Available online at www.sciencedirect.com
ScienceDirect
Virology 358 (2007) 472–484
VIROLOGY
www.elsevier.com/locate/yviro

Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect *Chlorella* NC64A

Lisa A. Fitzgerald [a], Michael V. Graves [b], Xiao Li [b], Tamara Feldblyum [c], William C. Nierman [c,d], James L. Van Etten [e,*]

[a] *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304, USA*
[b] *Department of Biological Sciences, University of Massachusetts-Lowell, Lowell, MA 01854, USA*
[c] *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA*
[d] *The George Washington University School of Medicine, Department of Biochemistry and Molecular Biology, Washington, DC 20037, USA*
[e] *Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE 68583-0722 and Nebraska Center for Virology, University of Nebraska-Lincoln, Lincoln, NE 68588-0666, USA*

Received 27 July 2006; returned to author for revision 17 August 2006; accepted 23 August 2006
Available online 4 October 2006

Abstract

Viruses NY-2A and AR158, members of the family *Phycodnaviridae*, genus *Chlorovirus*, infect the fresh water, unicellular, eukaryotic, chlorella-like green alga, *Chlorella* NC64A. The 368,683-bp genome of NY-2A and the 344,690-bp genome of AR158 are the two largest chlorella virus genomes sequenced to date; NY-2A contains 404 putative protein-encoding and 7 tRNA-encoding genes and AR158 contains 360 putative protein-encoding and 6 tRNA-encoding genes. The protein-encoding genes are almost evenly distributed on both strands, and intergenic space is minimal. Two of the NY-2A genes encode inteins, the large subunit of ribonucleotide reductase and a superfamily II helicase. These are the first inteins to be detected in the chlorella viruses. Approximately 40% of the viral gene products resemble entries in the public databases, including some that are unexpected for a virus. These include GDP-D-mannose dehydratase, fucose synthase, aspartate transcarbamylase, Ca^{++} transporting ATPase and ubiquitin. Comparison of NY-2A and AR158 protein-encoding genes with the prototype chlorella virus PBCV-1 indicates that 85% of the genes are present in all three viruses.

Keywords: Chlorella viruses; Phycodnaviridae; Virus NY-2A; Virus AR158; Genome sequence

Introduction

Members and prospective members of the family *Phycodnaviridae* constitute a genetically diverse but morphologically similar group of viruses that have eukaryotic algal hosts from both fresh and marine waters. The phycodnaviruses have dsDNA genomes that range in size from 170 kb to 560 kb and the viruses have hundreds of protein-encoding genes (Dunigan et al., 2006; Wilson et al., 2005). The phycodnaviruses, together with the poxviruses, iridoviruses, asfarviruses and the recently discovered 1.2-Mb Mimivirus, share a common evolutionary ancestor that may have arisen at the point of eukaryogenesis, 2 to 3 billion years ago (Iyer et al., 2006; Raoult et al., 2004; Villarreal, 2005; Villarreal and DeFilippis, 2000). These viruses share nine 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).

The most studied phycodnaviruses are the chlorella viruses that belong to the genus *Chlorovirus* (Van Etten, 2003; Yamada et al., 2006). The chloroviruses infect certain fresh water, unicellular, eukaryotic, chlorella-like green algae, which normally exist as endosymbionts in various protists, such as *Paramecium bursaria* (Kawakami and Kawakami, 1978; Van Etten et al., 1982), *Hydra viridis* (Meints et al., 1981) and *Acanthocystis turfacea* (Bubeck and Pfitzner, 2005). The prototype chlorella virus, *P. bursaria* chlorella virus (PBCV-1), has a 331-kb genome that contains 366 putative protein-encoding genes and a polycistronic gene that

* Corresponding author. Fax: +1 402 472 2853.
E-mail address: jvanetten@unlnotes.unl.edu (J.L. Van Etten).

0042-6822/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2006.08.033

encodes 11 tRNAs. PBCV-1 and the two viruses described in this report, NY-2A and AR158, infect *Chlorella* NC64A (NC64A viruses), an endosymbiont of *P. bursaria* that was originally isolated in North America. Other viruses infect *Chlorella* Pbi (Pbi viruses), an endosymbiont of *P. bursaria* that was isolated in Europe.

To investigate the diversity of the chlorella viruses, we are sequencing the genomes of several additional family members. The current manuscript describes the sequencing and annotation of the 369-kb genome from virus NY-2A and the 345-kb genome from virus AR158. NY-2A was chosen for sequencing for two reasons. First, it has the largest genome of the 36 partially characterized *Chlorella* NC64A viruses. Second, its genome is heavily methylated relative to that of the prototype virus PBCV-1 [45% of the cytosines are 5-methylcytosine (5mC) and 37% of the adenines are N^6-methyladenine (6 mA) versus 1.9% 5mC, 1.5% 6 mA]. Virus AR158 was chosen because it was the only NC64A virus that appeared to lack a gene encoding a 94 amino acid potassium ion channel protein (Kcv) that is believed to be involved in viral infection (e.g., Frohns et al., 2006). As reported here, AR158 encodes a truncated 33 amino acid K^+ channel protein. The following manuscript (Fitzgerald et al., this issue) describes the sequence and annotation of two viruses that infect *Chlorella* Pbi.

Results and discussion

As part of the chlorella virus genome sequencing effort, a project Web site has been created at http://www.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 NY-2A and AR158 genomes were assembled into contiguous sequences of 368,683-bp and 344,690-bp (Table 1), respectively, which agrees with their predicted sizes determined by pulse-field gel electrophoresis (unpublished results). Because the presumed hairpin termini were not sequenced, the left most nucleotide in the assembled sequences was designated 1.

To orient the NY-2A and AR158 genomes relative to the prototype virus PBCV-1, plots of PBCV-1 proteins and either the NY-2A or the AR158 proteins were compared. These alignments reveal a high degree of gene co-linearity between NY-2A, AR158 and PBCV-1 (Fig. 1). The average G+C content of the NY-2A and AR158 genomes is 40.7%, a concentration similar to the 40.0% G+C content of PBCV-1 (Van Etten et al., 1985).

Genes

A putative protein-encoding 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, 886 ORFs were identified in the 369-kb NY-2A genome and 815 ORFs were identified in the 345-kb AR158 genome. The ORF names were based on three criteria. First, the NY-2A ORF names begin with either a "B" for a major ORF (predicted to be a protein-encoding gene) or a "b" for a minor ORF (not considered a true protein-encoding gene). The names for the AR158 ORFs begin with either a "C" for a major ORF or a "c" 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 transcript runs either left-to-right or right-to-left, respectively. The letters "B" or "b" were chosen to name the NY-2A ORFs, which is the second NC64A virus genome sequenced, and "C" or "c" was chosen to name the AR158 ORFs, which is the third NC64A virus genome sequenced, thus avoiding confusion between the different chlorella viruses. The letters distinguish these virus ORFs from PBCV-1 ORFs (designated with an "A" or "a").

The 886 NY-2A ORFs and 815 AR158 ORFs were classified into major or minor ORFs based on the following criteria. When an ORF, of either the same or opposite polarity, resided within or significantly overlapped another ORF, the larger ORF was classified as a major ORF and 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 404 of the 886 NY-2A ORFs and 360 of the 815 AR158 ORFs probably encode proteins.

The Intein Database and Registry (InBase) was used to identify two inteins within the NY-2A ORFs, which are the first inteins identified in the chlorella viruses. The ribonucleotide reductase large subunit (B832R) contains a 337 amino acid intein that resembles an intein named CIV RIR1 (E value=5E-66) from Chilo iridescent virus (Amitai et al., 2004; Perler, 2002; Pietrokovski,

Table 1
NY-2A and AR158 genomes compared to the prototype chlorella virus, PBCV-1

Genome	General Characteristics				Similarity to PBCV-1		Similarity to NY-2A		Similarity to AR158	
	Size (bp)	Genes	tRNA Genes	G + C (%)	Homologous Genes[a]	Average a.a. Identity[b]	Homologous Genes[a]	Average a.a. Identity[b]	Homologous Genes[a]	Average a.a. Identity[b]
PBCV-1	330,743	366	11	40.0			87%	73%	87%	73%
NY-2A	368,683	404	7	40.7	84%	70%			96%	85%
AR158	344,690	360	6	40.7	91%	73%	98%	89%		

[a] Percentage of protein-coding genes that have at least one homolog (blastp E-value $\leq 10^{-5}$) in the indicated virus genome.
[b] Average amino acid identity (%) between homologous protein-coding genes.