Virology **285,** 313–321 (2001)
doi:10.1006/viro.2001.0935, available online at http://www.idealibrary.com on IDEAL®

Intron Conservation in the DNA Polymerase Gene Encoded by Chlorella Viruses[1]

Yuanzheng Zhang,* Byron Adams,† Liangwu Sun,*,[2] Dwight E. Burbank,*,[3] and James L. Van Etten*,[4]

*Department of Plant Pathology, University of Nebraska at Lincoln, Lincoln, Nebraska 68583-0722; and †Department of Entomology and Nematology, University of Florida, Gainesville, Florida 32611-0620

Received February 8, 2001; returned to author for revision March 8, 2001; accepted March 29, 2001

Previously we reported that 19 of 42 viruses that infect *Chlorella* strain NC64A (NC64A viruses) contain a short, nuclear-located, spliceosomal-processed intron in a pyrimidine dimer-specific glycosylase/apyrimidine lyase (*pdg*) gene. Surprisingly, the nucleotide sequence of the intron region is more conserved than the exon regions of the gene (L. Sun *et al.,* 2000, *J. Mol. Evol.* 50, 82–92). For comparative purposes, we determined the nucleotide sequence of a similar intron type and its flanking coding regions in the DNA polymerase (*dnapol*) gene from the same 42 NC64A viruses and also 5 viruses that infect *Chlorella* strain Pbi. Thirty-eight of the 42 NC64A viruses contained a 101-nucleotide intron and the remaining 4 had an 86-nucleotide intron located in the same position in *dnapol.* The 4 viruses with the smaller intron in *dnapol* also have a smaller intron in their *pdg* gene. There was no intron in the *dnapol* gene of the 5 Pbi viruses. Phylogenetic analyses indicate that the *dnapol* genes containing the 86-nucleotide intron represent the ancestral condition among the NC64A viruses. The intron in the *dnapol* gene is phase 0 (keeps codons intact), which differs from the phase 1 intron in the *pdg* gene. The intron in the *dnapol* gene, unlike the *pdg* intron, was conserved (83 to 100% identical) to about the same extent as the coding regions of the gene (78 to 100% identical).

Key Words: DNA polymerase; intron; dsDNA virus; chlorella viruses; *Phycodnaviridae.*

INTRODUCTION

Large (190 nm in diameter) icosahedral, plaque-forming, dsDNA-containing viruses that infect certain isolates of unicellular, eukaryotic chlorella-like green algae are common in freshwater collected throughout the world (Van Etten *et al.,* 1985a,c; Schuster *et al.,* 1986; Zhang *et al.,* 1988; Yamada *et al.,* 1991). Virions of the prototype chlorella virus, PBCV-1 (Family *Phycodnaviridae,* genus Chlorovirus), contain at least 50 proteins and a lipid membrane located inside the outer glycoprotein capsid (Skrdla *et al.,* 1984; Wang *et al.,* 1993; Yan *et al.,* 2000). The 330,744-bp PBCV-1 genome, a linear nonpermuted dsDNA molecule with covalently closed hairpin ends (Rohozinski *et al.,* 1989; Zhang *et al.,* 1994), encodes 700 open reading frames (ORF) 65 codons or larger, of which 375 are predicted to code for proteins (Van Etten and Meints, 1999).

One of the PBCV-1 encoded proteins resembles the bacteriophage T4 *denV* gene product (Furuta *et al.,* 1997). The *denV* gene encodes a well-characterized pyrimidine dimer-specific glycosylase/apyrimidine lyase, called endonuclease V (T4-PDG), that initiates repair of UV-induced pyrimidine dimers in DNA (Lloyd, 1998; McCullough *et al.,* 1999). The discovery of a PBCV-1-encoded homolog to T4-PDG (41% amino acid identity) led to structural and functional comparisons between the two enzymes (McCullough *et al.,* 1998; Garvish and Lloyd, 1999, 2000). The T4 enzyme cleaves *cis-syn* cyclobutane pyrimidine isomers, whereas the PBCV-1 enzyme (PBCV-1-PDG) cleaves both *cis-syn* and *trans-syn*-II cyclobutane pyrimidine dimers. In addition, PBCV-1-PDG has a stronger electrostatic attraction for DNA than the T4 enzyme; i.e., PBCV-1-PDG is more processive than T4-PDG.

The discovery of functional differences between the PBCV-1 and the T4 enzymes prompted us to characterize PDG homologs from 41 additional viruses that infect *Chlorella* NC64A (NC64A viruses); these viruses were isolated from diverse geographic regions. *Pdg* genes from 15 of these 42 viruses contain a 98-nucleotide pre-mRNA intron and another 4 viruses contain an 81-nucleotide intron; the *pdg* genes in the remaining 23 viruses lack an intron (Sun *et al.,* 2000). The introns, which have the properties of a nuclear-located, spliceosomal-processed intron, interrupt the coding region at identical positions in the *pdg* gene. Surprisingly, the nucleotide sequence of the 98-nucleotide intron is 100% conserved, regardless of the origin of the viruses. Moreover, three of the four 81-nucleotide introns are identical; the fourth differs by 1 nucleotide. In contrast, the exon nucleotide sequences of the *pdg* genes are less conserved. These findings contradict the dogma that intron

[1] This article has been assigned Journal Series No. 13,244, Agricultural Research Division, University of Nebraska.

[2] Current address: Department of Molecular Genetics and Biochemistry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261.

[3] Deceased April 2, 2000.

[4] To whom correspondence and reprint requests should be addressed. Fax: (402) 472-2853. E-mail: jvanetten@unlnotes.unl.edu.

0042-6822/01 $35.00

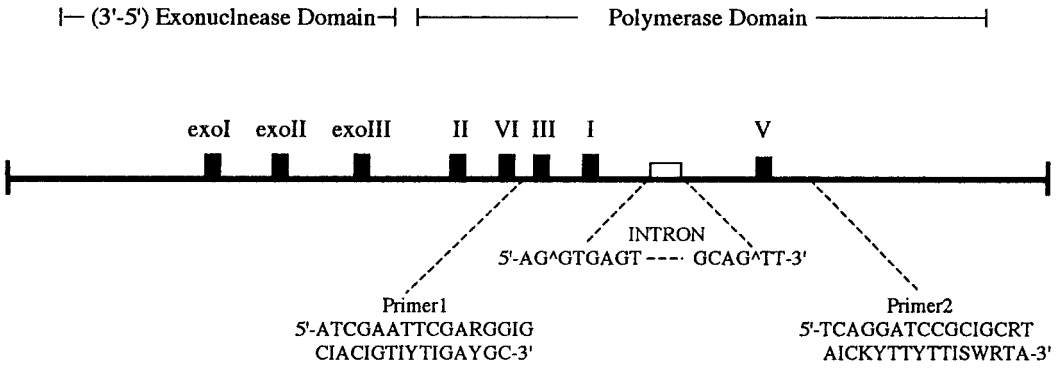

FIG. 1. A physical map of the *dnapol* gene from NC64A viruses PBCV-1 and NY-2A showing conserved motifs (dark boxes) and an intron (open box) located between motifs I and V. Primers 1 and 2 were used to PCR amplify a portion of the *dnapol* gene used in this study. The boundary nucleotide sequences of the intron are also shown.

sequences diverge faster than exon sequences (e.g., Lewin, 1997).

We previously reported that the DNA polymerase genes (*dnapol*) from two NC64A viruses, PBCV-1 and NY-2A, contain a 101-nucleotide and an 86-nucleotide intron, respectively (Grabherr *et al.,* 1992). Both introns are located at the same position in the *dnapol* genes and, like the *pdg* intron, appear to be nuclear-located spliceosomal-processed introns.

In this study we sequenced a portion of the *dnapol* gene from the same 42 NC64A viruses used in the *pdg* intron report, as well as 5 viruses that infect *Chlorella* Pbi (Pbi viruses), in order to answer five questions. (i) How common are introns in the *dnapol* gene? (ii) How conserved are introns in the *dnapol* gene? (iii) Does the *dnapol* intron phylogenetic tree resemble its exon phylogenetic tree? (iv) How similar is the *dnapol* intron to the *pdg* intron? (v) Does the *dnapol* phylogenetic tree resemble the *pdg* phylogenetic tree?

RESULTS AND DISCUSSION

An intron is common in chlorella virus *dnapol* genes

The 913-codon *dnapol* genes from two NC64A viruses, PBCV-1 and NY-2A, have been described (Grabherr *et al.,* 1992). The two polymerases contain all the conserved motifs in the 3′–5′ exonuclease and catalytic domains found in B family δ DNA polymerases (Fig. 1). The conservation of several domains in family A, B, and C DNA polymerases from prokaryotes, eukaryotes, and their viruses indicates that DNA polymerases probably evolved from a common ancestral gene (Jung *et al.,* 1987; Wang *et al.,* 1989; Braithwaite and Ito, 1993). It has been suggested that the DNA polymerase gene from the chlorella viruses resides near the root of the clade containing all the eukaryotic δ DNA polymerases, implying that these are ancient viruses (Villarreal, 1999; Villarreal and DeFilippis, 2000).

The protein-coding regions of the two NC64A virus *dnapol* genes have 76% nucleotide sequence identity and a deduced 88% amino acid sequence identity. The PBCV-1 *dnapol* contains a 101-nucleotide intron and the NY-2A gene contains an 86-nucleotide intron located in the same coding position of the gene (Fig. 1) (Grabherr *et al.,* 1992). Both genes have 5′ and 3′ splice-site sequences characteristic of nuclear-located, spliceosomal processed introns (Fig. 1).

To address the questions listed in the Introduction, we cloned and sequenced a portion of the *dnapol* gene from the same 42 NC64A viruses used in the *pdg* study (Table 1). *dnapol* genes from five Pbi viruses were also examined. Twenty-nine of the 42 NC64A viruses were isolated from water samples collected throughout the United States in 1983–1984 (Van Etten *et al.,* 1985a,c; Schuster *et al.,* 1986) and 7 were isolated from water collected in China in 1987 (Zhang *et al.,* 1988). We have grouped these 37 viruses into 16 classes by several criteria (Table 1) (Van Etten *et al.,* 1991). To increase the geographic diversity in this study, we included an additional virus from China, 2 viruses from Argentina, and 1 each from Israel and Australia; these 5 NC64A viruses were isolated in 1995–1997. The 5 Pbi viruses were isolated in Germany in 1986–1987 (Reisser *et al.,* 1988).

Using PBCV-1 and NY-2A *dnapol* as a guide, degenerate oligonucleotide primers were designed to amplify a portion of *dnapol,* including the intron (Fig. 1); the region matched PBCV-1 nucleotide position 95983 to 96744 (Accession No. U42580). The expected 660-nucleotide, amino-acid-encoding, polymerase chain reaction (PCR) products contain the most highly conserved amino acid sequence in δ DNA polymerases, the catalytic YGDTDS site in motif I (Wang, 1991).

PCR products from 38 of the 42 NC64A viruses were the same size (761 nucleotides) as the product from PBCV-1 (Table 1). PCR products from the remaining 4 viruses were identical in size (746 nucleotides) to the product from virus NY-2A. Using different primers (Materials and Methods), PCR products from the 5 Pbi viruses

TABLE 1

Chlorella NC64A Viruses Used in This Study and the Size of the PCR Product Using *dnapol* Primers

Virus	Virus isolated from	Date collected	Class[a]	Intron size in *pdg* gene (bp)[b]	PCR product size with *dnapol* primers (bp)
NE-8D	Nebraska, U.S.A.	Sept. 1984	1	N	761
NYb-1	New York, U.S.A.	Aug. 1984	1	N	761
CA-4B	California, U.S.A.	Nov. 1984	1	N	761
AL-1A	Alabama, U.S.A.	Oct. 1984	2	98	761
NY-2C	New York, U.S.A.	Aug. 1984	2	98	761
NC-1D	North Carolina, U.S.A.	Oct. 1983	2	98	761
PBCV-1	North Carolina, U.S.A.	1981	3	N	761
NC-1C	North Carolina, U.S.A.	Oct. 1983	3	98	761
CA-1A	California, U.S.A.	Nov. 1984	4	98	761
CA-2A	California, U.S.A.	Nov. 1984	4	N	761
IL-2A	Illinois, U.S.A.	Oct. 1983	4	98	761
IL-2B	Illinois, U.S.A.	Oct. 1983	4	N	761
IL-3A	Illinois, U.S.A.	Oct. 1983	4	98	761
IL-3D	Illinois, U.S.A.	Oct. 1983	4	N	761
SC-1A	South Carolina, U.S.A.	Oct. 1983	5	98	761
SC-1B	South Carolina, U.S.A.	Oct. 1983	5	98	761
NC-1A	North Carolina, U.S.A.	Oct. 1983	6	N	761
NE-8A	Nebraska, U.S.A.	Sept. 1984	7	N	761
AL-2C	Alabama, U.S.A.	Oct. 1984	7	N	761
MA-1E	Massachusetts, U.S.A.	Aug. 1984	7	98	761
NY-2F	New York, U.S.A.	Aug. 1984	7	N	761
CA-1D	California, U.S.A.	Nov. 1984	7	98	761
NC-1B	North Carolina, U.S.A.	Oct. 1983	7	N	761
NYs-1	New York, U.S.A.	Aug. 1984	8	81	746
IL-5-2s1	Illinois, U.S.A.	May 1984	9	N	761
AL-2A	Alabama, U.S.A.	Oct. 1984	9	98	761
MA-1D	Massachusetts, U.S.A.	Aug. 1984	9	N	761
NY-2B	New York, U.S.A.	Aug. 1984	9	81	746
CA-4A	California, U.S.A.	Nov. 1984	10	N	761
NY-2A	New York, U.S.A.	Aug. 1984	11	81	746
XZ-3A	Xuzhou, China	Mar. 1987	12	N	761
SH-6A	Shanghai, China	Mar. 1987	13	N	761
BJ-2C	Beijing, China	Mar. 1987	13	N	761
XZ-6E	Xuzhou, China	Mar. 1987	14	N	761
XZ-4C	Xuzhou, China	Mar. 1987	15	N	761
XZ-5C	Xuzhou, China	Mar. 1987	16	N	761
XZ-4A	Xuzhou, China	Mar. 1987	16	N	761
IS-10	Israel	Aug. 1996		N	761
CH-57	Baoding, China	Aug. 1997		N	761
AN69C	Canberra, Australia	Mar. 1995		98	761
AR158	Buenos Aires, Argentina	Aug. 1997		81	746
AR93-2	Buenos Aires, Argentina	Aug. 1997		98	761

[a] The first 37 viruses have been separated into 16 classes as indicated (Van Etten *et al.,* 1991). The last 5 viruses were isolated recently and were included to increase the geographic diversity of the viruses.

[b] This column indicates the size (in nucleotides) of the intron in the *pdg* gene from the viruses (Sun *et al.,* 2000). An N means no intron was present.

contained 645 nucleotides. Sequencing the PCR products from all the viruses revealed that the 38 NC64A viruses with the 761-nucleotide PCR product, like PBCV-1, contained a 101-nucleotide intron. The remaining 4 NC64A viruses had an 86-nucleotide intron. All the introns were located at the same position in the *dnapol* coding region. A previous study established that, as expected, the intron region was removed from *dnapol* mRNA isolated from cells infected with viruses PBCV-1 and NY-2A (Grabherr *et al.,* 1992). Consequently, we assume that the intron region is deleted in the mature mRNA of all the viruses. The *dnapol* genes from the 5 Pbi viruses lacked an intron in this position. No PCR products were obtained with host DNA.

Sequence identity in the *dnapol* exon region

The 660-nucleotide sequences in the *dnapol* coding regions from all the NC64A viruses were compared to a consensus *dnapol* (data available upon request). The

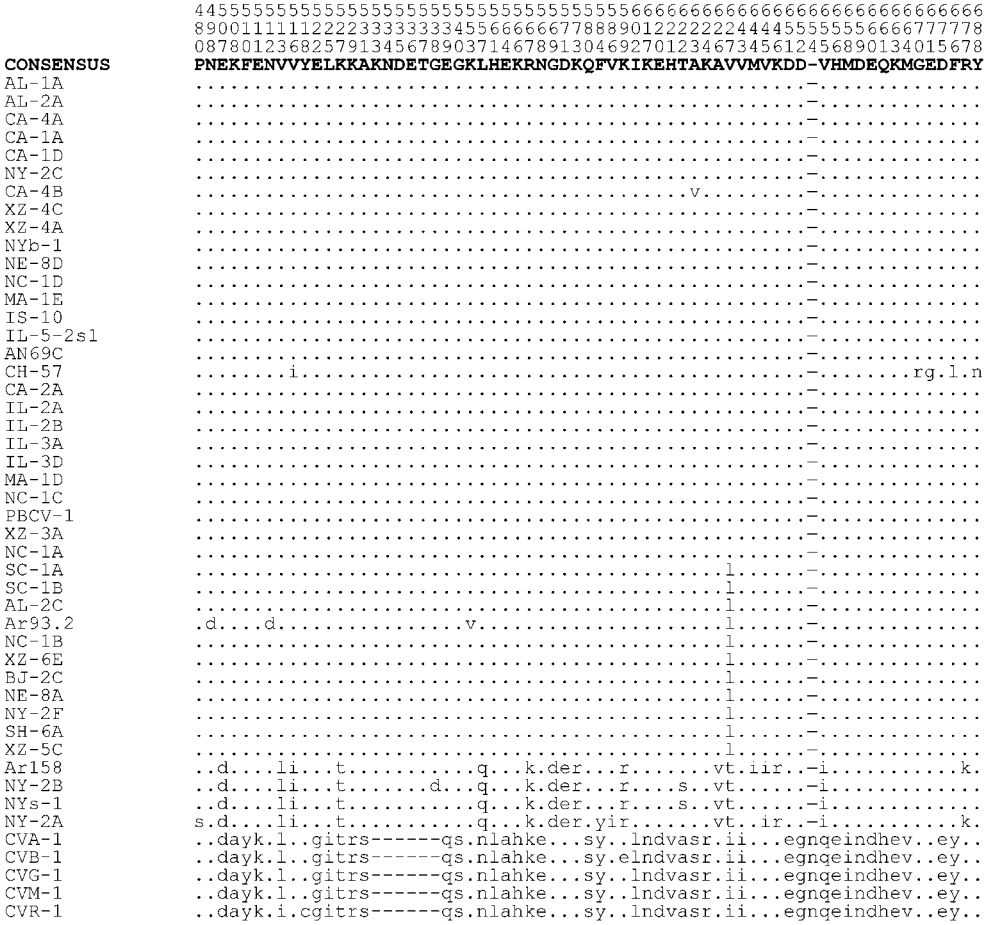

FIG. 2. Amino acid differences in 47 chlorella virus DNA polymerases. The numbers above the amino acids represent the position numbers of the amino acids. Only amino acids that differed from the consensus sequence (top) are included and indicated in lowercase. A (•) indicates no change in an amino acid. A (–) means no equivalent amino acid in that location.

nucleotide sequences from viruses NC-1C and XZ-3A were identical to that of PBCV-1, whereas the sequences from the remaining 35 101-nucleotide intron-containing viruses were 93 to 99.7% identical to PBCV-1. Nucleotide sequences of *dnapol* coding regions from the 4 86-nucleotide intron-containing viruses were 93 to 95% identical to one another and were 78 to 80% identical to the PBCV-1 sequence. Including all the NC64A viruses, 188 of the 660 nucleotides in the *dnapol* coding region differed from a consensus *dnapol.* One hundred forty-three, or 76%, of these 188 nucleotides differed in the third position of the codon. Nucleotide sequences of *dnapol* genes from the 5 Pbi viruses were ~70% identical to that of PBCV-1. Combining the NC64A and Pbi virus sequences, 295 of the 660 nucleotides in the coding region differed from a consensus *dnapol* and 198 or 67% of these differences were in the third position of the codon.

Not surprisingly, amino acid sequences of the DNA polymerases were more conserved. In fact, the predicted amino acid sequences from 24 of the 38 101-nucleotide intron-containing NC64A viruses were identical to that of PBCV-1, whereas the other 13 viruses were 97.7 to 99.5% identical to PBCV-1 (Fig. 2). The predicted amino acid sequences from the 4 86-nucleotide intron-containing NC64A viruses were 91 to 94% identical to that of PBCV-1 (Fig. 2). The DNA polymerase amino acid sequences from the 5 Pbi viruses, which lack an intron, were 88% identical to that of PBCV-1 DNA polymerase.

Ratios of synonymous and nonsynonymous nucleotide substitutions between 10 taxa were compared using the DIVERGE program (Table 2). As expected, synonymous substitutions were more frequent than nonsynonymous substitutions. However, the proportion of synonymous to nonsynonymous substitutions in the *dnapol* exon is greater than in the *pdg* exon.

Sequence identity of the *dnapol* introns

A comparison of the 101-nucleotide intron sequences revealed that viruses NC-1A, NC-1C, and XZ-3A were identical to PBCV-1 and the other viruses were 95 to 99% identical to PBCV-1 (Fig. 3). The sequences of the 86-nucleotide introns from viruses NY-2B, NYs-1, and Ar158

TABLE 2

Synonymous Substitutions and Nonsynonymous Substitutions among Representative *dnapol* Genes Estimated Using DIVERGE (GCG Version 10.1)[a]

	AL-1A	AN69C	Ar158	CH57	IS10	MA-1D	NE-8A	NY-2B	PBCV-1	XC-4C
AL-1A	—	0.45	175.99	0.00	0.00	5.84	28.92	141.04	2.86	0.00
AN69C	0.00	—	179.22	0.45	0.45	6.34	29.72	145.07	3.33	0.45
Ar158	5.61	5.61	—	175.29	175.99	167.61	175.86	10.69	164.63	175.99
CH57	1.05	1.05	6.38	—	0.00	5.29	28.24	139.31	2.86	0.00
IS10	0.00	0.00	5.61	1.05	—	5.84	28.92	141.04	2.86	0.00
MA-1D	0.00	0.00	5.61	1.17	0.00	—	28.15	143.18	5.97	5.84
NE-8A	0.18	0.18	5.66	1.35	0.18	0.18	—	144.09	27.52	28.92
NY-2B	4.92	4.92	1.30	5.68	4.92	4.92	4.97	—	133.93	141.04
PBCV-1	0.00	0.00	5.62	1.06	0.00	0.00	0.18	4.92	—	2.86
XZ-4C	0.00	0.00	5.61	1.05	0.00	0.00	0.18	4.92	0.00	—

[a] Upper triangle, synonymous substitutions per 100 synonymous sites. Lower triangle, nonsynonymous substitutions per 100 nonsynonymous sites.

were 97 to 99% identical to that of NY-2A, but only 83 to 84% identical to that of PBCV-1. The size difference between the 101- and the 86-nucleotide introns appears to result from a 15-nucleotide deletion or insertion. There are also single-nucleotide differences between the 101- and the 86-nucleotide introns (Fig. 3).

```
                1                                                  50 51                                                 100 101
CONSENSUS  GTGAGTTCAA AAAACAGCGA GCGGGGATTG TACAATTTTG TGATTTACAA CCCAGACTTC ACGTTTCTAA ACATTCCATC ACCTTCTGTT CCTTGTTGCA G
AL-1A      .......t.. ...g...... .......... .....c.... .......... .......... .......... .......... ..g....... ..a..c.... .
AL-2A      .......t.. ...g...... .......... .....c.... .......... .......... .......... .......... ..g....... ..a..c.... .
CA-4A      .......t.. ...g...... .......... .....c.... .......... .......... .......... .......... ..g....... ..a..c.... .
CA-1A      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
CA-1D      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
CA-2A      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
IL-2A      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
IL-2B      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
IL-3A      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
IL-3D      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
MA-1D      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
NY-2C      .......... .......... a......... .......... .......... .......... .......... .......... .......... .......... .
CA-4B      .......... .....g.... .......... .......... .......... .......... .......... .......... ..g....... .......... .
XZ-4C      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
XZ-4A      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
NYb-1      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
NE-8D      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
NC-1D      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
MA-1E      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
IS-10      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
IL-5-2s1   .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
AN69C      .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
CH-57      .......... .......... .......... .......... .......... .......... .......... .......... ....c..... .......... .
NC-1A      .......... .......... .......... .......... .......... .......... .......... .......... g.g....... .......... .
NC-1C      .......... .......... .......... .......... .......... .......... .......... .......... g.g....... .......... .
PBCV-1     .......... .......... .......... .......... .......... .......... .......... .......... g.g....... .......... .
XZ-3A      .......... .......... .......... .......... .......... .......... .......... .......... g.g....... .......... .
SC-1A      .....a.... ...g...... .......... ........a.. .a........ .......... .......... .......... ..g....... .......... .
SC-1B      .....a.... ...g...... .......... ........a.. .a........ .......... .......... .......... ..g....... .......... .
AL-2C      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
Ar93.2     .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
BJ-2C      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
NC-1B      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
NE-8A      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
NY-2F      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
SH-6A      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
XZ-5C      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
XZ-6E      .....a.... ...g...... .......... .......... .......... .......... .......... .......... ..g....... .......... .
Ar158      ........c. ........a. .a.a...... ..t..c.... ........g. t..t..t... ..a..c..g. .......--- ----.g..-- ------.... .
NY-2A      ........c. ........a. .a.a...... ..t..c.... ........g. t..t..t... ..a..c..gg .......--- ----.g..-- ------.... .
NY-2B      ........t. ........a. .a.a.a.... ..t..c.... ........g. t..t..t... ..a..c.... .......--- ----.g..-- ------.... .
NYs-1      ........t. ........a. .a.a.a.... ..t..c.... ........g. t..t..t... ..a..c.... .......--- ----.g..-- ------.... .
```

FIG. 3. The nucleotide sequences of 101-nucleotide introns (top 38) and 86-nucleotide introns (bottom 4) in the NC64A virus *dnapol* genes. Lowercase letters are nucleotides that are altered. A dash (–) indicates the absence of a nucleotide.

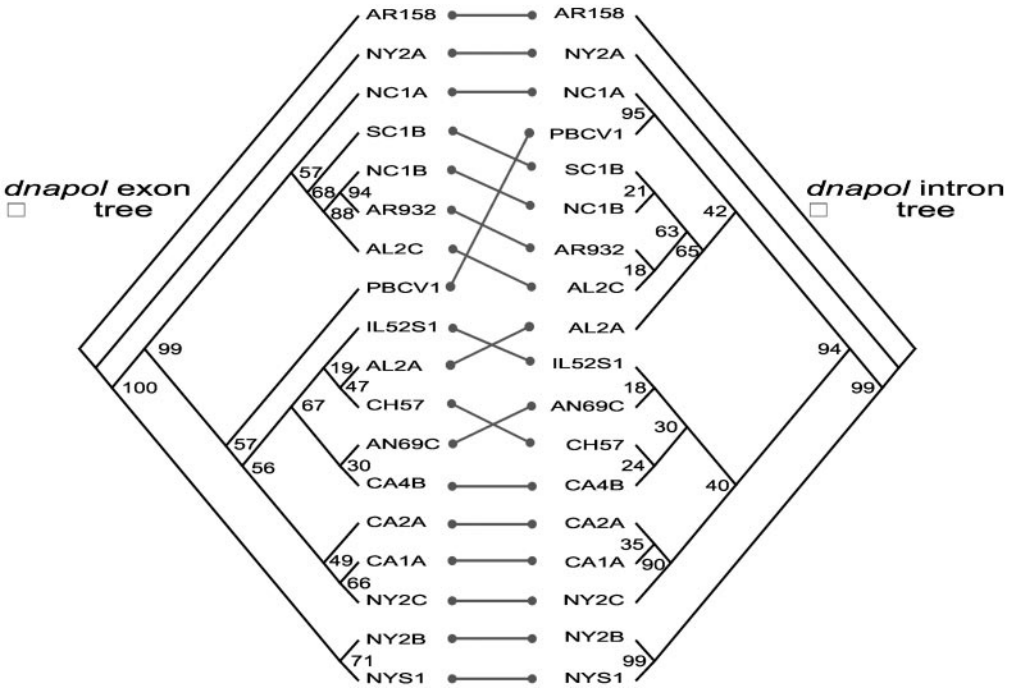

FIG. 4. Comparison of *dnapol* intron and exon phylogenies. Tree on the left is the *dnapol* exon gene tree, the tree on the right is the *dnapol* intron tree. Branch support indices appear at each node (percentage of bipartitions found in one or more trees, 10,000 quartet puzzling steps). Incongruent tracks depict intron/exon phylogenetic discordance and possible recombination events. Redundant exon genes (invariant sequences) and their intron counterparts were removed prior to mapping.

The A+T contents of the 101- and 86-nucleotide introns are 58 and 64%, respectively, whereas the *dnapol* coding regions of the NC64A viruses are about 53% A+T. The A+T contents for the entire coding region of the *dnapol* genes from viruses PBCV-1 and NY-2A are 56 and 54%, respectively (Grabherr *et al.,* 1992), which is slightly less than the 60% A+T content of NC64A virus genomes (Van Etten *et al.,* 1985b). Both introns contain three internal translational stop codons and neither intron encodes a significant ORF. The A+T content of the sequenced portion of the *dnapol* genes from the Pbi viruses is 52%, which is close to the 54% for the entire genome (Reisser *et al.,* 1988).

Phylogenetic analyses of the *dnapol* introns and exons

Intron and exon sequences were analyzed phylogenetically to determine the evolutionary relationships of *dnapol* from the 42 NC64A viruses. Maximum likelihood parameters were estimated, and trees were constructed, using quartet puzzling (Puzzle 4.0.2, Strimmer and von Haeseler, 1996; PAUP* 4.0b4, Swofford, 2000) based on models of evolution that "best fit" the two data sets as determined by likelihood ratio tests (Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998).

The likelihood ratio tests suggest a more complex pattern of nucleotide substitution for the exon compared to the intron. Evolution of the exon sequences is best described by a time-reversible model (Tamura and Nei, 1993) in which all six classes of substitution occur at different rates, but with some sites constrained so as not to evolve at all. The intron data best fit a simple Kimura two-parameter model of evolution (Kimura, 1980), consistent with a more neutral substitution pattern. However, the rate of evolution in the intron may be slightly slower than that of the exon. As calculated using PAUP* (Version 4.0b4a, Swofford, 2000), uncorrected "*P*" distances for the exon varied from 0.15 to 22.58%, whereas sequence divergence among the intron sequences ranged from 0 to 19.32%.

A partition homogeneity test (Farris *et al.,* 1994) suggested that the two data sets contained concordant phylogenetic signals and share a common evolutionary history ($P = 0.43$). The intron and exon trees were topologically similar. Where the two trees were incongruent, nodal support was dubious (Fig. 4). Discordant relationships between exons and their respective introns occurred for PBCV-1, AI-2A, AN69C, CH57, IL-5-2s1, NC-1B, AL-2C, SC-1B, and AR93-2. However, of these, only the AR93-2 and PBCV-1 genes show intron/exon discordance for nodes supported by greater than 75% puzzling support value, or bipartitions found in one or more trees (1000 quartet puzzling steps) (Fig. 4). Therefore, although mapping (TreeMap Version 1.0b, Page, 1995) relationships between exons and their introns suggest several recombination events, confidence in inferring such

events is compromised by poorly supported phylogenetic trees. Further evidence contradicting the notion of frequent recombination is the similarity between the *pdg* and the *dnapol* gene trees. The *pdg* and *dnapol* gene trees are congruent except for the position of PBCV-1 relative to NC-1A, a relationship not well resolved by either tree (this study; Sun *et al.,* 2000).

When rooted with orthologous *dnapol* genes from five Pbi viruses, the four NC64A viruses with the small 81-nucleotide intron (NY-2B, NYs-1, NY-2A, and AR158) appear to branch prior to the diversification of the remaining viruses. This agrees with phylogenetic analyses of the *pdg* gene tree from the same NC64A viruses (Sun *et al.,* 2000). This relationship suggests early acquisition of the short intron in the *dnapol* gene.

Comparison of *dnapol* and *pdg* introns

Splicesomal-processed introns are classified into four types based on their 5′ and 3′ splice-site borders and the small ribonucleoprotein particles used in the splicing process (Sharp and Burge, 1997). The types are U2-type GT-AG (the most common type encompassing ~95% of introns), U2-type AT-AC, U12-type GT-AG, and U12-type AT-AC. The U2-type GT-AG introns contain 5′-KAG-GTRAGT and 3′-YnYAG G splice-site sequences and a CTRAY branch-point sequence. The *dnapol* intron in the chlorella viruses have 5′-GAG^GTGAGT and 3′-(T/G)GCAG^TT sequences at their intron borders. Putative C(A/T)GA(C/T) branch-point sequences with lariat-forming adenine residues are 46 or 31 nucleotides upstream of the 3′ splice site for the 101-nucleotide or 86-nucleotide intron, respectively (Fig. 4). These sequences indicate that the *dnapol* intron belongs to the U2-type GT-AG class.

Introns in the *pdg* gene contain 5′-CAG^GTATGT and 3′-TTTGCAG^AA splice-site sequences. Putative CT(T/C)AA branch-point sequences with lariat-forming adenine residues are located 48 nucleotides upstream of the 3′ splice site (Sun *et al.,* 2000). These sequences indicate that the *pdg* intron, like the *dnapol* intron, belongs to the U2-type GT-AG class.

However, several differences exist between *pdg* and *dnapol* introns. (i) Although nucleotide sequences of the 5′ and 3′ splice-site borders in the *dnapol* and *pdg* introns are similar, the remainder of the *dnapol* intron is only ~37% identical to the *pdg* intron. (ii) The A+T content of the *dnapol* intron (57 and 64% for the long and short intron, respectively) is lower than that of the *pdg* intron (83 and 78% for the long and short intron, respectively). (iii) The sequence of the 101- and 86-nucleotide *dnapol* introns varies to about the same extent as that of the exon regions. In contrast, the 98- and 81-nucleotide *pdg* introns are more conserved than the coding regions (Sun *et al.,* 2000). (iv) Like most spliceosomal-processed introns (Long and Deutsch, 1999), the *dnapol* intron is in phase 0 (the codons remain intact), whereas the *pdg* intron is in phase 1. (v) All of the 42 NC64A viruses examined in this study have an intron in their *dnapol* gene, whereas only 19 of these 42 NC64A viruses contain an intron in the *pdg* gene.

CONCLUSIONS

A U2-type GT-AG pre-mRNA spliceosomal-processed intron is ubiquitous in the *dnapol* gene of the *Chlorella* NC64A viruses; the *dnapol* gene of Pbi viruses lacks this intron. The *dnapol* intron in NC64A viruses is less conserved than the intron in the *pdg* gene. The *dnapol* intron phylogenetic tree resembles the corresponding exon phylogenetic tree. Although both the *dnapol* and the *pdg* genes contain U2-type GT-AG introns, several differences exist between the two introns. The phylogenetic tree of the *dnapol* gene resembles the phylogenetic tree of the *pdg* gene. Interestingly, four NC64A viruses, NY-2B, NYs-1, NY-2A, and Ar158, have a smaller intron in both *dnapol* and *pdg* genes. Furthermore, phylogenetic analyses of both genes suggest that these four viruses represent the ancestral condition of the NC64A viruses. It will be interesting to determine if other chlorella virus genes exhibit this pattern.

MATERIALS AND METHODS

Viruses and host strains

The geographic sources of the 42 *Chlorella* NC64A viruses and the years they were isolated are listed in Table 1. The growth of the host algae, *Chlorella* strain NC64A on MBBM medium and *Chlorella* strain Pbi on FES medium, the plaque assay, the production of the viruses, and the isolation of virus DNAs have been described (Van Etten *et al.,* 1981, 1983a,b; Reisser *et al.,* 1988).

Polymerase chain reaction

Single plaques from chlorella viruses were transferred with sterile toothpicks to 200 μl of 50 mM Tris–HCl, pH 7.5. After soaking for 2 h, 50-μl aliquots were boiled for 10 min and the samples were used as templates for PCR.

Degenerate primers (primer 1, 5′-ATCGAATTCGARG-GIGCIACIGTIYTIGAYGC-3′, and primer 2, 5′-TCAGGATC-CGCIGCRTAICKYTTYTTISWRTA-3′) constructed to conserved regions of the PBCV-1 and NY-2A *dnapol* genes were used to amplify a portion of the *dnapol* gene from the NC64A viruses. This portion of the *dnapol* gene encodes the highly conserved motif II and motif V domains plus the intron region in the *dnapol* genes from viruses PBCV-1 and NY-2A (Fig. 1). Primers 1 and 2 contained an *Eco*RI and a *Bam*HI restriction site, respectively. The PCR products were cloned into the *Eco*RI and *Bam*HI site of pBluescript KS(+) (Stratagene, La Jolla, CA) before sequencing.

Primer 3 (5′-CGGAATTCAAGAAGGGAGCATACTTCA-CGC-3′) and primer 4 (5′-GCTCTAGACAAAATGTAAGGG-TAATAGATC-3′) were used to amplify an equivalent portion of the *dnapol* gene from the five Pbi viruses. Primer 3 contained an *Eco*RI restriction site and primer 4 contained an *Xba*I restriction site. The PCR products from primers 3 and 4 were cloned into the *Eco*RI/*Xba*I site of pBluescript KS(+) before sequencing.

DNA was amplified with *Taq* DNA polymerase (Sigma, St. Louis, MO) in 100-μl reactions which contained ~1 pg (2 μl) of virus DNA; 100 pM each primer; 0.2 mM each dATP, dGTP, dCTP, and dTTP; 10 μl of Mg^{2+}-free *Taq* DNA polymerase buffer (Sigma); and 0.25 mM Mg^{2+} by 35 cycles of heating and cooling: 1 min at 94°C for denaturing, 2 min at 45°C for annealing, 2 min at 72°C for elongation, and finally 7 min at 72°C.

To check for PCR-produced artifacts, at least two independent PCR products were sequenced from each viral DNA. If the two sequences differed, additional PCR products were sequenced.

Other procedures

Cloned PCR products were sequenced at the University of Nebraska Center for Biotechnology DNA sequencing core facility. DNA, RNA, and protein sequences were analyzed with the University of Wisconsin Genetics Computer Group Version 10.1 package of programs (Genetics Computer Group, 2000). Computer programs used for evolutionary and phylogenetic analyses are described under Results and Discussion. The 47 *dnapol* genes are deposited in GenBank under Accession Nos. AF344198 to AF344244.

ACKNOWLEDGMENTS

We thank James Gurnon for technical assistance and Les Lane and Mike Nelson for critical comments on the manuscript. The viruses from Argentina and Israel were provided by Mike Nelson. This investigation was supported in part by NIH Grant GM32441 (J.V.E.).

REFERENCES

Braithwaite, D. K., and Ito, J. (1993). Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res.* **21,** 787–802.

Farris, J. S., Kallersjo, M., Kluge, A. G., and Bult, C. (1994). Testing significance of incongruence. *Cladistics* **10,** 315–319.

Furuta, M., Schrader, J. O., Schrader, H. S., Kokjohn, T. A., Nyaga, S., McCullough, A. K., Lloyd, R. S., Burbank, D. E., Landstein, D., Lane, L., and Van Etten, J. L. (1997). Chlorella virus PBCV-1 encodes a homolog of the bacteriophage T4 UV damage repair gene *denV. Appl. Environ. Microbiol.* **63,** 1551–1556.

Garvish, J. F., and Lloyd, R. S. (1999). The catalytic mechanism of a pyrimidine dimer-specific glycosylase (pdg)/abasic lyase, chlorella virus-pdg. *J. Biol. Chem.* **274,** 9786–9794.

Garvish, J. F., and Lloyd, R. S. (2000). Active-site determination of a pyrimidine dimer glycosylase. *J. Mol. Biol.* **295,** 479–488.

Genetics Computer Group (2000). Wisconsin Package Version 10.1. Madison, WI.

Grabherr, R., Strasser, P., and 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.

Huelsenbeck, J. P., and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28,** 437–466.

Jung, G. H., Leavitt, M. C., Hsieh, J. C., and Ito, J. (1987). Bacteriophage PRD1 DNA polymerase: Evolution of DNA polymerases. *Proc. Natl. Acad. Sci. USA* **84,** 8287–8291.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120.

Lewin, B. (1997). "Genes VI," p. 690. Oxford Univ. Press, Oxford.

Lloyd, R. S. (1998). The initiation of DNA base excision repair of dipyrimidine photoproducts. *Prog. Nucleic Acids Res.* **62,** 155–175.

Long, M., and Deutsch, M. (1999). Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Mol. Biol. Evol.* **16,** 1528–1534.

McCullough, A. K., Dodson, M. L., and Lloyd, R. S. (1999). Initiation of base excision repair: Glycosylase mechanisms and structures. *Annu. Rev. Biochem.* **68,** 255–285.

McCullough, A. K., Romberg, M. T., Nyaga, S., Wei, Y., Wood, T. G., Taylor, J. S., Van Etten, J. L., Dodson, M. L., and Lloyd, R. S. (1998). Characterization of a novel *cis-syn* and *trans-syn*-II pyrimidine dimer glycosylase/AP lyase from a eukaryotic algal virus, *Paramecium bursaria* chlorella virus-1. *J. Biol. Chem.* **273,** 13136–13142.

Page, R. D. M. (1995). Parallel phylogenies: Reconstructing the history of host–parasite assemblages. *Cladistics* **10,** 155–173.

Posada, D., and Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14,** 817–818.

Reisser, W., Burbank, D. E., Meints, S. M., Meints, R. H., Becker, B., and Van Etten, J. L. (1988). A comparison of viruses infecting two different chlorella-like green algae. *Virology* **167,** 143–149.

Rohozinski, J., Girton, L. E., and Van Etten, J. L. (1989). Chlorella viruses contain linear nonpermuted double strand DNA genomes with covalently closed hairpin ends. *Virology* **168,** 363–369.

Schuster, A. M., Burbank, D. E., Meister, B., Skrdla, M. P., Meints, R. H., Hattman, S., Swinton, D., and Van Etten, J. L. (1986). Characterization of viruses infecting a eukaryotic chlorella-like green alga. *Virology* **150,** 170–177.

Sharp, P. A., and Burge, C. B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91,** 875–879.

Skrdla, M. P., Burbank, D. E., Xia, Y., Meints, R. H., and Van Etten, J. L. (1984). Structural proteins and lipids in a virus, PBCV-1, which replicates in a chlorella-like alga. *Virology* **135,** 308–315.

Strimmer, K., and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13,** 964–969.

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

Swofford, D. L. (2000). "PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)," Version 4. Sinauer, Sunderland, MA.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10,** 512–526.

Van Etten, J. L., Burbank, D. E., Kuczmarski, D., and Meints, R. H. (1983a). Virus infection of culturable chlorella-like algae and development of a plaque assay. *Science* **219,** 994–996.

Van Etten, J. L., Burbank, D. E., Schuster, A. M., and Meints, R. H. (1985a). Lytic viruses infecting a chlorella-like alga. *Virology* **140,** 135–143.

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

Van Etten, J. L., Lane, L. C., and Meints, R. H. (1991). Viruses and virus-like particles of eukaryotic algae. *Microbiol. Rev.* **55,** 586–620.

Van Etten, J. L., and Meints, R. H. (1999). Giant viruses infecting algae. *Annu. Rev. Microbiol.* **53,** 447–494.

Van Etten, J. L., Meints, R. H., Burbank, D. E., Kuczmarski, D., Cuppels, D. A., and 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., Schuster, A. M., Girton, L., Burbank, D. E., Swinton, D., and Hattman, S. (1985b). DNA methylation of viruses infecting a eukaryotic chlorella-like green alga. *Nucleic Acids Res.* **13,** 3471–3478.

Van Etten, J. L., Van Etten, C. H., Johnson, J. K., and Burbank, D. E. (1985c). A survey for viruses from fresh water that infect a eukaryotic chlorella-like green alga. *Appl. Environ. Microbiol.* **49,** 1326–1328.

Villarreal, L. P. (1999). DNA viruses: Their influence on host evolution. *In* "Origin and Evolution of Viruses" (E. Domingo, R. Webster, J. J. Holland, and T. Pickett, Eds.), pp. 391–420. Academic Press, London.

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

Wang, T. S. F. (1991). Eukaryotic DNA polymerases. *Annu. Rev. Biochem.* **60,** 513–552.

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

Wang, T. S., Wong, S. W., and Korn, D. (1989). Human DNA polymerase alpha: Predicted functional domains and relationships with viral DNA polymerase. *FASEB J.* **3,** 14–21.

Yamada, T., Higashiyama, T., and Fukuda, T. (1991). Screening of natural waters for viruses which infect chlorella cells. *Appl. Environ. Microbiol.* **57,** 3433–3437.

Yan, X., Olson, N. H., Van Etten, J. L., Bergoin, M., Rossmann, M. G., and Baker, T. S. (2000). Structure and assembly of large lipid-containing dsDNA viruses. *Nat. Struct. Biol.* **7,** 101–103.

Zhang, Y., Burbank, D. E., and Van Etten, J. L. (1988). Chlorella viruses isolated in China. *Appl. Environ. Microbiol.* **54,** 2170–2173.

Zhang, Y., Strasser, P., Grabherr, R., and Van Etten, J. L. (1994). Hairpin loop structure at the termini of the chlorella virus PBCV-1 genome. *Virology* **202,** 1079–1082.