Virology ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities

António Pagarete [a,*], Kanthida Kusonmano [b], Kjell Petersen [b], Susan A. Kimmance [c], Joaquín Martínez Martínez [d], William H. Wilson [c,d], Jan-Hendrik Hehemann [e], Michael J. Allen [c], Ruth-Anne Sandaa [a]

[a] Department of Biology, University of Bergen, Norway
[b] Computational Biology Unit, University of Bergen, Norway
[c] Plymouth Marine Laboratory, Plymouth, UK
[d] Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA
[e] Department of Civil & Environmental Engineering, Massachusetts Institute of Technology, USA

ARTICLE INFO

Article history:
Received 20 March 2014
Returned to author for revisions
23 April 2014
Accepted 18 May 2014

Keywords:
Coccolithovirus
Metagenome
Algae virus
Emiliania huxleyi virus
EhV
Emiliania huxleyi
Giant virus
Marine metagenomics
Virome
Genomic hyper-variable region
Bloom dynamics
Viral genomic diversity

ABSTRACT

Despite the global oceanic distribution and recognised biogeochemical impact of coccolithoviruses (EhV), their diversity remains poorly understood. Here we employed a metagenomic approach to study the occurrence and progression of natural EhV community genomic variability. Analysis of EhV metagenomes from the early and late stages of an induced bloom led to three main discoveries. First, we observed resilient and specific genomic signatures in the EhV community associated with the Norwegian coast, which reinforce the existence of limitations to the capacity of dispersal and genomic exchange among EhV populations. Second, we identified a hyper-variable region (approximately 21 kbp long) in the coccolithovirus genome. Third, we observed a clear trend for EhV relative amino-acid diversity to reduce from early to late stages of the bloom. This study validated two new methodological combinations, and proved very useful in the discovery of new genomic features associated with coccolithovirus natural communities.

Introduction

Viruses that infect phytoplankton play a key role in shaping the evolution and dynamics of the oceanic micro-scale ecosystem (Fuhrman, 1999; Sandaa, 2008; Suttle, 2005). Several studies have highlighted the role of viruses as major players in high phytoplankton turnover rates, a process termed the viral shunt (Wilhelm and Suttle, 1999). The interplay of viruses with their host communities is complex, and may assume different forms. Traditionally regarded as simple agents of mortality and catalysts for nutrient transformation (Suttle, 2005; Weinbauer and Rassoulzadegan, 2004), viruses are now also believed to play a fundamental role in controlling the biodiversity and functioning of their associated host communities (Frada et al., 2008; Thingstad, 2000; Thingstad and Lignell, 1997).

Emiliania huxleyi (Lohmann) Hay et Mohler, a single celled phytoplankton, is the most abundant and ubiquitous coccolithophore in extant marine systems (Brown and Yoder, 1994). *E. huxleyi* is an important species with respect to the past and present marine primary productivity, and in particular global carbon and sulphur cycles (Burkill et al., 2002; Westbroek et al., 1993). Poorly understood until recently, it is now clear that *E. huxleyi*-specific viruses (EhV, Coccolithoviridae) are closely involved in the control of their host's populations, a phenomenon better appreciated during the sudden crashes of vast *E. huxleyi* coastal and mid oceanic blooms (Bratbak et al., 1993; Jacquet et al., 2002; Schroeder et al., 2003; Wilson et al., 2002).

This long-established host–virus interaction (Coolen, 2011) will have driven the genomic evolution of both virus and host systems, leading to the development of infection/resistance strategies that

* Corresponding author.
E-mail address: antonio.pagarete@bio.uib.no (A. Pagarete).

http://dx.doi.org/10.1016/j.virol.2014.05.020
0042-6822/

are now fundamental to their ecology. An evolutionary consequence of the close intracellular interaction between the *E. huxleyi* and EhV systems is the high level of promiscuity between the two genomes that has enabled a series of horizontal gene transfer (HGT) events (Read et al., 2013). Some of these genes have potential implications for the infection strategy of these viruses and/or relate to their host's defence system (Monier et al., 2009; Pagarete et al., 2009; Vardi et al., 2009, 2012). At an ecological level we observe how the selection pressure imposed by these viruses is potentially linked to profound somatic consequences in *E. huxleyi*'s life cycle, namely the alternation between diploid and haploid phases as a key mechanism to evade infection (Frada et al., 2008).

Host–virus interaction analyses have commonly reported established phenomena where there is a significant decrease in EhV major capsid protein (MCP) diversity during the progression of bloom events (Martínez Martínez et al., 2007; Schroeder et al., 2003). Put simply, from an initial high diversity, a few dominant ecotypes eventually dominate as the bloom develops. The selection pressure acting upon the host–virus pairs is not trivial, especially when considering the ecological dynamics and consequences. For instance, if relative EhV MCP diversity is significantly reduced during a single bloom event, how can diversity be maintained between blooms? Despite the omnipresence of *E. huxleyi* cells in marine samples, the truth is that for the majority of cases (meaning non-blooming situations) these cells exist in low concentrations. For example, in the Norwegian coastal area studied here *E. huxleyi* concentrations are below 30 cells ml^{-1} for at least 6 months every year (unpublished data). The probability of an EhV virion finding a suitable *E. huxleyi* host outside the bloom windows is significantly decreased. Although this should in theory favour geographical isolation of EhV populations, infectivity experiments with EhV isolates and *E. huxleyi* strains has shown no increased capacity of EhV strains to more successfully infect either closely or distantly isolated host strains (Allen et al., 2007; Pagarete, 2010). The question on the capacity and relevance of dispersal and gene exchange of these oceanic viruses remains unanswered.

A host–virus system evolves under the guise of an arms race between two distinct genomes, the one of the host and the one of the virus (Stern and Sorek, 2011). Yet, in that arms race some genes, intra-gene regions, or even genomic regions will face different selective pressures. To date we have a poor understanding of how selection pressure influences giant EhV genomes, and consequently, the amino-acid composition of its associated proteins. It is currently unknown if selection is being homogeneously exerted on the whole of the EhV genome, or if in turn distinct conservation rates can be found for specific EhV genes or genomic regions. A recent study on Mimivirus, another giant virus, clearly showed a tendency for those viruses to endure significant and rapid genome reductions (through gene loss) after only 150 infection rounds (Boyer et al., 2011).

With these questions and issues in mind we used a new approach to study EhV metagenomic diversity within natural populations. Traditionally, studies of viral genomic diversity use sequence data from available isolates, but the large size of EhV genomes make it virtually impossible to isolate and sequence enough viral strains to comprehensively represent the EhV genetic diversity naturally existing. Hence we employed a new combination of DNA separation methodologies (based on single band sequencing from either pulsed field gel electrophoresis (PFGE) or CsCl gradient) with next generation sequencing (454 or Illumina technologies) to study genomic variability within a natural EhV community during an *E. huxleyi* bloom. Here we present two EhV metagenomes from the early and late stages of an induced *E. huxleyi* bloom with the aim of answering 3 specific questions: (1) what is the genomic resemblance of natural EhV populations to currently isolated EhV strains, (2) how is diversity and conservation distributed along the EhV metagenome, and (3) what is the progression of EhV metagenomic diversity during a bloom.

Building upon the answers to these questions, we then focus our analysis on two particular genes (ehv060 and ehv452) to demonstrate the potential, but also the intricacies, of the metagenomic approach presented for the identification of selective constraints and infection mechanisms acting in this virus–host system. The ehv060 gene encodes two domains with putative glycan binding function: a carbohydrate binding module (CBM) specific for sialic-acid residues in host glycans, and a C-type-lectin like domain. Both domains can be involved in glycan interactions, mediating viral attachment (Bowden et al., 2011; Jolly and Sattentau, 2013). Notably the ehv060 protein is present in the EhV virion (Allen et al., 2008). The ehv452 gene encodes a high mobility group (HMG) protein. HMG proteins are involved with chromatin structure, usually endowing the chromosome with nuclease sensitivity, and they also recruit transcription factors to bind to enhancers (Štros, 2010). The unexpected high levels of amino-acid diversity registered for these proteins justified their analysis in this manner.

Results

General bloom/infection dynamics

Initial *E. huxleyi* abundance at the start of the experiment (day 0) was approximately 2.1×10^2 cells ml^{-1}. Coccolithophore concentrations inside the mesocosm enclosure started increasing exponentially from day 6, reaching a maximum number of 1.7×10^5 cells ml^{-1} on day 12, followed by sharp decline (Fig. S1). The decline in *E. huxleyi* numbers coincided with the appearance and exponential increase of coccolithoviruses from day 11 onwards. A maximum concentration of 2.8×10^7 coccolithoviruses ml^{-1} was recorded on day 15. When samples were collected for metagenomic analysis, on days 11 and 15, EhV concentrations were 8.1×10^5 and 2.8×10^7 coccolithoviruses ml^{-1}, respectively. For an in-depth description and discussion of community dynamics during the mesocosm experiment refer to Kimmance et al. (2014), Pagarete et al. (2009, 2011) and Vardi et al. (2012).

Characteristics of the two metagenomes

Sample S11 was sequenced using 454 technology, generating 166,940 reads of 256 bp (on average) equivalent to nearly 0.043 Gb of sequence. Sample S15 was sequenced using Illumina technology, generating 11,576,462 paired-end reads of 51 bp (on average) equivalent to nearly 1.2 Gb of sequence. In both metagenomic datasets a significant percentage of the reads were identified as EhV sequences (approximately 18% and 68% for S11 and S15, respectively). Average level of sequence depth differed between the two metagenomes by an order of magnitude. Consequently, the conservative analysis of gene identification was carried out independently for each metagenome, with different thresholds of minimum average read depth and minimum DB coverage adopted for each metagenome (Table 1). In both metagenomes the identified proteins were homogeneously scattered around the EhV genome, with no obvious sign of sequencing bias towards specific genomic regions (Fig. 1). The Simpson index was chosen as an indicator of amino-acid diversity, after tests performed on a battery of diversity indexes, due to its status as the index least affected by read-depth, while also retaining high sensitivity to diversity. Nevertheless, for the calculation of the Simpson diversity index different minimum read depth thresholds were adopted for

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020

each of the two metagenomes to avoid biased diversity levels due to low read coverage (Table 1). This ultimately led to a different number of EhV proteins available from each sample for subsequent diversity analysis. For S15 we could confidently calculate diversity averages for 396 CDSs (approximately 85.2% of the EhV proteome). For S11 that value went down to 46 CDSs (approximately 10% of the EhV proteome) (Tables 1 and S1).

Table 1
Technical properties of the two sequenced metagenomes, along with the minimum thresholds applied to select CDSs for diversity analysis.

	S11 454 Seq.	**S15** Illumina Seq.
Average read depth (minimum threshold for identification) (minimum threshold for diversity analysis)	26 (≥ 10) (≥ 10)	2248 (≥ 50) (≥ 350)
Average CDS coverage (minimum threshold for identification) (minimum threshold for diversity analysis)	96% (≥ 50%) (≥ 80%)	99.9% (≥ 70%) (≥ 80%)
Total EhV CDSs identified (% total EhV genome)	300 (61%)	458 (93%)
CDSs available for diversity analysis (% total EhV genome)	46 (10%)	396 (85%)

Community resemblance to cultured viruses

Globally, the two metagenomic EhV community samples showed higher sequence homology to the EhV-99B1 than to EhV-86 (Fig. 1). Moreover, all the EhV proteins exclusive of the Fjord EhV-99B1 and EhV-163 isolates (Table 2) were found in our two metagenomes. On the contrary, only one putative membrane protein (ehv310) that exists in all the English Channel isolates but not the Fjord isolates was present in the metagenomes. An intein in gene ehv434 (which encodes an RNA polymerase subunit) that had been previously identified in EhV-99B1 and EhV-163 (Allen et al., 2011) was also identified in the current metagenome (Fig. S2).

Analysis of genomic diversity

Simpson diversity values per protein ranged between 0.005 and 0.070 (average = 0.013). These arranged in an inverted exponential distribution (Fig. 2) with only 5% of the proteins presenting diversity values higher than 0.028. The average diversity per protein was not homogeneously distributed along the genome sequence (Fig. 1). Most notably there was a hyper-variable region, ranging roughly between genes ehv263 and ehv296 (~21.350 bp). In that region we observed a concentration of genes with increased diversity values (Fig. 1). This region contained 33 genes, 25 of which are associated with a novel promoter (Allen et al., 2006c). Analyses

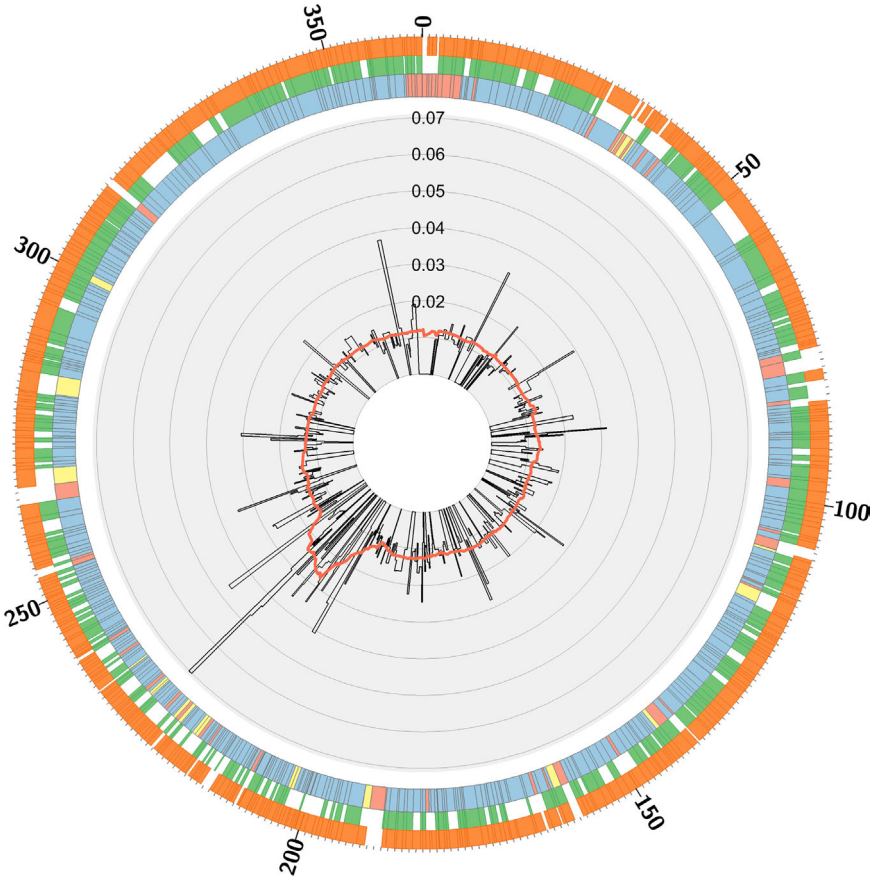

Fig. 1. Circular representation of the EhV genome with metagenome identity to reference strains and average amino-acid diversity per protein. The outside scale is numbered clockwise in kbp. Circle 1 (from inside out) is a combined representation of the CDSs identified in the reference genomes of EhV-86 and EhV-99B1. CDSs are color-coded by the presence/absence in each reference genome: blue, present in both genomes; rose, present only in EhV-86; yellow, present only in EhV-99B1. Circle 2 (green) represents the 300 EhV CDSs identified in metagenome S11. Circle 3 (orange) represents the 458 EhV CDSs identified in metagenome S15. Note that all yellow CDSs in Circle 1 overlap with CDSs in Circle 3, while most rose CDSs in Circle 1 are absent from Circle 3. In the interior histogram it is plotted average amino-acid diversity per EhV protein in metagenome S15. The trend line for that histogram is based on moving averages with period 35 (in red). Note the concentration of high diversity values between 237 and 258 kbp.

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020

Table 2
Presence (+) and absence (−) of specific genes in S11 and S15 metagenomes as compared to previously EhV isolates. Information for each isolate based on a combination of sequence and nucleotide microarray data.

EhV strains	English Channel isolates															Norwegian fjord isolates		S11 and S15 metagenomes	CDS annotation
	86	18	84	88	145	156	164	201	202	203	205	206	207	208	209	99B1	163		
EhV CDSs																			
ehv034	+	+	+	+	+	+	+	+	+	+	−	−	+	+	−	−	−	−	Putative membrane protein
ehv036A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv038A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv083	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative membrane protein
ehv088	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative membrane protein
ehv117	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative phosphate permease
ehv117A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Putative endonuclease
ehv129A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Putative transposase
ehv159	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative membrane protein
ehv161A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv188	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative membrane protein
ehv225A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv244A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv259	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Putative Membrane protein
ehv285A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Putative membrane protein
ehv286A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv296A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv296B	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein
ehv306A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Putative membrane protein
ehv310	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	Putative membrane protein
ehv316	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	Putative membrane protein
ehv344	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Hypothetical protein
ehv345	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	Hypothetical protein
ehv377A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Putative membrane protein
ehv396A	−	−	−	−	−	−	−	−	−	−	na	na	−	−	na	+	+	+	Hypothetical protein

na – no information available.

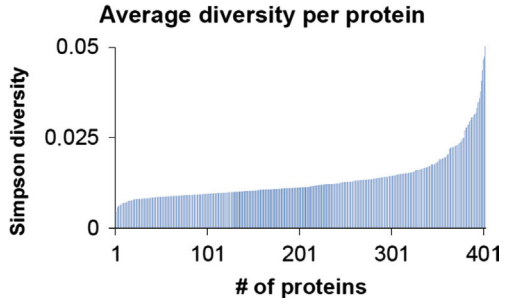

Fig. 2. Simpson diversity values per protein arranged in ascending order.

Table 3
Analyses of similarity (ANOSIM) between gene assemblages, based on Simpson diversity per amino-acid. Values that were statistically significant are in **bold**.

Factor	*R*	Significance (%)
KOG group	−0.064	96.5
NCLDV core gene set	−0.118	99.1
Hyper-variable region	0.582	**0.1**
Presence/absence of Family A promoters	0.859	**0.1**

of similarity on amino-acid Simpson diversity (ANOSIM) based on different factors (see description in "Materials and methods" section) revealed that diversity was significantly related ($p < 0.01$) to placement inside or outside the observed hyper-variable region and to the presence or absence of Family A promoters (Table 3). Student's *t*-test confirmed that both the amino-acid diversity and GC content of the CDSs in the hyper-variable region are significantly different ($p = 4.9 \times 10^{-6}$ and $p = 3.4 \times 10^{-9}$, respectively) from the rest of the genome. On a global EhV genome scale, the Spearman's rank correlation test did not show a significant correlation ($p = 0.43$) between GC content and amino-acid diversity.

A gene by gene analysis showed that the 10 CDSs with highest average amino-acid diversity encoded for 3 high mobility group (HMG) proteins, 3 membrane proteins and 4 proteins without functional prediction (Table 4). The most diverse of all was ehv293, one of the putative HMG CDSs which is placed inside the hyper-variable region mentioned above. Its diversity value (0.070 ± 0.008) was $5.24\times$ higher than the global average. On the other hand, the top most conserved genes retrieved from our analysis encoded: 2 RNA pol sub-units (encoded in genes ehv144 and ehv458), 2 membrane proteins, and 6 proteins currently without functional prediction (Table 4). Diversity analysis on the intein present in the gene ehv434 showed that the region corresponding to the intein sequence is extremely well conserved (Fig. S2). An analysis of read depth across ehv434 revealed an average difference between the intein region and its flanking regions from 2200 to 3040 reads per site, respectively. This difference could tentatively be regarded as a proxy for the presence of this intein in ~72% of the EhV viruses in that sample.

Analysis of diversity per amino-acid revealed that the distribution of different amino-acids per site ranged from 1 to 20, and presented a bell shape, with an average value of 5 ± 2.09 (Fig. S3). The modal average was 5 different amino-acid possibilities per site. The average diversity per amino-acid was also not homogeneously distributed along the genome, with a concentration of variability in the hyper-variable region previously mentioned. Among the proteins with the largest number of diverse amino-acid sites we observed two proteins that are present in the EhV virion: ehv060 and ehv085 which encode lectin and major capsid proteins, respectively (Allen et al., 2008).

Progression of diversity during a bloom

Due to limitations in coverage and read-depth, sample S11 yielded a subset of 46 CDSs suitable for a comparison of diversity

Table 4
Top 10 CDSs with highest (roman) and lowest (italics) average amino-acid diversity in S15.

Gene	Amino-acid diversity	Annotation	KOG category
EhV293	0.0702	Putative high mobility group protein	General functional prediction only
EhV292	0.0474	Hypothetical protein	Function unknown
EhV307	0.0466	Hypothetical protein	Function unknown
EhV291	0.0436	Hypothetical protein	Function unknown
EhV265	0.0407	Putative high mobility group protein	General functional prediction only
EhV452	0.0377	Putative high mobility group protein	General functional prediction only
EhV284	0.0359	Hypothetical protein	Function unknown
EhV278	0.0349	Putative membrane protein	Function unknown
EhV341	0.0347	Putative membrane protein	Function unknown
EhV032	0.0333	TRAM/LAG1/CLN8 domain containing protein	Intracellular trafficking, secretion, and vesicular transport
EhV051	*0.0070*	*Hypothetical protein*	*Function unknown*
EhV039	*0.0069*	*Hypothetical protein*	*Function unknown*
EhV098	*0.0069*	*Hypothetical protein*	*Function unknown*
EhV321	*0.0066*	*Hypothetical protein*	*Function unknown*
EhV080	*0.0066*	*Putative membrane protein*	*Function unknown*
EhV458	*0.0064*	*DNA-directed RNA polymerases I, II, and III*	*Transcription*
EhV208	*0.0062*	*Hypothetical protein*	*Function unknown*
EhV144	*0.0060*	*Putative RNA polymerase Rpb3/Rpb11 dimerisation domain*	*Transcription*
EhV255	*0.0057*	*Hypothetical protein*	*Function unknown*
EhV457	*0.0048*	*Putative membrane protein*	*Function unknown*

levels between the two samples (Table S1). The vast majority of the amino-acid sites (81.9%) displayed similar diversity levels from days 11 to 15. However, a large amount of sites changed from variable to conserved (17.8%) during this period. The opposite change (from conserved to variable) was clearly lower (0.3%). Among the six CDSs that changed the most between the two days the level of similarity between the consensus amino-acid sequences went as low as 37% (Table 5). All these CDSs have no predicted function. On the other hand, the 9 proteins that changed the least during the sampling period presented amino-acid conservation rates around 90% (Table 5). The highest among these (ehv047) is annotated as a putative nonsense-mediated mRNA decay (NMD) protein. The remaining 8 CDSs have no assigned or putative functions.

Table 5
Most conserved (roman) and most variable (italics) CDSs between days 11 and 15.

Gene	Amino-acid conservation (%)	Function
ehv047	96.8	Translation, ribosomal structure
ehv256	96.7	Function unknown
ehv161	93.0	Function unknown
ehv040	88.8	Function unknown
ehv092	88.4	Function unknown
ehv039	87.6	Function unknown
ehv413	85.9	Function unknown
ehv304	85.3	Function unknown
ehv072	82.9	DNA-binding protein
ehv096	*45.9*	*Function unknown*
ehv205	*45.1*	*Function unknown*
ehv294	*42.7*	*Function unknown*
ehv148	*42.1*	*Function unknown*
ehv351	*39.7*	*Function unknown*
ehv221	*37.0*	*Function unknown*

Specific analysis of intra-gene diversity – the case of a high mobility group protein

Three HMG encoding genes were among the 10 CDSs displaying the highest average amino-acid diversity. One of these, gene ehv452, encodes two HMG2 domains. The closest homologues to ehv452 in NCBI reference protein database included only eukaryotic organisms, namely two microalgae (Bacillariophyta), 9 green plants (Magnoliophyta), and 2 Opisthokonta. The alignment made with the two HMG-box conserved domains revealed that the sequence similarity between the gene ehv452 and the other HMG proteins was too small to allow a significant resolution of its placement in a phylogenetic tree (Fig. S4). Conversely, from the 13 HMG proteins analysed, only the two microalgae sequences presented the same domain organisation as ehv452, notably with two HMG2 encoding domains sequentially aligned in the 5′ end of the gene.

Comparison of the gene domain organisation with the amino-acid diversity levels obtained from the metagenomic analysis revealed that the high levels of diversity registered resulted from very strong amino-acid variation towards the 3′-end of the sequence, a region without domain prediction. The two HMG2 domains presented very high levels of conservation (Fig. 3). The same pattern was also observed for ehv265 and ehv293, the two other HMG encoding genes (data not shown).

Analysis of intra-gene diversity – the specific case of a glycan binding protein

The gene ehv060 was among the CDSs displaying high levels of amino-acid diversity. This CDS encodes a large protein of 1994 amino acids (210 kDa) that is present in the EhV virion. ehv060 encodes two domains with putative glycan binding function: a family 40 carbohydrate binding module (CBM, pdb id: 2V73), and a c-type lectin domain (PFAM id: PF00059). The presence of a transmembrane region could be identified spanning residues 1739–1769, suggesting that the CBM and lectin domains could have an extracellular location. The analysis of amino-acid diversity along ehv060 revealed an important concentration of very variable residues towards the N terminus of the CBM domain. Such high levels of amino-acid variation could not be identified in the lectin domain (Fig. 4).

Discussion

In this study we analysed the variability within natural populations of EhV in a coastal area of Western Norway (North Sea) using

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020

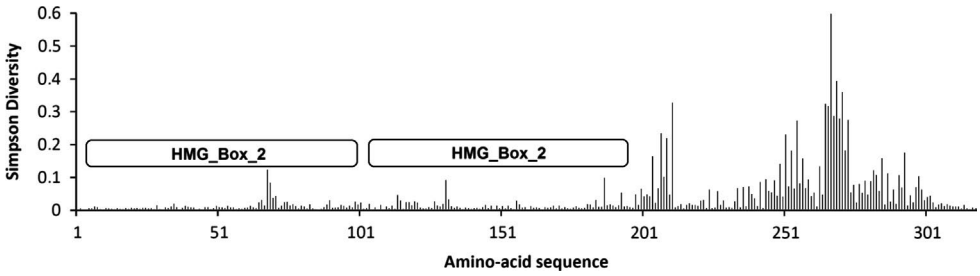

Fig. 3. Amino-acid diversity distribution in CDS ehv452 and position of the two HMG_Box_2 coding domains.

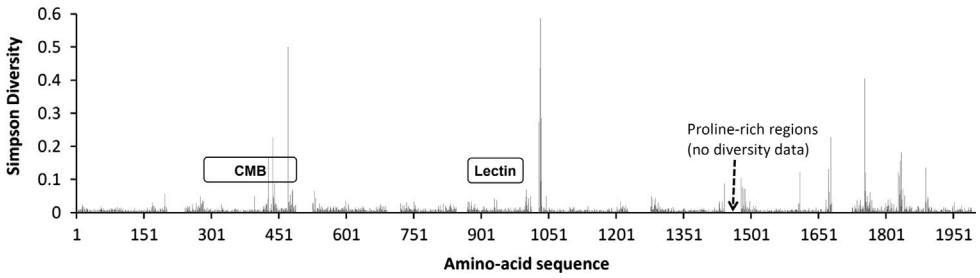

Fig. 4. Amino-acid diversity distribution along CDS ehv060 and the position of the two glycan binding domains.

a targeted metagenomic approach. Samples collected from two key stages of an induced *E. huxleyi* bloom allowed a snapshot of the total genomic diversity contained within a whole natural EhV community, but also a basic assessment of how that diversity changes with the progression of the bloom. We have deliberately chosen to conduct this analysis at the amino-acid level, as opposed to nucleotide level, to focus our study on the variability that has a functional meaning and hence is relevant to the ecology of these viruses and their interactions with the host *E. huxleyi*.

The present study is a culmination of two different, yet complementary, methods aiming to capture the natural metagenomic diversity of a specific oceanic virus gene pool, in this case coccolithoviruses. One method was a combination of pulsed-field gel electrophoresis-based DNA extraction with 454 pyrosequencing (Ray et al., 2012), the second was a combination of CsCl gradient-based DNA extraction with Illumina sequencing. With both methods we were able to recover a significant part of the EhV genome (61% and 93%, respectively), and importantly, we did not observe any bias towards amplification/sequencing of particular segments of that genome. With regards to sequence coverage, the read depth levels registered for the Illumina based method were consistently one to two orders of magnitude higher than the 454-based method. This was probably due to a combination of different factors: the 2 orders of magnitude difference in the amount of EhV particles present in the two water samples; the considerably different volumes of water taken with each method (10 l versus 150 l), or even intrinsic differences in the two sequencing methods. Regardless of the ultimate reason(s), our study allowed the analysis of diversity of 10% (454, day 11) and 93% (Illumina, day 15) of the EhV proteins encoded in EhV-86. 454 Pyrosequencing remains a more prudent approach as the longer 454 reads facilitate assembly. Having a reference genome (as was the case here) proved extremely useful for the analysis of the Illumina-based data.

Analysis of both generated metagenomes revealed their clear resemblance to the two EhV strains (EhV-99B1 and EhV-163) previously isolated from the same location, as opposed to the 15 EhV strains isolated in the English Channel. Previous comparisons of EhV isolates from these two locations had revealed that, despite the very high levels of sequence similarity ($>95\%$), there were a series of genomic features that could only be associated with strains from either location (Allen et al., 2007; Nissimov et al., 2011a, 2011b, 2012a, 2012b; Pagarete et al., 2012). Among others, those genomic differences included 46 gene insertion/deletion events. This study came to confirm that the majority (at least 89%) of those distinguishing traits had not resulted from an exploratory analysis of only two strains. Our results strongly indicate that those traits are widespread among the EhV populations in the sampled Norwegian coastal area, making them distinct from the EhVs in the English Channel. This is even more relevant if we consider that all current strains had been isolated approximately 10 years before these metagenomes were collected. During that period, and given the absence of any clear geographical barrier, one could expect the EhV populations to spread between these regions, eventually attenuating the genomic differences they might initially have. However, there appears to be a clear resilience of genetic traits and geographical segregation associated with these EhV populations.

These findings add to our emerging knowledge of the constraints associated with the geographic distribution of these oceanic viruses. On the one hand, there is no connection between virus–host range capacity and host origin, i.e. the distance between the isolation places of host and virus has no proven influence (Pagarete, 2010). On the other hand, there must be barriers to gene flow among Coccolithovirus populations that lead to the observed localised resilience of specific genomic traits. Those barriers could result from localised selective pressures imposed by their host populations and/or physiological adaptations of this host–virus system to different physico-chemical conditions of the environment. Altogether these results tell us that the answer to the question "*here a virus, there a virus, everywhere the same virus?*" (Breitbart and Rohwer, 2005) is more complex than a simple yes or no, and the same applies for the individual genetic elements they encode.

When analysing amino-acid diversity per protein we expected to find, among the most conserved, proteins involved in critical replication functions, such as nucleotide-sequence polymerases and proteins that directly interact with them. Therefore, it was without surprise that two out of the 10 most conserved proteins (encoded by ehv144 and ehv458) were identified as RNA polymerase subunits (Table 4). Of these, CDS ehv144 encodes an Rpb3 dimerisation domain and is essential to form a platform onto which the other subunits of the RNA polymerase assemble.

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020

CDS ehv458 encodes the Rpb6, which in eukaryotic systems forms a structure with at least two other subunits that stabilizes the transcribing polymerase on the DNA template. However, we have no functional prediction for the eight other most conserved proteins. Theoretically, the most conserved proteins, i.e. those under the strongest amino-acid conservation pressures, should be critical for the success of the "life" cycle of a virus. Yet, the function of the majority of these proteins in this host–virus system remains elusive. This is symptomatic of our significant ignorance regarding the functional potential encoded in the world's viromes. Currently, within metagenomic libraries, an average of 70% of the viral CDSs identified have no predicted function (Breitbart, 2012). Our results add to the necessity to find new strategies to prise open the black box that represents viral-encoded proteins.

The most unanticipated finding in this study was the existence of a hyper-variable region present in the EhV genome, roughly between genes ehv263 and ehv296. To our knowledge, this region (approximately 21 kbp in size, containing around 33 genes) is the first report of such a feature in a eukaryotic viral genome. The identification of the hyper-variable region through the metagenomic analysis of global amino-acid coding is intriguing as it overlaps with two other curious findings for that area of the EhV genome. First it coincides with the previously identified Family A repeat region (Wilson et al., 2005). Family A repeats are short noncoding segments (potential promoters) placed immediately upstream of 86 predicted CDSs within an overlapping 100 kbp region (Allen et al., 2006c). The CDSs in the Family A repeat region have few database homologues, and their origin and function are completely unknown (Allen et al., 2006c). Secondly, it is known that the 39 early transcripts, that are expressed within the first hour of EhV infection, are localised to a single region of the EhV genome, ranging from gene ehv218 to ehv366 (Allen et al., 2006a). It is noteworthy that 10 of these 39 early transcripts are located in the hyper-variable region.

At this stage, the role of this hyper-variable region in the genome of this large virus is uncertain. These proteins may represent crucial components in the host–virus recognition and/or interaction process, and their diversity a direct consequence of the evolutionary requirement to constantly change in the ceaseless arms race between host and virus. To that extent, it is worth noting the existence of a hyper-variable island in *Prochlorococcus* (cyanobacteria) genomes (Coleman et al., 2006). In the *Prochlorococcus*-virus system, host resistance-associated mutations seem confined to a hyper-variable region, known to be associated with viral attachment to the host cell surface and are linked with a fitness cost to the host (Avrani et al., 2011). Moreover, most genes found in genomic islands result from horizontal gene transfer and/or recombination events (Coleman et al., 2006; Lindell et al., 2007). The newly identified hyper-variable region in the EhV genome could also represent a hotspot for those events.

Previous reports have shown that the relative diversity of the MCP gene (ehv085) becomes significantly reduced during blooms as a few genotypes come to dominate the overall community structure (Martínez Martínez et al., 2007; Schroeder et al., 2003). Here, we observed this known ecological phenomenon occurring not just with an individual gene, but affecting around 20% of the amino-acid sites encoded in EhV (certainly for the 46 genes analysed). This is good evidence that the decrease in EhV diversity towards the end of the blooms is not limited to virion proteins such as MCP. However, this is not yet definitive since systematic bias from method/sequencing differences cannot be completely ruled out here. The question remains though, how do natural EhV communities maintain high diversity levels from one bloom to the next?

It is important to note that diversity in amino-acids within a protein might not always be associated with an obvious, or indeed actual, diversity in function, i.e. active sites and/or domains. An example of this is represented by the HMG proteins encoded in EhV genome. A close look at these HMG proteins reveals that the "variable" parts of these CDSs reside away from the very conserved HMG domains. A search for homologous domains retrieved sequences belonging to distant organisms, all of which belonged to the eukaryotic world. Of the closest 13 HMG domain containing proteins analysed, only two microalgal sequences (both diatoms) presented the same domain organisation as that present in ehv452. This, along with the absence of homology found in any other viral protein, leads us to believe that this might be yet another example of horizontal gene exchange between these giant viruses and single-celled organisms (Monier et al., 2007, 2009).

In conclusion, we described a successful approach to address genomic diversity of an aquatic virus population (coccolithoviruses). By creating a map of genomic variation within the coccolithovirus gene pool, we have gained a new level of understanding about the genes and genomic regions potentially linked to their ecology and evolution. In addition, this study has laid a marker for future studies on the molecular mechanisms by which viruses evolve. The surprising finding of a hyper-variable region, and high levels of diversity associated with a CBM motif in the lectin-like virion protein, ehv060, are but two clear examples of how we can use the information presented here to select and focus our future research efforts on the most interesting aspects of the giant EhV genome.

Materials and methods

Set-up of the mesocosm experiment and monitoring

The mesocosm experiment was conducted in the Raunefjorden, Western Norwegian coast (N 60°16′, E 5°13′) for a period of 17 days (5th–21st of June 2008). The mesocosm bags (11 m^3 each) were filled with unfiltered Fjord water pumped from 10 m depth adjacent to the raft. Homogeneous water masses within the enclosures were ensured by pumping water from the bottom of the bag to the surface. *E. huxleyi* and EhV concentrations were measured using flow cytometry (FCM) according to Jacquet et al. (2002) and Marie et al. (1999), respectively. Further details of host and virus population analyses can be found in Kimmance et al. (2014), Pagarete et al. (2009, 2011), Sorensen et al. (2009) and Vardi et al. (2012).

DNA preparation for high throughput sequencing

Two samples were collected for metagenomic characterisation of EhV communities. Samples were taken at early-mid and late (during bloom demise) stages of the bloom, on days 11 (S11) and 15 (S15), respectively. The water samples were collected from the surface of the mesocosm enclosure with 20 l carboys. The carboys were immediately brought back to the lab for sample filtration and concentration as follows.

Sample S11 – PFGE and 454 pyrosequencing

A 10 l seawater sample was collected, passed through a 200 μm mesh, and concentrated to a final volume of approx. 50 ml by tangential flow filtration using a Vivaflow 200 benchtop system with a 50,000 molecular weight cut off polyethersulfone membrane (Sartorius). PFGE was performed using a 1% w/v SeaKem GTG agarose (FMC, Rockland, Maine) gel in TBE gel buffer using a Bio-Rad DR-II CHEF Cell (Bio-Rad, Richmond CA, USA) electrophoresis unit with pulse-ramps at 8–30 s for 24 h at 14 °C according to the methods of Sandaa et al. (2010). Bands of approx. 410 kb, the

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020

same size of the EhV genome (407 kb), were excised and frozen at −80 °C. DNA was eluted from the PFGE agarose gel slices in 10,000 MWCO Spectra/Por, Regenerated Cellulose dialysis membranes (Spectrum Laboratories Inc., CA, USA) by electrophoresis in TAE buffer (40 mM Tris–HCl, 1 mM EDTA, 40 mM acetic acid, pH 8.0) for 3 h at 70 V, following the method of Ray et al. (2012). Further concentration of the DNA was performed using Vivaspin 500 columns (Millipore Corp) according to the manufacturer's protocol. Eluted DNA was amplified based on a linker–adaptor PCR method using the WGA1 and Genome Plex WGA reamplification kit from Sigma (Sigma Aldrich, St Louis, MO, USA). Six separate WGA reactions were run and pooled before further processing. The amplified products were purified using the GenElute PCR Clean-Up Kit (Sigma Aldrich) and stored at −80 °C until sequencing. Pyrosequencing was performed by the Broad institute of MIT & Harvard (Cambridge, USA) using the Roche/454 GS FLX Titanium pyrophosphate sequencing platform (Basel, Switzerland). Sequence reads were deposited in CAMERA (https://portal.camera.calit2.net/gridsphere/gridsphere) under project "CAM_PROJ_BroadPhage", sample name "Virome ME-08-2".

Sample S15 – CsCl gradient and Illumina sequencing

150 l of water were concentrated using a 30 kDa Sartocon Slice tangential flow filtration (TFF) system to 2 l, followed by a 30 kDa Midgee TFF system to 15 ml and applied to a CsCl gradient as previously described (Wilson et al., 2005). A clear band was visible which had a density identical to that of laboratory produced coccolithovirus (EhV-86). Following extraction and a second round of CsCl purification, only a single band corresponding to the coccolithovirus fraction of the natural community was obtained. DNA extraction of S15 produced 10 μg of material which was then subjected to paired-end Illumina sequencing. Sequence reads were deposited to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under the following project accession: PRJEB5540.

Bioinformatic analyses

EhV read calling and noise removal

Sequenced reads from S11 and S15 were mapped to a reference protein database including all the CDSs in EhV-86 (isolated in the English Channel) and EhV-99B1 (isolated at the same site at the Western Norwegian coast where the metagenomes were collected). Given the different sample characteristics and sequencing techniques used, which yielded dissimilar read quality levels, the mappings of S11 and S15 reads were performed by using BLASTX and Burrows–Wheeler Aligner (BWA) for short reads (Li and Durbin, 2009), respectively. To determine the presence and absence of proteins mapped to EhV-86 and EhV-99B1, the cut-off for S11 was with E-value $< 1e^{-3}$, similarity $\geq 40\%$. The minimum thresholds of average read depth and CDS coverage for S11 and S15 are shown in Table 1.

Community resemblance to cultured viruses

Information on EhV isolate diversity was retrieved from Uniprot and integrated with nucleotide microarray-based information (Allen et al., 2007). This was used to create a list of 25 EhV genes whose presence is distinctive between 15 isolates from the English Channel and 2 from the studied Norwegian Fjord (Table 2). The presence of the encoded 25 proteins was investigated in S11 and S15 metagenomes using BLASTP with cut-offs E-value $< 1e^{-4}$ and sequence similarity $\geq 80\%$.

Analysis of genomic diversity

Analysis of genomic diversity was performed on 396 EhV genes identified in S15. This corresponded to 85.2% of the currently predicted proteins in the combined EhV gene pool. For each of those genes, BLASTX was used to align the reads from BWA results to S15 consensus amino-acid sequences. After examination of different diversity indexes, the Simpson index was used to calculate amino-acid diversity per site and per protein. Analyses of similarity (ANOSIM, Primer6) were carried out to test the absence of significant diversity differences between groups of proteins depending on different factors: KOG functional group, belonging to a previously identified set of NCLDV core genes (Allen et al., 2006b), placement inside or outside the observed hypervariable region, and the presence or absence of Family A promoter (Allen et al., 2006c).

Progression of diversity during the bloom

A subset of 46 EhV CDSs was selected to evaluate changes in amino-acid diversity levels between the two metagenomic samples (Table S1). CDSs choice was based on minimum thresholds of read-depth and CDS coverage in each of the two metagenomes (Table 1). In total those 46 CDSs corresponded to 8679 amino-acid sites. In order to circumvent the structural differences in read-depth between the two datasets, a classification was adopted based on the percentage of dominant amino-acids present at each amino-acid site. In this classification, a site was considered "conserved" if its most dominant amino acid was represented in more than 90% of the reads for that site. Conversely, a site was considered "variable" if its most dominant amino acid was represented in less than 90% of the reads for that site.

The consensus sequences for the above-mentioned 46 CDSs were also compared between the two datasets. The percentages of same amino-acid conservation from S11 to S15 were calculated per protein.

Specific analysis of intra-gene diversity

For genes ehv060 and ehv452, protein blasts (BLASTP) were performed against RefSeq, the NCBI reference protein database. Hits with e-values < 0.001 were retained for phylogenetic analysis. The evolutionary histories of genes ehv060 and ehv452 were conducted in MEGA5 (Tamura et al., 2011) and inferred using the Neighbour-Joining method (Saitou and Nei, 1987). The bootstrap consensus trees inferred from 500 replicates were taken to represent the evolutionary history of the taxa analysed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed.

Acknowledgments

The authors would like to acknowledge sequencing support from the Broad Foundation and the Gordon and Betty Moore Foundation through the Virome Sequencing Project, the European Research Council (through project MINOS, No. 250254), the NSF (through grants EF0723730 and EF0949162), and the Natural Environment Research Council (through grants NE/D001455/1, NE/A509332/1 and MGF-SPG-271). The study also formed part of the NERC Oceans 2025 programme through which S.A.K. and M.J.A. were funded. Finally, the authors would like to thank the two anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Appendix A. Supplementary information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.05.020.

References

Allen, M.J., Forster, T., Schroeder, D.C., Hall, M., Roy, D., Ghazal, P., Wilson, W.H., 2006a. Locus-specific gene expression pattern suggests a unique propagation strategy for a giant algal virus. J. Virol. 80, 7699–7705.

Allen, M.J., Schroeder, D.C., Holden, M.T.G., Wilson, W.H., 2006b. Evolutionary history of the Coccolithoviridae. Mol. Biol. Evol. 23, 86–92.

Allen, M.J., Schroeder, D.C., Wilson, W.H., 2006c. Preliminary characterisation of repeat families in the genome of EhV-86, a giant algal virus that infects the marine microalga *Emiliania huxleyi*. Arch. Virol. 151, 525–535.

Allen, M.J., Martínez Martínez, J., Schroeder, D.C., Somerfield, P.J., Wilson, W.H., 2007. Use of microarrays to assess viral diversity: from genotype to phenotype. Environ. Microbiol. 9, 971–982.

Allen, M.J., Howard, J.A., Lilley, K.S., Wilson, W.H., 2008. Proteomic analysis of the EhV-86 virion. Proteome Sci. 6, 11.

Allen, M.J., Lanzén, A., Bratbak, G., 2011. Characterisation of the coccolithovirus intein. Mar. Genomics 4, 1–7.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., Lindell, D., 2011. Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature 474, 604–608.

Bowden, T.A., Jones, E.Y., Stuart, D.I., 2011. Cells under siege: viral glycoprotein interactions at the cell surface. J. Struct. Biol. 175, 120–126.

Boyer, M.l., Azza, S.D., Barrassi, L., Klose, T., Campocasso, A.l., Pagnier, I., Fournous, G., Borg, A., Robert, C., Zhang, X., Desnues, C., Henrissat, B., Rossmann, M.G., La Scola, B., Raoult, D., 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc. Natl. Acad. Sci. USA 108, 10296–10301.

Bratbak, G., Egge, J.K., Heldal, M., 1993. Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. Mar. Ecol. Prog. Ser. 93, 39–48.

Breitbart, M., 2012. Marine viruses: truth or dare. In: Carlson, C.A., Giovannoni, S.J. (Eds.), Annual Review of Marine Science, vol. 4. Annual Reviews, Palo Alto, pp. 425–448.

Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13, 278–284.

Brown, C.W., Yoder, J.A., 1994. Coccolithophorid blooms in the global ocean. J. Geophys. Res.: Oceans 99, 7467–7482.

Burkill, P.H., Archer, S.D., Robinson, C., Nightingale, P.D., Groom, S.B., Tarran, G.A., Zubkov, M.V., 2002. Dimethyl sulphide biogeochemistry within a coccolithophore bloom (DISCO): an overview. Deep-Sea Res. Part II – Top. Stud. Oceanogr. 49, 2863–2885.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., DeLong, E.F., Chisholm, S.W., 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311, 1768–1770.

Coolen, M.J.L., 2011. 7000 Years of *Emiliania huxleyi* viruses in the Black Sea. Science 333, 451–452.

Felsenstein, J., 1985. Confidence-limits on phylogenies – an approach using the bootstrap. Evolution 39, 783–791.

Frada, M., Probert, I., Allen, M.J., Wilson, W.H., de Vargas, C., 2008. The "Cheshire Cat" escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. Proc. Natl. Acad. Sci. USA 105, 15944–15949.

Fuhrman, J.A., 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399, 541–548.

Jacquet, S., Heldal, M., Iglesias-Rodriguez, D., Larsen, A., Wilson, W., Bratbak, G., 2002. Flow cytometric analysis of an *Emiliana huxleyi* bloom terminated by viral infection. Aquat. Microb. Ecol. 27, 111–124.

Jolly, C., Sattentau, Q., 2013. Attachment factors. In: Pöhlmann, S., Simmons, G. (Eds.), Viral Entry into Host Cells, vol. 790. Springer, New York, pp. 1–23.

Kimmance, S., Allen, M.J., Pagarete, A., Martínez Martínez, J., Wilson, W.H., 2014. Reduction in photosystem II efficiency during a virus-controlled *Emiliania huxleyi* bloom. Mar. Ecol. Prog. Ser. 495, 65–76.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., Kettler, G., Sullivan, M.B., Steen, R., Hess, W.R., Church, G.M., Chisholm, S.W., 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449, 83–86.

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

Martínez Martínez, J., Schroeder, D.C., Larsen, A., Bratbak, G., Wilson, W.H., 2007. Molecular dynamics of *Emiliania huxleyi* and their co-occurring viruses during two separate mesocosm studies. Appl. Environ. Microbiol. 73, 554–562.

Monier, A., Claverie, J.M., Ogata, H., 2007. Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. BMC Genomics 8 (1), 456.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.-M., Ogata, H., 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. 19, 1441–1449.

Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., Allen, M.J., 2011a. Draft genome sequence of the coccolithovirus EhV-84. Stand. Genomic Sci. 5, 1–11.

Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., Allen, M.J., 2011b. Draft genome sequence of the Coccolithovirus *Emiliania huxleyi* virus 203. J. Virol. 85, 13468–13469.

Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., Allen, M.J., 2012a. Draft genome sequence of the Coccolithovirus *Emiliania huxleyi* virus 202. J. Virol. 86, 2380–2381.

Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., Allen, M.J., 2012b. Draft genome sequence of four Coccolithoviruses: *Emiliania huxleyi* virus EhV-88, EhV-201, EhV-207, and EhV-208. J. Virol. 86, 2896–2897.

Pagarete, A., 2010. Functional Genomics of Coccolithophore Viruses. Université Paris VI, Paris.

Pagarete, A., Allen, M.J., Wilson, W., Kimmance, S., de Vargas, C., 2009. Host–virus shift of the sphingolipid pathway along an *Emiliania huxleyi* bloom: survival of the fattest. Environ. Microbiol. 11, 2840–2848.

Pagarete, A., Le Corguillé, G., Tiwari, B., Ogata, H., de Vargas, C., Wilson, W.H., Allen, M.J., 2011. Unveiling the transcriptional features associated with coccolithovirus infection of natural *Emiliania huxleyi* blooms. FEMS Microbiol. Ecol. 78, 555–564.

Pagarete, A., Lanzén, A., Puntervoll, P., Sandaa, R.A., Larsen, A., Larsen, J.B., Allen, M.J., Bratbak, G., 2012. Genomic sequence and analysis of EhV-99B1, a new Coccolithovirus from the Norwegian fjords. Intervirology 56, 60–66.

Ray, J., Dondrup, M., Modha, S., Steen, I.H., Sandaa, R.A., Clokie, M., 2012. Finding a needle in the virus metagenome haystack – micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. PLoS One 7, e34238.

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., Young, J., Aguilar, M., Claverie, J.M., Frickenhaus, S., Gonzalez, K., Herman, E.K., Lin, Y.C., Napier, J., Ogata, H., Sarno, A.F., Shmutz, J., Schroeder, D., de Vargas, C., Verret, F., von Dassow, P., Valentin, K., Van de Peer, Y., Wheeler, G., Allen, A.E., Bidle, K., Borodovsky, M., Bowler, C., Brownlee, C., Mark Cock, J., Elias, M., Gladyshev, V.N., Groth, M., Guda, C., Hadaegh, A., Debora Iglesias-Rodriguez, M., Jenkins, J., Jones, B.M., Lawson, T., Leese, F., Lindquist, E., Lobanov, A., Lomsadze, A., Malik, S.B., Marsh, M.E., Mackinder, L., Mock, T., Mueller-Roeber, B., Pagarete, A., Parker, M., Probert, I., Quesneville, H., Raines, C., Rensing, S.A., Riano-Pachon, D.M., Richier, S., Rokitta, S., Shiraiwa, Y., Soanes, D.M., van der Giezen, M., Wahlund, T.M., Williams, B., Wilson, W., Wolfe, G., Wurch, L.L., Dacks, J.B., Delwiche, C.F., Dyhrman, S.T., Glockner, G., John, U., Richards, T., Worden, A.Z., Zhang, X., Grigoriev, I.V., 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499, 209–213.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sandaa, R.-A., Short, S.M., Schroeder, D.C., 2010. Fingerprinting aquatic virus communities. In: Wilhelm, S.W., Weinbauer, M.G., Suttle, C.A. (Eds.), Manual of Aquatic Viral Ecology. ASLO, pp. 9–18.

Sandaa, R.A., 2008. Burden or benefit? Virus–host interactions in the marine environment. Res. Microbiol. 159, 374–381.

Schroeder, D.C., Oke, J., Hall, M., Malin, G., Wilson, W.H., 2003. Virus succession observed during an *Emiliania huxleyi* bloom. Appl. Environ. Microbiol. 69, 2484–2490.

Sorensen, G., Baker, A.C., Hall, M.J., Munn, C.B., Schroeder, D.C., 2009. Novel virus dynamics in an *Emiliania huxleyi* bloom. J. Plankton Res. 31, 787–791.

Stern, A., Sorek, R., 2011. The phage-host arms race: shaping the evolution of microbes. Bioessays 33, 43–51.

Suttle, C.A., 2005. Viruses in the sea. Nature 437, 356–361.

Štros, M., 2010. HMGB proteins: interactions with DNA and chromatin. Biochim. Biophys. 1799, 101–113.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Thingstad, T.F., 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol. Oceanogr. 45, 1320–1328.

Thingstad, T.F., Lignell, R., 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat. Microb. Ecol. 13, 19–27.

Vardi, A., Van Mooy, B.A.S., Fredricks, H.F., Popendorf, K.J., Ossolinski, J.E., Haramaty, L., Bidle, K.D., 2009. Viral glycosphingolipids induce lytic infection and cell death in marine phytoplankton. Science 326, 861–865.

Vardi, A., Haramaty, L., Van Mooy, B.A.S., Fredricks, H.F., Kimmance, S.A., Larsen, A., Bidle, K.D., 2012. Host–virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. Proc. Natl. Acad. Sci. USA 109, 19327–19332.

Weinbauer, M.G., Rassoulzadegan, F., 2004. Are viruses driving microbial diversification and diversity? Environ. Microbiol. 6, 1–11.

Westbroek, P., Brown, C.W., Vanbleijswijk, J., Brownlee, C., Brummer, G.J., Conte, M., Egge, J., Fernandez, E., Jordan, R., Knappertsbusch, M., Stefels, J., Veldhuis, M., Vanderwal, P., Young, J., 1993. A model system approach to biological climate forcing – the example of *Emiliania huxleyi*. Glob. Planet. Change 8, 27–46.

Wilhelm, S.W., Suttle, C.A., 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.

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

Wilson, W.H., Schroeder, D.C., Allen, M.J., Holden, M.T.G., Parkhill, J., Barrell, B.G., Churcher, C., Harnlin, N., Mungall, K., Norbertczak, H., Quail, M.A., 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.

Please cite this article as: Pagarete, A., et al., Dip in the gene pool: Metagenomic survey of natural coccolithovirus communities. Virology (2014), http://dx.doi.org/10.1016/j.virol.2014.05.020