Arch Virol (2006) 151: 525–535
DOI 10.1007/s00705-005-0647-1

Archives of Virology
Printed in Austria

Preliminary characterisation of repeat families in the genome of EhV-86, a giant algal virus that infects the marine microalga *Emiliania huxleyi*

M. J. Allen[1], **D. C. Schroeder**[2], and **W. H. Wilson**[1,2]

[1]Plymouth Marine Laboratory, Plymouth, U.K.
[2]Marine Biological Association, Plymouth, U.K.

Received June 9, 2005; accepted August 8, 2005
Published online September 30, 2005

Summary. EhV-86 is a large double stranded DNA virus with a 407,339 base pair circular genome that infects the globally important microalga *Emiliania huxleyi*. It belongs to a new genus of viruses termed the *Coccolithoviridae* within the algal virus family *Phycodnaviridae*. By plotting the EhV-86 genome against itself in a dot-plot analysis we revealed three families of distinctly different repeat sequences throughout its genome, designated Family A, B and C. Family A repeats are non-coding, found immediately upstream of 86 predicted coding sequences (CDSs) and are likely to play a crucial role in controlling the expression of the associated CDSs. Family B repeats are GC rich, coding and correspond to possible calcium binding sites in 22 proline-rich domains found in the protein products of eight predicted EhV-86 CDSs. Family C repeats are AT-rich, non-coding and are likely to form part of the origin of replication. We suggest that these repeat regions are of fundamental importance during virus propagation being involved with transcriptional control (Family A), virus adsorption/release (Family B) and DNA replication (Family C).

Introduction

The majority of algal viruses characterised to date fall within the family *Phycodnaviridae*. Members of this family share icosahedral morphology and have been distinguished by the taxonomic affiliation of their algal host into six genera (*Chlorovirus*, *Prasinovirus*, *Prymesiovirus*, *Phaeovirus*, *Coccolithovirus* and *Rhapidovirus*) [31]. These viruses infect marine and freshwater algae and have large double-stranded DNA genomes ranging from 150 kb to 560 kb. Three *Phycodnaviridae* genome sequences are known in their entirety, namely, the *Paramecium bursaria chlorella virus*, (PBCV-1) [25], *Ectocarpus siliculosus* *virus*, (EsV-1) [6] and the recently sequenced *Emiliania huxleyi virus* (EhV-86) [30]. *Emiliania huxleyi* is a marine coccolithophorid found throughout the world's oceans. It is well known for its vast coastal and mid-oceanic blooms at temperate latitudes and can cover 10,000 km^2 or more. EhV-86 is a virus strain isolated from a bloom in the English Channel in 1999 [29]. Initial characterisation revealed a 407,339 bp length genome, with a 40.2% G+C content which is predicted to contain 472 CDSs making it the largest member of the *Phycodnaviridae* sequenced to date [30].

Double-stranded DNA virus genomes have been shown to contain homologous or repetitive regions which are thought to play important roles in replication and transcription [10, 11]. Indeed, a characteristic feature of the EsV-1 genome is the presence of large blocks of repetitive elements comprising approximately 12% of the genome [6]. These authors suggested that many of the repeats in EsV-1 may serve as origins of replication and remain unwound and single-stranded for packaging into the virus capsid. Hence, explaining the reason for extensive single-strandedness in extracted EsV-1 DNA [15]. Here, we report the preliminary classification and description of homologous regions contained within the EhV-86 genome.

Materials and methods

The EhV-86 genome can be accessed via accession number AJ890364 in the GenBank database. To identify repetitive sequences within the EhV-86 genome, a dot-plot analysis was used (LBDotView Version 1.0, [14]). Such an analysis can compare one genome on the x axis against another genome, or in this case itself, on the y axis indicating the precise location and orientation of homologous sequences within the plot. Alignments were conducted using ClustalW [24].

Results and discussion

Full genome analysis by dot-plot [13] revealed the presence of three distinct families (designated A to C) of homologous regions contained within the EhV-86 genome (Fig. 1).

Family A

Family A homologous regions (Fig. 1) consist of regularly spaced, variable sized (30–300 bp) homologous regions found within a section of the EhV-86 genome from 200 kb to 304 kb. Wilson et al. (2005) found this region was unusual since it contained no gene homologues from the data base. The size of the homologous repeats appears to increase towards the centre of this 104 kb region, with the largest repeats found in the region 252 to 260 kb. The repeat units correspond to the non-coding regions found directly upstream of 86 of the 151 predicted CDSs annotated in this 104 kb region of the genome and are characterised by the presence of a 5′ conserved GTTCCC(T/C)AA nonamer, usually directly followed by a downstream ATG. Indeed, in 67 of 86 of these CDSs the proceeding ATG

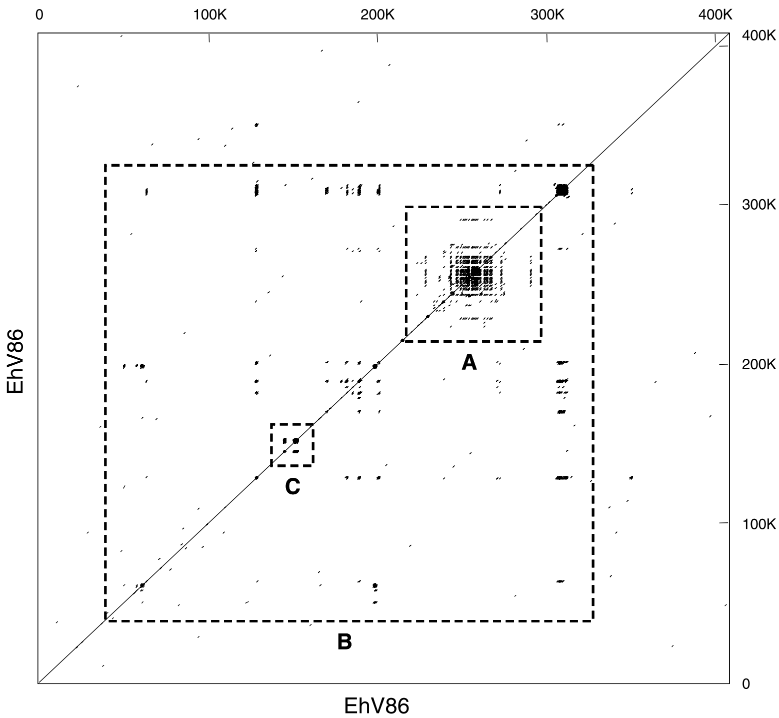

Fig. 1. Dot Plot analysis (IBDot) of the full length EhV-86 genome and grouping of repeat families. Family A are found at 105 locations between 204 kb and 304 kb, Family B are dispersed throughout the genome from 50 kb to 320 kb and Family C are found in 2 locations at 144 kb and 150 kb

is predicted to indicate the start of translation. A search of the entire genome for the sequence GTTCCC(T/C)AA revealed it was found at 106 locations, with all but one being located within the 104 kb region identified previously in Fig. 1. The precise location of the Family A homologous regions, i.e. immediately upstream of the ATG start codon, and their apparent non-coding nature suggests that they could function as promoter elements essential for transcription [1]. If this is the case, the highly conserved non-coding GTTCCC(T/C)AA nonamer would provide an excellent candidate for a specific binding site for a transcription factor.

ClustalW alignment of the 300 bp immediately upstream of each of the 86 CDSs (i.e. the promoter regions) associated with Family A repeats revealed a conserved sub group of sequences (Fig. 2). These correspond to the upstream regions of CDSs ehv294, ehv295, ehv296, and ehv297 which are located in the centre of the 104 kb region and match up to the longer central homologous regions identified. The apparent variation in size of homologous regions is presumably due to deviations, in the 5′ region, from this 'consensus' sequence. Furthermore, these highly conserved putative promoters appear to contain 2 more copies of the repeating nonamer further upstream separated by the 6 bp sequence ACGCCA (Fig. 2). Localisation of this family of repeats, corresponding to likely promoters, to a

```
ehV294   ----TGTCTTGTAAATTAACATGACCAATTAACGAAACCCATTAACGAAACCCATTAACG 56
ehV295   ------AACCCACTTTTAACGAAACCCATTAACGAAACCCATTAACGAAACCCATTAACG 54
ehV296   ----GACGAAACCCATTGACGAAACCCATTGACGAAACCCATTGACGAAACCCATTAACG 56
ehV297   -----TGGTAGAATTTCAATCCATGGGATTAACGAAACCAATTAACGAAACTCATTAACG 55
                        *  *        *** ******** *** ******* ********

ehV294   AAACCCATAGAAGGAATATTCCTCCGCGTCTCATTGCTCTCGGACCGAGATAGAATCCCG 116
ehV295   AAACCCA--GAAGGAATATTCCTCCGCGTCTCATTGCCCCCGGGCCGAGATAGAATCCCG 112
ehV296   AAACCCATGGAAGGAATATTCCTCCGCGTCTCATTGCTCCCGGACCGAGATAGAATCCCG 116
ehV297   AAACCCATAAAAGGAATATTCCTCCGCGTCTCATTGCTATCGGACCGAGATAGAATCCCG 115
         *******   ***************************   *** ****************

ehV294   ACGGCGGATTCTGTTGAGGACTTAGCAGAATTCAGCCAAGTC-TCGGTTCCCTAAACGCC 175
ehV295   ACGGCGGATTCTGTTGAGGACTTAGCAGAATTCAGCCAAGTC-TCGGTTCCCTAAACGCC 171
ehV296   ACGGCGGATTCTGTTGAGGACTTAGCAGAATTCAGCCAAGTC-TCGGTTCCCTAAACGCC 175
ehV297   ACGGCGGATTCTGTTGAGGACTTAGCAGAATTCAGCCCATTGGTCGGTTCCCTAAACGCC 175
         ************************************* * *  *****************

ehV294   AGTTCCCTAAACGGCTTAATATTAAATCGACTGATGAGCAAACGGAATATTCCATCGAGG 235
ehV295   AGTTCCCTAAAAGGCTTAATATTAAATCGACTGATGAGCAAACGGAATATTCCATCGAGG 231
ehV296   AGTTCCCTAAAAGGCTTAATATTAAATCGGCTGATGAGCAAACGGAATATTCCATCGAGA 235
ehV297   AGTTCCCTAAAAGGCTTAATATTAAATCGACTGATGAGCAAACGGAATATTCCATCGAGG 235
         *********** ***************** *****************************

ehV294   TCACAATCTCGCGACCCACGCTCGGTCGCATCAGCCAACTCCGCAACTCGCCACCAGTTC 295
ehV295   TCACAATCTCGCGACCCACGCTCGGTCGCATCAGCCAACTCCACAACTCGCAACCAGTTC 291
ehV296   TCACAATCTCGCGAGCCACGCTCGGACGCAGCGGCCAACTCCACAACTCGCCACCAGTTC 295
ehV297   TCACAATCTCGCGACCCACGCTCGGACGCGTCAACCAACTCCACAACTCGCCACCAGTTC 295
         ************** ********** ***    *  ******** ******** ********

ehV294   CCTAAATG---- 303
ehV295   CCTAAAAGGATG 303
ehV296   CCTAAATG---- 303
ehV297   CCTAAATG---- 303
         ****** *
```

Fig. 2. Alignment of the 300 bp immediately upstream from the predicted start of translation of ehv294, ehv295, ehv296 and ehv297. The ATG start of translation codon is also shown. An AT-rich region is denoted by grey shading. Nonamers are indicated in bold and are boxed. Stars (*) indicate conserved sequences

particular genomic region would suggest a form of highly coordinated expression of a particular sub-set of CDSs. Since EhV-86 encodes its own RNA polymerase [30], these sites may provide appropriate binding sites allowing the expression of transcripts during the early stages of virus infection. Indeed, homologous regions have been shown to be essential for transient gene expression and to act as *cis* activators of early genes in baculoviruses leading to their designation as 'super-enhancers' [5]. It is plausible that these repeat regions may indicate the location of CDSs expressed at a particular stage in the life cycle of this virus i.e. immediate early, early or late CDSs. Although there is no obvious TATA box, an AT-rich region can be found sandwiched between the two nonamers, at approximately −100 to the predicted start of transcription (Fig. 2). AT-rich regions 5′ to CDSs have previously been identified in PBCV-1 [21]. However, searches did not reveal any significant matches in the GenBank, EMBL or DDBJ database (BLASTN 2.2.11, 05/05).

Table 1. Family B repeat proteins

CDS	Size (amino acids)	Number of prolines	Expression confirmed[a]	Comments
ehv060	1994	334	no	similar to calcium/calmodulin binding protein from *Paramecium tetraurelia* (e^{-13})
ehv062	194	46	no	no database similarity
ehv137	516	178	no	no database similarity
ehv192	2873	422	yes	similar to calcium/calmodulin binding protein from *Paramecium tetraurelia* (e^{-14})
ehv204	621	351	yes	no database similarity
ehv207	430	183	yes	no database similarity
ehv364	2332	1014	yes	no database similarity
ehv416	403	97	no	similar to *Drosophila* sperm protein (e^{-08})

[a]Expression confirmed by Wilson et al. [31]

Family B

Family B repeat regions (Fig. 1) are found clustered at eight locations throughout the genome, contain multiple repeats of the nucleotides ‘CCN’ (typically in sets of 3–5 repeats) and correspond to proline rich regions of the predicted CDSs ehv060, ehv062, ehv137, ehv192, ehv204, ehv207, ehv364 and ehv416 (Table 1). There are 1170 copies of the repeating unit CCNCCNCCN in the EhV-86 genome of which 950 are found within these eight CDSs. These CDSs are commonly found to be open in at least 5 reading frames in the regions where the repeats occur. CDSs ehv204, ehv062, ehv416, ehv137 and ehv207 are predicted to encode proteins of 621 (351), 194 (46), 403 (97), 516 (178) and 430 (183) amino acids, respectively (numbers in brackets indicate the number of predicted prolines). CDSs ehv192, ehv364 and ehv060 are unusually long and contain 2873, 2332 and 1994 amino acids (each remarkably maintained in an open reading frame) containing 422, 1014 and 334 proline residues, respectively, distributed in 4 or more domains. The predicted proteins typically contain long stretches of three or four prolines interrupted by a serine or leucine. Using microarray analysis, Wilson et al. 2005 revealed that CDSs ehv204, ehv207, ehv364 and ehv192 are expressed during infection [30]. There is no expression data for the other CDSs.

Intriguingly, ehv192 and ehv364 flank the 104 kb central region, which contains the family A putative promoters and corresponding CDSs. It is difficult to determine the significance of this, however, it could be speculated that these flanking CDSs act as sites for recombination and was a mechanism for transporting this 104 kb region into the genome. Highly repetitive CDSs have previously been suggested to act as recombinational hotspots [12]. Transcriptomic analysis revealed that, during infection, some of the most highly expressed CDSs are from the central part of this family A repeats region [30].

The advantage of the family B repeats to the virus is unclear, particularly since BLAST searches reveal no obvious homologues for the majority of these CDSs (See Table 1; GenBank, 08/05). However, many proline-rich proteins have been described in the literature previously, having been found in a diverse range of organisms including ORF180 in the recently sequenced Mimivirus, the largest virus sequenced to date [18]. Human salivary secretions contain groups of proteins in which the proline content is typically from 20% to over 40% [16, 22]. Intriguingly, one group of these proline-rich proteins has been implicated in the inhibition of calcification [20]. Indeed, the evidence for interactions between proline-rich proteins and calcium is well documented. The crustacean DD4 protein (14% proline), which is expressed during calcification of the exoskeleton, also binds calcium [7]. While the role of DD4 in calcification remains to be elucidated, it may be involved in transport or storage of Ca^{2+} or the formation of calcium crystals. A photoreceptor cell protein found in drosophila, Calphotin, (20.6% proline) [17] and the mammalian calreticulin family of proteins (containing proline rich domains) [8] have also been identified as binding calcium. Furthermore, Both ehv060 and ehv192 show similarity to a calcium binding protein from *Paramecium tetraurelia* (Table 1). The wealth of evidence for the calcium-binding properties of proline-rich proteins is interesting since this virus infects a cell that actively sequesters calcium carbonate scales (coccoliths) onto its surface during active growth [28]. Calcification is clearly an important process in *E. huxleyi* and it is closely coordinated with the rest of cellular metabolism, including photosynthesis [4]. Whether these 8 virus-encoded proline rich proteins are involved during the initial infection of *E. huxleyi* or even during the later packaging and release of virions through the coccolith secretion pathway is unknown, but certainly warrants further investigation.

Family C

Family C repeats consist of non-coding AT-rich (approx. 74%) repeating units of approximately 324 bp. There are two clusters of units (designated α and β) which are separated by approximately 6 kb at two locations (144 kb and 150 kb, see Fig. 1) in the EhV-86 genome. Cluster α contains 2 copies of the repeating unit (1 complete, 1 incomplete), while cluster β comprises 7 copies (5 complete, 2 incomplete) (Fig. 3). The proximal 3′ repeat unit in each cluster contains a partial deletion in the 3′ end, βVII contains a 43 bp deletion and αII contains a

Fig. 3. Family C repeat sequences aligned using ClustalW. Palindromic sequences are marked in bold and boxed. * indicates the presence of conserved bases. α and β prefixes denote repeat elements found in the first (∼144 kb) and second (∼150 kb) clusters of repeat units respectively, Roman numeral suffixes denote the order the repeat units are found in within the clusters

```
αI       --ACACAGGGAAAAA-GTGCCGTTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 57
αII      ACACACAGGGAAAAA-GTGCCGTTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βI       --ATATAGGAAAAAA-GTGCCGTTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 57
βII      ACACACAGGAAAAAA-GTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βIII     ACACACAGGAGAAAT-GTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βIV      ACACACAGGAAAAAA-GTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βV       ACACACAGGAAAAAA-GTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βVI      ACACACAGGAAAAAA-GTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 59
βVII     ACACACAGGAAAAAAAGTGCCATTATTAGCACAATAAATAACTCTGGTAAAACGTACTTA 60
           * * ***  ***  ***** ***************************************

αI       GCAATTTTCGAGTTATTTATTTTCATTCGTGAAATTATGAACTTAAAATATTTTTTTTCA 117
αII      GCGATTTTCGAGTTATTTATTTTCATTCGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βI       GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 117
βII      GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βIII     GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βIV      GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βV       GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βVI      GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 119
βVII     GCAATTTTCGAGTTATTTATTTTCATTTGTGAAATTATGAACTTAAAATATTTTTTTTCA 120
         ** ************************ ********************************

αI       TTTTTGGGAATATGAAATTAAAGGGCTTTCCAAATCGTAACCAAAAATAATAAAGGACTT 177
αII      TTTTTGGGAATATGAAATTAAAGGGCTTTCCAAATCGTAACCAAAAATAATAAAGGACTT 179
βI       TTTTTAGGAAAAGTAAATTAAAGGGCTTTCCAAATCGTAACCAAAAATAATAAAGGACTT 177
βII      TTTTTAGGAAAAGCAAATCAAAGTGCTTTCAAAATCGTAACCAAAAATAATAAAGGACTT 179
βIII     TTTTTAGGAAAAGTAAATTAAAGGGCT--------------------------------- 148
βIV      TTTTTAGGAAAAGCAAATCAAAGTGCTTTCAAAATCGTAACCAAAAATAATAAAGGACTT 179
βV       TTTTTAGGAAAAGCAAATCAAAGTGCTTTCAAAATCGTAACCAAAAATAATAAAGGACTT 179
βVI      TTTTTAGGAAAAGCAAATCAAAGTGCTTTCAAAATCGTAACCAAAAATAATAAAGGACTT 179
βVII     TTTTTAGGAAAAGCAAATCAAAGTGCTTTCAAAATCGTAACCAAAAATAATAAAGGACTT 180
         ***** **** *  **** **** *****

αI       GAAAATCCACAAGGGGTCAAATAAATAACTTTTTTTGGCTATATACTTTTTACAGAGTTA 237
αII      GGAAATCTACGATGGGTCAAATAAATAACTTTTTTTGGCTGTATACTTTTTACAGAGTTA 239
βI       GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 237
βII      GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 239
βIII     ---------------------------------------------------------TTA 149
βIV      GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 239
βV       GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 239
βVI      GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 239
βVII     GAAATTCATCGGGGGGTCAAATAAATAACTTTTTTTTGCTGTATACTTTTTACAGAGTTA 240
                                                                  ***

αI       TTTATTCTGGTAACAATCAAGTTATTTATTCTAAGTGTTTTGTGAATTTATGAACAATTC 297
αII      TTTATTGTGGTAATAATCAAGTTATTTATTCTAAG------------------------- 274
βI       TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 297
βII      TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 299
βIII     TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 209
βIV      TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 299
βV       TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 299
βVI      TTTATTGTGGTAATAATCAAGTTATTTATTCTAAGTGTTTTATGAATTTGTGAACAATTC 299
βVII     TTTATTCTGGTAATAATCAAGTTATTTATTCTAAGTGTTTT------------------- 281
         ****** ****** ********************

αI       AAAATACATTCAAATATTTCTAGCT 322
αII      -------------------------
βI       AAAATACATTTAGATATTTCTAGCT 322
βII      AAAATACATTCAGATATTTCTAGCT 324
βIII     AAAATACATTTAGATATTTCTAGCT 234
βIV      AAAATACATTTAGATATTTCTAGCT 324
βV       AAAATACATTTAGATATTTCTAGCT 324
βVI      AAAATACATTTAGATATTTCTAGCT 324
βVII     -------------------------
```

50 bp deletion. A key structural feature of Family C repeats is the presence of a conserved 11 bp palindromic repeat (CAATAAATAAC) in the 5′ region, which is inverted and repeated twice (once perfectly and once with a 1 bp change) at the 3′ end of the repeated sequence.

GenBank and EMBL searches (05/05) for sequence similarity with the nucleotide sequences of the 324 bp repeat did not retrieve any sequence with significant similarity (data not shown). AT-rich repeat regions can be characteristic of virus genome origins of replication [9] and together with the presence of palindromic sequences suggest that the Family C homologous region may be the origin of replication (ori) for EhV-86. Indeed, the structure of this repetitive region in EhV-86 resembles the origin of replication (*oriP*) in Epstein-Barr Virus (EBV), another large dsDNA virus [19]. EBV is a human herpes virus that maintains its genome extrachromosomally in infected cells [23]. *OriP* is composed of two functional elements, the dyad symmetry (DS) and the family of repeats (FR). FR consists of two repeat elements separated by approximately 1.8 kb, containing 4 and 20 binding sites for EBNA-1, respectively [19]. EBNA-1 is a virally encoded protein which contributes to *oriP* synthesis and maintenance. The larger cluster contributes to transcriptional enhancement and maintenance, whereas the smaller cluster is the site at which DNA synthesis is initiated.

Genomes from algal virus strains EsV-1 and PBCV-1 both contain terminal repeats (identical 2.2 kb inverted repeats in PBCV-1 and 1.8 kb and 1.6 kb almost perfect inverted repeats in EsV-1), thought to play a crucial role in DNA replication [26]. It is common for large, linear, double-stranded DNA viral genomes to circularise via termini repeats to allow a rolling circle type of replication [2, 6, 27]. Initial characterisation and sequencing of the EhV-86 genome predicted a linear genome of 407,339 bp. However, the presence of a predicted origin of replication suggested to us that the EhV-86 genome may have a circular stage at some point during its life cycle. This was confirmed by PCR using primers annealing to the termini of the genome [30]. The genome appears to have either an A or a T (at equal ratio) single base pair overhang at each of the termini of its genome, indicating a possible method for circularisation of the genome. The EhV-86 genome contains a putative DNA polymerase (ehv030), topoisomerase (ehv444) and three helicases (ehv104, ehv356 and ehv430). In addition, the EhV-86 genome contains a DNA ligase (ehv158), which, intriguingly, is located in the 6 kb gap between the two AT rich clusters. This conformation of genes is characteristic of rolling circle type method of replication [3].

Closing discussion

The *in silico* analysis performed in this study has provided unique insights into many fundamental aspects of the EhV-86 life cycle. We have identified regions of the genome which we believe to be involved with DNA replication (Family C), transcriptional control (Family A) and virus adsorption/release (Family B).

Clearly, the mechanism of EhV-86 replication needs to be further elucidated, but the identification of a putative *ori* site and a possible method of circularising the genome are important first steps. The presence of eight large, repetitive, proline rich CDSs in one virus genome is intriguing. The calcium-binding properties of proline-rich proteins is well documented and due to the calcium carbonate nature of the host cell, their presence may provide a clue of vital importance to the mechanism of virus adsorption and/or release. This clearly warrants further investigation. However, the identification of a family of possible promoter elements localised to a 100 kb region of the genome is perhaps the most interesting product of this analysis. The transcriptional, functional and evolutionary relevance of this region is completely unknown but is likely to be crucial to the EhV-86 infection cycle.

Acknowledgements

We would like to thank Matthew Holden and Julian Parkhill (Wellcome Trust Sanger Institute, Cambridge, UK) who introduced us to the dot blot analysis. The research was supported by the Environmental Genomics community programme, funded by the Natural Environmental Research Council of the United Kingdom (NERC), through award number NE/A509332/1 to WHW. DCS is a Marine Biological Association of the UK (MBA) Research Fellow funded by grant in aid from the NERC. WHW is supported through the NERC-funded core strategic research programme of the Plymouth Marine Laboratory.

References

1. Allen MJ, Schroeder DC, Holden M, Wilson WH (2005) Evolutionary history of Coccolithoviridae. Mol Biol Evol, doi: 10.1093/molbev/msj010
2. Blum H, Zillig W, Mallok S, Domdey H, Prangishvili D (2001) The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses. Virology 281: 6–9
3. Boehmer PE, Lehman IR (1997) Herpes simplex virus DNA replication. Annu Rev Biochem 66: 347–384
4. Brownlee C, Nimer N, Dong LF, Merrett MJ (1994) Cellular regulation during calcification in *Emiliania huxleyi*. In: Green JC, Leadbeater BSC (eds) The haptophyte algae. Syst Assoc Special vol 51: 133–148, Clarendon Press, Oxford
5. Chen Y, Yao B, Zhu ZZ, Yi YZ, Lin X, Zhang ZF, Shen GF (2004) A constitutive super-enhancer: homologous region 3 of Bombyx mori nucleopolyhedrovirus. Biochem Biophys Res Commun 318: 1039–1044
6. Delaroque N, Muller DG, Bothe G, Pohl T, Knippers R, Boland W (2001) The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. Virology 287: 112–132
7. Endo H, Persson P, Watanabe T (2000) Molecular cloning of the crustacean DD4 cDNA encoding a Ca^{2+}-binding protein. Biochem Biophys Res Commun 276: 286–291
8. Fliegel L, Burns K, MacLennan DH, Reithmeier RA, Michalak M (1989) Molecular cloning of the high affinity calcium-binding protein (calreticulin) of skeletal muscle sarcoplasmic reticulum. J Biol Chem 264: 21522–21528
9. Galli I, Iguchi-Ariga SM, Ariga H (1992) The AT-rich tract of the SV40 ori core: negative synergism and specific recognition by single stranded and duplex DNA binding proteins. Nucleic Acids Res 20: 3333–3339

10. Gompels UA, Macaulay HA (1995) Characterization of human telomeric repeat sequences from human herpesvirus 6 and relationship to replication. J Gen Virol 76 (Pt 2): 451–458
11. Hayakawa T, Rohrmann GF, Hashimoto Y (2000) Patterns of genome organization and content in lepidopteran baculoviruses. Virology 278: 1–12
12. Hill CW (1999) Large genomic sequence repetitions in bacteria: lessons from rRNA operons and Rhs elements. Res Microbiol 150: 665–674
13. Huang Y, Zhang L (2004) Rapid and sensitive dot-matrix methods for genome analysis. Bioinformatics 20: 460–466
14. Huang Y, Zhang L (2004) Rapid and sensitive dot-matrix methods for genome analysis. Bioinformatics 20: 460–466
15. Klein M, Lanka S, Muller D, Knippers R (1994) Single-stranded regions in the genome of the *Ectocarpus siliculosus* virus. Virology 202: 1076–1078
16. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299: 2033–2033
17. Martin JH, Benzer S, Rudnicka M, Miller CA (1993) Calphotin: a Drosophila photoreceptor cell calcium-binding protein. Proc Natl Acad Sci USA 90: 1531–1535
18. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-megabase genome sequence of mimivirus. Science 306: 1344–1350
19. Ritzi M, Tillack K, Gerhardt J, Ott E, Humme S, Kremmer E, Hammerschmidt W, Schepers A (2003) Complex protein-DNA dynamics at the latent origin of DNA replication of Epstein-Barr virus. J Cell Sci 116: 3971–3984
20. Schlesinger DH, Hay DI (1986) Complete covalent structure of a proline-rich phosphoprotein, PRP-2, an inhibitor of calcium phosphate crystal growth from human parotid saliva. Int J Pept Protein Res 27: 373–379
21. Schuster AM, Graves M, Korth K, Ziegelbein M, Brumbaugh J, Grone D, Meints RH (1990) Transcription and sequence studies of a 4.3-kbp fragment from a ds-DNA eukaryotic algal virus. Virology 176: 515–523
22. Stahl L, Wright R, Castle J, Castle A (1996) The unique proline-rich domain of parotid proline-rich proteins functions in secretory sorting. J Cell Sci 109: 1637–1645
23. Sugden B (2002) In the beginning: a viral origin exploits the cell. Trends Biochem Sci 27: 1–3
24. Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
25. van Etten JL, Meints RH (1999) Giant viruses infecting algae. Annu Rev Microbiol 53: 447–494
26. Van Etten JL, Graves MV, Muller DG, Boland W, Delaroque N (2002) *Phycodnaviridae* – large DNA algal viruses. Arch Virol 147: 1479–1516
27. Vink C, Beuken E, Bruggeman CA (1996) Structure of the rat cytomegalovirus genome termini. J Virol 70: 5221–5229
28. Westbroek P, Dejong EW, Vanderwal P, Borman AH, Devrind JPM, Kok D, Debruijn WC, Parker SB (1984) Mechanism of calcification in the marine alga *Emiliania huxleyi*. Philos Trans Roy Soc Lond Ser B Biol Sci 304: 435–444
29. 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 Marine Biol Assoc UK 82: 369–377
30. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, Churcher C, Hamlin N, Mungall K, Norbertczak H, Quail MA, Price C, Rabbinowitsch E, Walker

D, Craigon M, Roy D, Ghazal P (2005) Complete genome sequence and lytic phase transcription profile of a coccolithovirus. Science 309: 1090–1092

31. Wilson WH, Van Etten JL, Schroeder DS, Nagasaki K, Brussaard C, Delaroque N, Bratbak G, Suttle C (2005) Family: *Phycodnaviridae*. In: Fauquet CM, Mayo MA, Maniloff J, Dusselberger U, Ball LA (eds) Virus taxonomy, VIIIth ICTV Report. Elsevier/Academic Press, London, pp 163–175

Author's address: Dr. William H. Wilson, Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, PL1 3DH, U.K.; e-mail: whw@pml.ac.uk