Virioplankton Community Structure in Tunisian Solar Salterns

Ines Boujelben, a,b Pablo Yarza, b Cristina Almansa, a,b Judith Villamor, a Sami Maalej, a Josefina Antón, b,d and Fernando Santos b

Département des Sciences de la Vie, Faculté des Sciences de Sfax, Université de Sfax, Tunisie a; Departamento de Fisiología, Genética y Microbiología, Universidad de Alicante, Alicante, Spain b; Servicios Técnicos de Investigación (SSTI), Universidad de Alicupsica, Universidad de Alicante, Alicante, Spain c; and Instituto Multidisciplinar para el Estudio del Medio Ramón Margalef, Universidad de Alicante, Alicante, Spain d

The microbial community inhabiting Sfax solar salterns on the east coast of Tunisia has been studied by means of different molecular and culture-dependent tools that have unveiled the presence of novel microbial groups as well as a community structure different from that of other coastal hypersaline environments. We have focused on the study of the viral assemblages of these salterns and their changes along the salinity gradient and over time. Viruses from three ponds (C4, M1, and TS) encompassing salinities from moderately hypersaline to saturated (around 14, 19, and 35%, respectively) were sampled in May and October 2009 and analyzed by transmission electron microscopy (TEM) and pulsed-field gel electrophoresis (PFGE). Additionally, for all three October samples and the May TS sample, viral metagenomic DNA was cloned in fosmids, end sequenced, and analyzed. Viral concentration, as well as virus-to-cell ratios, increased along the salinity gradient, with around 10^10 virus-like particles (VLPs)/ml in close-to-saturation ponds, which represents the highest viral concentration reported so far for aquatic systems. Four distinct morphologies could be observed with TEM (spherical, tailed, spindled, and filamentous) but with various proportions in the different samples. Metagenomic analyses indicated that every pond harbored a distinct viral assemblage whose G+C content could be roughly correlated with that of the active part of the microbial community that have constituted the putative hosts. As previously reported for hypersaline metaviromes, most sequences did not have matches in the databases, although some were conserved among the Sfax metaviromes. BLASTx, BLASTp, and dinucleotide frequency analyses indicated that (i) factors additional to salinity could be structuring viral communities and (ii) every metavirome had unique gene contents and dinucleotide frequencies. Comparison with hypersaline metaviromes available in the databases indicated that the viral assemblages present in close-to-saturation environments located thousands of kilometers apart presented some common traits among them in spite of their differences regarding the putative hosts. A small core metavirome for close-to-saturation systems was found that contained 7 sequences of around 100 nucleotides (nt) whose function was not hinted at by in silico search results, although it most likely represents properties essential for hyperhalophilic viruses.

Haloviruses infect halophilic microorganisms and are found in hypersaline environments. These systems harbor the highest density of viruses reported for aquatic systems, with concentrations reaching up to 10^9 virus-like particles (VLPs) per ml and virus-to-cell ratios up to 100 (15), a number that is considerably higher than ratios in most environments (from 3 to 10 viruses per microbial cell [41]). Furthermore, haloviruses may be the main biological factor controlling prokaryotic populations in hypersaline environments, since there are not usually eukaryotic predators at salinities above 25% (reference 52 and references therein). The study of haloviruses has depended traditionally on the isolation of the virus-host systems. Almost 70 haloviruses have been isolated, mainly from extremely halophilic Archaea (5, 27, 46). Most of the virus isolates correspond to Caudovirales (head-tailed double-stranded-DNA [dsDNA]-genome viruses), although spherical, “spindle-like,” or pleomorphic morphologies have also been obtained by cultivation (5, 13, 14). However, most isolated haloviruses infect hosts that frequently constitute minor components of natural prokaryotic communities, and genomic sequences from isolated haloviruses are only rarely related to viral sequences directly retrieved from the environment. In fact, there have been no published reports on the isolation of viruses infecting two of the most conspicuous groups of extremely halophilic prokaryotes, the archaeon Haloquadratum walsbyi (19, 20, 23) and the bacterium Salinibacter ruber (3, 4, 34), and closely related microbes.

Viral communities from solar salterns (La Trinitat and Bras del Port in Spain), the Dead Sea, and hypersaline lakes (the Retba, Great Salt, and Mono lakes) have been observed by transmission electron microscopy (TEM) (15, 17, 22, 30, 38, 50), unveiling the presence of four main morphologies in variable proportions: head-tailed, icosahedral, spindle like, and filamentous. Haloviral genomic diversity, in terms of genome size distribution, has also been analyzed by pulsed-field gel electrophoresis (PFGE) (25, 31, 47, 48, 50, 51). In some cases, while several morphotypes can be observed by TEM in a given sample, only a single DNA band is observed in PFGE gels, indicating that viral “species” present in the sample harbor a very similar genome size (28, 50, 51). Sequence data from hypersaline systems (viral metagenomes from the Mono and Reba lakes and Bras del Port and San Diego solar salterns) have also shown that the uncultured halovirus community is highly diverse and not related to the cultivable fraction (28, 45, 47, 51, 53).

We used culture-independent approaches to analyze the viral assemblage of selected ponds from Sfax multipond solar salterns...
in southeast Tunisia. During the past 4 years, this thalassohaline system has been the object of many studies that have allowed the isolation and description of new halophilic bacterial and archaeal strains (7, 56), as well as the description, using the rRNA approach, of the communities inhabiting pond waters (8, 21, 55), sediments (10), magnesium-rich brines (11), and salt crystals (9).

The prokaryotic community in Sfax crystallizers (ponds where sodium chloride precipitates) is dominated by Archaea, as revealed by fluorescence in situ hybridization (FISH) (21). Haloquadratum spp., Halorubrum spp., and uncultured Halobacteriaceae-related sequences were found to be predominant in both 16S rRNA gene libraries and denaturing gradient gel electrophoresis (DGGE) profiles (8, 21). However, studies carried out by combining cell sorting of Sybr green-stained cells and sequencing of 16S rRNA genes indicated that in the crystallizers nearly 70% of Haloquadratum sequences corresponded to the low nucleic acid content (LNA) fraction of cells (35), thus probably indicating that a significant percentage of the archaeal assemblage was not very active. Compared to that of crystallizers, the species composition from medium-salinity ponds (from 10 to 20% salinity) showed more fluctuations and was dominated by bacterial populations (mostly Bacteroidetes and Proteobacteria), which exceeded archaeal representatives (21). Apart from these 16S rRNA gene sequences, classically retrieved from hypersaline environments, Narasingarao et al. (36) have recently found in Sfax salterns (and other hypersaline settings) sequences from the new haloarchaeal group “Nanohaloarchaea.”

All the above-mentioned studies were focused on Sfax prokaryotic populations, while in our study the focus was on the viral assemblage and how it changes with salinity and time. For this purpose, the viral community present in three saltern ponds of increasing salinity was sampled in May and October 2009 and analyzed by transmission electron microscopy, pulsed-field gel electrophoresis, and metagenomic investigation of selected samples.

MATERIALS AND METHODS

Sampling. Samples from ponds C4, M1, and TS were taken in May 2009 and October 2009 from the Sfax solar salterns located on the central eastern coast of Tunisia (34°39’N, 10°42’E). Salinity of the six samples (named C4May, C4Oct, M1May, M1Oct, TSMay, and TSOct) was measured in situ with a hand refractometer (Eclipse). Temperature and pH were also measured in situ with a mercury glass thermometer graduated at 0.1°C and a pH meter, respectively.

DAPI, FISH, and Sybr gold counts. Water samples were fixed with formaldehyde (7% final concentration) during 16 h at 4°C and filtered by 0.2-μm GTTP filters (Millipore). Cells were stained with DAPI (4’-6-diamidino-2-phenylindole) after in situ hybridizations with EUB338 and Arc915 probes (1) as described by Antón et al. (2). Cells were visualized and counted in an epifluorescence microscope (Leica, type DM4000B; Vashaw Scientifics Inc., Norcross, GA).

For viral counts, 10 to 100 μl from each water sample was fixed with formaldehyde (4% final concentration) for 30 min at room temperature. From 1 to 100 μl of fixed samples was filtered using 0.02-μm Anodisc 25 filters (Whatman). Sybr gold (Sigma) was used for staining viral particles, which were then visualized under an epifluorescence microscope (Leica, type DM4000B; Vashaw Scientifics Inc., Norcross, GA) and low-salinity metagenomes from San Diego (26); 3 metagenomes from Santa Pola’s CR30 crystalizer pond (28, 32, 51); 1 metagenome from Lake Retba (53); and 1 metagenome from the Dead Sea (18). The BLASTn results were automatically parsed to calculate the amount of nonredundant nucleotides from an Sfax metavirome “A” that were shared with a second selected metagenome “B” by means of meaningful BLAST high-

Transmission electron microscopy. A 10–μl sample from each viral suspension was fixed with formaldehyde (4% final concentration) for 30 min at 4°C and stained for 5 min with uranyl acetate (0.5%) on Formvar-coated carbon grids (Electron Microscopy Sciences). Virus-like particles were observed in a Jeol JEM-2010 transmission electron microscope operating at 200 kV. To see the proportion of the different viral morphotypes, 160 TEM images were taken at the same magnification and more than 900 VLPs were counted.

Viral DNA purification. Virus concentrates were mixed with equal volumes of 1.6% low-melting-point agarose (Pronadisa), dispensed into 100-μl molds, and allowed to solidify at 4°C. Agarose plugs were incubated for 1 h with 4 units of Turbo DNase (Ambion) to digest dissolved DNA. The plugs were then incubated overnight at 50°C in ESP (0.5 M EDTA [pH 9.0], 1% N-laurylsarcosine, 1 mg/ml protease K) for disruption of viral capsids. Some plugs from each sample were then used for pulsed-field gel electrophoresis (see below), and the rest were used for viral DNA extraction. For DNA extraction, plugs were washed with TE-Pefabloc (10 mM Tris-HCl, 1 mM EDTA [pH 8.0], 3 mM Pefabloc) to inactivate protease K and incubated for 15 min at 65°C. The mixture of viral DNA and melted agarose was treated with β-agarase (New England BioLabs) for 1.5 h at 42°C (1 enzyme unit per 0.1 g of melted mixture). DNA was then concentrated and purified using Microcon YM-100 centrifugal filter devices (Millipore) and checked for quality on a 1% agarose gel.

PFGE and Southern analysis. Plugs containing viral DNA were subjected to PFGE in a 1% low-electroendosmosis (LE) agarose gel (FMC) in Tris-borate-EDTA (TBE) (0.5× buffer, using a Bio-Rad (Richmond, CA) Chef-DR-III system operating at 6 V/cm, with a 1- to 25-min pulse ramp, at 14°C for 24 h. A lambda low-range DNA size ladder (New England BioLabs) served as a molecular weight marker. The gel was visualized after staining with ethidium bromide (1 μg/ml) and washed with distilled water. DNA was then transferred onto a membrane for Southern hybridization against labeled viral DNA from the pond sample TSOct. The DIG High Prime DNA labeling and detection starter kit II (Roche) was used for Southern analysis according to the manufacturer’s protocol.

Fosmid library constructions and sequencing. Four viral metagenomic libraries were constructed from samples TSMay, C4Oct, M1Oct, and TSOct. From 0.5 to 1 μg of viral DNA was end-repaired and cloned using the CopyControl HTP Fosmid library production kit with the pCC2FOS vector (Epiction) according to the manufacturer’s recommendations. Fosmid clones were induced, and fosmid ends sequenced by the GATC Biotech service (Konstanz, Germany) using pCC2FOS sequencing primers (Epiction).

Sequence analyses. (i) Assembly. High-quality and manually revised sequences were assembled to search for groups of similar viruses with the SeqMan application of the Lasergene software (DNASTAR Corporation) and the following assembly parameters: minimum match size of 50 nucleotides, minimum identity of 95%, and minimum alignment length of 100 nucleotides. An assembly with the two sequences of each fosmid (parameters: minimum match size of 50 nucleotides, minimum identity of 100%, and minimum alignment length of 100 nucleotides) was carried out to search for inverted terminal repeats in the viral genomes.

(ii) BLASTn analysis. The 4 metaviromes reported in the present work were complemented with a selection of 24 additional metagenomes found in the literature, and all possible pairwise BLASTn comparisons were carried out. Additional data sets comprised metagenomes from C4Oct, M1Oct, TSOct, and TSMay (present work); 19 high-, medium-, and low-salinity metagenomes from San Diego (26); 3 metagenomes from Santa Pola’s CR30 crystalizer pond (28, 32, 51); 1 metagenome from Lake Retba (53); and 1 metagenome from the Dead Sea (18). The BLASTn results were automatically parsed to calculate the amount of nonredundant nucleotides from an Sfax metavirome “A” that were shared with a second selected metagenome “B” by means of meaningful BLAST high-

BLASTn analysis. The 4 metaviromes reported in the present work were complemented with a selection of 24 additional metagenomes found in the literature, and all possible pairwise BLASTn comparisons were carried out. Additional data sets comprised metagenomes from C4Oct, M1Oct, TSOct, and TSMay (present work); 19 high-, medium-, and low-salinity metagenomes from San Diego (26); 3 metagenomes from Santa Pola’s CR30 crystalizer pond (28, 32, 51); 1 metagenome from Lake Retba (53); and 1 metagenome from the Dead Sea (18). The BLASTn results were automatically parsed to calculate the amount of nonredundant nucleotides from an Sfax metavirome “A” that were shared with a second selected metagenome “B” by means of meaningful BLAST high-

This research was supported by the Ministry of Environment and Sustainable Development, Tunisia. We thank the Technical Services of the Sfax Salterns for providing场地; and the Dead Sea (18). The BLASTn results were automatically parsed to calculate the amount of nonredundant nucleotides from an Sfax metavirome “A” that were shared with a second selected metagenome “B” by means of meaningful BLAST high-

This research was supported by the Ministry of Environment and Sustainable Development, Tunisia. We thank the Technical Services of the Sfax Salterns for providing
scoring segment pairs (HSPs) (i.e., a maximum E-value of 0.001, a minimum of 80% sequence identity, and a minimum alignment length of 100 positions).

(iii) BLASTx searches. In parallel with automatic annotation through the Seed platform (6), we subjected our contigs to BLASTx runs against the nonredundant nucleotide database using the on-line BLAST tool at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). For each contig, the five best matches with a maximum E-value of 0.001 were automatically retrieved and manually inspected. For a given BLAST-analyzed sequence a functional category was assigned if four of the five best matches were related to the same protein. If not, the BLAST-analyzed sequence was assigned as coding for a “conserved hypothetical protein.”

(iv) BLASTp analysis. All 4 of the metaviromes were self-compared and compared with each other to ascertain the degree of functional conservation within the same sample across the salinity gradient and during different seasons. Open reading frames (ORFs) were automatically extracted from SEED annotation (6). The same procedure and parameters as in reference 51 were used, except that matches against the query sequence were assigned as coding for a conserved hypothetical protein.

(v) Dinucleotide frequencies. The statistical procedure for dinucleotide analysis was carried out as in reference 51. Although three principal components (PCs) were enough to explain 72.58% of the variance, four PCs (82.67% variance explained) provided a better explanation of the data and were selected for subsequent clustering approaches. Seven k-mean groups were modeled and proved for significance (at 0.05) with the analysis of similarity (ANOSIM) test (43).

(vi) Hypersaline core metavirome. All known metaviromes obtained from high-salinity ponds (see references above) were classified into four distinct new data sets according to distinct sample sites: (i) the SanDiego data set, made up of the 3 metaviromes of high salinity from the San Diego solar salterns in the United States; (ii) the CR30 data set, containing the 2 metaviromes from the Santa Pola CR30 crystallizer pond in Spain; (iii) the TSall data set, formed by joining the two TSMay and TSOct Sfax metaviromes reported in the present work; and (iv) the LakeRetba data set, representing the metavirome reported for this lake in Senegal. Pairwise BLASTn comparisons of the four data sets (which were performed following standard procedures cited above) allowed us to recognize the sequences showing meaningful HSPs in all four of the data sets. The size of the calculated “core” is given in terms of the shortest sequences found in a group of matching sequences.

Nucleotide sequence accession numbers. The sequences reported in this article have been deposited in the GenBank database under accession numbers JQ793966 to JQ794444.

RESULTS AND DISCUSSION

Abundance of prokaryotic cells. The three ponds analyzed in this study were distributed along a salinity gradient, from 13.8% to salt saturation (Table 1). Ponds C4 and M1 were considered to be of medium salinity, whereas pond TS was a crystallizer and therefore had salt concentrations close to saturation. The ponds did not undergo drastic changes in salinity before and after summer, as shown by the values obtained in May and October 2009 (Table 1).

The numbers of cells per milliliter ranged from 1.09 × 10⁸ (in M1Oct) to 4.92 × 10⁸ (in M1May), which were slightly higher than DAPI counts previously reported for samples taken from the same ponds in 2007 and 2008 (21) and very close to cytometric characterizations carried out by Trigui et al. (55) from M1 and TS samples also taken in October 2009. A temporal variation in the DAPI counts was clearly observed only in the M1 pond, where cell numbers in May were 4.5 times higher than in October. These data are in agreement with the above-mentioned work by Boujbel et al. (21), who also found larger variations in the M1 prokaryotic assemblage, both in numbers and diversity (21).

The proportion of cells detected by FISH was significantly different in TS crystallizer and medium-salinity ponds (Table 1). While Archaea were not detected in C4 and M1 samples, they dominated the prokaryotic community in TS (around 60% of the total DAPI counts), as happened with most of the salt-saturated waters studied formerly (2, 33, 35). Previous studies carried out in Sfax salterns (55) demonstrated that most cells in a sample (also taken in October 2009) from the TS crystallizer pond had a high nucleic acid (HNA) content and therefore were highly active. Thus, the high proportion of FISH-detected cells in TS could correspond to the HNA fraction, a result that is in agreement with the idea that FISH detects the active fraction of a community (37). In the same study by Trigui et al. (55), the sample taken in October 2009 from pond M1 harbored a large proportion (around 64%) of low-nucleic-acid-content (LNA) cells. Accordingly, the proportion of cells detected by FISH was considerably lower than in TS (Table 1), with Bacteria as the dominant component of the community.

Abundance of virus-like particles. The numbers of VLPs (Table 1) in Sfax ponds were always higher than the number of cells, reaching more than 10⁵⁰ particles per ml in the crystallizer TS. This is the highest number of viruses reported to date for aquatic systems. Virus-to-cell ratios in TS were in the range of values obtained for other salt-saturated environments (between 42 and 100 viruses per host), like the Dead Sea, the Santa Pola crystallizer CR30, the Great Salt Lake, and the San Diego crystallizers (for a discussion of virus-to-cell ratios see reference 52).

Transmission electron microscopy of VLPs. Transmission electron microscopy was used to ascertain the different viral morphotypes (Fig. 1A) and their relative abundance. More than 900 VLPs were analyzed in total, 53% of which were in the range between 50 and 100 nm, and 41.5% had a size smaller than 50 nm; some filamentous viruses above 100 nm were also observed. Spherical viruses (including some head-tailed viruses that could have lost their tails in the TEM preparations) dominated the viral community in Sfax salterns, reaching up to 80% in C4 samples (Fig. 1B). Although most spherical VLPs in Sfax salterns showed a
clear virus-like morphology, we cannot rule out the possibility that some of these particles corresponded to DNA-containing membrane vesicles, such as the ones described as hyperthermophilic *Archea* of the order *Thermococcales* (54).

The number of spherical viruses decreased with salinity, while the proportion of tailed and “spindle-like” (also known as “lemon-shaped”) viruses was higher in the M1 and TS ponds, which may indicate that this morphotype is better adapted to higher salinities (Fig. 1B). Indeed, lemon-shaped viruses can reach up to 25% in Mediterranean crystallizers (30) or even dominate the community in the close-to-salt-saturation Lake Retba (17). As the increase in salinity is normally accompanied by an increase of archaeal members, it has been suggested that lemon viruses would infect haloarchaea (30). In fact, His1 and His2 phages, isolated from *Haloarcula hispanica*, belong to this morphotype (13, 14).

**Pulsed-field gel electrophoresis of viral assemblages.** PFGE was used to evaluate the genomic size pattern of the most abundant members of Sfax virioplankton (viral genomes below a certain concentration cannot be detected by this technique). Genomic patterns included between 2 and 5 different bands with
study (31), results indicated a faster recycling of viral genotypes in the environment for at least a couple of years (50). In the Mono Lake regions) directly retrieved from a crystallizer remained in the case, the results suggested that a viral genome (or certain genomic terranean crystallizer pond and Mono Lake (31,50). In the first communities have also been carried out with samples from a Mediterranean crystallizer pond, although some viral genomes (or certain genomic regions) could remain in the crystallizer, at least for a period of several months. Similar hybridization experiments with viral assemblages in Mediterranean salterns and the Mono Lake, the San Diego and Santa Pola salterns, and Lake Retba by using different methodologies that include cloning in plasmids and fosmids, pyrosequencing, and combinations of both approaches (references 28 and 52 and references therein).

Here, a total of 4 metaviromes (Table 2) were constructed to (i) describe the viral diversity in the different ponds, (ii) compare the viral populations found along the salinity gradient (samples C4Oct, M1Oct, and TSOct), and (iii) study the viral recycling in the crystallizer after a 5-month period (samples TSMay and TSOct). The metagenomic libraries were constructed with the viral DNA corresponding to the fraction between 23 and 50 kb observed in the PFGE analysis, purified, and directly cloned into fosmids. As explained above, most of the DNA that could be detected by PFGE was included within this size range. The 4 metaviromes constructed from Sfax samples contained a total of around 12 M bp of cloned DNA (Table 2).

**Primary sequence data, contig assembly, and statistics.** Between 76 and 92% of the total available clones yielded good quality end sequences with both forward and reverse vector primers, while the remaining clones were successfully sequenced only with either one of the primers. An assembling step was carried out in order to avoid redundancy in subsequent analyses. The lowest percentage of unique sequences (corresponding to different viral genotypes) was found in the C4Oct sample (45%), suggesting that the increase in salinity is also accompanied by an increase in viral genotype diversity.

<table>
<thead>
<tr>
<th>Samples</th>
<th>TSMay09</th>
<th>C4Oct09</th>
<th>M1Oct09</th>
<th>TSOct09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cloned DNA (Mb)*</td>
<td>3.6</td>
<td>3.8</td>
<td>3.5</td>
<td>2.5</td>
</tr>
<tr>
<td>No. of sequences</td>
<td>159</td>
<td>181</td>
<td>153</td>
<td>116</td>
</tr>
<tr>
<td>Avg length (bp)</td>
<td>392</td>
<td>653</td>
<td>630</td>
<td>513</td>
</tr>
<tr>
<td>Total sequenced (kb)</td>
<td>94</td>
<td>118</td>
<td>96</td>
<td>59.5</td>
</tr>
<tr>
<td>Total assembled (kb)</td>
<td>73.5</td>
<td>84.3</td>
<td>85</td>
<td>57</td>
</tr>
<tr>
<td>G+C content (avg) (%)</td>
<td>61</td>
<td>56</td>
<td>59</td>
<td>58</td>
</tr>
<tr>
<td>No. of sequenced fosmids</td>
<td>90</td>
<td>93</td>
<td>87</td>
<td>63</td>
</tr>
<tr>
<td>Fosmids with unique sequences (%)</td>
<td>56.6</td>
<td>45</td>
<td>63.2</td>
<td>67.7</td>
</tr>
<tr>
<td>Most abundant fosmid (%)</td>
<td>6.6</td>
<td>8.5</td>
<td>7</td>
<td>6.5</td>
</tr>
<tr>
<td>BLASTx hits (% of contigs)</td>
<td>41.94</td>
<td>40.31</td>
<td>49.64</td>
<td>55.17</td>
</tr>
<tr>
<td>Conserved hypothetical proteins</td>
<td>54.61</td>
<td>31.78</td>
<td>30.66</td>
<td>33.62</td>
</tr>
<tr>
<td>Terminases</td>
<td>3.23</td>
<td>11.63</td>
<td>11.68</td>
<td>3.45</td>
</tr>
<tr>
<td>Nucleases</td>
<td>4.03</td>
<td>5.43</td>
<td>3.65</td>
<td>3.45</td>
</tr>
<tr>
<td>Methylases</td>
<td>1.61</td>
<td>0.00</td>
<td>0.73</td>
<td>0.86</td>
</tr>
<tr>
<td>Integrases</td>
<td>0.00</td>
<td>1.55</td>
<td>1.46</td>
<td>0.00</td>
</tr>
<tr>
<td>Others</td>
<td>0.00</td>
<td>7.75</td>
<td>2.92</td>
<td>2.59</td>
</tr>
</tbody>
</table>

* Value calculated assuming 40 kb as the average insert size.

a Assembling parameters: match size of 50 nucleotides in a minimal alignment of 100 nucleotides, with a minimum identity of 95%.
(Table 2). The different viral genotypes can infect not only hosts from different prokaryotic species but also hosts from different strains in a given species. In fact, crystallizers are hypersaline systems with a low diversity and a very high microdiversity (24, 32, 39, 40, 42).

Both forward and reverse sequences from a given insert were also analyzed to search for inverted terminal repeats (ITRs) in the viral genomes. These repeats have been found in several linear and double-stranded DNA viruses that contain 5’-covalently linked terminal proteins and whose replication takes place through a protein-primed mechanism (12). In Sfax salterns, ITRs were found in 16% (TSMay), 12% (C4Oct), 14% (M1Oct), and 11.3% (TSOct) of the analyzed fosmids.

**G+C content and tentative hosts.** Since the G+C content of a given virus is normally very close to that of its host (44), the G+C content of the Sfax cloned viral sequences was determined in order to ascertain their likely hosts among the microbial populations present in the system (8, 21, 55). Surprisingly, most of the viral sequences from the whole Sfax metavirome presented a high G+C content (Table 2). In previous work carried out with the metavirome of the CR30 crystallizer from the Santa Pola salterns (51), half of the viral metagenomic sequences (from both fosmid and plasmid libraries) had a low G+C content while the rest of the sequences showed G+C values above 50%. Dinucleotide frequency values indicated that the high-G+C-content sequences most likely included viruses infecting *Salinibacter ruber* (with a G+C content of around 70%) and high-G+C-content haloarchaea. On the other side, low-G+C-content viral sequences could belong to viruses infecting the low-G+C-content *Hgr. walsbyi*, the dominant free microbe in the CR30 crystallizer (2, 16, 29). In Sfax salterns, medium-salinity ponds are dominated by high-G+C-content prokaryotes (21), and accordingly, 70% of the viral sequences analyzed from C4Oct and up to 90% of sequences from M1Oct showed a high G+C content. However, the metaviromes from TS crystallizer presented a low proportion (around 10%) of low-G+C-content viral sequences, an unexpected finding for a system that seems to be dominated by the “square” archaeon (8). A possible explanation for this result could be attributed to a very high proportion of inactive squares in the TS crystallizer. As discussed by Trigui et al. (55), 40% of the cells in the TS crystallizer had low nucleic acid content and could be considered the inactive group, 60% of which were assigned to *Hgr. walsbyi*. In the active (or HNA) cluster, up to 70% of the cells were identified as *Halorsobacterium* spp. and uncultured *Halobacteriaeae*. Viruses need to infect active cells in order to complete their life cycles, which makes active cells more susceptible to infection than inactive or dormant ones. Therefore, the unexpectedly high proportion of high-G+C-content viruses in TS could be the direct consequence of the dominance of high-G+C-hosted viruses in the active part of the community.

**Comparison of Sfax metaviromes with themselves and with other hypersaline systems by BLASTn.** BLASTn comparisons among the 4 metaviromes under study revealed that viral diversity was significantly different in each pond, although medium-salinity sequences were more similar to each other than those from crystallizer samples themselves (see Table S1 in the supplemental material). Comparisons between Sfax sequences and additional metagenomes (see Materials and Methods) using BLASTn showed that most of the C4Oct hits occurred when BLAST analysis was performed against medium-salinity metaviromes from San Diego salterns, while samples M1 and TS more frequently matched high-salinity viral sequences from CR30, San Diego, and Lake Retba sequences. Although hits between M1Oct (a medium-salinity pond) and high-salinity samples from San Diego salterns were found, the percentage was lower than in the case of the TS crystallizer. The number of hits between the TS crystallizer samples and the Lake Retba and CR30 crystallizer samples from the Santa Pola salterns (close-to-salt-saturation environments) were also higher than in the case of sample M1 (for the C4Oct sample hits were not found).

In an attempt to identify the “viral core metagenome” of all the analyzed close-to-saturation environments, we looked for the sequences common to the high-salinity ponds from the San Diego salterns, Lake Retba, and the Santa Pola crystallizer CR30 and TS samples from the Sfax salterns. The search yielded a total of 7 sequences of around 100 nucleotides each that were present in all these samples and absent from the rest of metaviromes in the databases. BLASTx analyses indicated that these sequences coded for hypothetical proteins, which would be conserved among the hypersaline metaviromes, while no motifs suggesting their function or structure could be found. The small size of the core and the small amount of information it provides was due in part to the small size of some of the metaviromes analyzed and the short length of the 454 sequences that constitute some them, which hamper ORF prediction and protein-based comparison. In any case, the presence of identical sequences in all the analyzed data sets is remarkable and suggests that this core metagenome codes for functions that are ecologically relevant in hypersaline systems.

**Annotation of Sfax metaviromes.** BLASTx analysis showed that between 40 and 55% of the contigs did not have matches in the databases. Most C4 and M1 sequences produced bacterial hits, while a high proportion of both TSMay and TSOct sequences matched viral sequences (up to 22% of total searches in TSMay; Fig. 3A). This might be biased due to the previously sequenced CR30 viral metagenomes (28, 51), as ORFs annotated as “hypothetical proteins” were turned into “conserved.” Bacterial hits decreased along the salinity gradient while archaeal hits showed the opposite trend (Fig. 3B), indicating that most of the hosts in salt-saturated environments are *Archaea*.

Terminases and nuclease were the most abundant functions found in the 4 metaviromes (Table 2). Terminases are components of the molecular motor that translocates genomic DNA into empty capsids during DNA packaging in the head-tailed viruses, order *Caudovirales* (dsDNA viruses with head-tail morphology).

**Self-to-self comparisons of predicted proteins.** We performed BLAST analysis against themselves of all 520 translated ORFs from the 4 metaviromes. The outcome has been summarized in Fig. 4A. First, certain proteins (presumably with relevant functions for halo-viruses) appeared to be conserved in several contigs from the same pond. For example, locators 283 from the TS pond, which encodes a conserved hypothetical protein, was repeated nine times in other ponds of the molecular motor that translocates genomic DNA into empty capsids during DNA packaging in the head-tailed viruses, order *Caudovirales* (dsDNA viruses with head-tail morphology).
BLASTn results that did not show any similarity at the nucleotide level between C4 and TS.

Although most of these genes have unknown functions, the genes are conserved. This most likely indicates that they code for functions relevant for haloviruses within the same pond (e.g., locator 438, a conserved hypothetical protein, is shared by 6 distinct contigs that are all from TSMay), and even between distinct ponds (e.g., locator 188, which encodes a terminase, is shared by 3 M1 contigs and 4 TS contigs). These findings might reflect the different life strategies and adaptations to certain salinity conditions, as indicated by the overlap between the M1 and C4 ponds (e.g., locator 3, a conserved hypothetical protein, and locator 82, a terminase), or the absence of genes conserved between ponds C4 and TS. Additionally, seasonal variation between TSMay and TSOct is supported by different sets of conserved genes (e.g., TSOct-exclusive locators 294 and 364; TSMay-exclusive locators 438 and 316), which indicate viral recycling phenomena.

### Dinucleotide frequency analyses

Principal component (PC) analysis based on the dinucleotide abundances (see Materials and Methods) indicated that 72.58% of the global variance could be explained by the three principal components. This value was comparable to others reported for data sets based on multiple environments (57). The three PCs have been used to draw the scatter plots shown in Fig. 4B and C. By using four PCs (82.67% variance explained) we were able to discriminate seven groups of dinucleotide usage ($k_1$ to $k_7$) in the data set delivered by $k$-means and statistically supported by an ANOSIM test. These groups have been plotted with different symbols in Fig. 4B. For comparison, all sequences have also been plotted separately in Fig. 4C and labeled according to their sample of origin. The high overlap between samples might be caused by the low number of sequences analyzed from each pond and/or by the overrepresentation of certain viral genotypes that are common to the three ponds. However, there is an incipient separation of samples from low-to-medium salinity ponds (C4 and M1) from those from high-salinity ponds (TS). Indeed, $k$-means analysis supported this observation and allowed the following deductions: (i) groups $k_1$, $k_5$, $k_6$, and $k_7$ show the predominance of ponds C4 and M1; (ii) groups $k_3$ and $k_4$ show the predominance of pond TS; (iii) group $k_2$ is a mixture of all four ponds; and (iv) group $k_3$ shows a clear predominance of TSMay over TSOct.

When we compared dinucleotide usage with functional divergences (BLASTp analyses) among samples, we observed that sets of conserved ORFs typically belonged to sequences associated with certain dinucleotide usage groups, indicating a good correlation between the two methods. Thus, ORFs specific for sample C4Oct were associated with dinucleotide groups $k_1$, $k_5$, and $k_7$; while ORFs specific for M1Oct were mostly associated with groups $k_6$, $k_1$, and $k_2$. The ORFs found only in crystallizer TS (including both May and October samples in combination) were mostly associated with $k_3$, $k_4$, and $k_6$. Within these samples a slight temporal variation was also unveiled by dinucleotide analysis since conserved ORFs from TSOct were associated with $k_4$ and $k_6$, whereas conserved ORFs from TSMay were mostly associated with $k_3$ and $k_4$.

Thus, the different tools used to analyze the 4 Sfax metaviromes discussed here have provided similar results: all of them unveil changes in the viral assemblage along the salinity gradient and reveal a closer relationship between the viral communities in ponds M1 and TS than between those in ponds C4 and TS. Since the salinities of C4 and M1 are very close to each other, and much lower than that of TS, these results indicate that other factors besides the total salt concentration are structuring viral communities along salt gradients. In spite of the differences among all the hypersaline systems whose metaviromes have been analyzed to date, it is still possible to find some common traits among them, most likely reflecting the most important strategies for virus survival in close-to-saturation systems. However, the results pre-
sented here also point to a rather wide diversity among both the microbial and viral assemblages in hypersaline systems, reflecting the fact that every individual system has its particularities, and therefore caution should be exerted when general conclusions are being drawn from data retrieved from individual settings.

ACKNOWLEDGMENTS
This work was funded by projects A/018342/08 from the Spanish Agency for International Cooperation (AECI) and CGL2009-12651-C02-01 from the Spanish Ministry of Science and Innovation (MICINN), including FEDER funds.

We thank Inmaculada Meseguer and Manuel Martínez-García for technical assistance and encouraging discussions.

REFERENCES


