Reconstruction of Novel Cyanobacterial Siphovirus Genomes from Mediterranean Metagenomic Fosmids

Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Inmaculada Garcia-Heredia, Ana-Belen Martin-Cuadrado, Rohit Ghai

Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, San Juan de Alicante, Alicante, Spain

Cellular metagenomes are primarily used for investigating microbial community structure and function. However, cloned fosmids from such metagenomes capture phage genome fragments that can be used as a source of phage genomes. We show that fosmid cloning from cellular metagenomes and sequencing at a high coverage is a credible alternative to constructing metaviromes and allows capturing and assembling novel, complete phage genomes. It is likely that phages recovered from cellular metagenomes are those replicating within cells during sample collection and represent “active” phages, naturally amplifying their genomic DNA and increasing chances for cloning. We describe five sets of siphoviral contigs (MEDS1, MEDS2, MEDS3, MEDS4, and MEDS5), obtained by sequencing fosmids from the cellular metagenome of the deep chlorophyll maximum in the Mediterranean. Three of these represent complete siphoviral genomes and two represent partial ones. This is the first set of phage genomes assembled directly from cellular metagenomic fosmid libraries. They exhibit low sequence similarities to one another and to known siphoviruses but are remarkably similar in overall genome architecture. We present evidence suggesting they infect picocyanobacteria, likely *Synechococcus*. Four of these sets also define a novel branch in the phylogenetic tree of phage large subunit terminases. Moreover, some of these siphoviral groups are globally distributed and abundant in the oceans, comparable to some known myoviruses and podoviruses. This suggests that, as more siphoviral genomes become available, we will be better able to assess the abundance and influence of this diverse and polyphyletic group in the marine habitat.

Phages are the most abundant biological entities on the planet (1). However, phage biomass is difficult to retrieve from oligotrophic marine waters in amounts required to construct metaviromes. In comparison to the total DNA of a single microbial cell (typically 3 to 4 Mb), the genome size of most phages is very small (~50 kb). So even if there are 10 phage particles for every microbial cell in the sample, the total phage DNA is several times less than a prokaryotic genome. Moreover, there are several methodological biases inherent in extracting phage (2) or even bacterial DNA. This makes sequencing metaviromes challenging, and sometimes it is necessary to resort to biased approaches such as linker-amplified shotgun libraries or multiple-displacement amplification (3). Recently, however, novel methods, combining FeCl₃-based viral precipitation with improved protocols for linker-amplified shotgun library methods, have been developed to enhance phage DNA recovery from complex communities (4–6). On the other hand, the use of fosmid cloning from metaviromes to recover large phage genomic fragments that are more amenable to annotation and complete genome reconstruction has also been proposed (7). Besides, metagenomic studies have now shown that when a standard metagenomic (cellular) fosmid library from a marine sample is sequenced, a significant number of the assembled DNA fragments retrieved correspond to phages (8, 9). A possible explanation for this apparent paradox is that some cells in the sample are undergoing the lytic cycle of an infecting phage, and these cells contain large amounts of phage DNA that is retrieved in the cellular fraction. Regardless, this observation does provide the opportunity to use metagenomes to rescue genomes of phages that might be replicating actively within the population at the time of sample collection.

Phage DNA containing fosmids have been found to be particularly abundant in deep chlorophyll maximum (DCM) cellular fractions that have been analyzed worldwide (8–10). The DCM is the zone of maximal photosynthetic activity in the marine habitat

(11). It is a common seasonal feature in all temperate waters (~50 to 100 m deep) and a permanent one (typically at >100 m deep) in tropical seas. At these depths, light intensity and concentrations of nutrients are ideal, sustaining a dense growth of planktonic microbes. It is likely that the high productivity of this specific marine habitat increases the ratio of infected cells compared to other situations such as in bathypelagic waters.

Cyanophages belonging to three bacteriophage families, *Podoviridae*, *Myoviridae* and *Siphoviridae*, have been isolated (12–15). A total of 17 marine myoviral (T4-like) and 6 podoviral (T7-like) genomes, infecting cyanobacteria, are available currently (16–22). Though a large number of siphoviral genomes have been described (nearly half of all sequenced viral genomes), genomes of cyanosiphoviruses have been described only recently. The first genome of a cyanosiphovirus, P-SS2, isolated from Atlantic slope waters, infecting the *Prochlorococcus* host MIT9313, was described in 2009 (23). Additionally, four more new cyanosiphoviruses (S-CBS1, S-CBS2, S-CBS3, and S-CBS-4), infecting *Synechococcus*, are now available (24). Analysis of these phages revealed genetically diverse and polyphyletic genomes, and apart from some typical structural phage genes (e.g., terminase), no host photosystem-related genes that are frequently found in podoviruses and myoviruses infecting cyanobacteria were found, except for a single high-light-inducible (*hli*) gene in S-CBS2 (24).

Received 6 September 2012 Accepted 9 November 2012

Published ahead of print 16 November 2012

Address correspondence to Francisco Rodriguez-Valera, frvalera@umh.es.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/AEM.02742-12>.

Copyright © 2013, American Society for Microbiology. All Rights Reserved.
doi:10.1128/AEM.02742-12

The high similarity of the genomic fragments from sequenced fosmids to pure-culture-isolated viruses indicates that they are derived from viral genomes. This has been shown already for all cyanomyoviruses, cyanopodoviruses (9), and even cyanosiphoviruses (24). This suggests that it is possible to assemble, identify, and define closely related virus-derived contigs from a cellular metagenome. In this work, we have examined the possibility of reconstructing larger, more complete phage genomes from metagenomic fosmids. We show that (i) it is indeed possible to obtain a complete phage genome in a fosmid from a cellular metagenome, (ii) it is possible to identify highly syntenic and identical contigs representing consensus phage genomes, and (iii) this strategy can lead to the discovery of novel representatives in the vast and largely unknown world of marine phages. In all, we describe five sets of novel siphoviral contigs obtained from a fosmid library of a cellular metagenome.

MATERIALS AND METHODS

Sampling, filtration, and fosmid library construction. The complete protocols for sample collection, processing, and fosmid library construction have already been described in detail (9). Briefly, a 30-liter seawater sample was collected from the Mediterranean DCM (depth, 50 m) off the coast of Alicante, Spain, with a Niskin bottle. The sample was sequentially filtered through 5- μ m- and 0.22- μ m-pore-size filters. The planktonic cells retained on the 0.22- μ m filter were lysed, and DNA was extracted. The fosmid genomic library was constructed as described previously (9), generating a total of 12,192 clones.

Fosmid sequencing and assembly. The fosmid library from which the fosmids were selected for sequencing contains ca. 12,000 fosmids, and the sequencing of ~1,200 fosmids from this library (using Roche 454 pyrosequencing) has been described previously (9). A total of 130 new fosmids were selected randomly from this library (excluding those sequenced before). Fosmids were divided into 10 batches, with 13 or 14 fosmids in each. DNA was extracted individually for each fosmid (as described previously) (9), and the total DNA for each batch was pooled together. Each batch contained equal amounts of DNA for each fosmid clone. All batches were sequenced in a single lane of Illumina (35-bp paired ends) using 10 tags (GATC, Konstanz, Germany). Expected coverage for each fosmid was ca. 150 $\times$, considering an average fosmid size of 50 kb. Sequence output differed from batch to batch, ranging from 71 Mb to 200 Mb (total sequence output was 1,013 Mb). The sequence of the pCCI Fos vector was screened for all reads by BLAST. All reads with more than 95% identity and a minimum alignment length of 15 bp (out of 35 bp) were screened out from the data set. All batches were assembled independently using the IDBA assembler (25). The lengths of the contigs assembled for each batch are shown in Fig. S1 in the supplemental material.

Sequence analysis. Gene prediction was done using Prodigal (26), and the predicted protein sequences were compared to the NCBI-NR database using BLAST (E value, $<1e-5$). Domain predictions in the sequences were performed using the HMMER package (27) (against the Pfam database), the HHpred server (28), and the NCBI Conserved Domains Database (29). Contig annotation was manually inspected for identifying those with a probable viral origin. To further check the assembly, contigs that belonged to known and fully sequenced microbes from the habitat were aligned to known genomes. All-versus-all comparisons of the assembled contigs with all cyanophage and several reference viral genomes (obtained from NCBI Viral RefSeq) were performed using BLAST. For making the phylogenetic trees, alignments were created using T-Coffee (30) and manually inspected and trimmed when necessary, and maximum likelihood trees were constructed using the program FastTree2 (31). Bootstrapping was performed using the Seqboot program in the PHYLIP package (32). Fragment recruitment maps for the assembled contigs were prepared from the results of BLASTN. A hit was considered if it was at least 50 bp long and had an E value of $<1e-5$. For the bar charts depicting compar-

ative recruitment across metaviromes and metagenomes, an additional criterion of $>90\%$ identity was added. CRISPR repeats and spacers were identified using the CRISPRdb database (33) and PILER-CR (34).

Nucleotide sequence accession numbers. The contigs assembled and described in this work have been submitted to the NCBI GenBank database and can be accessed using the accession numbers JX519263, JX519264, JX519265, JX519266, JX519267, and JX536274.

RESULTS AND DISCUSSION

Batch fosmid sequencing approach. In a previous study (9) (sampling details are provided in Materials and Methods), involving sequencing and assembly of nearly 1,200 fosmids of the Mediterranean DCM (using 454) in batches of ~100 each (at a low coverage of ~10 $\times$), 24 contigs >30 kb in size were obtained (only 2% of the 1,200 fosmids), indicating a very small recovery of nearly complete or complete fosmids. When we considered only the long contigs (217 contigs >10 kb), nearly 35 appeared to be clearly of viral origin (~16% of long contigs). Although 10 $\times$ coverage is considered adequate while sequencing microbial genomes derived from pure cultures using Sanger sequencing, much more coverage is usually recommended (e.g., ~30 $\times$ for 454 and ~100 $\times$ for Illumina) with next-generation sequencing methods. Moreover, factors like differences in initial amounts of DNA from different fosmids (induced by variations in copy control regulation or in DNA quantification), especially while sequencing a large number of fosmids in a single batch, also likely contribute to a low yield.

In the current work, in order to retrieve more viral sequences, we have sequenced another 130 randomly selected fosmids (not sequenced before) from the same library, but at a far higher coverage (~150 $\times$) using Illumina, in 10 batches (see Materials and Methods). The use of a small number of fosmids per tag and sequencing multiple samples in a single Illumina lane allowed very efficient assembly of fosmids (see Fig. S1 in the supplemental material). In total, we assembled 34 contigs >30 kb in size, i.e., 26% of all fosmids were recovered nearly completely compared to the previous effort where only 2% of all the sequenced fosmids were retrieved (9). To examine the quality of the assembly, once we had annotated all contigs, we identified those belonging to well-known, fully sequenced microbes and made comparisons. Some examples of assembled contigs ascribed to *Prochlorococcus* and *Synechococcus* compared to the known genomes are shown in Fig. S2 in the supplemental material. Complete synteny and high levels of sequence identity suggested that the assembled contigs were not chimeric.

General features and classification of the viral contigs. From all the assembled contigs >10 kb in size, 16 revealed typical phage genes and were selected for detailed analysis. These contigs were compared to each other, to all viral genomes obtained from NCBI Virus RefSeq, and to previously described phage contigs from the previous 454 pyrosequencing study of the same DCM sample (9) to identify the closest related sequences. Based on sequence similarities and synteny, six of these contigs, along with 26 other highly related contigs assembled before (9), could be assigned to five different groups (see Table S1 in the supplemental material). The other remaining phage-like contigs found only once (singletons) were not studied any further because of their lower reliability. The presence of several tail proteins identifies the contigs in these groups clearly as tailed-bacteriophages of the order *Caudovirales*. Without exception, all these contigs lack tail baseplate wedge proteins, critical for the formation of a contractile tail in myoviruses

FIG 1 Overall architecture of the MEDS phages. The horizontal lines under some genes indicate that these have been moved from the other end of the sequence to improve comparison across all genomes. Contigs assembled in this work are labeled in red.

(20, 35), indicating that they are not myoviruses (phages with contractile tails). On the other hand, several contigs possessed tape measure proteins, which act as a scaffold to assemble the phage tail in both myoviruses and siphoviruses but not in podoviruses. Moreover, important core genes of T7-like podoviruses were also absent in these contigs, e.g., RNA polymerase and DNA ligase (20). Besides, none of these fosmids displayed any similarity to known podoviruses or myoviruses. On the other hand, a clear, albeit distant, relationship to recently described siphoviruses infecting *Synechococcus* (S-CBS1 and S-CBS3) was detected for some (see below). Taken together, the evidence indicates that these contigs probably belong to genomes of novel siphoviruses. There does not appear to be any obvious bias for/against any phage type in this methodology as in the previous data set where 1,200 fosmids were sequenced and contigs that could be ascribed to podoviruses, myoviruses, and also siphoviruses were found (9, 24). However, fortuitously, in this smaller set of fosmids, all the assembled contigs appeared to be of siphoviral origin. These groups are discussed in more detail below.

In total, we have five groups of metagenomic viral fosmids (designated MEDS1, MEDS2, MEDS3, MEDS4, and MEDS5; MEDS for *M*editerranean *S*iphovirus), each sufficiently distinct from the other but sharing an appreciable amount of gene content within themselves to be considered cohesive, clearly related sets. All of these groups (except one) have at least one long representative sequenced in this work. An overall view of all these contigs is shown in Fig. 1. Clearly, they are strikingly similar to each other in overall structure and organization.

Terminase phylogeny. The terminase gene is essential in all head-tail phages (36), encoding the molecular motor translocating DNA into empty capsids. There is a large diversity of terminases, and they can be used to resolve different phage groups (23, 24, 37). A phylogenetic analysis of all the large terminase subunits identified in this work is shown in Fig. 2. The groups MEDS1 to MEDS4 formed a novel, distinct uncharacterized siphoviral lineage. We have designated this clade as the MEDS1-like terminases. The similarity in the large terminase subunit indicates that replication and DNA packaging are quite similar in these groups. However, since the MEDS1-like terminases form a novel clade, the structure of the ends of the genome of these phages cannot be ascertained solely by this analysis.

However, the MEDS1-like terminases do provide a clue about their putative host. As has been observed before with cyanosiphoviruses P-SS2 and S-CBS2, whose terminases are phylogenetically related to large terminase subunits encoded in *Synechococcus* genomes (24), the closest relative of the MEDS1-like terminase genes was from a degraded phage locus from the genome of *Synechococcus* CC9605 (isolated from the coast of California) (see Fig. S3 in the supplemental material). This locus codes for several hypothetical genes, a recombinase gene, a resolvase, and the phage terminase large subunit. The terminase large subunit of the MEDS5 group was the only one which shared a close relationship with known siphoviruses, S-CBS1 and S-CBS3 (Fig. 2). Based on this finding, it is predicted that the genome of the MEDS5 cluster phages have 5'-extended COS ends.

FIG 2 Phylogenetic analysis of large subunit terminases (TerL). The tree was constructed using maximum likelihood. Filled black circles on nodes indicate a bootstrap support of >90%, filled gray circles on nodes indicate a bootstrap support of >75%, and white circles on nodes indicate a bootstrap support of >60%. All MEDS phage large terminases are shown in red.

MEDS1 to MEDS4. A total of seven assembled contigs were grouped together and designated the MEDS1 group (Fig. 3). The order of genes in these contigs appears to be conserved, with distinct clusters of tail, capsid, and DNA packaging proteins seen in all. The longest contig assembled in this group was MedDCM-OCT-S13-C2, 39.2 kb in size. We found a 33-bp repeated region on the ends of this contig (indicated by arrows in Fig. 3). Closer examination revealed that the repeats actually belonged to the same gene. The bifunctional DNA polymerase gene (at the 5' end of the contig) does not have a valid start codon. The valid 5' end of this gene is seen at the opposite end of the contig, where, however, the 3' end is truncated. This indicates we have cloned and sequenced a complete phage genome from a single fosmid. Here, we will refer to this particular phage genome as MEDS1. No genes

related to host metabolism were found in the MEDS1 phage genome.

The MEDS2 group of 11 contigs (see Fig. S4 in the supplemental material) share extremely high levels of nucleotide sequence identity within themselves (frequently >95%) in large contiguous segments, and this is the most cohesive set of contigs among all MEDS groups. Similarly to MEDS1, even with such high percent identity levels, some tail proteins have regions which have no identity with each other, indicating a very high level of divergence likely due to the host-specificity-determining nature of these regions. The comparison of the two new, longer contigs (MedDCM-OCT-S14-C1 and MedDCM-OCT-S19-C1), assembled in this work, shows that the two are nearly identical, and a cluster of genes nearly 10 kb in size (outlined in black in Fig. S4) is located at

FIG 3 MEDS1 group. The complete genome of the MEDS1 phage is shown on the top (within the gray box) as a reference genome. The arrows indicate the 33-bp repeat found at both ends of the contig. Genes belonging to different functional categories are indicated in different colors (bottom legend). Also shown is the comparison of all the contigs assigned to this group with each other and to the MEDS1 phage (taken as the representative genome of this set). Although all-versus-all comparisons (using TBLASTX) were made, only selected pairwise comparisons are shown for clarity. The level of similarity between different contigs is indicated in shades of gray (see legend on the right). Contigs assembled in this work are labeled in red.

opposite ends in both contigs. This suggests that these contigs derive from concatemers and allows the reconstruction of a complete and novel phage genome of approximately 45 kb. The MEDS3 group of five contigs (see Fig. S5 in the supplemental material) appeared to be similar to the MEDS2 group, although the overall sequence similarity was not very high. The host metabolism-related genes identified in these groups, MEDS2 (carrying a phosphoadenylyl sulfate [PAPS] reductase) and MEDS3 (carrying a dCTP deaminase) (see below), were also different. The presence of nearly identical genes at opposite ends in different contigs indicates that, like in the case of MEDS2, complete phage genomes must have been retrieved.

The MEDS4 group comprises six contigs, all of which were assembled in a previous study carried out by 454 pyrosequencing of fosmid clones (9), although, in this case, a very long contig (ca. 38 kb) could be assembled. These contigs (see Fig. S5 in the supplemental material) did not show any evidence to suggest that they represent a complete phage genome.

MEDS5. This group, which clustered with known cyanosiphoviruses in the terminase phylogeny, comprises four contigs (see Fig. S6 in the supplemental material), three of which were assembled before (9), although with the Illumina sequencing carried out here a single but much longer contig (30.9 kb) could be assembled. They are similar to the *Synechococcus* CBS1 and CBS3 cyanosiphoviruses (24). We could not identify any DNA polymerase in this group or any genes with clear links to host metabolism. It is likely that the MEDS5 contigs do not represent a complete phage genome.

Host genes. Several cyanophages (both podoviruses and myoviruses) have been shown to carry genes related to the photosynthetic cycle, e.g., photosystem I and II genes and high-light-inducible genes (20, 38–42). Of all known siphoviruses, only S-CBS2 has been shown to carry an *hli* gene (24). Among the phage groups described here, MEDS2 phages carry a PAPS reductase. This enzyme is vital for the sulfur assimilation pathway widely distributed across prokaryotes. A phylogenetic analysis of the MEDS2 PAPS reductase indicated its relatedness to PAPS reductases from several phages, e.g., staphylococcus phage SA1 and enterobacterial phage lambda (see Fig. S7 in the supplemental material). Sulfur, unlike iron, is quite abundant in the marine habitat as sulfate and is not expected to be a limiting factor. In freshwater systems, though, sulfur can be limiting. It might be speculated that under intracellular conditions during infection, when phage protein components are being rapidly synthesized, in addition to host protein biosynthesis, reduced sulfur is limiting.

The MEDS3 group carried another host metabolism-related gene, a dCTP deaminase (involved in nucleotide metabolism), which is also frequently carried by phages. A phylogenetic analysis indicated close relationships of the MEDS3 dCTP deaminase with that of *Synechococcus elongatus* and cyanosiphovirus S-CBS2 (see Fig. S7 in the supplemental material).

DNA polymerase. The DNA polymerases identified in MEDS1, MEDS2, and MEDS3 phage genomes are bifunctional primases/polymerases, with both primase and polymerase activities in a single gene. These polymerases, which use only deoxyribonucleotides (43, 44), are not very common. A very restrictive

FIG 4 Comparative fragment recruitment. Number of hits obtained per kilobase of phage sequence and per gigabase of the data set for MEDS phages and all known cyanophages to the indicated metagenomes (MG) and metaviromes (MV). The phages are grouped into myoviruses, podoviruses, and siphoviruses. The MEDS phage contigs are shown as a separate group. Only an abbreviated form of the name of the MEDS phages is shown (e.g., MEDS1-S13-C2 stands for MedDCM-OCT-S13-C2, belonging to the MEDS1 group).

search for sequences with exactly the same domain architecture revealed only four sequences in the NCBI-NR database. Interestingly, one of these belonged to an EBPR (enhanced biological phosphorus removal) siphovirus (45), two to cyanobacterial DNA polymerases, and one to a planctomycete. A subsequent phylogenetic analysis showed that all bifunctional primases/polymerases identified in the MEDS phages clustered together and were related to DNA polymerases in cyanobacterial genomes of *Synechococcus elongatus* and *Cyanothece*, providing another link between the MEDS phages and cyanobacteria.

***Synechococcus* as a putative host for the MEDS phages.** Evidence from a phylogenetic analysis of terminases, host metabolism genes, and DNA polymerases suggests that these phages are unclassified siphoviruses that could infect *Synechococcus* (from the Mediterranean). To our knowledge, there are currently only two *Synechococcus* strains that have been sequenced from the Mediterranean, *Synechococcus* sp. BL107, isolated from Blanes Bay, Spain, at a 100-m depth, and *Synechococcus* sp. RCC307, isolated from the western Mediterranean, between Mallorca and Sardinia, at a 15-m depth (46). However, we did not detect the presence of any terminase-like sequences in either of these genomes, so the specific *Synechococcus* host for these phages is as yet unknown. Several genomes contain the CRISPR system as a form of immunity against invading phages (47). The presence of a phage sequence fragment as a spacer in a microbial genome is clinching evidence of past infection. We searched all available *Synechococcus* genomes for CRISPR spacers and compared them with our collection of MEDS phages. We obtained no significant hit, indicating that the current genomes are not the hosts for these phages.

Metagenomic recruitment. Recruitments done before with different cyanophages against the global ocean sampling (GOS) data set have indicated that the cyanomyoviruses are most abundant, followed by cyanopodoviruses and then cyanosiphoviruses (24). This might partially be explained by the lack of host genes that are commonly found in the former. We have analyzed the recruitment of the putative phages described here against multiple metagenomes, including four metaviromes with data sets from the Sargasso Sea, Bay of Columbia, Gulf of Mexico, and the Arctic Ocean (3). The GOS data set, a cellular metagenome, was also included, as it is the largest marine metagenomic data set available by far. In addition, the Mediterranean DCM data set, which was sequenced by 454 from the same DNA which was used to generate the fosmids described in this study (9), was included. All known cyanophages were also used for comparison. The results reveal considerable variation in the reads recruited in different data sets (Fig. 4; see also Fig. S8 in the supplemental material). Clearly the MEDS phages are as abundant as myoviruses and podoviruses in the Mediterranean DCM data set. A similar result was obtained with the Sargasso metavirome. Moreover, the MEDS phages also appear to be more abundant than all other known cyanosiphoviruses everywhere within the data sets tested.

Among the groups described here, the MEDS2 group recruited the most reads from the DCM metagenome (Fig. 4). However, all other MEDS groups recruited many more reads in the Sargasso metavirome. Representative fragment recruitment plots of the MEDS1 and MEDS2 phages against the GOS data set and the DCM metagenome, respectively, are shown in Fig. 5. The MEDS phages recruit reads along their entire lengths. Finally, the recruit-

FIG 5 Fragment recruitments of MEDS1 compared to the global ocean sampling data set (A) and MEDS2 compared to the Mediterranean deep chlorophyll maximum (DCM) metagenome (B). A representation of the contig is shown on the top, and recruited reads are drawn as horizontal lines. The y axis indicates percent identity in nucleotides to the contig (BLASTN). Only alignments >50 bp are shown. The dashed horizontal line indicates 95% identity.

ment data shown here indicate that these new phages are abundant and globally distributed.

It may be argued that some of the phage groups or contigs described here are not true phages but prophages. However, there are some indications that this is not the case. First of all, the total amount of genomic DNA we have sequenced (130 fosmids, ~6.5 Mb in all) is only two genome equivalents; the possibility of capturing 6 long fragments that are nearly complete phage genomes is very low. Moreover, none of the genes found in these phages show any similarities to known bacterial or archaeal genomes that would be expected to be found if a fragment of a prophage were to be cloned in a fosmid.

Throughout the short history of metagenomics, fosmids have been an important tool providing long, natural contigs. Even now when some remarkable *de novo* assemblies are being achieved from short, high-coverage reads, fosmids have been instrumental in providing confirmation of assembled contigs and scaffolding (48). It has been shown that fosmids are also a very useful tool for assembling phage genomes from environmental DNA, directly from purified viral biomass (7). Here we illustrate the possibility of phage genomes being retrieved from cellular metagenomes. This observation has been made by other authors as well (8). We speculate that when a cell is undergoing a phage lytic cycle, the genome of the phage is present in large amounts, at least comparable to the genome of the cell itself, and is therefore easily retrieved and cloned in fosmids. Furthermore, the phage genome, being repeated many times as identical copies during replication, provides extra amplification, increasing the chances of getting a complete genome. Here we have used a strategy based on Illumina sequencing to sequence fosmids in an efficient manner. This has allowed the recovery of complete or nearly complete fosmid sequences and, hence, of phage genomes of previously undescribed marine siphoviruses. Some groups described here seem to be abundant in the Mediterranean DCM (at least in the single sample

studied). Furthermore, some of the MEDS phages may also be abundant worldwide.

ACKNOWLEDGMENTS

This work was supported by projects MAGYK (BIO2008-02444), MICROGEN (Programa CONSOLIDER-INGENIO 2010 CDS2009-00006), and CGL2009-12651-C02-01 from the Spanish Ministerio de Ciencia e Innovación and by projects DIMEGEN (PROMETEO/2010/089) and ACOMP/2009/155 from the Generalitat Valenciana. FEDER funds supported this project. R.G. was supported by a Juan de la Cierva scholarship from the Spanish Ministerio de Ciencia e Innovación.

We thank Nikole Kimes for a critical reading of the manuscript and the anonymous reviewers for their constructive suggestions and comments.

REFERENCES

- Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* 95:6578.
- Hurwitz BL, Deng L, Poulos BT, Sullivan MB. 2012. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* doi:10.1111/j.1462-2920.2012.02836.x.
- Angly F, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. 2006. Marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi:10.1371/journal.pbio.0040368.
- Duhaime MB, Deng L, Poulos BT, Sullivan MB. 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* 14:2526–2537.
- Duhaime MB, Sullivan MB. 2012. Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology.* doi:10.1016/j.virol.2012.09.036.
- John SG, Mendez CB, Deng L, Poulos B, Kauffman AK, Kern S, Brum J, Polz MF, Boyle EA, Sullivan MB. 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* 3:195–202.
- García-Heredia I, Martín-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Antón J, Rodríguez-Valera F. 2012. Reconstructing viral genomes

- from the environment using fosmid clones: the case of haloviruses. *PLoS One* 7:e33802. doi:10.1371/journal.pone.0033802.
8. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Rodriguez-Brito B, Chisholm SW, Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496–503.
 9. Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J, Verdú M, Deschamps P, Moreira D, López-García P, Rodríguez-Valera F. 2010. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* 4:1154–1166.
 10. Ngugi DK, Antunes A, Brune A, Stingl U. 2012. Biogeography of pelagic bacterioplankton across an antagonistic temperature-salinity gradient in the Red Sea. *Mol. Ecol.* 21:388–405.
 11. Estrada M, Marrase C, Latasa M, Berdalet E, Delgado M, Riera T. 1993. Variability of deep chlorophyll maximum characteristics in the north-western Mediterranean. *Mar. Ecol. Prog. Ser.* 92:289–289.
 12. Lu J, Chen F, Hodson RE. 2001. Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl. Environ. Microbiol.* 67:3285–3290.
 13. Suttle CA. 1993. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* 92:99–109.
 14. Waterbury JB, Valois FW. 1993. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.* 59:3393–3399.
 15. Wilson WH, Joint IR, Carr NG, Mann NH. 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. *Appl. Environ. Microbiol.* 59:3736–3743.
 16. Chen F, Lu J. 2002. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* 68:2589–2594.
 17. Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch H. 2005. The genome of S-PM2, a “photosynthetic” T4-type bacteriophage that infects marine *Synechococcus* strains. *J. Bacteriol.* 187:3188–3200.
 18. Millard AD, Zwirgmaier K, Downey MJ, Mann NH, Scanlan DJ. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. *Environ. Microbiol.* 11:2370–2387.
 19. Pope WH, Weigele PR, Chang J, Pedulla ML, Ford ME, Houtz JM, Jiang W, Chiu W, Hatfull GF, Hendrix RW. 2007. Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a “horned” bacteriophage of marine *synechococcus*. *J. Mol. Biol.* 368:966–981.
 20. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* 3:e144. doi:10.1371/journal.pbio.0030144.
 21. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe JA, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* 108:E757–E764.
 22. Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF, King J, Hatfull GF, Lawrence JG, Hendrix RW. 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ. Microbiol.* 9:1675–1695.
 23. Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, Sarracino D, Chisholm SW. 2009. The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial “mobilome.” *Environ. Microbiol.* 11:2935–2951.
 24. Huang S, Wang K, Jiao N, Chen F. 2012. Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage-host genetic exchanges. *Environ. Microbiol.* 14:540–558.
 25. Peng Y, Leung HC, Yiu SM, Chin FY. 2011. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27:i94–i101.
 26. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
 27. Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14:755–763.
 28. Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33:W244–W248.
 29. Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32:W327–331.
 30. Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205–217.
 31. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi:10.1371/journal.pone.0009490.
 32. Felsenstein J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* 5:164–166.
 33. Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8:172.
 34. Edgar R. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18.
 35. Kostyuchenko VA, Leiman PG, Chipman PR, Kanamaru S, van Raaij MJ, Arisaka F, Mesyanzhinov VV, Rossmann MG. 2003. Three-dimensional structure of bacteriophage T4 baseplate. *Nat. Struct. Biol.* 10:688–693.
 36. Maniloff J, Ackermann HW. 1998. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. *Arch. Virol.* 143:2051–2063.
 37. Casjens SR, Gilcrease EB, Winn-Stapley DA, Schickmaier P, Schmiegier H, Pedulla ML, Ford ME, Houtz JM, Hatfull GF, Hendrix RW. 2005. The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J. Bacteriol.* 187:1091–1104.
 38. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86–89.
 39. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U. S. A.* 101:11013–11018.
 40. Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* 424:741–741.
 41. Millard A, Clokie MRJ, Shub DA, Mann NH. 2004. Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *Proc. Natl. Acad. Sci. U. S. A.* 101:11007.
 42. Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N, Pinter RY, Partensky F, Koonin EV, Wolf YI, Nelson N, Béjà O. 2009. Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461:258–262.
 43. Halgasova N, Mesarosova I, Bukovska G. 2012. Identification of a bifunctional primase-polymerase domain of corynephage BFK20 replication protein gp43. *Virus Res.* 163:454–460.
 44. Lipps G, Weinzierl AO, von Scheven G, Buchen C, Cramer P. 2004. Structure of a bifunctional DNA primase-polymerase. *Nat. Struct. Mol. Biol.* 11:157–162.
 45. Skennerton CT, Angly FE, Breitbart M, Bragg L, He S, McMahon KD, Hugenholtz P, Tyson GW. 2011. Phage encoded H-NS: a potential Achilles heel in the bacterial defence system. *PLoS One* 6:e20095. doi:10.1371/journal.pone.0020095.
 46. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 40:D571–D579.
 47. Jansen R, Embden JDA, Gaastra W, Schouls LM. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* 43:1565–1575.
 48. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science* 335:587–590.