J. L. Van Etten, *Annu. Rev. Genet.* **37**, 153 (2003).
30. W. H. Wilson *et al.*, unpublished data.

31. This research was supported by a grant in aid from the Natural Environment Research Council (NERC) awarded to the Marine Biological Association (MBA) of the UK and by a grant awarded to W.H.W. from the NERC Environmental Genomics thematic program (ref. NE/A509332/1). We would like to acknowledge support from the EGTDC, Centre for Ecology and Hydrology, Oxford. We thank E. Gudmundsdottir and K. Vierlinger from Scottish Centre for Genomic Technology and Informatics (SCGTI) for help with probe design and initial microarray labeling experiments.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5737/1090/DC1
Materials and Methods
Figs. S1, S3, S4
Table S2
Database S5

4 April 2005; accepted 5 July 2005
10.1126/science.1113109