Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium* sp.

Yoshitake Takao, Kazuyuki Mise, Keizo Nagasaki, Tetsuro Okuno and Daiske Honda

The complete nucleotide sequence of the genomic RNA of a marine fungoid protist-infecting virus (*Schizochytrium* single-stranded RNA virus; SssRNAV) has been determined. The viral RNA is single-stranded with a positive sense and is 9018 nt in length [excluding the 3′ poly(A) tail]. It contains two long open reading frames (ORFs), which are separated by an intergenic region of 92 nt. The 5′ ORF (ORF1) is preceded by an untranslated leader sequence of 554 nt. The 3′ large ORF (ORF2) and an additional ORF (ORF3) overlap ORF2 by 431 nt and are followed by an untranslated region of 70 nt [excluding the 3′ poly(A) tail]. The deduced amino acid sequences of ORF1 and ORF2 products show similarity to non-structural and structural proteins of dicistroviruses, respectively. However, Northern blot analysis suggests that SssRNAV synthesizes subgenomic RNAs to translate ORF2 and ORF3, showing that the translation mechanism of downstream ORFs is distinct from that of dicistroviruses. Furthermore, although considerable similarities were detected by using a BLAST genome database search, phylogenetic analysis based on both the nucleotide and amino acid sequences of the putative RNA-dependent RNA polymerase (RdRp) and the RNA helicase suggests that SssRNAV is phylogenetically distinct from other virus families. Therefore, it is concluded that SssRNAV is not a member of any currently defined virus family and belongs to a novel, unrecognized virus group.

INTRODUCTION

It was revealed by the end of the 1980s that viruses are highly abundant in marine environments (Bergh *et al.*., 1989; Proctor & Fuhrman, 1990), and viruses and virus-like particles (VLPs) have been discovered in a variety of phytoplankton and bacteria (Suttle *et al.*, 1990, 1991; Van Etten *et al.*, 1991). Until today, more than 13 viruses infecting marine eukaryotic microalgae have been isolated and characterized (Brussaard, 2004). The majority are large (100–200 nm in diameter), icosahedral, double-stranded DNA (dsDNA) viruses that are included in the family *Phycodnaviridae* (Schroeder *et al.*, 2002; Brussaard, 2004; Brussaard *et al.*, 2004b; Nagasaki *et al.*, 2005; Wilson *et al.*, 2005) or are considered to probably belong to this family (Jacobsen *et al.*, 1996; Gastrich *et al.*, 1998; Sandaa *et al.*, 2001; Tarutani *et al.*, 2001).

In contrast, there are few reports concerning viruses infecting apochlorotic protists. Nagasaki *et al.* (1993, 1995) observed large VLPs in marine apochlorotic flagellates. Garza & Suttle (1995) isolated and characterized a dsDNA virus infecting *Bodo* sp. (*Bodoniidae*, *Kinetoplastida*, *Mastigophora*), with characteristics similar to those of phycodnaviruses in shape, size and site of propagation. Recently, an extremely large dsDNA virus (400 nm in diameter, 1–2 Mbp) was found infecting amoebae and named mimivirus (La Scola *et al.*, 2005) and a genomic analysis was performed (Raoult *et al.*, 2004). Based on phylogenetic analysis of the DNA polymerase amino acid sequence, the mimivirus formed a sister group with *African swine fever virus*, but not with viruses belonging to the family *Phycodnaviridae*. In the thraustochytrids, the protists belonging to the family Thraustochytriaceae in the class Labyrinthulomycetes, Kazama & Schornstein (1972, 1973) found herpes-type VLPs in *Thraustochytrium* sp. that were round, enveloped, 110 nm in diameter and predicted to have a
DNA genome. However, this study has not been followed by any detailed analysis because of unsuccessful isolation.

There are few reports describing RNA viruses infecting marine eukaryotic micro-organisms. At present, four RNA viruses infecting marine eukaryotic microalgae have been isolated and examined: Heterosigma akashiwo RNA virus (HaRNAV; Tai et al., 2003), Heterocapsa circularisquama RNA virus (HcRNAV; Tomaru et al., 2004), Rhizosolenia setigera RNA virus (RsRNAV; Nagasaki et al., 2004a) and Micromonas pusilla RNA virus (MpRNAV; Brussaard et al., 2004a). HaRNAV is infectious to one of the noxious bloom-forming phytoflagellates, Heterosigma akashiwo (Raphidophyceae), and contains an 8-6 kb single-stranded RNA (ssRNA) genome (Lang et al., 2004). The genomic RNA contains only one large open reading frame (ORF) (7-7 kb) encoding both structural and non-structural proteins. Molecular phylogenetic analysis of the deduced amino acid sequence of the RNA-dependent RNA polymerase (RdRp) shows that HaRNAV is a distinct species related to the family Dicistroviridae (Lang et al., 2004). The family Dicistroviridae is a newly recognized virus family (split from the family Picornaviridae) that includes a marine virus that causes a disease of shrimp, the Taura syndrome virus (TSV) (Mari et al., 2002). HcRNAV is a small, ssRNA virus infectious to the bivalve-killing dinoflagellate Heterocapsa circularisquama (Tomaru et al., 2004). The genomic RNA is 4-4 kb in length and contains two ORFs encoding replicases and a structural protein, respectively (Nagasaki et al., 2006). Field surveys with regard to the ecological relationship between HcRNAV and its host alga revealed that viral infection is one of the most significant factors controlling the dynamics of host algal blooms (Nagasaki et al., 2004b; Tomaru & Nagasaki, 2004). RsRNAV is a small, ssRNA virus infectious to the bloom-forming diatom Rhizosolenia setigera (Nagasaki et al., 2004a). The major nucleic acid extracted from RsRNAV particles is an ssRNA molecule of 11-2 kb, with smaller RNA molecules (0-6, 1-2 and 1-5 kb) occasionally observed. MpRNAV is the only algal virus with a double-stranded RNA (dsRNA) genome presently isolated. It harbours 11 segments of dsRNA in the viral genome at 25-5 kb (Brussaard et al., 2004a). Among these four RNA viruses, intensive genomic analysis has been conducted only for HaRNAV.

There is a single report of RNA viruses infecting protists, which described the basic characteristics of Schizochytrium single-stranded RNA virus (SssRNAV) (Takao et al., 2005). Its host organism, Schizochytrium sp., is a marine fungal protist belonging to the thraustochytrids, which are cosmopolitan, heterotrophic micro-organisms playing important roles as decomposers, particularly in coastal ecosystems (Raghukumar, 1996, 2002). SssRNAV particles are icosahedral, lacking a tail and approximately 25 nm in diameter. They have a single-stranded, positive-sense RNA genome of 10-2 kb in length. Based on a number of similarities, such as morphological features, cytoplasmic assembly in an infected cell, number and size of the structural proteins and partial genome sequence, SssRNAV is predicted to be related closely to viruses belonging to the family Dicistroviridae and the marine microalgal virus HaRNAV (Takao et al., 2005).

To test this prediction, we describe the complete genome sequence and the genome organization of SssRNAV and consider the phylogenetic classification of this virus. To our knowledge, this is the first report describing the genome organization of an ssRNA virus infecting a marine fungoid protist.

**METHODS**

**Purification of virus.** A fresh suspension of SssRNAV (1-5 ml) was inoculated into a 1:5:1 vigorously growing culture of Schizochytrium sp. (strain NIBH N1-27) at an m.o.i. >1 and incubated at 20 °C for 48 h. Virus particles were purified from the culture lysate. The lysate was centrifuged at 14,000 g, 4 °C for 15 min to remove cellular debris, using an RS-20BH centrifuge and angle rotor BH-9 (Tomy Seiko Co.). Polyethylene glycol 6000 (Wako) was added to the supernatant at 10% (w/v) and incubated at 4 °C overnight. The preparation was centrifuged at 3600 g, 4 °C for 1 h using a CP56GII ultracentrifuge and angle rotor P42A (Hitachi Koki Co.). The supernatant was centrifuged at 120 ml 10 mM phosphate buffer (10 mM Na2HPO4, 10 mM KH2PO4 in distilled water, pH 7-2) and centrifuged again at 100,000 g, 4 °C for 2 h using a CP56GII ultracentrifuge and swing rotor P40ST. The pelleted virus was resuspended in 60 ml 10 mM phosphate buffer and centrifuged again at 3600 g, 4 °C for 15 min. The supernatant was centrifuged at 100,000 g, 4 °C for 2 h. This centrifugation process (3600 g, 4 °C for 15 min and 100,000 g, 4 °C for 2 h) was repeated four times to purify the virus particles. Finally, the purified virus pellet was resuspended in 500 µl distilled, deionized water.

**RNA isolation.** Two hundred microlitres of extraction buffer [0-33 M glycine, 0-33 M NaCl, 3-3 mM EDTA, 3-3% SDS and 8-3 mg bentonite ml-1 (pH 9-99)] was added to 500 µl purified virus suspension. The aqueous phase was extracted twice with 500 µl phenol/chloroform/isoamyl alcohol (25: 24: 1). The nucleic acids were precipitated with ethanol, dried and suspended in 30 µl RNase-free water.

**Synthesis of cDNA, cloning and sequencing.** The viral RNA was used to synthesize cDNA as template. First- and second-strand synthesis was performed by using the SuperScript Choice system for cDNA synthesis (Invitrogen) according to the manufacturer’s instructions, using both oligo(dt)12-18 primers and random hexamers. The cDNA products were ligated with EcoRI linkers and inserted into the EcoRI site of the alkaline phosphatase-treated pBluescript II SK(−) vector (Stratagene) by using a TaKaRa DNA Ligation kit (TaKaRa Bio Inc.). Then, the resultant plasmids were transformed into Escherichia coli DH5α competent cells as described previously (Hanahan et al., 1995).

A 5′ RACE (rapid amplification of cDNA ends) analysis was performed to determine the 5′-end sequence of the SssRNAV genome RNA, as described previously (Iwamoto et al., 2001). The viral RNA was reverse-transcribed by using SuperScript II (Invitrogen) with SssRNAV-specific primer R5 (5′-CAAAACAGTCTAATACTCGGC-3′; Fig. 1b), derived from the results of the above experiment, at 37 °C for 1 h. After purifying the product by using a SUPREC-02 column (TaKaRa Bio Inc.), the first-strand cDNAs were polyadenylated by using terminal deoxynucleotidyltransferase (TaKaRa Bio Inc.) at 37 °C for 5 min. Then, the second-strand cDNAs were synthesized by using Ex Taq polymerase (TaKaRa Bio Inc.) with an anchor primer (Anchor sequence 5′-GGCCACCAGCTCAGTCTGCAGTAC-3′) + Poly(T)
with one round of the following cycle parameters: denaturation at 95 °C (15 s), annealing at 55 °C (1 min) and extension at 72 °C (2 min). The double-stranded cDNAs were amplified with Ex Taq polymerase (TaKaRa Bio Inc.) using primer 5'-'GGCCACGGTGCAGTACAGTAC-3' (Iwamoto et al., 2001) and the SsRNase-specific primer R7 (5'-ATCAAGTGCTGGTGTTG-3'; Fig. 1b) according to the following cycle parameters: denaturation at 95 °C (40 s), annealing at 55 °C (1 min) and extension at 72 °C (2 min). Following 30 rounds of amplification, the resultant PCR products were ligated into the pGEM-T Easy vector (Promega). Then, the resultant plasmids were transformed into E. coli DH5α competent cells as mentioned above. DNA sequencing was conducted by using a DNA auto-sequencer (model 310; Applied Biosystems); fragmented sequences were assembled by using DNASIS Mac software (Hitachi Software Engineering). To verify the sequence of a region with low redundancy (≤2), RT-PCR was performed by using the primers R6 (5'-TCCCT-AATAGGGGAAA-3'; Fig. 1b) and F15 (5'-CAATCTGTCACCAAGTC-3'; Fig. 1b) and the amplicons were sequenced as described above.

**Protein sequencing.** Purified SsRNase virions were electrophoresed by using a denaturing SDS-PAGE gel, blotted onto a PVDF membrane (Millipore Immobilon-P; catalogue no. IPVH00010) and stained with Coomassie brilliant blue R-250 (Nacalai Tesque). Each stained protein band was excised and N-terminal amino acid sequencing was conducted by Edman degradation (Edman, 1950) for seven rounds using an Applied Biosystems Procise 492 cLC protein sequencer.

Total RNA was extracted from SsRNase-infected cells of *Schizochytrium* sp. NIBH N1-27 at 5 h post-infection by using an RNeasy Plant mini kit (Qiagen). Total RNA extracted similarly from uninfected cells served as a control. RNA size markers (3-9 and 0-7 kb) were prepared as follows. Briefly, the cDNA fragments corresponding to nt 5057–9018 and 8234–9018 of the SsRNase genome RNA were ligated into pBlueScript II SK(−) vector in an antisense orientation behind the T7 promoter using appropriate restriction enzymes; then, each RNA marker was transcribed in *vitro* by using T7 RNA polymerase (Roche Molecular Biochemicals). The RNAs were fractionated on a formaldehyde/agarose gel (1-5% agarose, 0-5% formaldehyde) and transferred onto a Hybond-N+ membrane (Amersham Biosciences); they were then cross-linked on the blots with UV illumination at 1200 × 100 μJ cm⁻² using a Funa-UV Cross-linker (Funakoshi). The membranes were incubated in hybridization buffer [50% formamide, 5 × SSC (1 × SSC: 150 mM NaCl, 15 mM sodium citrate, pH 7-0), 2% blocking reagent (Roche Molecular Biochemicals), 0-1% sarcosine and 0-2% SDS] for >1 h at 68 °C. The membranes were hybridized at 68 °C for 16 h with the RNA probes specific for the SsRNase genome sequence described below. Hybridized membranes were washed twice in 2 × SSC containing 0-1% SDS at room temperature for 5 min, then washed three further times in 0-1 × SSC containing 0-1% SDS at 68 °C for 15 min. After washing, the viral RNA was detected immunologically by using anti-digoxigenin (DIG)–alkaline phosphatase Fab fragments (Roche Molecular Biochemicals) and CDP-Star (Roche Molecular Biochemicals) according to the manufacturer’s protocols. We examined the preparation by using a luminescence image analyser (LAS 1000 Plus; Fuji Photo Film).

RNA probes to detect the ORF1 (probe 1) and ORF2 (probe 2) were prepared by cloning the cDNA fragments corresponding to nt 2062–3652 and 6950–8234 of the SsRNase genome RNA, respectively (Fig. 1d), into the pBlueScript II SK(−) vector in an antisense orientation behind the T7 promoter by using appropriate restriction enzymes. The RNA probe to detect ORF3 (probe 3) was prepared by cloning the cDNA fragment corresponding to nt 8234–9018 (Fig. 1d) into the pBlueScript II SK(−) vector in an antisense orientation behind the T3 promoter. After linearization, each riboprobe was transcribed in *vitro* by using the appropriate RNA polymerase in the presence of DIG-dUTP (Roche Molecular Biochemicals) according to the manufacturer’s recommendations.

**Computer analysis of the sequences.** Computer analyses were performed on both the nucleotide sequence and the amino acid sequence of the two non-structural protein domains: the putative RNA-dependent RNA polymerase (RdRp) and the putative RNA helicase. The potential coding region in the SsRNase genome RNA was predicted by using DNASIS Mac software. Database searches were performed by using BLASTX (Altschul et al., 1997).

Alignments of datasets were prepared by using the following procedures. First, the amino acid sequences were aligned automatically by using CLUSTAL_X (Thompson et al., 1997) and refined manually. Next, the nucleotide sequences were aligned according to the amino acid alignment data. The positions of the third base in each codon, ambiguous bases and gaps were removed for subsequent phylogenetic analysis and compared with this list of organisms (GenBank/ DDBJ accession numbers are shown): *Aichi virus* (AIV), AB010145;
Phylogenetic analyses. Generally, it is considered problematic to infer basal evolutionary relationships between RNA viruses based only on single-gene phylogenies (Zanotto et al., 1996); thus, in most cases, phylogenetic trees are constructed by using concatenated sequences of genes, which are treated as an undivided long unity. However, analyses of combined sequences do not explicitly take into account the differences of tempo and mode of evolution among the different genes. Therefore, we selected the total-evaluation method, using maximum-likelihood analyses of multiple genes by using the TotalML program in the MOLPHY version 2.3 package (Adachi & Hasegawa, 1996). Analyses of combined sequences do not explicitly take into account the differences of tempo and mode of evolution among the different genes. Therefore, we selected the total-evaluation method, using maximum-likelihood analyses of multiple genes by using the TotalML program in the MOLPHY version 2.3 package (Adachi & Hasegawa, 1996). The total support for a particular tree can be evaluated simply by summing up the estimated log-likelihoods of each individual gene, and the total log-likelihoods for different trees can then be compared (Adachi & Hasegawa, 1996). Using this method, the program must calculate the likelihood values for given topologies, so the selection of topologies is very important for evaluation within a reasonable time. First, the constraint tree was used to search for the best topology. The consensus tree (Fig. 2) was generated from eight trees that were constructed by using neighbour-joining (NJ; Saitou & Nei, 1987) and maximum-likelihood (ML; Kishino & Hasegawa, 1989) methods for both amino acid and nucleotide sequences for each single gene: the RdRp and RNA helicase. The bootstrap values from the NJ method were estimated as the local bootstrap probabilities by using NucML and TotalML programs from 4000 topologies that contained each of the nucleotide sequences of two genes derived by analysis using PAUP version 4.0b10 (Swofford, 2003) and PHYLIP version 3.6a3 (Felsenstein, 2002), respectively. The bootstrap values from the ML method were estimated as the local bootstrap probabilities by using NucML and ProtML.

Assessing the confidence limits of the phylogenetic trees. Selected topologies were assessed [the rejection limit was set at 10% (P value > 0.1)] for confidence by the Shimodaira–Hasegawa (SH) test (Shimodaira & Hasegawa, 1999) based on nucleotide and amino acid sequences by using a combination of NucML, ProtML and CONSEL version 0.1b (Shimodaira & Hasegawa, 2001). Bootstrap probabilities. The statistic for the nodes of topologies was evaluated by using bootstrap probabilities (Felsenstein, 1985) on the concatenated sequences. The bootstrap values from the NJ method were estimated by using the nucleotide and amino acid sequences of two genes derived by analysis using PAUP version 4.0b10 (Swofford, 2003) and PHYLIP version 3.6a3 (Felsenstein, 2002), respectively. The bootstrap values from the ML method were estimated as the local bootstrap probabilities by using NucML and ProtML.

RESULTS AND DISCUSSION
cDNA clones and sequence analysis of SssRNAV RNA

The nucleotide sequence of the viral genome RNA was determined by assembling four overlapping cDNA clones and five RT-PCR clones. The 3′-terminal nucleotide sequence was determined by comparing the sequences of seven clones (Fig. 1c). Consequently, the SssRNAV genome was shown to have a 3′ poly(A) tail and was estimated to be 9018 nt in length, excluding the poly(A) tail (Fig. 1a). Thus, the genome was determined to be smaller than the previously estimated size by means of formaldehyde/agarose gel electrophoresis (10.2 kb; Takao et al., 2005), which may reflect the additional length of the 3′ poly(A) tail. The base composition was A, 24.1%; C, 26.1%; G, 23.7%; U, 26.1%. The A + U content of SssRNAV was 50.2%; this is not as high as those of other viruses within the family Dicistroviridae (56.8–64.1%; Table 2).

Coding and non-coding regions of SssRNAV genome RNA

Computer-assisted analysis shows that the genome RNA contains two large ORFs encoded in different reading frames, nt 555–5240 (ORF1) and nt 5333–8749 (ORF2). Another ORF, nt 8319–8948 (ORF3), was also identified, which overlaps ORF2 by 382 nt (Fig. 1a). These three ORFs account for 93% of the SssRNAV genome [excluding the poly(A) tail]; the other 7% consists of non-coding or untranslated regions (UTRs). These include the 5′ UTR of 554 nt, an intergenic region of 92 nt separating ORF1 and ORF2, and the 3′ UTR of 70 nt, followed by the poly(A) tail.

**Fig. 2.** Consensus tree generated from the eight trees. The tree was constructed by using PHYLIP ver. 3.6a3 by the strict-rule consensus-tree method from eight phylogenetic trees in Table 1 (phylogenetic trees 4–11).
Table 1. Comparison between TotalML and other phylogenetic analysis topologies by SH test

Abbreviations: \(-\ln L\), log-likelihood; Diff \(-\ln L\): difference of log-likelihood; SH, \(P\) value of SH test. Nucleotide – NJ distances were estimated on the HKY85 model (Hasegawa et al., 1985) generated by PAUP ver. 4.0b10 (Swofford, 2003). Nucleotide – ML: a transition/transversion (Ti/Tv) parameter (0-55 – RdRp, 0-54 – helicase) for the HKY85 model was estimated on the ML criterion generated by PAUP ver. 4.0b10. Amino acid – NJ distances were estimated on the JTT model (Jones et al., 1992) generated by PHYLIP ver. 3.6a3 (Felsenstein, 2002). Amino acid – ML: the ML criterion was estimated on the JTT model generated by PHYLIP ver. 3.6a3. Topologies are shown in the case rooted with TSV. Abbrevations of taxa in topologies: A, AiV; B, BQCV; Bo, BoCV; Bp, BPMV; C, CrPV; Cp, CPSMV; D, DCC; Dw, DWV; H, HaRNAV; N, NV; P, PV; Py, PYFV; R, RTSV; S, SssRNAV; Sb, SBV; T, TrV; Ts, TSV.

<table>
<thead>
<tr>
<th>Phylogenetic tree</th>
<th>Topologies</th>
<th>Nucleotide sequences</th>
<th>Protein sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Rank</td>
<td>(-\ln L)</td>
</tr>
<tr>
<td>1. TotalML nucleotide – a (Fig. 5a)</td>
<td>(Ts,(((S,((H,((P,A),(N,Bo))),((R,Py),(Bp,Cp))),((Sb,Dw)),(T,B)),(D,C))));</td>
<td>1</td>
<td>12149-1</td>
</tr>
<tr>
<td>2. TotalML nucleotide – b (Fig. 5b)</td>
<td>(Ts,(((S,((H,((P,A),(N,Bo))),((R,Py),(Bp,Cp))),((Sb,Dw)),(T,B)),(D,C))));</td>
<td>1</td>
<td>12149-1</td>
</tr>
<tr>
<td>3. TotalML amino acid (Fig. 5c)*</td>
<td>(Ts,(((H,((P,A),(N,Bo)),(Sb,Dw))),(T,B)));</td>
<td>5</td>
<td>12155-4</td>
</tr>
<tr>
<td>4. RdRp nucleotide – NJ</td>
<td>(Ts,((C,D),(B,T),(S,(((H,((A,P),(N,Bo))),((Py,R),(Cp,Bp)),(Dw,Sb))),(T,B)));</td>
<td>4</td>
<td>12151-0</td>
</tr>
<tr>
<td>5. RdRp nucleotide – ML</td>
<td>(Ts,(((B,T),(S,(((A,(H,P)),(N,Bo)),((Py,R),(Cp,Bp)),(Dw,Sb))),(D,C))));</td>
<td>3</td>
<td>12150-1</td>
</tr>
<tr>
<td>6. RdRp amino acid – NJ</td>
<td>(Ts,(((S,B,T)),((H,((A,P),(N,Bo))),((Py,R),(Cp,Bp)),(Dw,Sb))),(T,B)));</td>
<td>8</td>
<td>12162-3</td>
</tr>
<tr>
<td>7. RdRp amino acid – ML</td>
<td>(Ts,(((S,B,Dw)),((Bp,Cp),(R,Py)),((H,((Bo,N),(P,A))),((S,T),(B,D))),(T,B)));</td>
<td>6</td>
<td>12160-2</td>
</tr>
<tr>
<td>8. Helicase nucleotide – NJ</td>
<td>(Ts,(((B,T),(S),(((A,P),(N,Bo)),(Dw,Sb)),(D,C))));</td>
<td>7</td>
<td>12160-8</td>
</tr>
<tr>
<td>9. Helicase nucleotide – ML</td>
<td>(Ts,(((B,T),(S,(((A,H,P),(N,Bo)),((Py,R),(Cp,Bp)),(Dw,Sb))),(T,B),(C,D))));</td>
<td>10</td>
<td>12173-6</td>
</tr>
<tr>
<td>10. Helicase amino acid – NJ</td>
<td>(Ts,(((A,H,P),(N,Bo)),(Dw,Sb)),(B,D));</td>
<td>9</td>
<td>12172-5</td>
</tr>
<tr>
<td>11. Helicase amino acid – ML</td>
<td>(Ts,(((A,H,P),(V,Bo)),(S,B,Dw)),((R,Py),(Sb,Bp))),(T,B),(C,D));</td>
<td>11</td>
<td>12199-4</td>
</tr>
</tbody>
</table>

*TotalML amino acid tree topology is consistent with the NJ trees deduced from the amino acid and nucleotide sequence of tandemly combined domains of helicase and RdRp.
Table 2. Base usage (%) of SssRNAV and dicistroviruses

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
<th>A+U</th>
</tr>
</thead>
<tbody>
<tr>
<td>SssRNAV</td>
<td>24·09</td>
<td>26·14</td>
<td>23·66</td>
<td>26·00</td>
<td>50·09</td>
</tr>
<tr>
<td>Dicistroviridae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCV</td>
<td>29·93</td>
<td>16·24</td>
<td>20·37</td>
<td>33·44</td>
<td>63·37</td>
</tr>
<tr>
<td>CrPV</td>
<td>32·60</td>
<td>18·42</td>
<td>20·89</td>
<td>28·07</td>
<td>60·67</td>
</tr>
<tr>
<td>BQCV</td>
<td>29·20</td>
<td>18·53</td>
<td>21·63</td>
<td>30·61</td>
<td>59·81</td>
</tr>
<tr>
<td>TrV</td>
<td>28·70</td>
<td>16·11</td>
<td>19·81</td>
<td>35·37</td>
<td>64·07</td>
</tr>
<tr>
<td>TSV</td>
<td>27·99</td>
<td>20·20</td>
<td>23·03</td>
<td>28·76</td>
<td>56·75</td>
</tr>
</tbody>
</table>

There are three AUG initiation codons within the first 100 nt of ORF1. The first and second AUG codons are located at nt 555–557 and 578–580, respectively, and the sequences around them (UGUAUGC and UAUAGUC) are not in agreement with the most common initiation sequence Kozak consensus, ACCAUUG (Kozak, 1991). In contrast, the sequence around the third AUG codon (nt 627–629) coincides with the Kozak consensus (Fig. 1a). Although there are three and two AUG initiation codons within the first 150 nt of ORF2 and ORF3, respectively, no Kozak consensus is identified. Currently, we do not have experimental data to demonstrate which AUG codon works as the initiation codon for each ORF.

Alignment of the amino acid sequence for non-structural proteins

The deduced amino acid sequence of ORF1 showed high similarity to proteins involved in the replication of picorna-like viruses, especially from the family Dicistroviridae (e value $> 1 \times 10^{-25}$). Further analysis revealed that it contains the core motifs of the picornavirus 2C RNA helicase, 3C cysteine protease and 3D RNA-dependent RNA polymerase (RdRp) (Koonin & Dolja, 1993). The first conserved motif of RNA helicase, GXXGXXGK (motif A), was found at aa 343–349, assuming that the first AUG (nt 555–557) is the initiation codon (Fig. 3a); the second conserved motif, QX₃DD (motif B), was not identified; the third domain, KGX₃SX₃STN (motif C), was found as KGX₃PX₃DTN at aa 437–452 (Fig. 3a). The cysteine protease motif, GXCG, was found at aa 933–936 (Fig. 3b). The conserved domains of RdRp, LKDE (motif I), SGX₃TX₃N (motif V), YGDD (motif VI) and LKR (motif VII), were found at aa 1208–1211, 1350–1359, 1405–1408 and 1459–1461, respectively (Fig. 3d).

Alignment of the amino acid sequence for the capsid proteins

The deduced amino acid sequence of ORF2 shows similarities to the capsid proteins of picorna-like viruses, especially those from the family Dicistroviridae (e value $\geq 2 \times 10^{-36}$).

SDS-PAGE analysis revealed that SssRNAV has three major structural proteins (37, 34 and 32 kDa) and one minor protein (18 kDa) (Takao et al., 2005). We designated them VP1, VP2, VP3 and VP4 in decreasing order of their molecular masses. The N-terminal sequence of VP3 was revealed to be SKPLV by using Edman degradation (data not shown). This coincides with the deduced amino acid sequence from ORF2 (nt 7046–7060). This amino acid sequence is consistent with the N-terminal sequence of major capsid proteins from dicistroviruses, HaRNAV and picorna-like viruses (Fig. 3c). The N-terminal sequence of the other capsid proteins could not be determined by using Edman degradation, presumably due to an N-terminal block.

Northern blot analysis and translation strategy of ORF2 and ORF3

Northern blot analysis of the total RNA extracted from SssRNAV-infected cells using probe 1 (specific for ORF1) revealed one RNA band of the viral genome size (Fig. 4, lane 1). In contrast, two bands (5·7 and 4·9 kb) in addition to the genome-size RNA were detected in the analysis using probe 2 (specific for ORF2) (Fig. 4, lane 2). When using probe 3 (specific for ORF3), a dense band of 0·85 kb and two paler bands similar to those identified by using probe 2 were observed (Fig. 4, lane 3).

The genome arrangement of SssRNAV is similar to that of dicistroviruses and of some calciviruses where the replication proteins and capsid proteins are coded in distinct ORFs; in the case of the calciviruses, the intergenic region is only 3–16 bases in length and the sequence consistency with SssRNAV is very low (e value $\geq 8 \times 10^{-8}$). ORF2 of SssRNAV, encoding the capsid proteins, is located 92 nt downstream of ORF1, encoding the replication proteins (Fig. 1a). Generally, the downstream ORF can only be translated by using specific mechanisms mentioned below.

There are three conceivable ways to translate ORF2 and ORF3 of SssRNAV: (i) a read-through translation and/or ribosomal frameshift may occur between ORF1 and ORF2 and between ORF2 and ORF3; (ii) translation of ORF2 and/or ORF3 is initiated independently by an internal ribosome entry site (IRES); or (iii) the subgenomic RNA may be synthesized during replication that includes ORF2 or ORF3. Although we cannot exclude the first possibility, the second method is used by the family Dicistroviridae (Mayo, 2002). The intergenic region (IGR) of these viruses forms stem–loop structures that act as an IRES (Sasaki & Nakashima, 2000). The ribosomes bind this structure directly and initiate translation of the downstream ORF without a universal AUG initiation codon (Sasaki & Nakashima, 2000; Wilson et al., 2000). Because we failed to determine the N-terminal sequence of VP1, we could not identify the IGR precisely; thus, its secondary structure was not determined. Although the IRESs of dicistroviruses share some similarity at the nucleotide-sequence level, no sequence showing similarity to IRESs of dicistroviruses was identified within the SssRNAV genome. In general, when the IGR contains an IRES, only one type of messenger-sense RNA containing both ORFs is produced, and the ribosomes bind directly to the IRES and translate each ORF independently (Sasaki &
Fig. 3. Multiple alignments of conserved regions in putative replicases of SssRNAV and picorna-like viruses. The number to the left of each sequence indicates amino acid position relative to the corresponding replicase protein, and those within sequences represent the number of omitted amino acids. Residues identical in more than three viruses are shaded. (a) Alignments of putative helicase domains, in which conserved regions (Koonin & Dolja, 1993) are marked Hel-A, -B and -C. (b) Alignments of protease domains where asterisks indicate amino acid residues that may be essential for protease activity. (c) Alignment of partial regions of viral capsid proteins of CrPV, DCV, HaRNAV and SssRNAV. The vertical bar indicates the cleavage site for VP3 and VP4. (d) Alignments of RdRp domains where conserved regions (Koonin & Dolja, 1993) are marked I–VIII.
The synthesis of subgenomic RNA is another tactic commonly found in many positive-sense ssRNA viruses to translate downstream ORFs (Miller & Koev, 2000). Northern blot analysis shows that smaller-sized RNAs also accumulated during SssRNAV replication (Fig. 4), suggesting the production of subgenomic RNAs. Further study is required to determine whether the detected RNAs are the subgenomic RNAs of SssRNAV.

Detection of a 0·85 kb band (Fig. 4, lane 3) strongly suggests that the protein encoded by ORF3 is also expressed from the subgenomic RNA that is synthesized during replication. Independent synthesis of the ORF3 subgenomic RNA is a remarkable feature of SssRNAV that caliciviruses do not have (Koopmans et al., 2005). No significant similarity was found in GenBank for the ORF3 product. Further study is necessary to determine the function of ORF3.

**Phylogenetic analysis**

The total evaluation of RdRp and RNA helicase with the ML analyses provided three topologies that contained two equally good topologies based on the nucleotide sequences and one best topology based on the amino acid sequences (Fig. 5). SssRNAV was located in a different phylogenetic position in each topology. Three of these trees showed the following monophyletic relationships, which were also supported by high bootstrap probabilities: *Caliciviridae* and *Picornaviridae; Comoviridae and Sequiviridae; all members of the family *Dicistroviridae*, excluding the genus *Iflavirus*. However, SssRNAV did not form a monophyletic group with any of the described families. In the TotalML tree based on nucleotide sequence, SssRNAV and HaRNAV formed a monophyly; however, this relationship was supported by a low bootstrap probability (Fig. 5b).

The results of the SH test in the TotalML trees and other phylogenetic trees are shown in Table 2. The best topologies selected by TotalML analysis were also evaluated as the best by means of the SH test. The three topologies considering the RNA helicase gene (nucleotide, ML; amino acid, NJ and ML) were rejected. That is, it is necessary to consider all of the trees of the RdRp and a tree of the nucleotide sequences of the RNA helicase by NJ analysis as the possible phylogeny, as well as the three best TotalML trees.

These phylogenetic analyses suggest that all of the five established families are probably natural taxa, and that SssRNAV and HaRNAV have an independent lineage that evolved from a deep internal branch in the picorna-like viruses. Therefore, each of these marine RNA viruses should be recognized as a distinct family-level taxon, even if the number of members of the taxon is small. Moreover, if SssRNAV is classified in the family *Dicistroviridae* based only on the similarity of the genome, it is clear that the reliable taxonomic criterion will be lost.

**Conclusions**

In this study, we have determined and analysed the genome sequence of SssRNAV infecting the marine fungoid protist *Schizochytrium* sp. This is, to our knowledge, the first report describing the genome organization of ssRNA viruses infecting marine fungoid protists. The sequence analysis showed some similarity of SssRNAV to members of the family *Dicistroviridae* in genome size and arrangement, and also the amino acid sequence of structural and non-structural proteins (Figs 3). However, Northern blot data suggest that SssRNAV does not utilize IRES-mediated translation for downstream ORFs. Furthermore, phylogenetic analyses indicated that SssRNAV and dicistroviruses do not have a close relationship and they should be classified into distinct groups. Additionally, the A+U content of SssRNAV is not as high as those of members of the family *Dicistroviridae*. In comparing SssRNAV and HaRNAV, both infecting marine eukaryotic micro-organisms, their hosts are classified within the same category, the stramenopiles, which are characterized by common ultrastructure (Patterson, 1989); however, these two viruses are clearly different in genome size, genome structure and, presumably, in replication mechanism. From these results, we conclude that SssRNAV is a species distinct from those belonging to the family *Dicistroviridae* and other virus families, and that SssRNAV is a member of a previously unrecognized virus group.

Few reports for picorna-like viruses isolated from marine environments have been made, although there should exist a
Fig. 5. Phylogenetic trees selected by using the TotalML program. Best results based on nucleotide sequence (a, b) and amino acid sequence (c). The branch lengths were estimated by using the ML method based on concatenated nucleotide or the amino acid sequence of RdRp and helicase, respectively. Numbers at each internal branch represent the bootstrap values. Nuc, Nucleotide sequence; Prot, protein; LBP, local bootstrap probability.
huge number of unrecognized and highly diversified RNA viruses. This is based on the analysis of the RdRp gene sequence amplified from the natural marine communities, where Culley et al. (2003) postulated that marine RNA viruses are highly diverse. Hence, for understanding the classification and evolutionary history of RNA viruses more precisely, we should note that the ocean is a fascinating field where findings of undiscovered RNA viruses are strongly expected; thus, more efforts should be thrown into it.

**ACKNOWLEDGEMENTS**

This work was supported by the Kato Memorial Bioscience Foundation and the Asahi Glass Foundation. The authors are grateful to Dr. Toru Nakahara, Dr. Toshio Yokohi (National Institute of Advanced Industrial Science and Technology, Japan) and Ms. Rinka Yokoyama (Konan University, Japan), who kindly provided the thraustochytrid strains. We wish to thank Dr. Hironori Mizumoto and Mr. Masahiro Tsutsui (Kyoto University, Japan) for their technical assistance in the molecular biology experiments. We are also grateful to Dr. Hitoshi Shimodaira (Tokyo Institute of Technology, Japan), who kindly gave help in the phylogenetic analyses.

**REFERENCES**


