2-20-2007

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

Lisa A. Fitzgerald
University of Nebraska-Lincoln, lisa.fitzgerald@nrl.navy.mil

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

Willaim C. Nierman
The Institute for Genomic Research, Rockville, MD, wnierman@tigr.org

See next page for additional authors

Published in *Virology* 358:2 (February 20, 2007), pp. 472-484. doi:10.1016/j.virol.2006.08.033 http://www.sciencedirect.com/science/journal/00426822 Submitted July 27, 2006; revised August 17, 2006; accepted August 23, 2006; published online October 4, 2006.

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], and James L. Van Etten[e, *]

[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] The George Washington University School of Medicine, Department of Biochemistry and Molecular Biology, Washington, DC 20037
[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
* Corresponding author. Email: jvanetten@unlnotes.unl.edu

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 ds-DNA genomes that range in size from 170 kb to 560 kb and the viruses have hundreds of protein-encoding genes (Dunigan *et al.*, 2006 and 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 and 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 nucleo-cytoplasmic 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 and 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 and 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 encodes 11 tRNAs. PBCV-1 and the two viruses described in this report, NY-2A and AR158, infect *Chlorella* NC64A (NC64A viruses), an en-

dosymbiont 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 nonredundant, 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,

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.

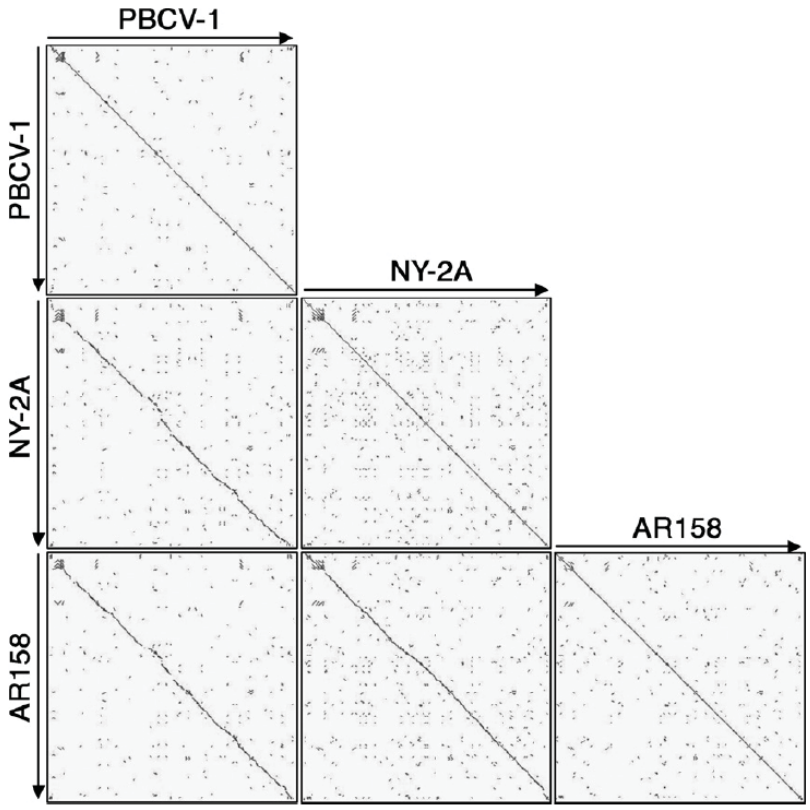

Figure 1. Comparison of the three sequenced *Chlorella* NC64A viruses, PBCV-1, NY-2A and AR158, major ORFs with blastp dot plots. The dots represent ORF homology between two viruses with an E-value of less than 0.001.

Perler, 2002 and Pietrokovski, 1998). A second 384 amino acid intein, Lpe Helicase (E value = 1E-31) from *Listonella pelagia* phage phiHSIC (Paul *et al.*, 2005 and Perler, 2002), exists in a putative helicase (B508R).

In addition, several introns exist in the NY-2A and AR158 genomes. The DNA polymerase genes (*b249r* and *c230r*) contain an identically located 86-nucleotide splicesome processed intron with 5′-AG/GUGAGU and 3′UGCAG/UU splice site sequences, as well as a predicted branch point UCAC sequence (Grabherr *et al.*, 1992). The DNA polymerase genes from 38 other NC64A viruses also either have an 86 or 101 nucleotide intron at the same position (Zhang *et al.*, 2001).

The pyrimidine dimer-specific glycosylase genes *b076l* and *c064l* have an identical 81-nucleotide splicesomal processed intron that is also present in two other NC64A viruses (Sun *et al.*, 2000). NY-2A contains three additional introns. A 402 nucleotide self-splicing group I intron is found in a putative transcription factor TFIIS gene (*b175l*); this intron is 99% identical to the intron found in PBCV-1 TFIIS (*a125l*) (Li *et al.*, 1995). A related intron (75.6% identity to the TFIIS intron) also occurs in NY-2A gene *b496l* and AR158 gene *c437l*; the function of this gene product is unknown. This self-splicing intron occurs in one or more genes from 80 other NC64A viruses (Nishida *et al.*, 1998 and Yamada *et al.*, 1994). Finally, a small intron (13 nucleotides) occurs in the NY-2A $tRNA^{Tyr}$; this intron and tRNA are absent in AR158.

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 the promoter region, the frame in which the putative protein is encoded, the number of amino acids in the encoded protein, the predicted protein molecular weight and the isoelectric points. These properties are listed in Supplements 1 and 2. Fig. 2 reports some general characteristics for both the NY-2A and AR158 major ORFs, including the relative orientation of the ORFs (Figs. 2A and B). The directions in which the ORFs are encoded are slightly skewed in the reverse (~ 54%) orientation for both viruses. The average size of all the putative NY-2A proteins is 279 amino acids (Fig. 2C), whereas the average size of all the putative AR158 proteins is 287 amino acids (Fig. 2D); nearly 45% of the proteins are 65 to 200 amino acids long. The predicted pIs of the proteins are depicted in Figs. 2E and 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 presumably help neutralize the positively charged genomic DNA. However, the functions of the proteins that have pIs in the 4.5 range vary [*e.g.*, an exonuclease (B214R and C200R), a SKP-1 protein (B068L and C056L), a sliding clamp processivity factor protein (PCNA) (B261L and C241L) and an ornithine/arginine decarboxylase (B278R and C256R)]. Figs. 2G and H indicate the intergenic space between the major ORFs. Approximately 65% of the ORFs in both viruses are separated by less than 100 nucleotides.

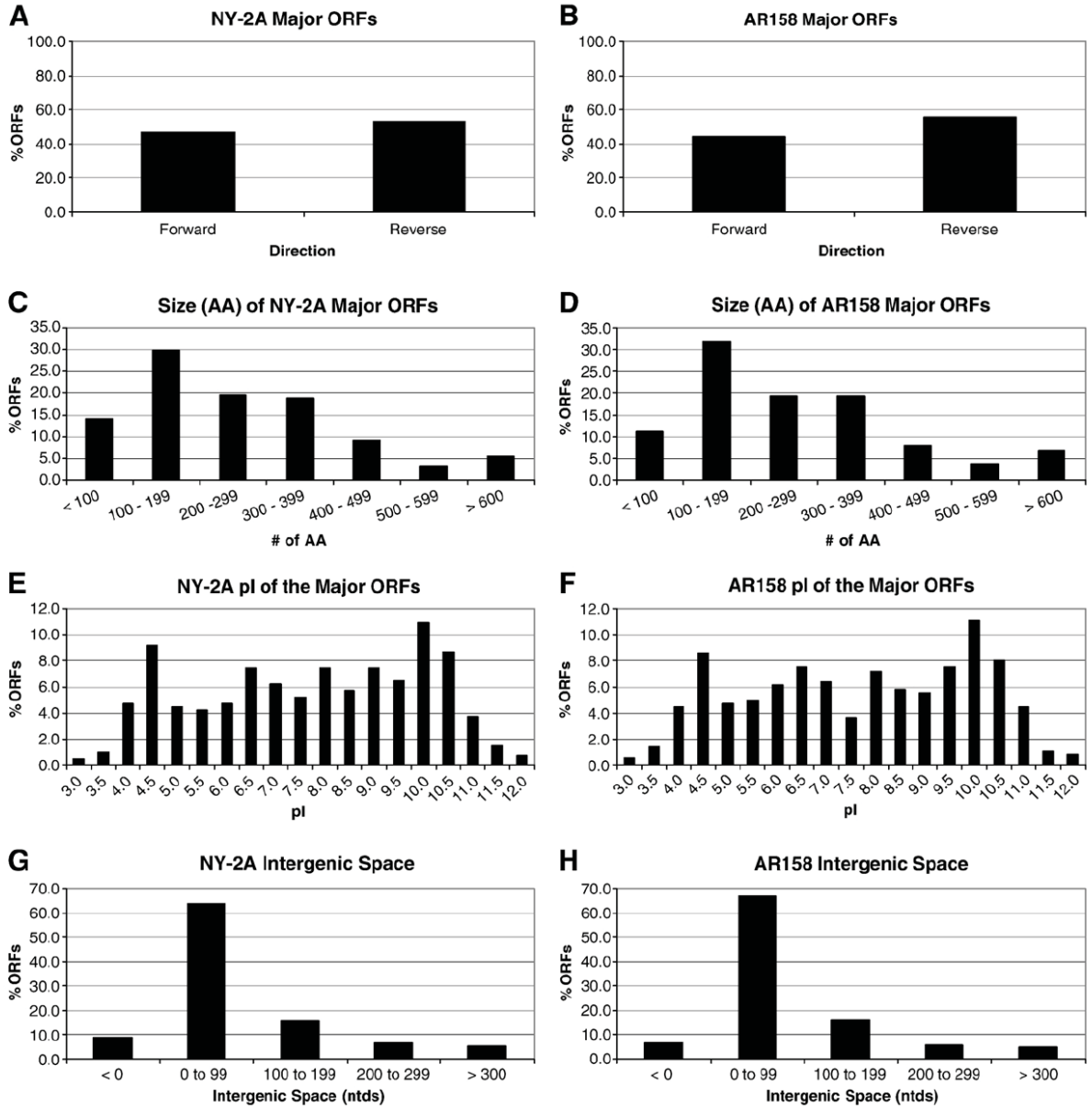

Figure 2. General characteristics of NY-2A and AR158 major ORFs. Direction of the NY-2A (A) and AR158 (B) ORFs. Size of the major NY-2A (C) and AR158 (D) ORFs. Predicted isoelectric points of the major NY-2A (E) and AR158 (F) ORFs. Intergenic space between major NY-2A (G) and AR158 (H) ORFs.

Annotation of NY-2A and AR158 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 the COG databases were used to identify conserved domains and proteins in the NY-2A and AR158 ORFs (Supplements 1 and 2). Gene maps of the NY-2A and AR158 genomes illustrate the location of the putative genes (Fig. 3) and some of the ORFs are listed by their predicted metabolic function (Table 2). Only a few NY-2A and AR158 gene products have been tested for activity (see below). However, we assume that any NY-2A and AR158 encoded proteins that have functional PBCV-1 homologs are also functional.

Eighty-four to ninety-eight percent of the major ORFs are homologous between any two of the three NC64A viruses, *i.e.*, NY-2A, AR158 and PBCV-1. This finding suggests that the majority of major ORFs from the NC64A viruses are essential for virus replication in nature. The average amino acid identity between homologous proteins from PBCV-1 and either NY-2A or AR158 is 73%. There is an average of 87% amino acid identity between NY-2A and AR158 homologs. NY-2A and AR158 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

NY-2A and AR158 have 13 ORFs that are involved in either DNA replication, recombination or repair, such as δ-DNA

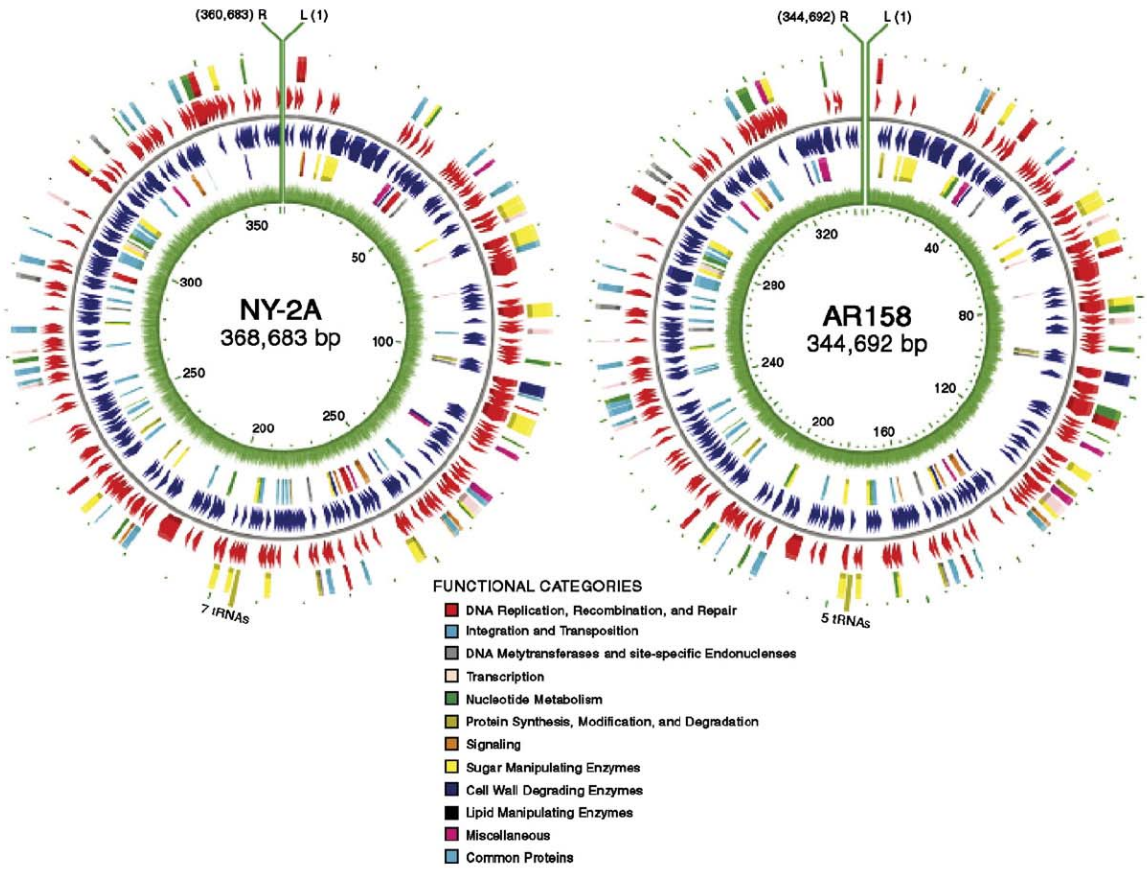

Figure 3. Map of the NY-2A and AR158 genomes arranged in a circle. However, the genomes are linear molecules and the ends are depicted at the top of the figures as green lines L and R represent the left and right ends of the genomes, respectively. The predicted ORFs are shown in both orientations and colored according to their fuctional category. The inner green circle represents the percent A + T across the genome using a 25-bp window.

polymerase (B249R and C230R), type II DNA topoisomerase (B781L and C707L), superfamily III helicase (B623L and C562L), DNA primase (B633R and C573R), RNase H (B547R and C491R), exonuclease (B214R and C200R), ATP-dependent DNA ligase (B734R and C658R), two ORFs that resemble sliding clamp processivity factor (PCNA) proteins (B261L, B767L, C241L and C694L), replication factor C protein (B571L and C519L) and pyrimidine dimer–DNA glycosylase (B076L and C064L) (Table 2).

The NY-2A and AR158 type II DNA topoisomerases have approximately 40% amino acid identity with type II topoisomerases from several eukaryotic organisms. The enzyme is ATP dependent and functions by threading a double-stranded DNA segment through a transient double-strand break in the DNA (Roca, 1995). PBCV-1 encodes one of the smallest known type II DNA topoisomerases (1061 amino acids) (Lavrukhin *et al.*, 2000). The PBCV-1 enzyme cleaves dsDNAs approximately 30 times faster than the human type II DNA topoisomerase (Fortune *et al.*, 2001). The NY-2A type II DNA topoisomerase has 90% amino acid identity and is the same size as its PBCV-1 homolog (1061 amino acids). The AR158 type II DNA topoisomerase has 1062 amino acids and has 89% identity to its PBCV-1 homolog; however, it has 98% amino acid identity to its NY-2A homolog.

Like PBCV-1, NY-2A and AR158 encode two proteins that resemble PCNA proteins. The NY-2A (B261L and B767L) and the AR158 (C241L and C694L) 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 not only involved in DNA replication but also DNA repair and post-replicative processing, such as DNA methyltransferases and DNA transposases (Warbrick, 2000). Because the chlorella viruses encode proteins involved in both DNA repair and DNA methylation, the 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, including NY-2A and AR158. 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. NY-2A and AR158 encode at least four putative transcription factor-like elements: TFIIB (B154L and C142L), TFIID (B743R and C669R), TFIIS (B175L and C167L) and VLTF2-type transcription fac-

Table 2
Putative ORFs group by their functional categories for NY-2A, AR158 and the prototype PBCV-1 virus

	PBCV-1		NY-2A		AR158	
DNA Replication, Recombination and Repair						
	ORF	AA	ORF	AA	ORF	AA
δ-DNA polymerase	A185R	913	B249R	913	C230R	913
Archaeo-eukaryotic primase	A468R	443	B633R	443	C573R	442
PCNA	A193L	262	B261L	262	C241L	262
	A574L	264	B767L	245	C694L	245
Replication factor C	A417L	429	B571L	434	C519L	432
RNase H	A399R	194	B547R	193	C491R	195
Helicase-Superfamily III	A456L	654	B623L	654	C562L	654
DNA Topoisomerase II	A583L	1061	B781L	1061	C707L	1062
ATP-dependent DNA ligase	A544R	298	B734R	323	C658R	323
ATPase (PP-loop)	A554/ 556/557L	498	B744L	504	C670L	504
ATPase (DNA packaging)	A392R	258	B536R	254	C479R	254
Pyrimidine dimer-specific glycosylase	A050L	141	B076L	116	C064L	141
Exonuclease	A166R	268	B214R	268	C200R	268
Transcription						
	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF		ORF	AA
Transcription factor TFIIB	A107L	290	B154L	338	C142L	314
Transcription factor TFIID	A552R	270	B743R	317	C669R	317
Transcription factor TFIIS	A125L	180	B175L	192 + A57	C167L	180
VLTF2-type transcription factor	A482R	215	B647R	215	C586R	216
Superfamily II helicase	A153R	459	B203R	460	C191R	459
	A241R	725	B316R	725	C292R	725
	A363R	811	B508R	1612	C449R	1225
mRNA guanylyltransferase	A103R	330	B148R	321	C137R	327
RNA triphosphatase	A449R	193	B612R	188	C556R	188
Histone H3, Lys 27 methylase	A612L	119	B268L	190	C245L	190
			B813L	119	C731L	119
SWI/SNF chromatin remodeling complex	A189/ 192R	1299	B258R	1278	C239R	1369
SWI/SNF helicase	A548L	458	B738L	458	C663L	458
RNase III	A464R	275	B628R	268	C568R	268
Cytosine deaminase	A200R	118	B271R	145	C246R	119
Sugar Manipulation						
	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
D-lactate dehydrogenase	A053R	363				
GDP-D-mannose dehydratase	A118R	345	B163R	343	C155R	343
Fucose synthase	A295L	317	B395L	320	C344L	320
UDP-glucose 6-dehydrogenase	A609L	389	B465R	379	C413R	379
					C729L	414
Glucosamine synthetase	A100R	595	B143R	595	C132R	595
Hyaluronan synthase	A098R	568				
Chitin synthase			B139R	508	C128R	508
			B472R	527	C418R	527
Cellulase precursor					C134L	482
Polysaccaride deacetylase			B469L	403	C415L	373
Glycosyltransferase	A064R	638	B159R	867	C150R	860
	A111/114R	860	B618R	270	C265R	634
	A219/222/226R	677	B736L	405	C559R	270
	A473L	517			C661L	396
	A546L	321				

Table 2 (*continued*)

	PBCV-1		NY-2A		AR158	
Lipid Manipulation						
	ORF	AA	ORF	AA	ORF	AA
Acetyltransferase	A254R	158				
N-acetyltransferase	A654L	197	B853L	197	C767L	197
Glycerophosphoryl diesterase	A049L	219	B075L	228	C063L	228
Lipoprotein lipase	A402R	227	B550R	232	C494R	232
Lysophospholipase	A271L	159	B354L	264	C315L	264
Patatin-like phospholipase	A173L	288	B226L	279	C208L	279
Cell Wall Degradation						
	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Chitinase	A181/182R	830	B239R	807	C220R	807
	A260R	484				
Chitosanase	A292L	328	B393L	322	C342L	326
β & α 1,4 linked glucuronic lyase	A215L	321	B288L	306	C263L	306
			B468R	251	C414R	294
β-1,3-glucanase	A094L	364	B137L	320	C126L	320
Signaling						
	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Potassium channel protein	A250R	94	B336R	94	C305aR	33
Ligand-gated channel protein	A163R	433				
Glutamate receptor	A162L	411				
Ca2+ transporting ATPase					C785L	870
Dual specificity phosphatase	A305L	204	B430L	181	C378L	182
Serine/Threonine protein kinase	A034R	308	B331R	309	C051R	320
	A248R	308	B365L	605	C300R	288
	A277L	303	B368L	501	C321L	604
	A278L	610	B388L	268	C324L	506
	A282L	569	B816L	492	C337L	271
	A289L	283	B818R	317	C733L	509
	A614L	577			C735R	319
	A617R	321				
Integration and Tranposition						
	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Transposase	A625R	433	B080L	645	C068R	548
			B334R	432	C303R	432
			B378L	645	C636L	379
			B702L	432	C745R	432
			B711L	395		
			B829R	432		
Resolvase			B083L	228	C069R	186
			B381L	228		
			B715L	228		
Tlr 6Fp DNA mobile protein	A121R	97	B168R	104	C160R	104
GIY-YIG endonuclease	A134L	165	B039L	418	C036L	418
	A287R	251	B185L	164	C157R	417
	A315L	246	B206L	279	C175L	165
	A351L	358	B246R	202	C236R	370
	A495R	221	B286R	363	C281R	368
	A539R	173	B346L	249	C339R	233
	A651L	230	B389R	291	C438L	250
			B433L	273	C440L	345
			B440L	272	C547L	214
			B497L	251	C569R	257
			B499L	345	C641L	263
			B602L	214	C651R	173

(continued on next page)

Table 2 (*continued*)

Integration and Tranposition	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
GIY-YIG endonuclease			B629R	287		
			B724R	173		
			B850L	391		
HNH endonuclease	A087R	456	B022R	401	C018R	400
	A267L	314	B061R	399	C203R	405
	A354R	237	B133R	444	C372L	351
	A422R	342	B165R	459	C392R	412
	A478L	310	B173L	414	C504L	352
	A490L	310	B199R	343	C540L	393
			B218R	400	C580L	310
			B324L	495	C678L	320
			B370L	349	C806L	400
			B424L	351		
			B446R	412		
			B598L	399		
			B718L	320		
			B747L	309		
			B753L	380		
			B798R	313		
			B805R	340		
			B878L	400		

Protein Synthesis, Modification and Degradation	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Translation elongation factor-3	A646L	141			C788L	918
	A666L	918				
Prolyl-4-hydroxylase	A085R	242	B126R	230	C118R	230
Thiol oxidoreductase	A465R	118	B630R	117	C570R	118
Protein disulfide isomerase	A448L	106	B611L	106	C554L	106
SKP-1 protein	A039L	151	B068L	153	C056L	153
Ubiquitin			B699L	78		
Ring finger ubiquitin ligase	A481L	224	B645L	224	C584L	224
Ubiquitin C-terminal hydrolase	A105L	284	B150L	290	C140L	284
Zn metallopeptidase	A521L	392	B685L	202	C621L	202
	A604L	134	B803L	115	C726L	115

Nucleotide Metabolism	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Aspartate transcarbamylase	A169R	323	B222R	321	C204R	321
Ribo. Reductase (small subunit)	A476R	324	B641R	324	C579R	324
Ribo. Reductase (large subunit)	A629R	771	B832R	1103	C748R	767
Deoxynucleoside kinase	A416R	188				
dCMP deaminase	A596R	142	B795R	142	C718R	142
dUTP pyrophosphatase	A551L	141	B741L	141	C667L	141
Thymidylate synthase X	A674R	216	B865R	216	C793R	216
Glutaredoxin	A438L	78	B592L	78	C538L	78
Thioredoxin	A427L	119	B581L	119	C528L	119

Miscellaneous	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Ornithine/Arginine decarboxylase	A207R	372	B278R	372	C256R	372
Agmatine iminohydrolase	A638R	359	B844R	359	C759R	359
N-carbamoylput. amidohydrolase	A078R	298	B116R	298	C104R	299
Homospermidine synthase	A237R	518	B305R	507	C286R	507
Histidine decarboxylase	A598L	363	B796L	366	C719L	382

Table 2 (*continued*)

Miscellaneous	PBCV-1		NY-2A		AR158	
	ORF	AA	ORF	AA	ORF	AA
Monoamine oxidase	A217L	394	B289L	382	C264L	383
Amidase	A284L	279	B371L	279	C327L	279
Cu/Zn-superoxide dismutase	A245R	187				
O-methyltransferase	A061L	209				
FkbM Methyltransferase			B183L	224	C173L	198
ABC transporter protein	A445L	462	B606L	462	C551L	462
ATPase (AAA+ class)	A044L	599	B073L	603	C061L	603
Fibronectin binding protein	A180R	108				

tor (B647R and C586R). However, none of these proteins are packaged in the PBCV-1 virion (Dunigan *et al.*, manuscript in preparation) and are unlikely to be packaged in the NY-2A or AR158 virions. NY-2A and AR158 encode two proteins that are involved in creating a mRNA cap structure, a mRNA capping enzyme (B148R and C137R) and an RNA triphosphatase (B612R and C556R). NY-2A and AR158 also encode an RNase III enzyme (B628R and C568R) that is presumably involved in processing viral mRNAs and/or tRNAs.

In the immediate-early phase of infection, the host is reprogrammed to transcribe viral RNAs, which in PBCV-1 begins 5–10 min p.i. 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 that contains a SET domain (named vSET) that di-methylates Lys^{27} in histone 3 (Manzur *et al.*, 2003). vSET is packaged in the PBCV-1 virion and accumulating evidence indicates that vSET could be involved in repressing host transcription after PBCV-1 infection (Manzur *et al.*, manuscript in preparation). NY-2A and AR158 each contain a vSET homolog (B813L and C731L). Furthermore, unlike PBCV-1, NY-2A and AR158 encode a second protein that contains a SET domain (B268L and C245L); B268L and C245L are 190 amino acids long and have 25% amino acid identity to the smaller vSET proteins. In addition to these histone methyltransferases, NY-2A and AR158 encode a putative SWI/SNF family helicase (B738L and C663L) and a SWI/SNF chromatin remodeling complex protein (B258R and C239R). Both proteins are also implicated in chromatin remodeling (Kim and Clark, 2002).

Finally, NY-2A and AR158, as well as all the chlorella viruses, encode a putative cytosine deaminase (B271R and C246R). This observation suggests that either some of the viral transcripts or host transcripts may undergo post-transcriptional editing (Gerber and Keller, 2001).

Protein synthesis, modification and degradation

PBCV-1 was the first virus to encode a translation elongation factor (EF) (Yamada *et al.*, 1993). The PBCV-1 protein has about 45% amino acid identity to an EF-3 protein from fungi (Belfield and Tuite, 1993 and Chakraburtty, 2001). The fungal protein stimulates EF-1 GTP-dependent binding of an amino acyl-tRNA to the ribosome A site. Like fungal EF-3 proteins, the virus-encoded proteins have an ABC transporter family signature and two ATP/GTP-binding site motifs. AR158 encodes a 918 amino acid protein (C788L) that has 93% amino acid identity to PBCV-1 EF-3. However, NY-2A lacks an EF-3-encoding gene.

NY-2A and AR158 have genes that encode proteins involved in post-translational modifications, including prolyl-4-hydroxylase (B126R and C118R), protein kinases (see below) and glycosyltransferases (see below). NY-2A and AR158 also encode a protein disulfide isomerase (B611L and C554L), a SKP-1 protein (B068L and C056L) and a thiol oxidoreductase (B630R and C570R). Additionally, the two viruses encode proteins involved in protein degradation including an ubiquitin C-terminal hydrolase (B150L and C140L), a ring finger ubiquitin ligase (B645L and C584L) and two Zn metallopeptidases (B685L, B803L, C621L and C726L). NY-2A is the first chlorella virus to encode a 78 amino acid ubiquitin protein (B699L).

tRNAs

The NY-2A and AR158 genomes were analyzed for tRNAs using the tRNAscan-SE program (Lowe and Eddy, 1997). NY-2A is predicted to encode 7 tRNAs: 2 for Leu and 1 each for Arg, Asn, Lys, Tyr and Val (Table 3). These 7 tRNAs are clustered in a region of the NY-2A genome, nucleotide sequence 194,698 to 195,554. AR158 is predicted to encode 6 tRNAs: 2 for Leu and 1 each for Arg, Asn, Ile and Val (Table 3). These 6 tRNAs are also clustered in the AR158 genome, nucleotide sequence 172,099 to 172,780. Presumably, the tRNAs are transcribed as a large precursor RNA and processed via intermediates to mature RNAs as they are in chlorella virus CVK2 (Nishida *et al.*, 1999). Only five of the eleven tRNAs encoded by PBCV-1 are found in both NY-2A and AR158; neither NY-2A nor AR158 encode a unique tRNA. Although the orientation of the tRNA genes is the same in all three genomes, their order varies between the viruses (Table 3). 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 from NY-2A, $tRNA^{Tyr}$, is predicted to contain a 13-nucleotide intron. The insertion of a small intron in the tyrosine tRNA (anti-codon GTA) also occurs in PBCV-1; however, the intron and tRNA are absent in AR158. Codon usage analyses of viral-encoded proteins indicate a strong correlation between the abundance of the viral-encoded tRNAs and their usage in viral proteins.

Nucleotide metabolism

NY-2A and AR158 encode eight enzymes involved in nucleotide metabolism. These enzymes are important because the DNA concentration in viral-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. NY-2A and AR158 each encode the small (B641R and C579R) and large (B832R and C748R) subunits of ribonucleotide reductase, aspartate transcarbamylase (B222R and C204R), dUTP pryophosphatase (B741L and C667L), deoxycytidylate (dCMP) deaminase (B795R and C718R), glutaredoxin (B592L and C538L), thioredoxin (B581L and C528L) and thymidylate synthase X (B865R and C793R).

PBCV-1 was the first virus found to encode a functional aspartate transcarbamylase, the key regulatory enzyme in the de novo biosynthesis of pyrimidines (Landstein *et al.*, 1996). NY-2A and AR158 each encode an aspartate transcarbamylase that has 84–85% amino acid identity to the PBCV-1 homolog. The NY-2A and AR158 aspartate transcarbamylases have 98% amino acid identity to each other.

Two NY-2A and AR158-encoded enzymes, dUTP pyrophosphatase and dCMP deaminase, produce dUMP, the substrate for thymidylate synthetase. The chlorella viruses, including NY-2A and AR158 lack a traditional thymidylate synthetase A. Instead, they encode a protein that is a member of a new family of flavin-dependent thymidylate synthetases called ThyX (Graziani *et al.*, 2004 and Myllykallio *et al.*, 2002).

Protein kinases, phosphatases and channel proteins

Both NY-2A and AR158 encode several Ser/Thr protein kinases (Table 2) and a protein that resembles a dual-specificity phosphatase (B430L and C378L). 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 important for virus replication. AR158, but not NY-2A or PBCV-1, also encodes an 870 amino acid Ca^{2+} transporting ATPase (C785L).

The chlorella viruses are the first viruses to encode K^+ channel proteins (called Kcv) and Kcv genes from 40 NC64A viruses, including NY-2A (B336R), have been expressed successfully in *Xenopus* oocytes (Kang *et al.*, 2004 and Plugge *et al.*,

Table 3
Comparison of the PBCV-1, NY-2A and AR158 tRNA genes

tRNA	Anti-codon	PBCV-1			NY-2A			AR158		
		tRNA#[a]	Start	End	tRNA#[a]	Start	End	tRNA#[a]	Start	End
Leu	CAA	1	164,916	164,996	2	194,921	195,005	4	172,430	172,513
Ile	TAT	2	165,022	165,094	–			2	172,209	172,281
Leu	TAA	3	165,118	165,200	1	194,698	194,778	1	172,099	172,183
Asn	GTT	4	165,224	165,297	4	195,201	195,272	3	172,307	172,378
Lys	CTT	5	165,301	165,373	6	195,382	195,456	–		
Asn	GTT	6	165,398	165,471	–			–		
Lys	CTT	7	165,475	165,547	–			–		
Tyr[b]	GTA	8	165,572	165,658	5	195,295	195,379	–		
Lys	TTT	9	165,681	165,753	–			–		
Arg	TCT	10	165,777	165,851	3	195,029	195,100	5	172,537	172,608
Val	AAC	11	165,885	165,957	7	195,482	195,554	6	172,708	172,780

[a] Order of the tRNA genes in the genome.
[b] tRNA gene contains an intron.

2000). AR158 is unusual in that it only has a 33 amino acid Kcv (C305aR) homolog. The AR158 Kcv homolog lacks the N-terminal portion of the protein including the predicted transmembrane 1 and pore helix domains (Fig. 4). The truncated protein has 100% amino acid identity to the C-terminal portion of NY-2A Kcv.

Sugar- and lipid-manipulating proteins

PBCV-1, NY-2A and AR158 encode several proteins with high identities to enzymes involved in either manipulating sugars, synthesizing polysaccharides or transferring sugars to proteins. Two of the viral encoded enzymes, GDP-d-mannose dehydratase (GMD) (B163R and C155R) and fucose synthase (B395L and C344L), comprise a three-step pathway that converts GDP-d-mannose to GDP-l-fucose. Unexpectedly, the PBCV-1 GMD differs from other GMDs because in addition to the dehydratase activity, the protein also has a strong stereospecific NADPH-dependent reductase activity that produces GDP-d-rhamnose (Tonetti *et al.*, 2003). Both fucose and rhamnose are present in the glycans attached to the PBCV-1 major capsid protein (Nandhagopal *et al.*, 2002 and Wang *et al.*, 1993).

NY-2A and AR158 also encode a glucosamine synthetase (B143R and C132R). NY-2A encodes one UDP-glucose dehydrogenase (B465R) whereas the AR158 has two genes (C413R and C729L) that encode the enzyme; thus, AR158 is similar to the chlorella virus CVK2 (Ali *et al.*, 2005). Unlike PBCV-1, neither NY-2A nor AR158 have a hyaluronan synthase homolog. Instead, NY-2A and AR158 encode two chitin synthases (B139R, B472R, C128R and C418R) like some other NC64A viruses (Ali *et al.*, 2005) chitin, rather than hyaluronan, is formed on the external surface of cells infected with viruses that encode chitin synthases (Ali *et al.*, 2005 and Kawasaki *et al.*, 2002).

NY-2A encodes three (B159R, B618R and B736L) and AR158 encodes four (C150R, C265R, C559R and C661L) glycosyltransferases, which are probably involved in glycosylation of the virus major capsid proteins (Graves *et al.*, 2001). NY-2A and AR158 also encode a putative polysaccharide deacetylase (B469L and C415L) which is absent in PBCV-1.

NY-2A and AR158 encode several proteins involved in lipid metabolism including lipoprotein lipase (B550R and C494R), lysophospholipase (B354L and C315L), *N*-acetyl-transferase (B853L and C767L), glycerophosphoryl diesterase (B075L and C063L) and patatin phospholipase (B226L and C208L).

Cell wall-degrading enzymes

NY-2A and AR158 encode five proteins that may be involved in degrading *Chlorella* NC64A cell walls either during virus infection or virus release. These proteins include a chitinase (B239R and C220R), a chitosanase (B393L and C342L), a β-1, 3-glucanase (B137L and C126L) and two polysaccharide lyases that cleave chains of either β- or α-1, 4 linked glucuronic acids (B288L, B468R, C263L and C414R) (Sugimoto *et al.*, 2004). The larger polyglucuronic acid lyases (B288L and C263L) are 306 amino acids in length and have 72% amino acid identity to a PBCV-1 A215L homolog. In contrast, the smaller (251 amino acids) polyglucuronic acid lyases (B468R and C414R) are 100% identical to each other but only have 36% amino acid identity to the PBCV-1 A215L homolog.

Restriction–modification enzymes

Chlorella viruses contain different concentrations of 5mC and 6 mA in their genomes (Van Etten *et al.*, 1991). Therefore, it is not surprising that these viruses encode multiple 5mC and 6 mA DNA methyltransferases (MTase). The NY-2A genome is heavily methylated (45% 5mC and 37% 6 mA). The level of methylation in the AR158 genome is unknown. NY-2A encodes 18 DNA MTases (Table 4): eleven are 6 mA MTases and 7 are 5mC MTases. AR158 encodes 16 MTases (Table 4): nine are 6 mA MTases and 7 are 5mC MTases.

The DNA sequences methylated by some of the NY-2A MTases are known (Table 4) (Chan *et al.*, 2006 and Zhang *et al.*, 1998). NY-2A also encodes two companion site-specific endonucleases, CviQI (B542R) and Nt.CviQII (B361R). CviQI creates double-stranded breaks 5′ to the T in the palindromic sequence G/TAC (Xia *et al.*, 1987), whereas Nt.CviQII makes single-stranded breaks 5′ to the A in R/AG sequences (Chan *et al.*, 2006 and Zhang *et al.*, 1998). Both endonucleases are inhibited when 6 mA is present in their cleavage sites.

Nothing is known about the specificity of the AR158 DNA MTases and possible site-specific endonucleases. However, all 16 of the AR158 MTases are homologous to one of the NY-2A MTases and so one can predict some of the sequences they methylate. For example, ORF C486R has 100% amino acid identity with the NY-2A CviQI restriction endonuclease and so it cleaves G/TAC sequences. In contrast, the NY-2A nicking enzyme Nt.CviQII only has 37% amino acid identity with AR158 ORF C618L. However, C618L has 94% amino acid identity with another chlorella virus nicking enzyme, Nt.CviPII, which cleaves at/CCD sequences (Chan *et al.*, 2004 and Xia *et al.*, 1988). Therefore, C618L is predicted to encode a nicking endonuclease, probably cleaving at CCD sequences.

Integration and transposition enzymes

NY-2A has 6 ORFs that resemble bacterial transposases and AR158 has 4 such ORFs (Table 2). Three of the NY-2A

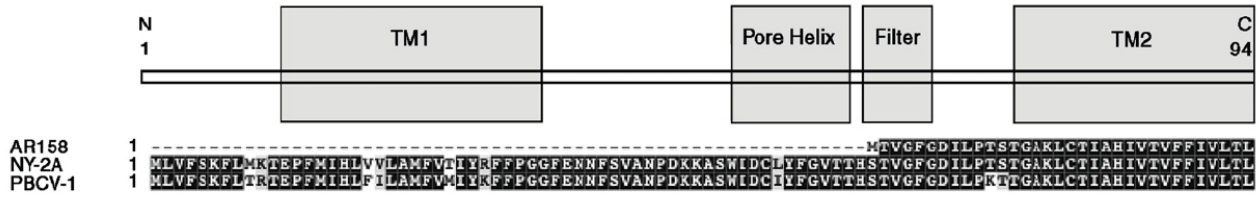

Figure 4. Alignment and putative structure domains of PBCV-1, NY-2A and the truncated AR158 Kcv proteins.

Table 4
DNA methyltransferases and site-specific endonucleases coded by viruses NY-2A, AR158 and PBCV-1 that infect *Chlorella* NC64A

ORFs			Recognition sequence
NY-2A	AR158	PBCV-1	
N^6-Adenine methyltransferases			
B010R	–	–	Unknown
B016L	C011L	–	TCG^{m}A*
B230L	C212L	–	TGC^{m}A*
B236L	C217L	A251R	C^{m}ATG*
B359R	–	–	^{m}AG*
B399R	C346R	–	G^{m}ANTC*
B416R	C363R	–	Unknown
B418R	C365R	–	Unknown
B543L	C487L	–	GT^{m}AC*
B567L	C513L	–	Unknown
B774R	C701R	A581R	G^{m}ATC
5-Methyl cytosine methyltransferases			
B008R	C006R	–	Unknown
B088R	C075L	–	Unknown
B411L	C358L	–	Unknown
B566R	C512R	–	Unknown
B681L	C619L	–	^{m}CCD
B697R	C632R	A530R	Unknown
B769R	C696R	–	Unknown
–	–	A517L	Unknown
–	–	A683L	RG^{m}CY
Site-specific endonucleases			
B361R	–	–	R/AG*
B542R	C486R	–	G/TAC*
–	–	A252R	C/ATG*
–	–	A579L	/GATC*
–	C618L	–	/CCD*

* The recognition sequence has been established for this protein or its homolog.

transposases have internal resolvases whereas only one AR158 transposase has an internal resolvase. NY-2A and AR158 also encode an ORF that resembles a Tlr 6Fp DNA mobile protein (B168R and C160R). This protein is a member of a family of genetic elements limited to *Tetrahymena thermophila* (Wuitschick *et al.*, 2002).

In addition to the transposases and resolvases, NY-2A contains 33 ORFs and AR158 contains 21 ORFs with motifs found in homing endonucleases (Table 2). Homing endonucleases are rare DNA-cleaving enzymes typically encoded by introns and inteins. These endonucleases are classified into four families (Belfort and Roberts, 1997). Fifteen of the NY-2A ORFs are members of the GIY-YIG family, and eighteen are members of the HNH family. The AR158 genome codes for 12 ORFs that are members of the GIY-YIG family, and 9 are members of the HNH family. Thus, collectively, the NY-2A and AR158 viruses have many protein-encoding genes that could facilitate DNA rearrangements either within or between viruses, and possibly with its hosts. These genes are distributed throughout both virus genomes. However, it is not known if any of these virus-encoded transposases, resolvases and homing endonucleases, including ones encoded by PBCV-1, are functional.

Polyamine biosynthetic enzymes

PBCV-1 was the first virus reported to encode polyamine biosynthetic enzymes, including two pathways to synthesize putrescine. All four PBCV-1 genes, which encode functional enzymes (Kaiser *et al.*, 1999, Morehead *et al.*, 2002 and Baumann *et al.*, submitted for publication), are also present in NY-2A and AR158. These enzymes are ornithine/arginine decarboxylase (ODC) (B278R and C256R), agmatine iminohydrolase (B844R and C759R), *N*-carbamoylputrescine amidohydrolase (B116R and C104R) and homospermidine synthase (B305R and C286R). 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 enzyme decarboxylates arginine more efficiently than ornithine (Shah *et al.*, 2004). NY-2A and AR158 each encode a 372 codon ODC, with 83–86% amino acid identity to its PBCV-1 homolog. The product of arginine decarboxylation is agmatine; agmatine iminohydrolase and *N*-carbamoylputrescine amidohydrolase convert agmatine to putrescine.

Homospermidine synthase synthesizes homospermidine, a rare polyamine, from two molecules of putrescine (Kaiser *et al.*, 1999). The NY-2A and AR158 homospermidine synthase homologs have 93% amino acid identity with the PBCV-1 enzyme.

Polyamines putrescine, spermidine and spermine are common in cells and also are structural components of many viruses, where they help to neutralize viral nucleic acids (Tyms *et al.*, 1990). PBCV-1 virions, as well as uninfected and virus-infected *Chlorella* NC64A cells, contain putrescine, cadaverine, spermidine and homospermidine. However, it is unlikely that these polyamines are important in neutralizing the chlorella virus DNA because the number of polyamine molecules per virion could only neutralize ~ 0.2% of the DNA phosphate residues (Kaiser *et al.*, 1999).

NY-2A and AR158 encode two additional enzymes involved in amine metabolism, monoamine oxidase (B289L and C264L) and histidine decarboxylase (B796L and C719L). The finding that PBCV-1, NY-2A and AR158 each encode these six proteins suggests that they must serve some important role(s) in the replication of the viruses.

Miscellaneous proteins

NY-2A and AR158 also encode several other putative proteins, including an ABC transporter protein (B606L and C551L), an amidase (B371L and C327L) and an AAA+ class ATPase (B073L and C061L). Finally, NY-2A and AR158 each encode a FkbM methyltransferase (B183L and C173L) that has some similarity to an enzyme from several *Mycobacterium* species that methylates rhamnose (Jeevarajah *et al.*, 2002).

NY-2A and AR158 structural proteins

NY-2A ORF B585L and AR158 ORF C532L have the highest amino acid identity (96%) and are the same size as the PBCV-1 major capsid protein, A430L. Therefore, we assume

that B585L and C532L are the NY-2A and AR158 major capsid proteins, respectively. However, NY-2A and AR158 have additional ORFs that have significant amino acid sequence identity to the PBCV-1 major capsid protein. For example, B617L and C558L have 94% amino acid identity and are the same size as A430L. Other NY-2A and AR158 ORFs, including B059R, B529R, B748L, B825L, C048R, C470R, C675L and C741L, have 33 to 41% amino acid identity to A430L.

Identification of gene families

A total of 131 of the NY-2A ORFs resemble 1 or more other NY-2A 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. However, this number is somewhat misleading because some of these ORFs are grouped as families because they contain a common conserved domain, *e.g.*, ankyrin 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, 4 families have three members, 3 families have four members, 4 families have five members, 2 families have six members, 2 families have seven members, 1 family has eight members, 1 family has 10 members and 1 family has seventeen members.

A similar analysis indicates that 125 of the AR158 ORFs resemble 1 or more other AR158 ORFs. A total of 17 families have two members, 1 family has four members, 3 families have five members, 1 family has six members, 3 families have seven members, 1 family has nine members, 1 family has twelve members and 1 family has twenty-four members.

Conclusions

The 368,683-bp NY-2A genome and the 344,690-bp AR158 genome, the largest chlorella viral genomes sequenced to date, are predicted to encode 404 and 360 proteins as well as 7 and 6 tRNAs, respectively. The putative protein-encoding 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 others resemble eukaryotic proteins. Approximately 85% of the NY-2A, AR158 and PBCV-1 protein-encoding genes are homologous, suggesting that these proteins are important in NC64A virus replication. However, there are some interesting exceptions in which a gene is only present in one virus, *e.g.*, ubiquitin is only encoded by NY-2A, Ca^{++} transporting ATPase is only encoded by AR158, and Cu/Zn superoxide dismutase is only encoded by PBCV-1. Consequently, the total number of genes in the chlorella virus gene pool exceeds that of a single isolate.

Materials and methods

Viral DNA isolation and sequencing

Plaque-forming viruses NY-2A and AR158 were isolated from fresh-water samples collected in New York state (August, 1984) and Buenos Aires, Argentina (August, 1997), respectively. The NY-2A and AR158 host, *Chlorella* NC64A, was grown on MBBM medium (Van Etten *et al.*, 1983). The NY-2A and AR158 viruses were produced, purified and the viral DNAs were isolated using methods and protocols developed for PBCV-1 (Van Etten *et al.*, 1981 and Van Etten *et al.*, 1983). Both DNAs were sequenced to 8-fold coverage and assembled at The Institute for Genomic Research (TIGR).

Genomic sequence analysis

A potential protein-encoding region, or ORF, 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://www.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 NY-2A and AR158 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 $< 10^{-3}$ are presented. Putative tRNA genes were identified using the tRNAscan-SE program developed by Lowe and Eddy at Washington University School of Medicine in 1997 (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 NY-2A and AR158 DNA sequences have been deposited in the GenBank database (accession number DQ491002 and DQ491003, respectively) and the sequences can also be found at http://greengene.uml.edu/.

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 Po-

sition, A.A. length, Peptide Mw, pI, 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 b002R – b797R
Appendix B: Gene Names B001L – B886R
Appendix C: Gene Names c001R – c814L
Appendix D: Gene Names C006R – C815L

Acknowledgments

We thank James Gurnon for preparing the virus DNAs and Shmuel Pietrokovski for identifying the inteins in NY-2A. This investigation was supported in part by the National Science Foundation grant EF-0333197 (MG and JVE), by the National institutes of Health grant GM32441 (JVE) and by the Center of Biomedical Research Excellence program of the National Center for Research Resources Grant P20-RR15635 (JVE).

References

Ali, M.M.A., Kawasaki, T., Yamada, T., 2005. Genetic rearrangements on the chlorovirus genome that switch between hyaluronan synthesis and chitin synthesis. Virology 342, 102–110.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Amitai, G., Dassa, B., Pietrokovski, S., 2004. Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. J. Biol. Chem. 279, 3121–3131.

Baumann, S., Sander, A., Gurnon, J.R., Yanai-Balser, G., Van Etten, J.L., Piotrowski, M., submitted for publication. *Chlorella* viruses encode a complete polyamine biosynthesis pathway. Virology.

Belfield, G.P., Tuite, M.F., 1993. Translation elongation factor 3: a fungus-specific translation factor? Mol. Microbiol. 9, 411–418.

Belfort, M., Roberts, R.J., 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Res. 25, 3379–3388.

Bubeck, J.A., Pfitzner, A.J.P., 2005. Isolation and characterization of a new type of chlorovirus that infects an endosymbiotic chlorella strain of the heliozoon *Acanthocystis turfacea*. J. Gen. Virol. 86, 2871–2877.

Chakraburtty, K., 2001. Translational regulation by ABC systems. Res. Microbiol. 152, 391–399.

Chan, S.-H., Zhu, Z., Van Etten, J.L., Xu, S.-Y., 2004. Cloning of CviPII nicking and modification system from chlorella virus NYs-1 and application of Nt. CviPII in random DNA amplification. Nucleic Acids Res. 32, 6187–6199.

Chan, S.H., Zhu, Z., Dunigan, D.D., Van Etten, J.L., Xu, S.Y., 2006. Cloning of Nt.CviQII nicking endonuclease and its cognate methyltransferase: M. CviQII methylates AG sequences. Protein Expression Purif. 48, 138–150.

Davis, R.H., Morris, D.R., Coffino, P., 1992. Sequestered end products and enzyme regulation: the case of ornithine decarboxylase. Microbiol. Rev. 56, 280–290.

Dunigan, D.D., Fitzgerald, L.A., Van Etten, J.L., 2006. Phycodnaviruses: a peek at genetic diversity. Virus Res. 117, 119–132.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L., Bateman, A., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247–D251.

Fitzgerald, L.A., Graves, M.V., Li, X., Feldblyum, T., Hartigan, J., Van Etten, J.L., this issue. The sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect Chlorella Pbi. Virology 358, 459–471.

Fortune, J.M., Lavrukhin, O.V., Gurnon, J.R., Van Etten, J.L., Lloyd, R.S., Osheroff, N., 2001. Topoisomerase II from chlorella virus PBCV-1 has an exceptionally high DNA cleavage activity. J. Biol. Chem. 276, 24401–24408.

Frohns, F., Kasmann, A., Kramer, D., Schafer, B., Mehmel, M., Kang, M., Van Etten, J.L., Gazzarrini, S., Moroni, A., Thiel, G., 2006. Potassium ion channels of chlorella viruses cause rapid depolarization of host cells during infection. J. Virol. 80, 2437–2444.

Gerber, A.P., Keller, W., 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem. Sci. 26, 376–384.

Grabherr, R., Strasser, P., Van Etten, J.L., 1992. The DNA polymerase gene from chlorella viruses PBCV-1 and NY-2A contains an intron with nuclear splicing sequences. Virology 188, 721–731.

Graves, M.V., Bernadt, C.T., Cerny, R., Van Etten, J.L., 2001. Molecular and genetic evidence for a virus-encoded glycosyltransferase involved in protein glycosylation. Virology 285, 332–345.

Graziani, S., Xia, Y., Gurnon, J.R., Van Etten, J.L., Leduc, D., Skouloubris, S., Myllykallio, H., Liebl, U., 2004. Functional analysis of FAD-dependent thymidylate synthase ThyX from *Paramecium bursaria* chlorella virus-1. J. Biol. Chem. 279, 54340–54347.

Iyer, L.M., Aravind, L., Koonin, E.V., 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. 75, 11720–11734.

Iyer, L.M., Balaji, S., Koonin, E.V., Aravind, L., 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117, 156–184.

Jeevarajah, D., Patterson, J.H., McConville, M.J., Billman-Jacobe, H., 2002. Modification of glycopeptidolipids by an O-methyltransferase of *Mycobacterium smegmatis*. Microbiology 148, 3079–3087.

Kaiser, A., Vollmert, M., Tholl, D., Graves, M.V., Gurnon, J.R., Xing, W., Lisec, A.D., Nickerson, K.W., Van Etten, J.L., 1999. Chlorella virus PBCV-1 encodes a functional homospermidine synthase. Virology 263, 254–262.

Kang, M., Moroni, A., Gazzarrini, S., DiFrancesco, D., Thiel, G., Severino, M., Van Etten, J.L., 2004. Small potassium ion channel proteins encoded by chlorella viruses. Proc. Natl. Acad. Sci. U.S.A. 101, 5318–5324.

Kawakami, H., Kawakami, N., 1978. Behavior of a virus in a symbiotic system, *Paramecium bursaria*-zoochlorella. J. Protozool. 25, 217–225.

Kawasaki, T., Tanaka, M., Fujie, M., Usami, S., Sakai, K., Yamada, T., 2002. Chitin synthesis in chlorovirus CVK2-infected chlorella cells. Virology 302, 123–131.

Kim, Y., Clark, D.J., 2002. SWI/SNF-dependent long-range remodeling of yeast HIS3 chromatin. Proc. Natl. Acad. Sci. U.S.A. 99, 15381–15386.

Landstein, D., Mincberg, M., Arad, S., Tal, J., 1996. An early gene of the chlorella virus PBCV-1 encodes a functional aspartate transcarbamylase. Virology 221, 151–158.

Lavrukhin, O.V., Fortune, J.M., Wood, T.G., Burbank, D.E., Van Etten, J.L., Osheroff, N., Lloyd, R.S., 2000. Topoisomerase II from chlorella virus PBCV-1: characterization of the smallest known type II topoisomerase. J. Biol. Chem. 275, 6915–6921.

Li, Y., Lu, Z., Burbank, D.E., Kutish, G.F., Rock, D.L., Van Etten, J.L., 1995. Analysis of 43 kb of the chlorella virus PBCV-1 330 kb genome: map position 45 to 88. Virology 212, 134–150.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved

detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Manzur, K.L., Farooq, A., Zeng, L., Plotnikova, O., Koch, A.W., Sachchidanand, Zhou, M.-M., 2003. A dimeric viral SET domain methyltransferase specific to Lys27 of histone H3. Nat. Struct. Biol. 10, 187–196.

Meints, R.H., Van Etten, J.L., Kuczmarski, D., Lee, K., Ang, B., 1981. Viral infection of the symbiotic chlorella-like alga present in *Hydra viridis*. Virology 113, 698–703.

Morehead, T.A., Gurnon, J.R., Adams, B., Nickerson, K.W., Fitzgerald, L.A., Van Etten, J.L., 2002. Ornithine decarboxylase encoded by chlorella virus PBCV-1. Virology 301, 165–175.

Myllykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P., Liebl, U., 2002. An alternative flavin-dependent mechanism for thymidylate synthesis. Science 297, 105–107.

Nandhagopal, N., Simpson, A.A., Gurnon, J.R., Yan, X., Baker, T.S., Graves, M.V., Van Etten, J.L., Rossmann, M.G., 2002. The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. Proc. Natl. Acad. Sci. U.S.A. 99, 14758–14763.

Nishida, K., Suzuki, S., Kimura, Y., Nomura, N., Fujie, M., Yamada, T., 1998. Group I introns found in chlorella viruses: biological implications. Virology 242, 319–326.

Nishida, K., Kawasaki, T., Fujie, M., Usami, S., Yamada, T., 1999. Aminoacylation of tRNAs encoded by chlorella virus CVK2. Virology 263, 220–229.

Paul, J.H.,Williamson, S.J., Long, A., Authement, R.N., John, D., Segall, A.M., Rohwer, F.L., Androlewicz, M., Patterson, S., 2005. Complete genome sequence of phiHSIC, a pseudotemperate marine phage of *Listonella pelagia*. Appl. Environ. Microbiol. 71, 3311–3320.

Perler, F.B., 2002. InBase: the intein database. Nucleic Acids Res. 30, 383–384.

Pietrokovski, S., 1998. Identification of a virus intein and a possible variation in the protein-splicing reaction. Curr. Biol. 8, R634–R635.

Plugge, B., Gazzarrini, S., Nelson, M., Cerana, R., Van Etten, J.L., Derst, C., DiFrancesco, D., Moroni, A., Thiel, G., 2000. A potassium channel protein encoded by chlorella virus PBCV-1. Science 287, 1641–1644.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350.

Roca, J., 1995. The mechanisms of DNA topoisomerases. Trends Biochem. Sci. 20, 156–160.

Shah, R., Coleman, C.S., Mir, K., Baldwin, J., Van Etten, J.L., Grishin, N.V., Pegg, A.E., Stanley, B.A., Phillips, M.A., 2004. *Paramecium bursaria* chlorella virus-1 encodes an unusual arginine decarboxylase that is a close homolog of eukaryotic ornithine decarboxylases. J. Biol. Chem. 279, 35760–35767.

Sugimoto, I., Onimatsu, H., Fujie, M., Usami, S., Yamada, T., 2004. vAL-1, a novel polysaccharide lyase encoded by chlorovirus CVK2. FEBS Lett. 559, 51–56.

Sun, L., Li, Y., McCullough, A.K., Wood, T.G., Lloyd, R.S., Adams, B., Gurnon, J.R., Van Etten, J.L., 2000. Intron conservation in a UV-specific DNA repair gene encoded by chlorella viruses. J. Mol. Evol. 50, 82–92.

Tonetti, M., Zanardi, D., Gurnon, J.R., Fruscione, F., Armirotti, A., Damonte, G., Sturla, L., De Flora, A., Van Etten, J.L., 2003. *Paramecium bursaria* Chlorella virus 1 encodes two enzymes involved in the biosynthesis of GDP-L-fucose and GDP-D-rhamnose. J. Biol. Chem. 278, 21559–21565.

Tyms, A.S., Clarke, J.R., Ford, L., Eloranta, T., 1990. Polyamines and the growth of pathogenic viruses. In: Goldenberg, S.H., Algranti, J.D. (Eds.), The Biology and Chemistry of Polyamines. IRL Press, Oxford, pp. 121–133.

Van Etten, J.L., 2003. Unusual life style of giant chlorella viruses. Annu. Rev. Genet. 37, 153–195.

Van Etten, J.L., Meints, R.H., Burbank, D.E., Kuczmarski, D., Cuppels, D.A., Lane, L.C., 1981. Isolation and characterization of a virus from the intracellular green alga symbiotic with *Hydra viridis*. Virology 113, 704–711.

Van Etten, J.L., Meints, R.H., Kuczmarski, D., Burbank, D.E., Lee, K., 1982. Viruses of symbiotic chlorella-like algae isolated from *Paramecium bursaria* and *Hydra viridis*. Proc. Natl. Acad. Sci. U.S.A. 79, 3867–3871.

Van Etten, J.L., Burbank, D.E., Xia, Y., Meints, R.H., 1983. Growth cycle of a virus, PBCV-1, that infects chlorella-like algae. Virology 126, 117–125.

Van Etten, J.L., Burbank, D.E., Joshi, J., Meints, R.H., 1984. DNA synthesis in a chlorella-like alga following infection with the virus PBCV-1. Virology 134, 443–449.

Van Etten, J.L., Schuster, A.M., Girton, L., Burbank, D.E., Swinton, D., Hattman, S., 1985. DNA methylation of viruses infecting a eukaryotic chlorella-like alga. Nucleic Acids Res. 13, 3471–3478.

Van Etten, J.L., Lane, L.C., Meints, R.H., 1991. Viruses and viruslike particles of eukaryotic algae. Microbiol. Rev. 55, 586–620.

Villarreal, L.P., 2005. The dilemma of the big transition in evolution: the eukaryotes. Viruses and the Evolution of Life. ASM Press,Washington, DC, pp. 101–141.

Villarreal, L.P., DeFilippis, V.R., 2000. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J. Virol. 74, 7079–7084.

Wang, I.N., Li, Y., Que, Q., Bhattacharya, M., Lane, L.C., Chaney, W.G., Van Etten, J.L., 1993. Evidence for virus-encoded glycosylation specificity. Proc. Natl. Acad. Sci. U.S.A. 90, 3840–3844.

Warbrick, E., 2000. The puzzle of PCNA's many partners. BioEssays 22, 997–1006.

Wilson, W.H., Schroeder, D.C., Allen, M.J., Holden, M.T., Parkhill, J., Barrell, B.G., Churcher, C., Hamlin, N., Mungall, K., Norbertczak, H., Quail, M.A., Price, C., Rabbinowitsch, E., Walker, D., Craigon, M., Roy, D., Ghazal, P., 2005. Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. Science 309, 1090–1092.

Wuitschick, J.D., Gershan, J.A., Lochowicz, A.J., Li, S., Karrer, K.M., 2002. A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. Nucleic Acids Res. 30, 2524–2537.

Xia, Y., Narva, K.E., Van Etten, J.L., 1987. The cleavage site of the RsaI isoschizomer, CviII, is G/TAC. Nucleic Acids Res. 15, 10063.

Xia, Y., Morgan, R., Schildkraut, I., Van Etten, J.L., 1988. A site-specific single strand endonuclease activity induced by NYs-1 virus infection of a chlorellalike green alga. Nucleic Acids Res. 16, 9477–9487.

Yamada, T., Fukuda, T., Tamura, K., Furukawa, S., Songsri, P., 1993. Expression of the gene encoding a translational elongation factor 3 homolog of chlorella virus CVK2. Virology 197, 742–750.

Yamada, T., Tamura, K., Aimi, T., Songsri, P., 1994. Self-splicing group I introns in eukaryotic viruses. Nucleic Acids Res. 22, 2532–2537.

Yamada, T., Onimatsu, H., Van Etten, J.L., 2006. Chlorella viruses. In: Maramorosch, K., Shatkin, A. (Eds.), Advances in Virus Research, vol. 66. Elsevier Inc., pp. 293–336.

Zhang, Y., Nelson, M., Nietfeldt, J., Xia, Y., Burbank, D., Ropp, S., Van Etten, J.L., 1998. Chlorella virus NY-2A encodes at least 12 DNA endonuclease/methyltransferase genes. Virology 240, 366–375.

Zhang, Y., Adams, B., Sun, L., Burbank, D.E., Van Etten, J.L., 2001. Intron conservation in the DNA polymerase gene encoded by chlorella viruses. Virology 285, 313–321.