The enigmatic genome of Chara australis virus

Adrian J. Gibbs,† Marjo Torronen,‡ Anne M. Mackenzie,§ Jeffery T. Wood,∥ John S. Armstrong,† Hideki Kondo,¶ Tetsuo Tamada‖ and Paul L. Keese#

Research School of Biological Science, Australian National University, Canberra, ACT 0200, Australia

Most of the genomic sequence of Chara australis virus (CAV), previously called Chara corallina virus, has been determined. It is a ssRNA molecule of 9065 nt with at least four ORFs. At its 5’ end is an ORF encoding a protein of 227 kDa, distantly homologous to the multifunctional replicases of benyviruses and rubiviruses. Next is an ORF encoding a protein of 44 kDa, homologous to the helicases of pestiviruses. The third ORF encodes an unmatched protein of 38 kDa that is probably a movement protein. The fourth and 3’-terminal ORF encodes a protein of 17.7 kDa homologous to the coat proteins of tobamoviruses. The short methyltransferase region of the CAV replicate matches only the C-terminal motif of benyvirus methyltransferases. This and other clues indicate that approximately 11 % and 2 % of the 5’ and 3’ termini of the complete CAV genome, respectively, are missing from the sequence. The aligned amino acid sequences of the CAV proteins and their nearest homologues contain many gaps but relationships inferred from them were little affected by removal of these gaps. Sequence comparisons show that three of the CAV genes may have diverged from the most closely related genes of other viruses 250–450 million years ago, and the sister relationship between the genes of CAV and those of benyviruses and tobamoviruses, mirroring the ancient sister relationship between charophytes (i.e. the algal host of CAV) and embryophytes (i.e. the plant hosts of tobamoviruses and benyviruses), is congruent with this possibility.

INTRODUCTION

A large charophyte alga growing in the Murrumbidgee River near Canberra, Australia was reported to be infected with a tobamovirus (Gibbs et al., 1975; Skotnicki et al., 1976). The host was identified as Chara corallina var. corallina K. ex Wild., but more recently this species has been reclassified and named Chara australis var. nobilis (A. Braun) R. D. Wood (http://data.gbif.org/search/chara%20australis; http://www.algaebase.org/search/species/detail?species_id=27145). This species is rightly called ‘nobilis’ as individual plants are up to 2 m in length, and their coenocytic internodal cells are up to 100 mm in length!

The virus was reported to be a tobamovirus as its virions closely resembled those of tobacco mosaic tobamovirus (Zaitlin, 2000) except that their modal length was approximately 530 nm, rather than the 300 nm of all reported tobamoviruses. Moreover, preparations of its virions reacted with high-titre antiserum prepared against the virions of tobacco mosaic and odontoglossum ringspot viruses, but not those of several other tobamoviruses. The new virus was named Chara corallina virus, but with the updated nomenclature of the host, we now call it Chara australis virus (CAV).

Up to 22 % of the naturally infected Chara plants first found contained CAV virions, and some infected cells contained birefringent inclusions, like those found in tobacco mosaic virus (TMV)-infected plants (Sheffield, 1934; Steere & Williams, 1953). Healthy C. australis plants became infected when injected with virions, but none of 14 species of standard angiosperm virus test plants was infected by sap inoculation. No similar virions were found in several charophyte species sampled in the 1980s from rivers and lakes of highland and coastal south-east Australia and northern and central UK, nor in a large worldwide collection of charophyte species kept in Lubbock, Texas by V. W. Proctor.

When CAV was first found in the 1970s, C. australis was dominant in large stretches of the slower parts of the Murrumbidgee and Shoalhaven rivers within 100 km of
Canberra. However, since the 1980s and especially over the past decade, this plant has disappeared as the rivers have become more eutrophic, a consequence of increasing drought and siltation, and the use of fertilizers on irrigated alfalfa crops draining into the rivers. Nonetheless, bulk frozen *C. australis* samples from the earlier period, together with small samples of fresh infected material from a region of the river (149.0756°E: 36.1747°S) upstream of Cooma, provided sufficient material to determine most of the genomic sequence of CAV, and it is reported here.

**RESULTS**

The sequence of the CAV RNA (GenBank accession number JF824737; accession numbers for viruses used in this study are listed in Supplementary Table S1, available in JGV Online) is 9065 nt long, and probably represents most of the genome of CAV, as physical estimates of its size (Skotnicki et al., 1976) indicated that it is from 9800 to 11 900 nt long.

The CAV RNA has a nucleotide composition of 14.7 mol% guanine, 26.8 mol% adenine, 30.3 mol% cytosine and 28.1 mol% uracil. The sequence has 233 ORFs greater than 60 nt (20 codons) long, more than half of them in the complementary strand. Only the 11 longest ORFs are noticeably larger than expected from the trend line of a graph of the logarithms of the ranked ORF lengths (data not shown), and ten of these are in the messenger strand; Supplementary Table S2 (available in JGV Online) lists the positions and sizes of the 12 longest ORFs. ORFs 2, 7, 8 and 10 probably encode proteins as all four, but none of the others, have the ‘RNY’ pattern of nucleotide usage (Supplementary Fig. S1) shown by protein-encoding ORFs (Jukes, 1996; Shepherd, 1981); bioinformatic analysis alone will not show whether any of the other ORFs are active, overlapping genes (Keese & Gibbs, 1992; Rancurel et al., 2009).

A search of the international gene sequence databases using SWeBLAST (Fourment et al., 2008) with BLASTX and BLASTN found that ORFs 2, 7 and 10 have regions that significantly match genes, and especially the proteins they encode, of other viruses. The protein sequences matched with much smaller *E* values than the gene sequences, indicating that the relationships were distant, and so our analyses focused on the encoded amino acid sequences for all the reasons discussed by Oppendoes (2009).

The protein encoded by ORF 2, the longest, had regions that matched those of the replicases of several viruses of the alphavirus supergroup (Goldbach & Wellink, 1988; van der Heijden & Bol, 2002), especially those of benyviruses and alphaviruses. The ORF 10 protein matched the coat proteins of tobamoviruses, which are from the same supergroup. Regions of the ORF 7 protein matched helicases of flaviviruses and potyviruses, which are from the flavivirus and picornavirus supergroup (Goldbach, 1987; Goldbach & Wellink, 1988; Koonin et al., 2008). The largest of the other nine ORFs, ORF 8, is adjacent to, and on the 5′ side of, the tobamovirus-like coat protein gene, and so it is positioned where a single movement protein gene (Melcher, 2000) is found in the genomes of tobamoviruses and many other plant-infecting viruses of the alphavirus supergroup. The positions of these four ORFs in the sequence and the probable functions of their encoded proteins are summarized in Fig. 1.

**ORF 2**

ORF 2 covers two-thirds of the CAV sequence in the +3 reading frame and encodes a protein at least 1999 aa long. In a BLASTP search of the sequence databases, parts of the ORF 2 protein sequence closely matched regions of the replicase proteins of benyviruses: beet soil-borne mosaic virus (BSBMV), beet necrotic yellow vein virus (BNYVV) and rice stripe necrosis virus (RStNV) (*E* values $1 \times 10^{-45}$ to $3.2 \times 10^{-41}$); and also various hepeviruses ($1 \times 10^{-17}$).

The three regions of the CAV and BNYVV replicases that matched other proteins were used to search the Pfam (Protein Families) Database (Finn et al. 2010). Two gave clear matches: amino acids 450–690 of the CAV replicase together with the corresponding region of the BNYVV replicase matched viral UvrD-helicases (*E* values of 9.2 $\times 10^{-17}$–5.2 $\times 10^{-33}$); and amino acids 1710–2030 of the CAV replicase and the corresponding region of the BNYVV replicase matched RNA-dependent RNA polymerase (RdRp)-2 (*E* values $1 \times 10^{-21}$–2.5 $\times 10^{-22}$) (Fig. 2). A third region of the CAV replicase, amino acids 28–170, gave no direct matches with the Pfam database, but showed significant similarity with the C-terminal region of eight benyvirus replicase sequences in a BLASTP search (*E* values $2 \times 10^{-10}$–2 $\times 10^{-5}$), and those regions matched viral methyltransferases (*E* value 1.9 $\times 10^{-10}$) in Pfam database searches. Therefore we concluded that this region of the CAV replicase, like that of the benyvirus replicase, is part of a viral methyltransferase (MTase).

The regions between the matching regions of the CAV and BNYVV replicases were also checked separately in the Pfam and GenBank protein sequence databases. The C-terminal part of amino acids 1200–1750 unequivocally matched the BNYVV papain-like endopeptidase C36 (*E* value $3.3 \times 10^{-69}$), and a 50 aa region of its N-terminal region gave weak matches with an ovarian tumour-like cysteine protease (OTU; *E* value 5.8 $\times 10^{-3}$) and a gibberellin-regulated cysteine-rich protein (GASA; *E* value $1.2 \times 10^{-3}$) (Supplementary Fig. S2).

Three lines of evidence suggest that the CAV sequence is missing the 5′-terminal part of its MTase region. First, the methionine residue closest to the N terminus of the protein is encoded by codon 70 of the ORF; yet amino acids 30–69 also clearly match the adjacent N-terminal region of the benyvirus replicase. Secondly, the BNYVV replicase has a molecular mass of 237 kDa, whereas that encoded by the CAV sequence is only 227 kDa. Thirdly, codons 33–36 of
the CAV MTase region have the sequence `-DVGY-`, which is also the conserved core of motif IV (codons 406–409, nucleotides 1372–1383) of the MTase region of the BNYVV (Liu et al., 2009; Rozanov et al., 1992), and is also found in the replicases of BSBMV and RStNV. We therefore conclude that our CAV sequence is not the complete genome, and its 5' terminus is missing approximately 1300 nt, and these encode the six missing N-terminal motifs of its MTase (Liu et al., 2009; Rozanov et al., 1992).

The broader phylogenetic relationships of the CAV replicase were assessed by using the sequences that matched it most closely in BLASTP searches of the GenBank protein database (March 2011). As the benyviruses are part of the alphavirus cluster (Goldbach & Wellink, 1988; van der Heijden & Bol, 2002), sequences of replicases of representative genera were added to obtain a set of 40 sequences. The replicases ranged in length from 790 aa (alfalfa mosaic alfamovirus) to 3582 aa (citrus tristeza closterovirus) (mean 1867 aa). The collated sequences, together with that of CAV, were aligned, and checked for evidence of recombination, and none was found. Approximately 70% of the individual sites in the aligned sequences were gaps. Many of the insertions or deletions (indels) resulted from
the inclusion of the replicases of bromovirid species; therefore, as they are unequivocally a sister group to the closteroviruses and only distantly related to the benyviruses, they were omitted and the remaining sequences realigned.

Although the relative rates of accumulation of amino acid substitutions during evolution have been well studied (Opperdoes, 2009), the relative rate at which indels are acquired is less certain. Therefore the effect on tree topology of removing, from aligned sequences, all homologous positions that contained one or more indels was checked. Such trees would then be based on clock-like changes of amino acids, not confounded by the indels required to align them. The 36 replicase sequences produced an alignment 6654 positions long, but after all positions containing gaps had been removed (degapped), only 709 (10.7 %) remained. All regions of the gapped sequences contributed some positions to the degapped sequences, but approximately three-quarters were from the RdRp region. The phylogeny of the 36 aligned sequences, gapped and degapped, was inferred, and the topologies of the resulting trees found to be closely similar (Fig. 3). It can be seen that in both trees the CAV replicase is sister to the replicases of benyviruses and more distantly related to those of the rubiviruses, and this cluster is fully supported by bootstrap sampling of the sequences. The topology of the other parts of the trees was also closely similar, and there were three more bootstrap-supported clusters. One cluster was of the replicases of the alphaviruses, closterovirids and virgavirids, those from all tymovirids formed another, and those from hepeviruses formed the third. The only difference between the gapped and degapped trees was in the arrangement of the four major clusters, and none of the nodes involved in defining those arrangements had significant bootstrap support in either tree. The three matching regions of the CAV replicase all matched the corresponding regions of benyvirus sequences most closely.

The close similarity of the gapped and ungapped sequence trees is also shown by the close linear relationships of the patristic distances in the trees (Fig. 4). A Procrustes rotation (Mardia et al., 1979) of the patristic distances of either tree on to the other accounts for 98 % of the inertia about their centroids. Thus, although the pairwise patristic distances of the degapped sequences are approximately two-thirds those of the corresponding gapped sequences (a mean of 3.28 substitutions per site compared with 5.74 substitutions per site), the relative positions of different nodes are very similar. Furthermore, tests using the various substitution models and search algorithms available in the MAFFT (Multiple Alignment with Fast Fourier Transform) and PhyML programs, all produced trees with closely similar topologies.

**ORF 7**

The protein encoded by ORF 7 is 397 aa long, and its first methionine is only 5 nt to the 3’ side of the termination codon of ORF 2 (Supplementary Table S2). In the Pfam database its sequence (amino acids 50–290) matched those of proteins of the DEAD-like helicase C group and, in the GenBank database, the helicas of hepacviruses (E value 3 × 10^{-7}), pestiviruses, arthropod-borne flaviviruses and also the cylindrical inclusion (CI) proteins of potyviruses.

The 33 representative sequences that matched the ORF 7 protein most closely, together with that of the ORF 7 protein, gave an alignment 680 sites positions long including 21 % gaps, but only 299 positions (44 %) remained after all positions that contained gaps had been removed. **BLASTP** searches of the GenBank database using the complete ORF 7 sequence gave matches with E values of approximately 3 × 10^{-7}, whereas those with the degapped sequences were approximately 2 × 10^{-7}. Fig. 5 shows part of the alignment of the CAV ORF 7 helicase with selected pestiviruses and hepacviruses, and boxes indicate the position of the six conserved motifs (Stapleton et al., 2011).

Phylogenetic trees inferred from gapped (Fig. 6) and degapped sequences differed only in their within-cluster topology and bootstrap support. In both trees the CAV sequence was sister to the pestivirus helicas, but this relationship only had 100 % bootstrap support in the gapped sequence tree; 36 % in the degapped sequence tree. Four groupings had full bootstrap support in both trees: the hepacviruses, including the GB hepatitis viruses (Ohba et al., 1996; Stapleton et al., 2011), the pestiviruses, the flaviviruses and the potyvirids. The mean patristic distance between the gapped aligned sequences (3.45 substitutions per site) was one-fifth more than that between the degapped sequences (2.79 substitutions per site) and the pairwise patristic distances in these trees were closely and linearly related; a Procrustes rotation (Mardia et al., 1979) of the patristic distances of either tree on to the corresponding distances of the other accounts for 99.4 % of the inertia about their centroids.

**ORF 8**

ORF 8 has a clear ‘RNY’ signal and encodes a protein of 357 aa that shows no statistically significant **BLASTP** match to any other protein in the international sequence databases, although its C-terminal quarter resembles the domain of unknown function (DUF) 663 family of ‘uncharacterized eukaryotic proteins’, and a search of the Pfam database finds a non-significant similarity to the Spc97/98 spindle pole body proteins from the microtubule-organizing centre of yeast.

ORF 8 is adjacent to, and on the 5’ side, of ORF 10 (Fig. 1), which encodes a protein that is clearly homologous to tobamovirus coat proteins, and thus ORF 8 is positioned where tobamoviruses have a single movement protein (MP) gene (Leisner, 1999). MPs specifically assist the systemic movement of viruses through the plasmodesmata that link the cells of higher plants (Faulkner & Maule, 2011; Lucas, 2006; Taliantsky et al., 2008), and the cells of charophytes are also linked by structurally similar plasmodesmata (Brecknock et al., 2011).
Of the viruses with rod-shaped virions, only tobamoviruses, tobraviruses, furoviruses and *Nicotiana velutina* mottle virus (NVMV) (Randles & Rohde, 1990) have single MP genes; all the others have ‘triple gene block’ MPs (Verchot-Lubicz, 2003). MP genes are among the least conserved of viral genes (Koonin et al., 1991; Melcher, 1990, 2000; Mushegian & Koonin, 1993). Thus, if the ORF 8 protein is an MP then it is not surprising that it matches no other known protein.

A comparison of all such MPs (Melcher, 2000) found that their predicted secondary structures were similar. However, neither the sequence of the ORF 8 protein nor its predicted secondary structure share any significant similarity with individual or aligned MPs of tobamoviruses, nor with an MP consensus sequence (Fig. 3 in Melcher, 2000), whether aligned by MAFFT, compared by ‘dotplot’ (Gibbs & McIntyre, 1970) or by using the DISOPRED and DOMPRED programs (see Methods).

**Fig. 3.** The phylogenetic relationships of CAV replicase (ORF 2) protein and 35 viral replicase proteins that matched its sequence most closely. Maximum likelihood (ML) trees were inferred from the aligned amino acid sequences (a), and from the same sequences after removing all positions with gaps in one or more sequences (b). Nodes supported in more than 90 of 100 bootstrap samplings are circled. Virus acronyms and the GenBank accession numbers of the sequences are listed in Supplementary Table S1. The names of the supra-specific taxa are from Fauquet *et al.* (2005) and Anonymous (2010). Bars, 1 amino acid substitution per site.

**Fig. 4.** Comparison of the pairwise patristic distances of the aligned (x-axis) and corresponding degapped (y-axis) sequences of 36 replicase proteins in the ML trees in Fig. 3. Units are substitutions per site.

**ORF 10**

Of the 12 longest ORFs in the CAV sequence, ORF 10 is closest to its 3′ terminus and, at its 5′ end, only 23 nt separate it from ORF 8 (Fig. 1, Supplementary Table S2). ORF 10 encodes a protein of 154 aa in length. It terminates just 54 nt from the 3′ end of the CAV sequence. The ORF 10 gene sequence does not significantly match any sequence in the international databases, but in BLASTP searches its encoded amino acid sequence significantly matches the coat proteins (CPs) of several tobamoviruses, though it shares only approximately 16% identity with those sequences; the CPs of cucumber green mottle mosaic and cucumber mottle tobamoviruses are the closest (E values approximately $8 \times 10^{-6}$).

The sequence of the CAV CP was aligned with those of 19 tobamoviruses and 23 other viruses that have rod-shaped
virions with similar diameters, most of which are virgavirids [i.e. grouped by the International Committee of Taxonomy of Viruses (ICTV) as members of the family Virgaviridae; Anonymous, 2010], but included the benyviruses and NVMV (Randles & Rohde, 1990), which are both unassigned by the ICTV. The alignment was 274 aa sites in length including 33 % sequence gaps, but only 130 sites (47 %) remained after all positions with gaps had been removed.

Trees of the 43 CP protein sequences, gapped and degapped, were of almost identical topology, and a Procrustes rotation (Mardia et al., 1979) of the patristic distances in either of these trees onto the corresponding distances of the other accounts for 99.1 % of the inertia about the centroid. The mean patristic distance within the gapped tree was 3.21 substitutions per site and for the degapped tree it was 2.96 substitutions per site (92 %). The 43 CP proteins form three distinct lineages (Fig. 7) that are congruent with traditional generic groupings of the viruses. The CPs of CAV and the tobramoviruses are sister taxa in one lineage, those of the hordeiviruses, pecluviruses and tobraviruses form another lineage, and, in a third, the CP of NVMV is the sister taxon to the CPs of benyviruses, pomoviruses and furoviruses. These groupings are strongly supported by bootstrap sampling (Fig. 7), though the sister relationship of the gapped CPs of CAV and the tobramoviruses had only 67 % bootstrap support; and 81 % in the degapped tree. The CAV CP gene, like those of NVMV, the tobramoviruses, pecluviruses, hordeiviruses and tobraviruses, does not have a ‘read-through’ protein gene at its 3‘ end.

The predicted secondary structure of CAV CP is also virgavirid-like. PSIPRED found that the CAV CP sequence, like that of TMV, is dominated by long α-helical regions (45 % of the residues) interspersed with many short coiled regions (50 %), and only 5 % of its residues were in β-sheet structures. The linear pattern of these structures in the CAV CP sequence was significantly congruent with those of the tobramoviruses CPs (Supplementary Fig. S3).
Attempts to sequence the UTRs of the CAV RNA using standard RACE methods were unsuccessful. Approximately 1300 nt of the genome, including the UTR, is probably missing from the 5' end of the sequence. The fact that the CAV replicase protein contains a MTase-like region suggests that the CAV genome has the same m7GpppG cap found in the genomes of other related viruses (Fauquet et al., 2005).

The complete sequence of the CAV 3' UTR is also uncertain. This is only 68 nt long, to the 3' side of the CP ORF, whereas those of tobamoviruses and benyviruses are 190–260 nt long. There was no evidence of a primable poly-A tail, like the genomic RNAs of BNYVV, or of a 3'-terminal region with 65–70 mol% G+C that can fold into a pseudoknot (Olsthoorn et al., 1999). Perhaps the 5' and 3' regions of the CAV genome have stable secondary structures that foiled sequencing attempts, or they may have been degraded during purification, as although untreated sap of individual infected plants mostly contained modal length virions (Skotnicki et al., 1976), the particles in purified preparations were mostly of submodal length.

The assembly of TMV virions involves the assembly of double disks of 32 CP monomers, initially on a nucleation region in the genome about 1000 nt from its 3' end (Butler, 1999). The sequence of the TMV assembly nucleation region is -AAGAAGUCG-. The triplet of repeated
guanine residues with no more than one cytosine in the sequence is critical to its function. The TMV nucleation sequence is not present in the CAV sequence, but similar repetitive sequences are. The CAV sequence contains no \((XXG)_6\) sequences, where \(X\) is cytosine, adenine or uracil, but it has one \((XXG)_3\) region, five \((XXG)_4\) regions and 31 \((XXG)_5\) regions. The \((XXG)_5\) sequence includes four cytosines clustered around the second and third guanines; however, of the \((XXG)_5\) regions, that at nucleotides 6414–6426 has the sequence UAGUAGUGAUG and that at nucleotides 8293–8305 has the sequence AAGUUG-CUGUUUG, and so the latter is most similar to the nucleation region of TMV in terms of both its sequence and its position.

**DISCUSSION**

The CAV genomic sequence has at least four ORFs, three of which encode proteins that have significant but distant sequence similarities to proteins of other viruses. However, although two of these are very distantly related, the third is, unexpectedly, unrelated to the other two. The ORF 2 replicase is most closely related to the replicases of benyviruses and, more distantly, rubiviruses (Koonin & Dolja, 1993; Ward, 1993); the ORF 7 helicase is most closely related to the helicases of pestiviruses, and the ORF 10 CP is most closely related to the CPs of tobamoviruses – a combination of replicase and helicase lineages that seems not to have been reported before. Significantly, each of the three CAV proteins are the taxonomic sister taxon of the lineages containing its position. This indicates that the phylogenetic links between the CAV proteins and their nearest relatives are more ancient than each group of relatives.

The ML trees linking CAV and related proteins have large within-tree patristic distances: their maxima range from 7.7 substitutions per site for one pair of gapped replicase sequences to 4.2 substitutions per site for one pair of degapped helicase sequences. This reinforces the view that the CAV genes and those of their relatives have ancient linkages.

Several disparate lines of evidence have previously suggested that the tobamoviruses are ancient and have co-diverged with their hosts (Gibbs, 1980; Gibbs et al., 2008; Holmes, 1950; Larney et al., 1996), though studies of extant populations suggest that they may be only tens of thousands of years old (Pagan et al., 2010). This discordance may merely confirm that virus populations undergo active birth–death or ‘Red Queen’ processes (Clarke et al., 1994; Holmes, 2009). Taxonomies derived from the CP sequences of non-recombiant tobamoviruses place them in two groups (Gibbs et al., 2008; Larney et al., 1996): the primary natural hosts of one group are mostly eudicotyledonous plants of the asterid lineage, and those of the other are plants of the rosid lineage (Jansen et al., 2007; Moore et al., 2010). If this correlation reflects co-divergence (Gibbs, 1999; Gibbs et al., 1999, 2008), then the age of the asterid/rosid divergence of tobamoviruses and hosts may be similar, and the divergence of CAV and an aboriginal tobamovirus even earlier (Fig. 7).

The earliest angiosperm fossils are found in early Cretaceous rocks dating from approximately 125 million years BP (MYBP) (Frisch et al., 2010; Sun et al., 2002; Wang & Zheng, 2009), though gene sequence comparisons suggest an origin approximately 140–180 MYBP (Chaw et al., 2004; Smith et al., 2010) and, hence, a divergence of asterids from rosids, and of the two primary tobamovirus lineages, approximately 100–135 MYBP. In the gapped replicase sequence tree (Fig. 3) the mean patristic distance between the two ‘asterid-favouring’ tobamoviruses [TMV and youcai mosaic virus, (YMV)] and two ‘rosid-favouring’ tobamoviruses [sunn-hemp mosaic virus (SHMV) and cucumber green mottle mosaic virus (CGMMV)] is 1.32 substitutions per site, whereas the mean distance between the CAV replicase sequence and the three benyvirus RdRp sequences is 3.76 (+0.12) substitutions per site. This suggests that the CAV/benyvirus split is 2.83 times more ancient than the tobamovirus split, namely 396–509 MYBP. Likewise, using degapped sequences, the patristic distances are 0.69 and 2.04 substitutions per site, respectively, and hence the relative position of the CAV split is 2.98 times more ancient, namely 417–536 MYBP. Thus the mean CAV/benyvirus replicase divergence was 464 MYBP in the range 396–536 MYBP. Similarly the patristic distances of the gapped and degapped sequences of CAV and tobamovirus CPs suggest that their divergence was between 1.73 and 1.70 times more ancient than the primary tobamovirus divergence, namely 274 MYBP in the range 238–311 MYBP. These estimates are unlikely to be of similar accuracy; the replicase sequences are much longer and more variable than those of the CPs, but they are in separate virus lineages and may have different evolutionary constraints, whereas, in the CP trees, CAV and tobamoviruses are sister lineages in the same major lineage and so perhaps have evolved at similar rates.

The known relationships of the viruses and their hosts are also congruent with dates deduced from sequence comparisons. The sister relationship between the genes of CAV and those of benyviruses and tobamoviruses mirrors the ancient sister relationship between charophytes, which include the algal host of CAV, or some other streptophyte alga, and land plants (i.e. the hosts of tobamoviruses and benyviruses) (McCourt et al., 2004; Mishler et al., 1994; Turmel et al., 2003; Wodniok et al., 2011). The oldest fossilized charophyte oogonia (i.e. gyrogonites) are found in rocks of the Upper Silurian (420 MYBP) (Feist et al., 2005), which is in the same range as the estimated length of time that these viruses and their hosts may have co-diverged.

If the genes of CAV are even older than 250–450 MYBP, from whence did they come? This question has been discussed for many other viruses since it was first found that some proteins of seemingly unrelated viruses of plants
and animals were unequivocally related (Ahlquist et al., 1985; Argos et al., 1984; Fransen et al., 1984; Haseloff et al., 1984; Kamer & Argos, 1984). The replicases divided the viruses with RNA genomes into three superfamilies (SG1 alpha-like, SG2 flavi-like and SG3 picorna-like) (Koonin & Dolja, 1993), but their helicases placed them into three incongruent superfamilies (SF1 picorna-like, SF2 flavi-like and SF3 alpha-like) (Goldbach, 1987; Goldbach & Wellink, 1988; Gorbalenya et al., 1989, 1990; Kadaré & Haenni, 1997; Koonin & Dolja, 1993). Thus CAV has an SG3 replicase and an SF2 helicase, a unique combination. Replicase genes are only found in viruses with RNA genomes; however, the viral helicases and other viral proteins have clear sequence or structural homologies with proteins encoded by cellular organisms.

These facts have stimulated attempts to place the genomes and genes of viruses and their hosts into a single evolutionary scenario (Forterre, 2006; Forterre & Prangishvili, 2009a, b; Joyce, 2002; Koonin, 2009; Koonin & Martin, 2005; Koonin et al., 2008; Villarreal & Witzany, 2010). Most is known of the large number of picorna-like viruses, especially those withicosahedral virions constructed from ‘jelly-roll’ capsid proteins (Koonin et al., 2008). It seems most likely that these acquired their ‘hallmark’ genes from a primordial pool of genes more ancient than the current kingdoms of hosts, perhaps even from the era before the last universal common ancestor (Doolittle, 2000; Glansdorff et al., 2008) established the three domains of life (Woese et al., 1990). It is clear that CAV and its genetic relatives may have a similar ancient origin, but less is known of viruses, like CAV, that have tubular virions constructed from four helix bundle (4HB) capsid proteins. The best studied are a few plant viruses, notably TMV (Kendall et al., 2008; Namba et al., 1989; Wang et al., 1997). 4HB proteins are also found in the filamentous virions of some bacteriophages (Forterre & Prangishvili, 2009a; Goulet et al., 2009; Szymczyna et al., 2009) and in cellular organisms (Harris et al., 1994; Kurochkina, 2010), but they show no sequence similarity and have an inter-helix U-fold, rather than the Z-fold found in TMV CP.

**METHODS**

**Virion preparation.** Unsorted *C. australis* plants, collected from the river in batches of 1–5 kg, were homogenized in PUE buffer (50 mM sodium phosphate, 100 mM urea and 5 mM Na2EDTA; pH 7.5) using 0.5–1.0 l kg−1 of alga. Large solids were removed by filtering through muslin, and the fluid phase cooled to 4°C. The supernatant was resuspended and blended into approximately 5% of the volume of the original buffer mixture and centrifuged again. The supernatants were pooled, stirred for 1 h at room temperature with 8% n-butanol, and then emulsified with 10% chloroform and centrifuged again at 3000 g for 30 min. The aqueous phase was collected and then centrifuged for 3 h at 20 000 g; the pellets were resuspended in PUE buffer to give a virion preparation concentrated at least 1000-fold from the starting material. The use of borate or TRIS extraction buffers, and polyethylene glycol (4% in 2% sodium chloride) as a precipitant, did not produce more concentrated or less contaminated virion preparations.

**Gene sequencing.** Standard methods were used to extract nucleic acid from virions and to sequence it (Maniatis et al., 1982; Sambrook et al., 1989). cDNA was synthesized from CAV RNA by using random hexamer nucleotides and avian myeloblastosis reverse transcriptase, and was transcribed into dsDNA by the RNase H method (Gubler & Hoffman, 1983). The dsDNA was hydrolysed using several four-base-pair recognition site restriction endonucleases, and the resulting fragments separated and sequenced. A series of overlapping clones were obtained using the primers listed in Supplementary Table S3 (available in JGV Online) with their positions shown in Fig. 1. The 5′- and 3′-terminal sequences were determined by using standard RACE methods (Maniatis et al., 1982). Each nucleotide of the sequence was determined from at least two, usually more, independent overlapping clones in both orientations.

**Sequence analysis.** Sequences were manipulated by using BioEdit (Hall, 1999) and other programs. ORFs were predicted by using the MolQuest-Softberry viral gene detector, and also by plotting the ranked logarithmic lengths of all ORFs to identify those noticeably larger than the trend. The GenBank sequence database was searched by using its BLAST (N, P and X) facilities and SwiBLAST (Fourment et al., 2008), and protein relationships were checked by using the Pfam database (Finn et al., 2010). Sequences were aligned by using MAFFT (Katoh et al., 2002) with its ‘local pairs’ option. Sequence alignments of nucleotides and amino acids were checked by using KFDA (Martin et al., 2005) and GENECONV (Sawyer, 1989), and probable recombinants removed. Trees were inferred by using the neighbour-joining option (Saitou & Nei, 1987) of CLUSTAL_X (Jeanmougin et al., 1998) and PhyML (Guindon & Gascuel, 2003) with the BLOSUM 62 substitution matrix. Positions in aligned amino acid sequences that contained gaps were removed from the alignments by using POSORT, a program available from http://192.55.98.146/_resources/e-texts/README-POSORT.pdf. The patristic distances in phylogenetic trees were compared by using PATRISTIC (Fourment & Gibbs, 2006). The secondary structures of proteins were predicted from their amino acid sequences by using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/), and their patterns of those structures compared by using the Z-test using the Yates correction (Moroney, 1951).

**ACKNOWLEDGEMENTS**

We thank Gerald Stubbs for information about the CPs of helically constructed viros, Vernon Proctor for aiding the search for CAV-like viruses in other charophytes and also numerous colleagues who risked their lives helping to collect *Chara australis* plants.

**REFERENCES**


algae' and 'bryophytes'.


http://vir.sgmjournals.org

2689


