Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect *Chlorella Pbi*

Lisa A. Fitzgerald  
*University of Nebraska-Lincoln, fitz918@msn.com*

Michael V. Graves  
*University of Massachusetts–Lowell, Michael_Graves@uml.edu*

Xiao Li  
*University of Massachusetts–Lowell*

Tamara Feldblyum  
*The Institute for Genomic Research, Rockville, MD*

James Hartigan  
*Agencourt Bioscience Corporation, Beverly, MA*

*See next page for additional authors*

Follow this and additional works at: [http://digitalcommons.unl.edu/virologypub](http://digitalcommons.unl.edu/virologypub)

Part of the [Virology Commons](http://digitalcommons.unl.edu/virologypub)

This Article is brought to you for free and open access by the Virology, Nebraska Center for at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Virology Papers by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect *Chlorella* Pbi

Lisa A. Fitzgerald\(^a\), Michael V. Graves\(^b\), Xiao Li\(^b\), Tamara Feldblyum\(^c\), James Hartigan\(^d\), and James L. Van Etten\(^c\, f\,*

\(^a\)Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588-0304
\(^b\)Department of Biological Sciences, University of Massachusetts–Lowell, Lowell, MA 01854
\(^c\)The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850
\(^d\)Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, MA 01915
\(^e\)Department of Plant Pathology, University of Nebraska–Lincoln, Lincoln, NE 68583-0722
\(^f\)Nebraska Center for Virology, University of Nebraska, Lincoln, NE 68588-0666

*Corresponding author. Email: jvanetten@unlnotes.unl.edu

Abstract: Viruses MT325 and FR483, members of the family Phycodnaviridae, genus *Chlorovirus*, infect the fresh water, unicellular, eukaryotic, chlorella-like green alga, *Chlorella* Pbi. The 314,335-bp genome of MT325 and the 321,240-bp genome of FR483 are the first viruses that infect *Chlorella* Pbi to have their genomes sequenced and annotated. Furthermore, these genomes are the two smallest chlorella virus genomes sequenced to date, MT325 has 331 putative protein-encoding and 10 tRNA-encoding genes and FR483 has 335 putative protein-encoding and 9 tRNA-encoding genes. The protein-encoding genes are almost evenly distributed on both strands, and intergenic space is minimal. Approximately 40% of the viral gene products resemble entries in public databases, including some that are the first of their kind to be detected in a virus. For example, these unique gene products include an aquaglyceroporin in MT325, a potassium ion transporter protein and an alkyl sulfatase in FR483, and a dTDP–glucose pyrophosphorylase in both viruses. Comparison of MT325 and FR483 protein-encoding genes with the prototype chlorella virus PBCV-1 indicates that approximately 82% of the genes are present in all three viruses.

Keywords: Chlorella viruses, Phycodnaviridae, Virus MT325, Virus FR483, Genome sequence

Introduction

Members and prospective members of the family Phycodnaviridae consist of a genetically diverse, but morphologically similar, group of large dsDNA-containing viruses (170–560 kb) that infect eukaryotic algae from both fresh and marine waters (Kang *et al.*, 2005 and Wilson *et al.*, 2005b). The phycodnaviruses are among the virioplankton recognized as important ecological elements in aqueous environments (Suttle, 2005). They, along with other viruses, play significant roles in the dynamics of algal blooms, nutrient cycling, algal community structure, and possibly gene transfer between organisms. The discovery phase of aquatic viruses, including the phycodnaviruses, is just beginning with new viruses continually being discovered as more environmental samples are examined. The phycodnaviruses, together with the poxviruses, iridoviruses, asfarviruses, and the 1.2-Mb Mimivirus are believed to have a common evolutionary ancestor (Iyer *et al.*, 2001, Iyer *et al.*, 2006 and Raoult *et al.*, 2004). All of these viruses share 9 gene products and at least two of these viral families encode an additional 41 homologous gene products (Iyer *et al.*, 2006). Collectively, these viruses are referred to as nucleocytoplasmic large DNA viruses (NCLDV) (Iyer *et al.*, 2001).

Currently, the phycodnaviruses are grouped into 6 genera, based initially on host range and subsequently supported by sequence comparison of their DNA polymerases (Wilson *et al.*, 2005b). Members of the genus *Chlorovirus* (chlorella viruses) infect fresh water algae, whereas members of the other five genera (*Coccolithovirus, Phaeovirus, Prasinovirus, Prymnesiovirus*, and *Raphdovirus*) infect marine algae. The genomes...
of members of three Phycodnaviridae genera have been sequenced (Delaroque et al., 2001, Li et al., 1997 and Wilson et al., 2005a). Comparative analysis of the three genomes has revealed more than 1000 unique genes with only 14 genes in common among the three genera (Dunigan et al., 2006). Thus the gene diversity in the phycodnaviruses is enormous.

The chlorella viruses infect certain unicellular, eukaryotic chlorella-like green algae that normally exist as endosymbionts in various protists, such as Paramecium bursaria (Kawakami and Kawakami, 1978 and Van Etten et al., 1982), Hydra viridis (Meints et al., 1981 and Van Etten et al., 1981) and Acanthocystis turfaeae (Bubeck and Pfützner, 2005). P. bursaria chlorella virus (PBCV-1) is the type member of the group and has a 331-kb genome that was sequenced about 10 years ago. The virus contains 366 putative protein-encoding genes and a polycistronic gene that encodes 11 tRNAs (Li et al., 1997).

To investigate the diversity of the chlorella viruses, we are sequencing the genomes of several additional family members. The proceeding manuscript described the sequence and annotation of the 369-kb genome from virus NY-2A and the 345-kb genome from virus AR158, which, like PBCV-1, infect Chlorella NC64A (NC64A viruses). The current manuscript describes the sequence and annotation of the 314-kb genome from virus MT325 and the 321-kb genome from virus FR483 that infect another chlorella species, Chlorella Pbi (Pbi viruses). Virus MT325 was sequenced first and its sequence revealed a gene encoding a water channel, an aquaglyceroporin (AQPV), and a gene encoding a potassium ion channel protein (Kcv) whose predicted structure differed from Kcv proteins encoded by the NC64A viruses (Gazzarrini et al., 2006). Hybridization of the MT325 aqpv and kcv genes to DNA from 47 additional Pbi viruses produced signals with 46 of them for both genes. However, a very weak signal was obtained for both genes with virus FR483 DNA, suggesting that the DNA sequences were very different for both genes, or that they might be absent in FR483. Therefore, FR483 was chosen for sequencing. As reported here, FR483 lacks both aqpv and kcv genes but it does encode a putative potassium ion transporter protein.

Results and discussion

As part of the chlorella virus genome sequencing effort, a project website has been created at http://greengene.uml.edu/. This site contains the genomic DNA sequence assemblies as well as the predicted amino acid sequences of all virus-encoded ORFs and is viewable in text format or through a graphical genome browser. This database also contains the complete annotation for each chlorella virus-encoded ORF. The supplemental data files referenced below are also available at this site.

Description of the viral genomes

The MT325 and FR483 genomes were assembled into contiguous sequences of 314,335 bp and 321,240 bp (Table 1), respectively, which agrees with their predicted sizes determined by pulse-field gel electrophoresis (unpublished results). Since the presumed hairpin termini were not sequenced, the left most nucleotide of the assembled sequences was designated 1.

To orient the MT325 and FR483 genomes relative to the prototype virus PBCV-1, plots of PBCV-1 proteins and either the MT325 or the FR483 proteins were compared. Unlike the dot plots that compared the three Chlorella NC64A-infecting viruses (Fitzgerald et al., 2007), there is only slight co-linearity between the prototype PBCV-1 and the two Pbi genomes (Fig. 1). However, there is a high degree of co-linearity between MT325 and FR483.

The average G + C content of the MT325 and FR483 genomes is 45.3% and 44.6%, respectively, higher than those in the NC64A viruses, PBCV-1, NY-2A, and AR158 (40–41%) (Fitzgerald et al., 2007).

Genes

A putative protein-coding region, or open-reading frame (ORF), was defined as a continuous stretch of DNA that translates into a polypeptide that is initiated by an ATG translation start codon and extends for 64 or more additional codons. Using this criterion, 845 ORFs were identified in the 314-kb genome of MT325 and 849 ORFs were identified in the 321-kb genome of FR483. The ORF names were based on three criteria. First, the MT325 ORF name begins with either an “M” for a major ORF (predicted to be a protein-encoding gene) or an “m” for a minor ORF (unlikely to be a protein-encoding gene). The names for the FR483 ORFs begin with either an “N” for a major ORF or an “n” for a minor ORF. Second, the ORFs were numbered consecutively in the order in which they appeared in the genome after alignment with the PBCV-1 genome. Third, the letter R or L following the ORF number indicates that the putative transcript runs either left-to-right or right-to-left, respectively. The letters “M” or “m” were cho-

Table 1
MT325 and FR483 genomes compared to the prototype chlorella virus, PBCV-1

<table>
<thead>
<tr>
<th>Genome</th>
<th>General Characteristics</th>
<th>Similarity to PBCV-1</th>
<th>Similarity to MT325</th>
<th>Similarity to FR483</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Methods</td>
<td>Genes</td>
<td>Genes</td>
<td>(%)</td>
</tr>
<tr>
<td>PBCV-1</td>
<td>330,743</td>
<td>366</td>
<td>11</td>
<td>40.0</td>
</tr>
<tr>
<td>MT325</td>
<td>314,335</td>
<td>331</td>
<td>10</td>
<td>45.3</td>
</tr>
<tr>
<td>FR483</td>
<td>321,240</td>
<td>335</td>
<td>9</td>
<td>44.6</td>
</tr>
</tbody>
</table>

a Percentage of protein-coding genes that have at least one homolog (blastp E-value ≤10^-5) in the indicated virus genome.

b Average amino acid identity (%) between homologous protein-coding genes.
sen to name the MT325 ORFs, which was the first Pbi virus genome sequenced and “N” or “n” was chosen to name the FR483 ORFs, which was the second Pbi virus genome sequenced, thus avoiding confusion between the different chlorella virus genomes. The letters serve to distinguish these virus ORFs from those that infect Chlorella NC64A [i.e., PBCV-1, NY-2A, and AR158 (designated with upper and lower case “A,” “B,” and “C,” respectively)] (Fitzgerald et al., 2007).

The 845 MT325 ORFs and 849 FR483 ORFs were classified as either major or minor ORFs based on the following criteria. When an ORF, of either the same or opposite polarity, reside within or significantly overlapped another ORF, the larger ORF was classified as a major ORF; the smaller ORFs were classified as minor. All of the ORFs were analyzed using the non-redundant, Pfam, and COG databases and ORFs predicted to encode a functional protein were classified as major. These conditions led to the prediction that 331 of the 845 MT325 ORFs and 335 of the 849 FR483 ORFs probably encode proteins.

The NC64A viruses have three types of introns in their protein-encoding genes. PBCV-1 and NY-2A have a self-splicing intron in a transcription factor TFIIS-like gene (Fitzgerald et al., 2007, Li et al., 1997 and Yamada et al., 1994). A spliceosomal-processed intron is present in the DNA polymerase gene from all three sequenced NC64A viruses (Grabherr et al., 1992 and Zhang et al., 2001) and an 81-nucleotide spliceosomal-processed intron exists in the pyrimidine dimer-specific glycosylase gene from some of the viruses (Fitzgerald et al., 2007 and Sun et al., 2000). However, initial analyses of the MT325 and FR483 genomes did not reveal any introns in their protein-encoding genes including the DNA polymerase, TFIIS, and pyrimidine dimer-specific glycosylase gene homologs. Like PBCV-1, one of the MT325 and one of the FR483 tRNA genes are predicted to contain an intron (see below). No inteins were detected in the two Pbi viruses.

GCG software was used to determine several general characteristics and properties for each ORF, including the nucleotide composition of the ORF, the A + T content of the 50 nucleotides upstream of the ORF that is likely to contain promoter elements, the frame in which the putative protein is encoded, the number of amino acids in the encoded protein, the predicted protein molecular weights, and the isoelectric points. These properties are listed in Supplements 1 and 2. Fig. 2 reports some general characteristics of both the MT325 and FR483 major ORFs, including the relative orientation of the ORFs (Figs. 2A, B). The directions in which the MT325 ORFs are encoded are slightly skewed in the reverse (54%) orientation. In contrast, the FR483 genes are evenly distributed on both strands. The average size of all putative MT325 proteins is 284 amino acids (Fig. 2C), while the average size of all putative FR483 proteins is 290 amino acids (Fig. 2D); nearly half of these proteins are in the 65 to 200-amino-acid size range. The predicted isoelectric points of the proteins are depicted in Figs. 2E, F. Despite a trend for the proteins to have a pI in the 10–11 pH range, a peak also occurs at pH 4.5. Basic proteins are probably associated with the virion where they pre-
sumably help neutralize the positively charged genomic DNA. However, the functions of the proteins that have pI's in the 4.5 range vary [e.g., the small subunit of ribonucleotide reductase (M653L and N651L), a Ser/Thr protein kinase (M221L and N223L), and a β-1,3 glucanase (M124L and N119L)]. Figs. 2G and H indicate the intergenic space between the major ORFs. In both MT325 and FR483, approximately 70% of the ORFs are separated by less than 100 nucleotides.

Annotation of the MT325 and FR483 genomes

Every ORF was compared with the non-redundant database at NCBI using the criteria described in the Materials and methods section. The Pfam and COG databases were used to identify conserved domains and proteins within the MT325 and FR483 ORFs (Supplements 1 and 2). Gene maps of the MT325 and FR483 genomes illustrate the location of the putative genes (Fig. 3) and some of the ORFs are listed by their predicted metabolic function (Table 2). With the exception of the MT325-encoded potassium ion channel protein and the aquaglyceroporin protein (see below), none of the MT325 and FR483 proteins have been tested for function. However, we assume that any of the MT325- and FR483-encoded proteins that have functional PBCV-1 homologs are probably functional.

Seventy-nine to eighty-four percent of the major ORFs are homologous between any two of the Pbi viruses and the prototype virus, PBCV-1. However, approximately 95% of the genes are homologous between MT325 and FR483. This find-

Figure 2. General characteristics of MT325 and FR483 major ORFs. Direction of the MT325 (A) and FR483 (B) ORFs. Size of the major MT325 (C) and FR483 (D) ORFs. Predicted isoelectric points of the major MT325 (E) and FR483 (F) ORFs. Intergenic space between major MT325 (G) and FR483 (H) ORFs.
ing suggests that most of the major ORFs from the chlorella viruses are essential for virus replication in nature and the genes are more conserved between the Pbi viruses. The average amino acid identity between homologous proteins from PBCV-1 and either MT325 or FR483 is 50%. There is an average of 86% amino acid identity between MT325 and FR483 homologs. MT325 and FR483 contain the 9 genes that are shared by all the NCLDV viruses (Iyer et al., 2001 and Iyer et al., 2006).

DNA replication and repair-associated proteins

MT325 and FR483 have several ORFs that are involved in either DNA replication, recombination, or repair, such as δ-DNA polymerase (M019L and N023L), superfamily III helicase (M674R and N674R), two ORFs resembling sliding clamp processivity factor (PCNA) proteins (M014R, M697L, N019R, and N687L), DNA primase (M664L and N662L), RNase H (M570L and N575L), type II DNA topoisomerase (M546R and N549R), replication factor C protein (M430L and N441L), and pyrimidine-specific glycosylase (M627L and N629L). Interestingly, both MT325 and FR483 lack an ATP-dependent ligase that is present in the NC64A viruses (Fitzgerald et al., 2007).

The MT325 and FR483 type II DNA topoisomerases have approximately 40% amino acid identity with type II topoisomerases from several eukaryotic organisms. This enzyme is ATP-dependent and functions by threading a double-stranded DNA segment through a transient double-stranded break in the DNA (Roca, 1995). PBCV-1 encodes one of the smallest known type II DNA topoisomerases of 1061 amino acids (Lavrukhin et al., 2000). It cleaves double-stranded DNAs approximately 30 times faster than the human type II DNA topoisomerase (Fortune et al., 2001). However, the smallest characterized topoisomerase II enzyme is encoded by Pbi virus CVM-1 (Dickey et al., 2005). The CVM-1 type II enzyme is 1058 amino acids in length and displays DNA cleavage activity that is approximately 50-fold faster than the human topoisomerase II. The MT325 and FR483 type II DNA topoisomerases are the same size as the CVM-1 enzyme (1058 amino acids) and have 96% and 67% amino acid identities to the CVM-1 and PBCV-1 topoisomerases, respectively.

Like PBCV-1, MT325 and FR483 encode two proteins that resemble PCNA-like proteins from other organisms. The MT325 and FR483 proteins are more similar to their homologs from other organisms than they are to each other. This finding suggests that the viral PCNA genes did not arise recently by gene duplication. PCNA interacts with proteins involved in not only DNA replication but also in DNA repair and post-replicative processing, such as DNA methylases and
<table>
<thead>
<tr>
<th>DNA Replication, Recombination and Repair</th>
<th>PBCV-1</th>
<th>MT325</th>
<th>FR483</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-DNA polymerase</td>
<td>A185R</td>
<td>913</td>
<td>M019L</td>
</tr>
<tr>
<td>Archacocaculeptic primase</td>
<td>A468R</td>
<td>443</td>
<td>M664L</td>
</tr>
<tr>
<td>PCNA</td>
<td>A193L</td>
<td>262</td>
<td>M014R</td>
</tr>
<tr>
<td>Replication factor C</td>
<td>A417L</td>
<td>429</td>
<td>M430L</td>
</tr>
<tr>
<td>RNase H</td>
<td>A399R</td>
<td>194</td>
<td>M570L</td>
</tr>
<tr>
<td>Helicase-Superfamily III</td>
<td>A459L</td>
<td>654</td>
<td>M824L</td>
</tr>
<tr>
<td>DNA Topoisomerase II</td>
<td>A583L</td>
<td>1061</td>
<td>M546R</td>
</tr>
<tr>
<td>ATP-dependent DNA ligase</td>
<td>A544R</td>
<td>298</td>
<td></td>
</tr>
<tr>
<td>ATPase (CP-loop)</td>
<td>A58455R</td>
<td>498</td>
<td>M389L</td>
</tr>
<tr>
<td>ATPase (DNA packaging)</td>
<td>A392R</td>
<td>258</td>
<td>M586L</td>
</tr>
<tr>
<td>Pyrimidine dimer-specific glycosylase</td>
<td>A505L</td>
<td>141</td>
<td>M627L</td>
</tr>
<tr>
<td>Exonuclease</td>
<td>A166R</td>
<td>268</td>
<td>M215R</td>
</tr>
</tbody>
</table>

**Transcription**

| transcription factor TFIIH                | A105L  | 290   | M139L | 312   | N135L | 312   |
| transcription factor TFIIH               | A525L  | 270   | M266L | 273   | N271R | 273   |
| transcription factor TFIIH               | A125L  | 180   | M163L | 180   | N165L | 180   |
| VLF2-type transcription factor           | A482R  | 215   | M635R | 215   | N640R | 215   |
| Superfamily II helicase                  | A153R  | 459   | M201R | 454   | N206R | 454   |
| RNA polymerase                           | A241R  | 725   | M252L | 715   | N227L | 712   |
| histone H3, Lys 7                        | A363R  | 811   | M372R | 1131  | N383R | 1124  |
| DNA synthesis                           | A103R  | 230   | M333R | 319   | N328R | 319   |
| RNA triphosphatase                      | A449R  | 193   | M399L | 183   | N416L | 183   |
| Histone H3, Lys 7                       | A012L  | 119   | M727L | 119   | N719L | 119   |
| DNA methyltransferase                    | A189192R| 1299  | M015L | 1286  | N020L | 1290  |
| SWI/SNF chromatin remodeling complex    | A548L  | 548   | M527L | 459   | N277L | 555   |
| SWI/SNF helicase                        | A464R  | 275   | M672L | 266   | N670L | 266   |
| RNase III                               | A200R  | 118   | M101L | 119   | N014L | 119   |

**Sugar Manipulation**

<table>
<thead>
<tr>
<th>PBCV-1</th>
<th>MT325</th>
<th>FR483</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-lactate dehydrogenase</td>
<td>A053R</td>
<td>363</td>
</tr>
<tr>
<td>GTP-D-mannose dehydrogenase</td>
<td>A111R</td>
<td>345</td>
</tr>
<tr>
<td>Fructose synthase</td>
<td>A295L</td>
<td>317</td>
</tr>
<tr>
<td>UDP-glucose</td>
<td>A609L</td>
<td>389</td>
</tr>
<tr>
<td>6-deoxyglucose</td>
<td>A100R</td>
<td>595</td>
</tr>
<tr>
<td>Glucosamine synthetase</td>
<td>A099R</td>
<td>568</td>
</tr>
<tr>
<td>Cellulase precursor</td>
<td>M354R</td>
<td>438</td>
</tr>
<tr>
<td>Glucose oxidase</td>
<td>A064R</td>
<td>638</td>
</tr>
<tr>
<td>M663L, 234</td>
<td>456</td>
<td>M069R</td>
</tr>
<tr>
<td>N663L, 234</td>
<td>456</td>
<td>M069R</td>
</tr>
<tr>
<td>N663L, 234</td>
<td>456</td>
<td>M069R</td>
</tr>
</tbody>
</table>
DNA transposases (Warbrick, 2000). Because the chlorella viruses encode proteins involved in both DNA repair and DNA methylation, their two PCNAs may serve different functions in their respective viral life cycles.

**Transcription-associated proteins**

No recognizable RNA polymerase or RNA polymerase components have been detected in any of the chlorella viruses that have been sequenced to date, including MT325 and FR483. This observation supports the idea that infectious viral DNAs are targeted to the nucleus and that host RNA polymerase(s) initiates viral transcription, possibly in conjunction with virion-packaged transcription factors. MT325 and FR483 encode at least four putative transcription factor-like elements: TFIIB (M139L and N135L), TFIID (M266R and N271R), TFIIS (M163L and N165L), and VLT2-type transcription factor (M635R and N640R). However, none of these factors are packaged in the PBCV-1 virion (Dunigan et al., manuscript in preparation) and are unlikely to be packaged in the MT325 or FR483 virions. MT325 and FR483 encode two proteins that are involved in creating a mRNA cap structure, a mRNA guanylyltransferase (M133R and N128R) and an RNA triphosphatase (M399L and N416L). These two viruses also encode an RNase III (M672L and N670L) that presumably is involved in the processing of viral mRNAs and/or tRNAs.

In the immediate-early phase of infection, the host is reprogrammed to transcribe viral RNAs, which in the case of PBCV-1, begins 5–10 min p.i. (Schuster et al., 1986). It is not known how this process occurs, but histone methylation may be involved in inhibiting host transcription. PBCV-1 encodes a 119-amino acid protein with a SET domain (named vSET) that di-methylates Lys27 in histone 3 (Manzur et al., 2003). vSET is packaged in the PBCV-1 virion and accumulating evidence indicates that vSET may be involved in repressing host transcription after PBCV-1 infection (Manzur et al., manuscript in preparation). MT325 and FR483 each contain a PBCV-1 vSET gene homolog (M727L and N719L). In addition to the histone methyltransferase, MT325 and FR483 encode a putative SWE/NF family helicase (M272L and N277L) and a SWI/SNF chromatin-remodeling complex subunit, OSA2 (M015L and N020L). Both proteins are also implicated in chromatin remodeling (e.g., Kim and Clark, 2002).

Finally, MT325 and FR483, as well as all the chlorella viruses, encode a putative cytidine deaminase (M010L and N014L). This observation suggests that either some of the viral transcripts or host transcripts may undergo some post-transcriptional editing (Gerber and Keller, 2001).
proteins have an ABC transporter family signature and two ATP/GTP-binding site motifs.

MT325 contains two ORFs that encode proteins homologous to EF-3 from chlorella viruses, whereas the FR483 genome contains one EF-3-like ORF. The 901-amino-acid MT325 and FR483 EFs (M742R and N733R) are similar in size and have 65% amino acid identity to their PBCV-1 homolog. MT325 also encodes a smaller EF (M375R, 179 amino acids) that is similar to the N-terminus of a chlorella virus CVK2 EF-3 homolog (39% amino acid identity) (Yamada et al., 1993). This small ORF does not resemble any other proteins in the databases, including the MT325 M742R ORF and its function is unknown.

MT325 and FR483 have genes that encode proteins similar to those involved in post-translational modification, including prolyl-4-hydroxylase (M401R and N418R), protein kinases (see below), and glycosyltransferases (see below). MT325 and FR483 also encode a protein disulfide isomerase (M403R and N420R), a SKP-1 protein (M807R and N799R) and a thiol oxidoreductase (M670R and N668R). Additionally, the two viruses encode proteins involved in protein degradation including an ubiquitin C-terminal hydrolase (M137L and N132L), a ring finger ubiquitin ligase (M629L and N633L), and a Zn metallopeptidase (M496L and N506L).

tRNAs

The MT325 and FR483 genomes were analyzed for tRNAs using the tRNAscan-SE program (Lowe and Eddy, 1997). MT325 is predicted to encode 10 tRNAs: 2 for Asn and 1 each for Arg, Gly, Ile, Leu, Lys, Phe, Thr, and Tyr (Table 3). These 10 tRNAs are clustered in a region of the MT325 genome, nucleotide sequence 134,513 to 135,567. FR483 is predicted to encode 9 tRNAs: 2 for Asn and 1 each for Arg, Gly, Ile, Leu, Lys, Thr, and Tyr (Table 3). These 9 tRNAs are also clustered in the FR483 genome, nucleotide sequence 141,848 to 142,809. Presumably, the tRNAs are transcribed as a large polycistronic RNA and processed via intermediates to mature RNAs as they are in chlorella virus CVK2 (Nishida et al., 1999). MT325 and FR483, with the exception of tRNA-Phe in MT325, code for the same tRNAs which are present in the same order. Only seven of the eleven tRNAs encoded by PBCV-1 are found in both MT325 and FR483. Therefore, three unique tRNAs are encoded by the Pbi viruses; tRNA-Gly and tRNA-Phe in MT325 and FR483 and one tRNA unique to MT325, tRNA-Phe (Table 3). Even though the orientation of the tRNA genes is the same in all five sequenced chlorella genomes, their order varies between the viruses. None of the tRNAs have a CCA sequence at the 3′ end of the acceptor stem. Typically, these three nucleotides are added post-transcriptionally.

One tRNA, tRNA-Tyr, encoded by both the MT325 and FR483 genomes, contains a predicted 13-nucleotide intron from nucleotide 135,213 to 135,225 and 142,453 to 142,465, respectively. The insertion of a small intron in the tyrosine tRNA (anti-codon GTA) also occurs in NC64A viruses, PBCV-1 and NY-2A; however, tRNA-Tyr is absent from the AR158 genome (Fitzgerald et al., 2007). Codon usage analyses of a few viral-encoded proteins indicate a strong correlation between the abundance of the viral-encoded tRNAs and their usage in viral proteins.

Nucleotide metabolism

MT325 and FR483 encode eleven enzymes involved in nucleotide metabolism. These enzymes are important given that the DNA concentration in PBCV-1-infected cells increases at least four-fold following infection (Van Etten et al., 1984). Therefore, large quantities of dNTPs must be synthesized to support viral DNA replication. MT325 and FR483 encode the small (M653L and N651L) and large (M777L and M766L) subunits of ribonucleotide reductase, three thioredoxins (M445L, M448L, M449L, N453L, N457L, and N458L), two glutaredoxins (M241L, M423R, N241L, and N436R), dUTP pyrophosphatase (M264L and N269L), deoxycytidylate (dCMP) deaminase (M530L and N546L), and thymidylate synthase X (M034L and N034L). Interestingly, the two Pbi viruses MT325 and FR483 lack the gene for aspartate transcarboxypeptidase.

### Table 3

Comparison of the PBCV-1, MT325, and FR483 tRNA genes

<table>
<thead>
<tr>
<th>tRNA</th>
<th>Anticodon</th>
<th>PBCV-1</th>
<th>MT325</th>
<th>FR483</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>tRNA #</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>Ile</td>
<td>TAT</td>
<td>2</td>
<td>165,022</td>
<td>165,094</td>
</tr>
<tr>
<td>Leu</td>
<td>TAA</td>
<td>3</td>
<td>165,118</td>
<td>165,200</td>
</tr>
<tr>
<td>Phe</td>
<td>GAA</td>
<td>3</td>
<td>165,718</td>
<td>134,790</td>
</tr>
<tr>
<td>Arg</td>
<td>TCT</td>
<td>10</td>
<td>165,777</td>
<td>165,851</td>
</tr>
<tr>
<td>Gly</td>
<td>TCC</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Asn</td>
<td>GCT</td>
<td>4</td>
<td>165,224</td>
<td>165,297</td>
</tr>
<tr>
<td>Asn</td>
<td>GAT</td>
<td>6</td>
<td>165,398</td>
<td>165,471</td>
</tr>
<tr>
<td>Tyr</td>
<td>GTA</td>
<td>8</td>
<td>165,572</td>
<td>165,658</td>
</tr>
<tr>
<td>Lys</td>
<td>CTT</td>
<td>7</td>
<td>165,547</td>
<td>165,547</td>
</tr>
<tr>
<td>Thr</td>
<td>CGT</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Leu</td>
<td>CAA</td>
<td>1</td>
<td>164,916</td>
<td>164,996</td>
</tr>
<tr>
<td>Lys</td>
<td>CTT</td>
<td>5</td>
<td>165,301</td>
<td>165,373</td>
</tr>
<tr>
<td>Lys</td>
<td>TTT</td>
<td>9</td>
<td>165,681</td>
<td>165,753</td>
</tr>
<tr>
<td>Val</td>
<td>AAC</td>
<td>11</td>
<td>165,885</td>
<td>165,957</td>
</tr>
</tbody>
</table>

* a Order of the tRNA genes in the genome.
* b tRNA gene contains an intron.
bamylase that is present in the three NC64A viruses (Fitzgerald et al., 2007).

Two MT325 and FR483-encoded enzymes, dUTP pyrophosphatase and dCMP deaminase, produce dUMP, the substrate for thymidylate synthetase. The chlorella viruses, including MT325 and FR483, lack a traditional thymidylate synthetase A. Instead, they encode a protein that is a member of a newly recognized family of flavin-dependent thymidylate synthetases, ThyX (Graziani et al., 2004 and Myllykallio et al., 2002).

Unlike the NC64A viruses that encode a single glutaredoxin and thioredoxin, MT325 and FR483 each have two glutaredoxins, with 28–30% amino acid identity between the two glutaredoxins encoded by the same virus; however, M423R and N436R have 93% amino acid identity while M241L and N241L have 95% amino acid identity to one another. The two Pbi viruses each encode three thioredoxins with 26–34% amino acid identities between the thioredoxins encoded by the same virus. However, there is 98% amino acid identity between MT325 and FR483 thioredoxin homologs.

**Protein kinases, phosphatases, and channel proteins**

Both MT325 and FR483 encode several Ser/Thr protein kinases (Table 2) and a protein that resembles a dual-specificity phosphatase (M313L and N318L). The large number of viral-encoded proteins involved in phosphorylation/dephosphorylation suggests that they are involved in one or more signal transduction pathways that are vital for viral replication.

The algal viruses are the first viruses to encode K⁺ channel proteins (called Kcv) (Kang et al., 2004). The MT325 Kcv (M183R) conducts a current when expressed in *Xenopus* oocytes (Gazzarrini et al., 2006). It is interesting, however, that the predicted structure of the 95-amino acid MT325 Kcv differs from the 94-amino acid PBCV-1 Kcv (Gazzarrini et al., 2006), yet both are capable of conducting currents in heterologous systems. FR483 lacks a Kcv homolog. However, FR483 encodes a putative potassium ion transporter protein (N110R).

MT325 is unique among the chlorella viruses because it also encodes a functional aquaglyceroporin channel (called AQPV, M030R) (Gazzarrini et al., 2006). Genes encoding AQPV are absent in the three sequenced NC64A viruses and in Pbi virus FR483. MT325 also encodes an 871-amino-acid Ca²⁺-transporting ATPase (M535L). To date, a gene encoding this protein has only been detected in one NC64A virus, AR158 (Fitzgerald et al., 2007).

**Sugar- and lipid-manipulating proteins**

MT325, and FR483 encode several proteins that have high identities to enzymes involved in manipulating sugars, synthesizing polysaccharides, or transferring sugars to proteins (Yamada et al., 2006). Therefore, phycodnaviruses are unusual in that, of the five chlorella viruses sequenced, they all encode enzymes involved in sugar metabolism, but not always the same enzymes. Like PBCV-1, MT325 and FR483 each encode three enzymes, glucosamine synthetase (M037R and N035R), UDP-glucose 6-dehydrogenase (M719L and N712L), and hyaluronan synthase (M128R and N124R), which are involved in the synthesis of hyaluronan, a linear polysaccharide composed of alternating β-1,4 glucuronic acid and β-1,3-N-acetylglucosamine residues (DeAngelis et al., 1997 and Landstein et al., 1998). MT325 and FR483 also have four putative glycosyltransferases (Table 2) that are presumably involved in glycosylating their major capsid proteins.

MT325 and FR483 encode one enzyme, dTDP–glucose pyrophosphorylase (M174L and N177L), that is absent in the three NC64A viruses. MT325 and FR483 lack the two enzymes involved in converting GDP-d-mannose to GDP-1-fucose, GDP-d-mannose 4-dehydratase and GDP-4-keto-6-deoxy-d-mannose epimerase/reductase that exist in the three NC64A viruses (Fitzgerald et al., 2007). Furthermore, a d-lactate dehydrogenase is encoded by MT325 (M026L) but is absent in FR483.

Finally, MT325 and FR483 encode 3 enzymes that are predicted to be involved in lipid metabolism, N-acetyl-transf erase (M758R and N747R), patatin phospholipase (M219L and N222L) and lipoprotein lipase (M564L and N570L). These same enzymes are encoded by PBCV-1; however, none of them have been investigated for enzymatic activity. FR483 encodes an alky sulfatase (N003L), which is not present in the other four sequenced chlorella virus genomes.

**Cell wall-degrading enzymes**

MT325 has five ORFs and FR483 has four ORFs that may be involved in degrading *Chlorella* Pbi cell walls either during virus infection or virus release. These proteins include two chitinases (M085R and M791R) in MT325 and one chitinase (N779R) in FR483, a chitosanase (M091R and N087R), a β-1,3-glucanse (M124L and N199L), and a novel polysaccharide lyase that cleaves chains of either β- or α-1,4 linked glucuronic acids (M289R and N293R). All of the cell wall-degrading enzymes have functional homologs in PBCV-1 (Yamada et al., 2006). Unlike most algae, *Chlorella* Pbi contains chitin in its cell wall (Kapaun et al., 1992 and Kapaun and Reisser, 1995).

**Restriction–modification enzymes**

Chlorella viruses contain different levels of 5-methylcytosine (5 mC) and N⁰-methyladenine (6 mA) in their genomes (Van Etten et al., 1991). Therefore, it is not surprising that these viruses encode 5 mC and 6mA DNA methyltransferases, e.g., NY-2A and PBCV-1 encode 18 and 5 DNA methyltransferases, respectively (Fitzgerald et al., 2007). MT325 and FR483 only encode a few DNA methyltransferases. MT325 encodes one methyltransferase (M359L), which methylates cytosines. FR483 encodes three methyltransferases; one (N751L) is a 6 mA methyltransferase and two (N369L and N748L) are 5 mC methyltransferases.

**Integration and transposition enzymes**

Homing endonucleases are rare DNA-cleaving enzymes that are typically encoded by introns and inteins and cleave DNA infrequently. Homing endonucleases are classified into
four families (Belfort and Roberts, 1997). MT325 encodes 14 homing endonucleases; ten are members of the GIY-YIG family and four are members of the HNH family. FR483 encodes 15 homing endonucleases; nine are members of the GIY-YIG family and six are members of the HNH family (Table 2). It is unknown if the putative MT325- and FR483-encoded homing endonucleases are functional and/or if they have an essential role in the viral life cycle. Unlike the NC64A viruses, the two Pbi viruses do not encode transposases.

Polyamine biosynthetic enzymes

PBCV-1 was the first virus to encode polyamine biosynthetic enzymes, including two pathways for the synthesis of putrescine. Three of the four PBCV-1 enzymes, which are known to be functional (Van Etten, 2003), are also encoded by MT325 and FR483. These enzymes are an ornithine/arginine decarboxylase (ODC) (M307L and N312L), an N-carbamoylputrescine amidohydrolase (M103L and N095L), and a homospermidine synthase (M233L and N232L). However, agmatine iminohydrolase (M766L) is only encoded by MT325. ODC catalyzes the first and rate-limiting step in polyamine biosynthesis, the decarboxylation of ornithine to putrescine (Davis et al., 1992). The PBCV-1 ODC is the smallest ODC characterized to date (372 amino acids) (Morehead et al., 2002). Unexpectedly, the PBCV-1 genome decarboxylates arginine more efficiently than ornithine (Shah et al., 2004). MT325 and FR483 each encode a 372-amino-acid ODC that has 63% amino acid identity to its PBCV-1 homolog. The product of arginine decarboxylation is agmatine. PBCV-1 and MT325 each encode two enzymes, agmatine iminohydrolase and N-carbamoylputrescine amidohydrolase, which convert agmatine to putrescine (Nakada et al., 2001) (Baumann et al., submitted for publication). FR483 encodes N-carbamoylputrescine amidohydrolase, but lacks agmatine iminohydrolase.

Homospermidine synthase synthesizes homospermidine, a rare polyamine, from two molecules of putrescine (Kaiser et al., 1999). The MT325 and FR483 homospermidine synthase homologs have a 69% amino acid identity to the functional PBCV-1-encoded enzyme.

Both PBCV-1 and the two Pbi viruses, MT325 and FR483, encode two other putative enzymes involved in amine metabolism, monoamine oxidase (M283R and N289R) and histidine decarboxylase (M601L and N600L). The finding that all of the chlorella viruses sequenced to date encode five of these six proteins suggests that they must serve an important role(s) in the viral life cycle.

Miscellaneous proteins

MT325 and FR483 also encode several other putative proteins, including a Cu/Zn superoxide dismutase (M099R and N091R), an amidase (M101R and N093R), a fibronectin-binding protein (M789L and N777L), and an ABC transporter protein (M404R and N421R). The Cu/Zn superoxide dismutases have 76% amino acid identity with a PBCV-1-encoded homolog. This enzyme functions to convert superoxide radicals into molecular oxygen and hydrogen peroxide (Bannister et al., 1987). Presumably, the MT325- and FR483-encoded enzyme reduces light-induced superoxide accumulation. Finally, MT325 and FR483 each encode an enzyme (M177L and N180L) that has some similarity to an enzyme from several Mycobacterium species that methylates rhamnose (Jeevarajah et al., 2002).

MT325 and FR483 structural proteins

MT325 ORF M463L and FR483 ORF N470L have the highest amino acid identities (96–99%) and are approximately the same size as the characterized major coat protein, Vp49, from Pbi virus CVG-1 (Plugge et al., 1999). Therefore, we assume that M463L and N470L are the MT325 and FR483 major capsid proteins, respectively. Four additional MT325 ORFs (M078R, M269L, M381R, and M748L) and 5 FR483 ORFs (N074R, N254L, N274L, N395R, and N373L) have 31–43% amino acid identities to Vp49 from CVG-1.

Identification of gene families

A total of 68 of the MT325 ORFs resemble 1 or more other MT325 ORFs based on a blastp search with an E-value of less than 10−10, suggesting that they might be either gene families or gene duplications. This number is somewhat misleading since some of these ORFs are grouped as families because they contain a common conserved domain, e.g., ankryin repeats or a PAPK repeat, even though the amino acid sequence similarity of the rest of the protein is small. A total of 13 families have two members, 3 families have three members, 1 family has five members, 1 family has six members, 2 families have seven members, and 1 family has eight members.

A similar analysis indicates that 68 of the FR483 ORFs resemble 1 or more other FR483 ORFs. A total of 10 families have two members, 3 families have three members, 2 families have four members, 1 family has six members, 1 family has seven members, 1 family has eight members, and 1 family has ten members.

Conclusions

The 314,335-bp MT325 genome and the 321,240-bp FR483 genome, the smallest chlorella viral genomes sequenced to date and the first viruses to be sequenced that infect Chlorella Pbi, are predicted to encode 331 and 335 proteins as well as 10 and 9 tRNAs, respectively. The putative protein-coding genes are relatively evenly distributed on both strands and intergenic space is minimal. Approximately 40% of the gene products have been identified; some have prokaryotic characteristics whereas other resemble eukaryotic proteins. Approximately 82% of these Pbi protein-coding genes have homologs in PBCV-1, suggesting that these proteins are important in viral replication. However, some of the gene products are unique to MT325 (e.g., aquaglyceroporin), unique to FR483 (e.g., alkyl sulfatase and potassium ion transporter) or unique to the Pbi viruses (e.g., dTDP-glucose pyrophosphorylase). Finally, genes encoding several proteins in the three NC64A viruses are absent in the two Pbi viruses in-
including, an ATP-dependent DNA ligase, aspartate transcarbamylase, GDP-d-mannose dehydratase, fucose synthase, glycero-phosphoryl diesterase and lysophospholipase.

Materials and methods

Viral DNA isolation and sequencing

Plaque-forming viruses MT325 and FR483 were isolated from fresh-water samples collected in Montana state (August, 1996) and France (June, 1997), respectively. The MT325 and FR483 host, *Chlorella* Pbi, was grown on FES medium (Reisser et al., 1986). The MT325 and FR483 viruses were produced, purified, and the viral DNAs were isolated using methods and protocols developed for PBCV-1 (Van Etten et al., 1981, Van Etten et al., 1983 and Van Etten et al., 1991). MT325 genomic DNA was sequenced to 8-fold coverage and assembled at The Institute for Genomic Research (TIGR). The sequence of the FR483 genome was determined and assembled at Agencourt Biosciences Co., Beverly, MA. The FR483 genome was sequenced to a 8-fold coverage with a minimum quality value of 20 per base.

Genomic sequence analysis

A potential protein-encoding region was defined as a continuous stretch of DNA that translated into a polypeptide initiated by an ATG translation start codon and extended for 64 or more codons using the standard genetic code. The ORF Finder program (http://bioinformatics.org/sms/orf_find.html) was used to identify all potential ORFs that met this criterion. The ORFs were numbered consecutively starting at the beginning of the genome (as determined by alignment with the PBCV-1 genome). The letter R or L following the number indicates that the orientation of the putative ORF is either left-to-right or right-to-left, respectively.

Dot plots of the virus major ORFs were created to determine the orientation of the MT325 and FR483 genomes relative to the PBCV-1 genome. Every major ORF was individually plotted against the PBCV-1 major ORFs using blastp (protein vs. protein). Similarities between the two ORFs with E-values of < 10^{-3} are presented. Transfer RNA genes were identified using the tRNAscan-SE program developed by Lowe and Eddy at Washington University School of Medicine (Lowe and Eddy, 1997). Gene families were identified when a major ORF had an E-value of less than 10^{-10} to another ORF within the same genome.

Analysis with public databases

Each ORF identified was used in a search for homologs using the protein–protein BLAST (blastp) program (Altschul et al., 1990) against the non-redundant (NR) protein databases at NCBI. The criterion used to search the NR database was as follows: Scoring matrix = blosum62. Each putative identified ORF was scanned for potential functional attributes using Pfam version 18.0 (Finn et al., 2006). Every identified ORF was additionally scanned to determine if it belonged to a particular COG. In each of the analyses the top 10 results were recorded regardless of the E-values.

Nucleotide sequence accession number

The MT325 and FR483 DNA sequences have been deposited in the GenBank database (accession numbers DQ491001 and DQ890022, respectively) and the sequences can also be found at http://greengene.uml.edu/

Acknowledgments

We thank James Gurnon for preparing the MT325 DNA and Ming Kang for preparing the FR483 DNA. This investigation was supported in part by National Science Foundation grant EF-0333197 (M.G. and J.V.E.), by National institutes of Health grant GM32441 (J.V.E.) and by the Center of Biomedical Research Excellence program of the National Center for Research Resources Grant P20-RR15635 (J.V.E.).

Supplementary data

Supplementary data associated with this article is archived in this repository as 4 separate files: Appendices A–D. Each document, in spreadsheet format, shows Gene Name, Genome Position, A.A. length, Peptide MW, pl, CDD Hit Number, COGs, COG Definition, Bit Score, E-value, % Identity, % Positive, Query from-to, Hit from-to, BLASTp Hit Number, Hit Accession, BLASTp Definition, Bit Score, E-value, % Identity, % Positive, Query from-to, and Hit from-to.

Appendix A: Gene names m002R through m843L
Appendix B: Gene names M001L through M807R
Appendix C: Gene names n001L through n849R
Appendix D: Gene names N003L through N847R

References


