Journal of General Virology (2000), **81**, 1473–1484. Printed in Great Britain

Gill-associated virus of *Penaeus monodon* prawns: an invertebrate virus with ORF1a and ORF1b genes related to arteri- and coronaviruses

Jeff A. Cowley, Christine M. Dimmock, Kirsten M. Spann and Peter J. Walker

Co-operative Research Centre for Aquaculture, CSIRO Tropical Agriculture, Long Pocket Laboratories, PMB3, Indooroopilly 4068, Australia

A 20089 nucleotide (nt) sequence was determined for the 5′ end of the (+)-ssRNA genome of gill-associated virus (GAV), a yellow head-like virus infecting *Penaeus monodon* prawns. Clones were generated from a ~ 22 kb dsRNA purified from lymphoid organ total RNA of GAV-infected prawns. The region contains a single gene comprising two long overlapping open reading frames, ORF1a and ORF1b, of 4060 and 2646 amino acids, respectively. The ORFs are structurally related to the ORF1a and ORF1ab polyproteins of coronaviruses and arteriviruses. The 99 nt overlap between ORF1a and ORF1b contains a putative AAAUUUU 'slippery' sequence associated with −1 ribosomal frameshifting. A 131 nt stem–loop with the potential to form a complex pseudoknot resides 3 nt downstream of this sequence. Although different to the G/UUUAAAC frameshift sites and 'H-type' pseudoknots of nidoviruses, *in vitro* transcription/translation analysis demonstrated that the GAV element also facilitates read-through of the ORF1a/1b junction. As in coronaviruses, GAV ORF1a encodes a 3C-like cysteine protease domain located between two hydrophobic regions. However, its sequence suggests some structural relationship to the chymotrypsin-like serine proteases of arteriviruses. ORF1b encodes homologues of the 'SDD' polymerase, which among (+)-RNA viruses is unique to nidoviruses, as well as metal-ion-binding and helicase domains. The presence of a dsRNA replicative intermediate and ORF1a and ORF1ab polyproteins translated by a −1 frameshift suggests that GAV represents the first invertebrate member of the Order *Nidovirales*.

Introduction

Gill-associated virus (GAV) is an enveloped, rod-shaped virus that has been associated with mortalities in farmed *Penaeus monodon* (black tiger prawns) in Australia (Spann *et al.*, 1997). In morphology and pathology, GAV closely resembles yellow head virus (YHV), which in Thailand and other Asian countries has caused massive production losses in farmed prawns (Limsuwan, 1991; Boonyaratpalin *et al.*, 1993; Chantanachookin *et al.*, 1993). A non-pathogenic form of GAV, lymphoid organ virus (LOV), has also been identified in hypertrophied cell foci in the lymphoid organ of healthy *P. monodon* in Australia (Spann *et al.*, 1995). Limited sequence comparisons indicate that LOV and GAV are minor variants of the same virus and that YHV is a distinct but closely related topotype (Cowley *et al.*, 1999, 2000).

Author for correspondence: Jeff Cowley.
Fax +61 7 32142881. e-mail Jeff.Cowley@tag.csiro.au

The sequence described in this paper has been deposited in the GenBank database under accession no. AF227196.

There has been considerable confusion about the taxonomic classification of these viruses. Based primarily on its size and enveloped rod-shaped appearance, YHV was originally proposed to be a baculovirus (Chantanachookin *et al.*, 1993). Subsequent evidence that the YHV genome comprised ssRNA led, however, to the suggestion that it may be a rhabdovirus or coronavirus (Wongteerasupaya *et al.*, 1995). P. C. Loh and colleagues identified four major structural proteins (~ 200, 135, 67 and 22 kDa) in purified YHV, and obtained evidence that the 135 kDa protein is glycosylated and that the ~ 22 kb ssRNA genome may have negative polarity (Loh *et al.*, 1997; Nadala *et al.*, 1997). This led them to classify YHV provisionally in the *Rhabdoviridae*. More recently, Tang & Lightner (1999) used strand-specific probes and PCR to demonstrate that YHV virions derived from haemolymph contain a (+)-ssRNA genome. It is evident that these conclusions are drawn from relatively few data and that additional information on the genome sequence, organization and replication strategy is required to establish the appropriate taxonomic classification of YHV and GAV.

In this paper, we describe the 20089 nucleotide (nt) 5′-

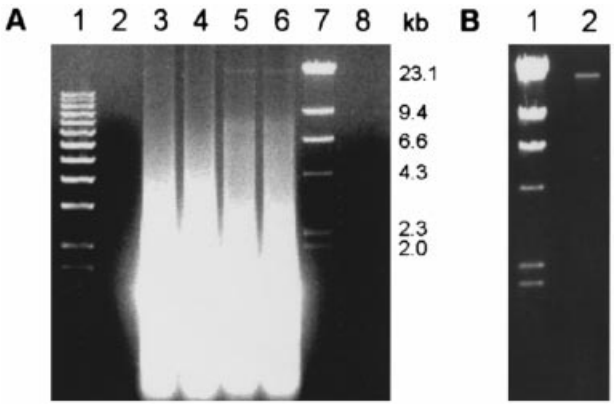

Fig. 1. (A) Neutral 0·8% agarose gel of lymphoid organ total RNA isolated from uninfected and GAV-infected *P. monodon* either untreated (lanes 1, 3, 5, 7) or digested with DNase I (lanes 2, 4, 6, 8). 1 kb DNA ladder (lanes 1, 2), RNA from uninfected prawns (lanes 3, 4) and infected prawns (lanes 5, 6), *Hin*dIII-cut λ DNA (lanes 7, 8). (B) Neutral 0·8% agarose gel of gel-purified ~ 22 kb dsRNA (lane 2), *Hin*dIII-cut λ DNA (lane 1).

terminal sequence of the GAV RNA genome. The sequence comprises a single gene encoding two long overlapping open reading frames (ORFs). The proteins encoded in these ORFs display homologies to conserved functional domains of the ORF1a and ORF1ab polyproteins of members of the *Coronaviridae* and *Arteriviridae*. We also show that the ORF1a/1b overlap contains an RNA pseudoknot that functions *in vitro* as a −1 ribosomal frameshift site. The structural organization and expression strategy of this genome region, together with the close genetic relationship of GAV to YHV (Cowley *et al.*, 1999), indicate that they represent the first invertebrate viruses of the Order *Nidovirales*.

Methods

■ **Virus.** GAV used in this study originated from a pool of two diseased *P. monodon* prawns from a farm in Queensland in 1996 (Spann *et al.*, 1997). Healthy *P. monodon* were infected with GAV by tail muscle injection (5 µl/g) of a 0·2 µm filtrate of homogenized head tissue as described previously (Spann *et al.*, 1997). Lymphoid organs were collected 6 days post-infection, frozen on dry ice and stored under liquid nitrogen.

■ **RNA isolation and analysis.** Total RNA was isolated from 0·4 g pooled lymphoid organs from uninfected and GAV-infected *P. monodon* using 8 ml TRIzol-LS (Gibco-BRL). RNA was resolved (10 µg/lane) in 0·8% agarose–TAE gels containing 0·5 µg/ml ethidium bromide. To determine the nature of a discrete ~ 22 kb band, 10 µg RNA was treated with 1 U RQ1-DNase (Promega) for 15 min at 37 °C. DNase digestion was monitored using 1 µg 1 kb DNA ladder (Gibco-BRL) and 0·5 µg *Hin*dIII-cut bacteriophage λ DNA. RNase sensitivity was determined by incubating the gel in NTE pH 7·4 containing 50 µg/ml RNase A until ribosomal RNA had been digested. To isolate the ~ 22 kb dsRNA, 400 µg total RNA was resolved in a 0·6% neutral LMP–agarose (Gibco-BRL) gel, the band excised and dsRNA recovered by β-agarase (Boehringer Mannheim) digestion.

■ **RT–PCR and cloning.** Sequences representative of the GAV genome were amplified using a random RT–PCR method (Froussard, 1992) modified as described previously (Cowley *et al.*, 1999). Briefly, cDNA was synthesized from 10 ng gel-purified ~ 22 kb dsRNA using 150 ng UNI-primer-dN_6 [5′ GCCGGAGCTCTGCAGAATTC$(N)_6$ 3′], 200 U Superscript II (Gibco-BRL) reverse transcriptase (RT) and incubation at 42 °C for 1 h followed by 99 °C for 5 min. Second-strand cDNA was synthesized using 8 U Klenow polymerase (Promega) and incubation at 37 °C for 30 min. Primer was removed using a Sephadex S400 HR spin column (Pharmacia). PCR amplification of cDNA (1 µl) utilized the UNI-primer 5′ GCCGGAGCTCTGCAGAATTC 3′ and Qiagen *Taq* DNA polymerase and buffer containing 1·5 mM $MgCl_2$ (Froussard, 1992). PCR products were purified using a QIAquick column (Qiagen), ligated into 25 ng pGEM-T (Promega) and transformed into competent *E. coli* DH5α cells (Gibco-BRL) by standard methods (Sambrook *et al.*, 1989).

Additional clones were generated similarly using lymphoid organ total RNA and GAV-specific primers, Superscript II RT and either RT–PCR using various conditions or long distance RT–PCR using the Expand Long Template PCR System (Boehringer Mannheim). Alternatively, clones were generated from cDNA synthesized using GAV-specific primers followed by second-strand synthesis randomly primed with UNI-primer-dN_6 and extended using Klenow DNA polymerase (Froussard, 1992). cDNAs of variable length were amplified by PCR (Froussard, 1992) using both the virus-specific primer and UNI-primer and *Taq* DNA polymerase and buffer (Promega) containing 2·5 mM $MgCl_2$. RT–PCR products were either gel purified or processed directly using QIAquick columns and cloned into either pGEM-T vector or *Eco*RV-cut pBluescript II KS(+).

■ **Analysis of the ORF1a/1b −1 ribosomal frameshift element.** Functional activity of the GAV −1 ribosomal frameshift site at the ORF1a/1b overlap was demonstrated by protein expression *in vitro* using a T7 TnT coupled transcription/translation system (Promega). A 693 nt sequence encompassing the ORF1a/1b overlap was amplified by PCR from clone P68-15 using primers GAV82 (5′ GCAATGG-ATCCATCCCTGTCGG 3′) and GAV83 (5′ TTAGCGGATCCATG-CAACACATC 3′) spanning *Bam*HI-like sites in ORF1a and ORF1b, respectively. Nucleotides modified to generate *Bam*HI sites are underlined. PCR utilized *Pwo* DNA polymerase (Boehringer Mannheim) and 25 cycles of 95 °C/45 s, 62 °C/45 s, 72 °C/45 s. The 0·7 kb PCR product was purified using a QIAquick column, digested with *Bam*HI, similarly repurified and cloned into a plasmid (pBG) containing the bovine ephemeral fever virus (BEFV) G gene (Walker *et al.*, 1992) downstream of a T7 RNA promoter in pBluescript II KS(+). The ORF1a/1b overlap region was cloned into pBG by insertion into a *Bam*HI site, at position G^{1176}–C^{1181} of the BEFV G gene, so that ORF1a and ORF1b were in-frame with the N- and C-terminal ORFs of the G protein, respectively. Clones with the insert in the correct (pBG-F1) and reverse orientations (pBG-R10) were identified and high quality DNA was prepared with a Plasmid Maxi Kit (Qiagen).

Plasmid DNA (0·5 µg) was combined with TnT Rabbit Reticulocyte Lysate reagents (Promega), 0·4 µl [^{35}S]methionine (1000–1500 Ci/mmol, Amersham) and 5 U T7 RNA polymerase (Promega) in a 10 µl reaction and protein was synthesized by incubating at 30 °C for 2 h. An aliquot (5 µl) of the TnT reaction was added to 20 µl SDS–PAGE buffer (125 mM Tris–HCl pH 6·8, 2% SDS, 5% 2-mercaptoethanol, 20% glycerol, 0·01% bromophenol blue) and heated at 99 °C for 5 min. Proteins were electrophoresed in an 8% SDS–polyacrylamide gel, fixed, impregnated with Amplify (Amersham) and dried. [^{35}S]Methionine-labelled proteins were detected by fluorography by exposure to Fuji RX film at −70 °C. Radioactivity in proteins was quantified using a Bio-Rad phosphor-imaging system.

■ **5′-RACE.** The putative 5′ terminus of the GAV genomic RNA was determined by a primer extension anchor–PCR method for the random

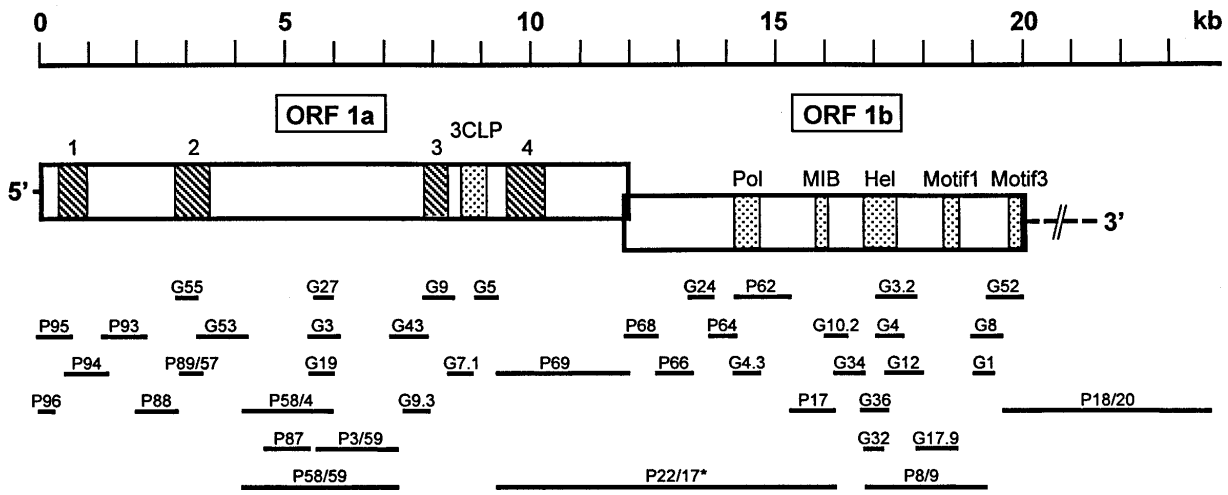

Fig. 2. Clones used to determine the sequence of the 5′-terminal ~ 20 kb region of the (+)-ssRNA genome of GAV containing the ORF1a and ORF1b coding regions. Clones were generated by random RT–PCR from purified dsRNA (G) or RT–PCR from total RNA using either one (e.g. P66) or two (e.g. P8/9) GAV-specific primers. At least three independent clones of specifically-primed RT–PCR products were sequenced. A ~ 6·8 kb RT–PCR product, P22/17*, spanning clones G5 and G34, was generated but not sequenced. The relative positions of the hydrophobic regions (1–4), 3CLP, polymerase, MIB, helicase and Motif 1 and 3 domains in ORF1a and ORF1b are indicated.

ORF1a →
M E

5′-ACGUUACGUUCCACGUACUUAUCCCACUGCAUCAAAUAAAUCUUCCCAGUUUUCCAGCGUUUCCGACC aug gag...3′

Fig. 3. Sequence of the putative 68 nt untranslated region upstream of the ORF1a coding sequence to the putative 5′ terminus of the GAV genome.

amplification of cDNA ends (5′-RACE) (Dumas *et al.*, 1991) using complementary anchor–PCR primers (156 and 2668) described previously (Walker *et al.*, 1994). Briefly, cDNA was synthesized using any of four GAV antisense primers including GAV94 (5′ CGTGGTGTGATCATAGTCCTT 3′) and GAV95 (5′ CTAAGCTCTGGAGGTTGATCAT 3′) at different positions toward the N terminus of ORF1a. cDNA was synthesized using either 1 µg lymphoid organ total RNA from GAV-infected prawns or RNA isolated from GAV semi-purified from clarified prawn haemolymph (15000 ***g*** for 5 min) by ultracentrifugation at 200000 ***g***, 45 min at 4 °C. The reaction mixture containing RNA, 150 ng primer and 200 U Superscript II RT was incubated at 42 °C for 1 h. RNA was digested with 1·5 U RNase H at 37 °C for 1 h and primer was removed using a Sephadex S400-HR column. Primer 156 (0·2 µg), 5′-phosphorylated using [γ-^{32}P]ATP (Bresatec) and T4 polynucleotide kinase (Promega), was ligated to 3′ cDNA ends using T4 RNA ligase (NEB) (Walker *et al.*, 1994). Primer was removed as above and cDNA was amplified by PCR using primer 2668 in combination with GAV-specific primers GAV95 or GAV96 (5′ TCTGACGGCAGACAGAACAGT 3′), *Taq* DNA polymerase and buffer (Promega) containing 2·5 mM $MgCl_2$ and 40 cycles of 95 °C/30 s, 60 °C/30 s and 72 °C/30 s. PCR products were purified using QIAquick columns, cloned into pGEM-T and sequenced.

■ **DNA sequencing and analyses.** Plasmid DNA was prepared using the RPM (BIO 101) or PerfectPREP (5 Prime–3 Prime) kits and sequenced using either the Thermo-Sequenase (Amersham) or Big Dye (ABI) dye-terminator systems. Sequencing utilized T7, SP6, universal pUC forward and reverse primers or GAV-specific oligonucleotides. Sequences were determined using an ABI Model 377 automated sequencer at the Australian Genome Reference Facility. SeqEd 1.0.3 and MacVector 6.0.1 were used for sequence alignments, ORF determinations and sequence translations. Database sequence similarity searches were done with the BLAST 2.0 program (Altschul *et al.*, 1997). Multiple sequence alignments were undertaken using Clustal W 1.7 (Thompson *et al.*, 1994). Phylogenetic trees were constructed using the maximum parsimony (PROTPARS) and neighbour-joining (NEIGHBOR) methods in Phylip 3.5 (Felsenstein, 1993). The COMPARE and GAP programs in the Genetics Computer Group (GCG) package were used for ORF1a and ORF1b sequence comparisons. RNA folding structures were predicted using MFOLD and PLOTFOLD (Zuker, 1989).

Results

Cloning and sequence analysis of the GAV genome

As continuous cell cultures are not available for propagating prawn viruses, GAV RNA was obtained from the lymphoid organs of infected prawns. Analysis in agarose gels revealed a discrete ~ 22 kb band in lymphoid organ total RNA from infected but not uninfected prawns (Fig. 1). The band was resistant to digestion by both DNase I and RNase A, indicating that it comprised dsRNA (Fig. 1 and data not shown). The ~ 22 kb dsRNA was gel-purified and used to prepare a library of GAV clones by random RT–PCR. After preliminary sequence analysis, additional overlapping clones generated by primer-specific RT–PCR were used to fill gaps in the sequence (Fig. 2).

The putative 5′ end of the GAV genome was deduced by 5′-RACE (Fig. 3). RT–PCR with different near-terminal primers generated single products of sizes indicating a common 3′

cDNA end (data not shown). Moreover, identical products were amplified from either total RNA from lymphoid organ tissue or semi-purified viral RNA. Of 20 clones sequenced, five initiated at 5′ GGAC, ten at 5′ GAC, four at 5′ AC and one at 5′ C. Similar analyses for putative sub-genomic RNAs initiating in downstream intergenic regions identified 5′ termini with double and single G residues not matching the genomic sequence (J. A. Cowley and others, unpublished). Taken together with the fact that the terminal addition of non-template-directed nucleotides is a common feature of reverse transcriptases (Peliska & Benkovic, 1992; Patel & Preston, 1994), we reason that the likely 5′ terminus of the GAV genome commences at 5′ AC.

The 20089 nt sequence from the putative 5′ AC terminus of the GAV genome contains a single gene encoding two long overlapping ORFs (Fig. 2). The first, ORF1a, commences 2 nt from the 5′ end and extends to a position 12248 nt downstream. The first in-frame initiation codon ACCAUGG is located 68 nt from the 5′ end and is in a favourable context for translation initiation (Fig. 3) (Kozak, 1986). Assuming initiation at this site, the ORF1a polypeptide comprises 4060 aa. The second, ORF1b, comprises 7941 nt [2646 amino acids (aa)] and overlaps ORF1a by 99 nt (33 codons).

ORF1a sequence

The ORF1a polypeptide has a deduced molecular mass of 460273 Da, a pI of 6·63, 43 potential *N*-linked glycosylation sites and four significant hydrophobic regions comprising putative multiple membrane-spanning domains (Fig. 2). The two C-terminal hydrophobic regions (aa 2575–2722 and 3176–3434) are also evident in nidovirus ORF1a polyproteins (Boursnell *et al.*, 1987; Lee *et al.*, 1991; Herold *et al.*, 1993). Comparisons of the predicted ORF1a polyprotein of GAV with those of bird (avian infectious bronchitis virus, IBV) and mammalian (mouse hepatitis virus, MHV; human coronavirus, HCV; transmissible gastroenteritis virus, TGEV) coronaviruses using the GAP program identified levels of amino acid sequence similarity (40·4–41·9%) and identity (15·7–17·2%) that were significantly lower than those evident among the coronavirus proteins (50·7–62·4% and 27·3–42·3%, respectively).

Of the putative functional motifs, including the papain-like Cys protease, chymotrypsin-like Ser protease (CSP), 3C-like Cys protease (3CLP) and growth factor/receptor-like domain, identified in nidovirus ORF1a polyproteins (Gorbalenya *et al.*, 1989; Boursnell *et al.*, 1987; Lee *et al.*, 1991; Herold *et al.*, 1993; Eleouet *et al.*, 1995; den Boon *et al.*, 1991), only a 3CLP homologue was readily identified in GAV. As in coronaviruses, the GAV 3CLP domain is encompassed by two highly hydrophobic regions located in similar positions in ORF1a (Fig. 2). Although the GAV 3CLP possesses a Cys catalytic nucleophile as in coronaviruses, alignment of upstream Thr and downstream His and Gly residues, important to formation of the substrate-binding pocket, suggests a structural similarity to arterivirus CSPs (Fig. 4A) (Godeny *et al.*, 1993; Snijder & Spaan, 1995; Snijder *et al.*, 1996). This was also supported by BLAST searches, in which the most significant alignments occurred with the related NIa Cys protease of tobacco etch virus (TEV) and other plant potyviruses (Carrington & Dougherty, 1987; Snijder *et al.*, 1996). The GAV ORF1ab polyprotein contains numerous Q(G/S/A) and E(G/S/A) dipeptides representing potential coronavirus 3CLP and arterivirus CSP cleavage sites, respectively. However, further studies are required to determine which of these or related sites are utilized.

ORF1b sequence

The hypothetical polypeptide comprising the ORF1b coding sequence has a deduced molecular mass of 302534 Da, a pI of 8·67, 16 potential *N*-linked glycosylation sites and no significant hydrophobic regions. The length of the GAV ORF1b (2646 aa) is more similar to the cognate ORFs of coronaviruses (2652–2733 aa) than that of the equine torovirus (ETV)-Berne (2291 aa) or arteriviruses (1410–1463 aa). Comparisons of ORF1b coding regions using the GAP program indicated comparable levels of sequence similarity (42·1–43·8%) between GAV and either coronaviruses (MHV, IBV, HCV and TGEV), ETV-Berne or the arterivirus EAV. These similarities were considerably below those among coronaviruses (69·9–83·8%) and similar to those between coronaviruses and either ETV-Berne (45·8–47·4%) or EAV (43·9–45·5%).

Visual scanning of the GAV ORF1b coding sequence and alignments using the GAP and COMPARE programs identified homologues of the conserved polymerase, metal ion binding (MIB), helicase, and the Motif 1 and Motif 3 (C-terminal) domains of nidoviruses (de Vries *et al.*, 1997; Snijder & Spaan, 1995; Lai & Cavanagh, 1997). Although the motifs are ordered similarly, the regions linking the polymerase–

Fig. 4. Clustal W 1.7 multiple sequence alignment of GAV functional motifs identified in (A) ORF1a (3CLP) and (B) ORF1b (polymerase, MIB, helicase, Motif 1 and C-terminal Motif 3). GAV sequences are aligned to avian (IBV-M42), murine (MHV-A59), human (HCV-229E) and porcine (TGEV) coronaviruses, to equine torovirus (ETV-Berne) and to equine arterivirus (EAV). (A) GAV is aligned to coronaviruses and EAV and (B) GAV and ETV-Berne are aligned to coronaviruses. Amino acid numbering (parentheses) is from the beginning of the ORF in which the motif resides. Highly significant amino acids representing potentially active residues are shown in bold. These include the absolutely conserved amino acids, including the SDD motif, in the eight domains (I–VIII) described for Supergroup 1 (+) RNA virus polymerases (Koonin, 1991), conserved Cys and His residues of the MIB domain and conserved amino acids in the purine NTP-binding motifs of the six helicase domains (I–VI) described by Hodgman (1988).

(A)

3C-like protease

```
EAV  (1078-1208)  FVAGSSYGT-GSV-WTRNNEVVVLTASHVVGR 19 NGDFAEAVTTQS 39 AWTTSGDSGSAVV-QGDAVV-----------GVH-TG-SNTSGVAY
                  :*    **  *.:    ..: : :.  *:..    : * :  :..:*      *..**.** :. : . **           *.* .* *  . *
GAV  (2855-2995)  IVRV--YGERGDLNGFLSGKSLHFPR-HIFDT 26 EYDATPFIKVES 40 ISTKDGDCGSIIFDHLGNVV-----------GAHIVGISCIPPVNG
                  :* *  *     ***:  .. :  **.*:: .     .   : .       *   * ***  :  :  .              . * .* .
IBV  (2796-2959)  IVSV-SYRG-NNLNGLWLGDTIYCPR-HVLGK 35 LKGAVLILQTAV 44 ASFLAGACGSVGFNIEKGVVNFFYMHHLELPNALH-TGTDL-MGEFY
MHV  (3350-3515)  IVSV-TYGN-MTLNGLWLDDKVYCPR-HVICS 35 MQGCQLVLTVTL 44 GSFLCGSCGSVGYVLTGDSVRFVYMHQLELSTGCH-TGTDF-SGNFY
HCV  (2982-3146)  VVRV-CYGN-TVLNGLWLGDIVYCPR-HVIAS 36 MHGVTLKIKVSQ 44 GSFINGACGSPGYNLKNGEVEFVYMHQIELGSGSH-VGSSF-DGVMY
TGEV (2895-3059)  IVRV-SYGN-NVLNGLWLGDEVICPR-HVIAS 34 YKGVNLVLKVNQ 44 GSFIAGTCGSVGYVLENGILYFVYMHHLELGNGSH-VGSNF-EGEMY
                  :* *  * .   ******.* : ***.**: .     :*  * : .      .**: *:*** *:  : . : *.***::**  ..* .*:.:  * :*
```

(B)

Polymerase

```
                        I                II                III               IVa              IVb               V
GAV  (646-928)    IPKISIQPVDKA-LRSIFIGPAFMNDVYRCFNTA 10 YNTVLIGFKDTHCG-INKLIN 13 WISQDYPKFDTCVDTMAQ 20 RGLCQL 12 SILIRKLHGVSSGDGATAIKNSHCN 53
                    * :*.   *   *::   . : . . * ::       .*:** .. : *  : * .     :. **.* * .: .: :     * :     .   *  *.:***.:**  **  *
IBV  (560-855)    NLKYAIS--AKNRARTVAGVSILSTMTNRQFHQK  9 NASVVIGTTKFYGGWDNMLRN  9 LMGWDYPKCDRAMPNLLR 21 YRLYNE 12 GGIYVKPGGTSSGDATTAYANSVFN 57
MHV  (550-845)    NLKYAIS--AKNRARTVAGVSILSTMTGRMFHQK  9 GVPVVIGTTKFYGGWDDMLRR  9 LMGWDYTKCDRAMPNILR 21 YRLANE 12 GCYYVKPGGTSSGDATTAFANSVFN 57
HCV  (534-829)    NLKYAIS--GKERARTVGGVSLLATMTTRQFHQK  9 NATVVIGTTKFYGGWDNMLKN  9 LMGWDYPKCDSAMPSMIR 21 YSLSNE 12 GGFYFKPGGTTSGDATTAYANSVFN 57
TGEV (535-830)    NLKYAIS--GKARARTVGGVSLLSTMTTRQYHQK  9 NATVVIGSTKFYGGWDNMLKN  9 LMGWDYPKCDRALPNMIR 21 YRLSNE 12 GGFYFKPGGTTSGDGTTAYANSAFN 57
                    *:*:*  .* *****.. *::::   *  *:     .   :** :*:   :..:*:    ::* **.*** ::*  :*    * * **    *    *****:***.***.:*: :*
ETV  (467-744)    ITKFALS--AKARARTVSSCSFIASTIFRFAHKP 12 GGFCLIGVSKYGLKFSKFLKD  9 VFGSDYTKCDRTFPLSFR 19 Y-L-NE 12 GMLLNKPGGTSSGDATTAHSNTFYN 50
```

Polymerase (cont.)

```
                        VI              VII          VIII
GAV               SRCATLSDDTLAIIN 31 EPYEFTSRY  8 YNAPLIERVF
                       **** :.  *     *:** *::       *  .*::
IBV               FSLMILSDDGVVCYN 40 GPHEFCSQH 12 LPYPDPSRIL
MHV               FSMMILSDDGVVCYN 40 GPHEFCSQH 12 LPYPDPSRIL
HCV               FSMMILSDDSVVCYN 40 GPHEFCSQH 12 LPYPDPSRII
TGEV              FSMMILSDDGVVCYN 40 GPHEFCSQH 12 LPYPDPSRIL
                  : : :****  . :.      .**** *    **    .*::
ETV               YFLNFLSDD-SFIFS 35 HIEEFCSAH 10 LPS--RGRLL
```

Motif 3 (C-terminal)

```
GAV  (2528-2600)  TIISDLY 18 LKEHGHFLFKVTNKF 17 IYKAPTNTVTSELWVA
                    ::**:*    *   * . .*:* :      : : .*: :** ::
IBV  (2532-2617)  LVISDMY 32 LALGGSFAVKVTETS 17 FCTA-VNASSSEAFLI
MHV  (2560-2641)  LIISDMY 28 LALGGSVAIKITEFS 17 FCTN-VNASSSEGFLI
HCV  (2511-2592)  LLISDMY 28 LAIGGSIAIKVTEYS 17 FCTS-VNTSSSEAFVV
TGEV (2506-2587)  LLVSDLY 28 LSLGGSVAIKITEFS 17 FCTS-VNTSSSEGFLI
                  *::**:*    *::**:...* ** *    * *  **:**** *::
ETV  (2139-2208)  LIVSDIY 16 LALGGTIVFKTTESS 17 FFTAGVNTSSSEVFVV
```

Metal ion binding (MIB) domain

```
GAV  (1215-1294) ICFACGGHGQFECYTCASNGYPFTFCNGSTCLRTHFTECEHYSYSLASRKRYIIQCA--DCEEMDIRKMYVN-NDFRCARHLH
                  *..*..:    *  *    *.  *   *.  .*      .   *:  *:  .*   *: *::::.  .: *  *
IBV  (953-1026)  VCVVCNSQTILRCGNCIR--KPF-LCC--KCCYDHVMHTDHKNVLSINP--YI--CSQLGCGEADVTKLYLGGMSYFCGNHKP
MHV  (943-1016)  ACVVCSSQTSLRCGSCIR--KPL-LCC--KCAYDHVMSTDHKYVLSVSP--YV--CNSPGCDVNDVTKLYLGGMSYYCEDHKP
HCV  (927-1000)  LCVVCGSQTVLSCGDCLR--SPM-LCT--KCAYDHVFGTDHKFILAITP--YV--CNTSGCNVNDVTKLYLGGLNYYCVDHKP
TGEV (928-1001)  MCVVCGSQTVLRCGDCLR--RPL-LCT--KCAYDHVMGTKHKFIMSITP--YV--CSFNGCNVNDVTKLFLGGLSYYCMNHKP
                  *. * . :  * .*    ** **   * * **. :.*. :      :  *   *.  ::::*::   .  * :
ETV  (850-918)   VCFCCPNPAVSVCEECYV---PLPLCA--YCYYVHVVISNHSKVEDK----FK--CF---CGQDNISELYMVLNNSICMYQCK
```

Helicase

```
                        Ia          Ib              II                 III              IV              V              VI
GAV  (1488-1778)  IQGPPGTGKTYT 20 ASSHSAVD 40 FATLQSTRGILCPSVEYMIIDEFSMA 16 IIFTGDPCQLSTQHVY 22 FLDITYRMGSKIN 68 DIQIPIYTVHTSQGR 18 PNIVNVAVSRFRFQCIC
                  :*****:**::     *.**:***    * *:::          : :::**.**      :: *** **.:  .    **   **   :*      : :   **.::**      * .***::* :   :
IBV  (1229-1525)  VQGPPGSGKSHF 16 ACSHAAVD 41 FSTINALPEVSC---DILLVDEVSML 16 VVYVGDPAQLPAPRTL 24 FLAKCYRCPKEIV 78 MLGLNVQTVDSSQGS 18 INRFNVALTRAKRGILV
MHV  (1218-1512)  VQGPPGTGKSHL 16 AASHAAVD 41 FTTINALPELVT---DIIVVDEVSML 16 YVYIGDPAQLPAPRVL 25 FLGTCYRCPKEIV 75 VLGLQTQTVDSAQGS 18 VNRFNVAITRAKKGILC
HCV  (1204-1498)  MQGPPGSGKSHC 16 ACSHAAVD 41 FSTVNALPEVNA---DIVVVDEVSMC 16 IVYVGDPQQLPAPSVL 25 FLHKCYRCPAEIV 75 LLGLQTQTVDSAQGS 18 ANRFNVAITRAKKGIFC
TGEV (1205-1499)  IQGPPGSGKSHC 16 ACSHAAVD 41 FCTVNALPEASC---DIVVVDEVSMC 16 IVYVGDPQQLPAPRTL 25 FLHKCYRCPAEIV 75 LLGLQTQTVDSAQGS 18 VNRFNVAITRAKVGILC
                  : ****:**:      * :*  *.    : * *:**        :::.****:      * :*** ** :* .     :*  ***** :*:     ** :. *:**:**:     **. *. :*:.   :
ETV  (1099-1372)  VMGPPGTGKTTF 18 APTHRLVG 34 LCTHNTLPFIKS---AVLIADEVSLI 15 VVLLGDPFQL-SP-VY 22 YLTACYRCPPQIL 67 GLG-DVTTIDSSQGT 18 VNRVIVGCSRST-THLV
```

Motif 1

```
GAV  (2086-2199)  RNGPNKCHVDDCPNHP-IWYTEHPCCAIHIDGVIQAFYNPYMVDIHGLTCPRCTSDEFTY-SQT------EDSIN---CHRCYPPTNTSFNTR---AKLQIMHDAVCPDKHGAAHDSGADAAMTLCIF
                  *    :   *   . .       . . .      ::          *  * ..  .: *        ... .    .  * *   .::     ..*   *.  *   *. .* :. ** ** *:
IBV  (1713-1828)  R--IVQMLADNLCNVSDCVVFVTWCHGLELT--TLRYFVKIGKEQV---CS-CGSRATTFNSHTQAYACWKHCLG---FDFVYNPLLVDIQQWGYSGNLQFNHDLHCN-VHGHAHVASVDAIMTRCLA
MHV  (1703-1819)  R--IVQMLSDHLVDLADSVVLVTWAASFELT--CLRYFAKVGREVV---CSVCTKRATCFHSRTGYYGCWRHSYS---CDYLYNPLIVDIQQWGYTGSLTSNHDPICS-VHKGAHVASSDAIMTRCLA
HCV  (1686-1800)  R--IVQMIADFLAGSSDVLVFVLWAGGLELT--TMRYFVKIGAVKH---CQ-CGTVATCYNSVSNDYCCFKHALG---CDYVYNPYVIDIQQWGYVGSLSTNHHAICN-VHRNEHVASGDAIMTRCLA
TGEV (1688-1803)  R--IVQMVCDYFDGLSDILIFVLWAGGLELT--TMRYFVKIGRPQK---CE-CGKSATCYSSSQSVYACFKHALG---CDYLYNPYCIDIQQWGYTGSLSMNHHEVCN-IHRNEHVASGDAIMTRCLA
                  *  :.::: *   . .    :: :. . :*.   ::::.:..       *. *  *         * * . . .    . : *   :*:*:     .*   *.  *. .*   * *  **:**:**
ETV  (1496-1608)  R--LFKFIIDTTFGQP---WFILYSCSNDLK--SLKFYVEFDTCYF---CS-CGEMAIVL-MRDGNYKCRN-CYGGMLISKLVNCKYLDVQKE--RVKLQDAHDAICQQFHGDSHEALCDAVMTKCLY
```

Fig. 4. For legend see facing page.

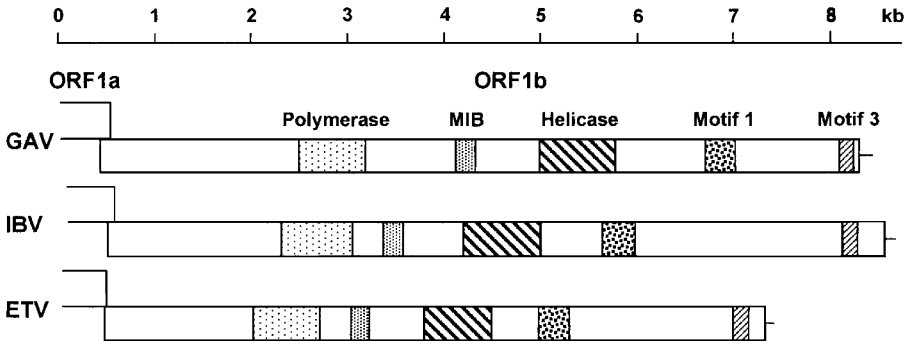

Fig. 5. Schematic diagram showing the relative positions of functional motifs in the ORF1b coding regions of GAV, avian coronavirus IBV and equine torovirus ETV-Berne. These include the 'SDD' polymerase, MIB domain, helicase and Motif 1 and C-terminal Motif 3 domains (de Vries *et al*., 1997).

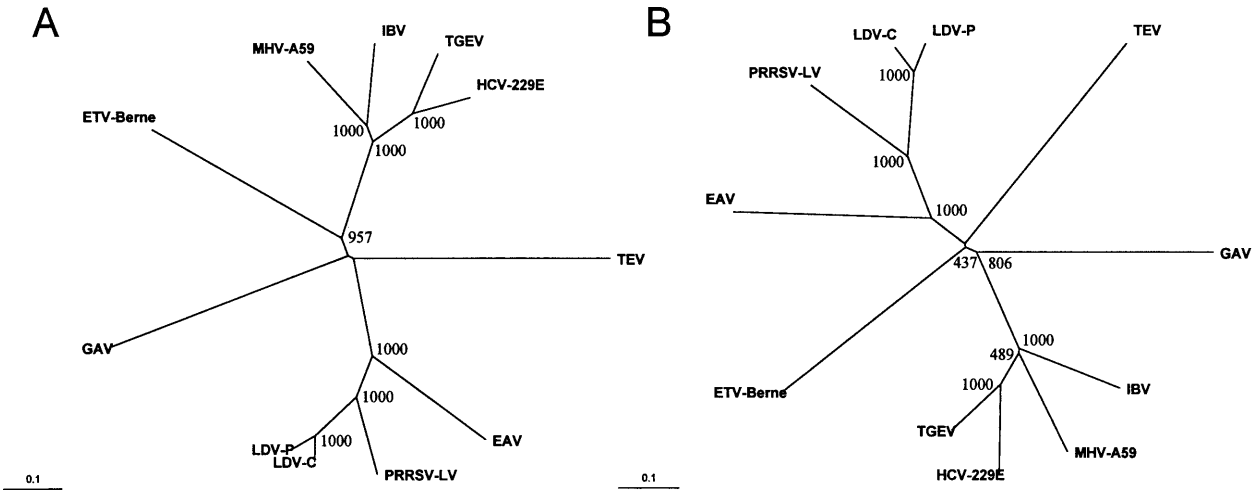

Fig. 6. Phylogenetic relationships of the (A) polymerase and (B) helicase domains of ORF1ab polyproteins of GAV, coronaviruses MHV-A59, IBV, HCV-229E and TGEV, the torovirus ETV-Berne and arteriviruses EAV, LDV-C and -P strains and PRRSV-LV strain in addition to the cognate domains of the plant potyvirus TEV. Unrooted, bootstrapped trees were generated from Clustal W multiple alignments using the neighbour-joining method.

MIB and helicase–Motif 1 domains are considerably longer than in corona- or toroviruses (Fig. 5) (Boursnell *et al*., 1987; Bredenbeek *et al*., 1990; Snijder *et al*., 1990*a*). Sequence relationships of the GAV domains to those of coronaviruses (IBV, MHV-A59, HCV-229E, TGEV) and ETV-Berne were determined using Clustal W multiple alignments (Fig. 4B). Similar comparisons of the ORF1b domains of corona- and torovirus with those of arteriviruses are described elsewhere for the arteriviruses equine arteritis virus (EAV) (den Boon *et al*., 1991), lactate dehydrogenase elevating virus (LDV) (Godeny *et al*., 1993) and porcine reproductive and respiratory syndrome virus (PRRSV) (Meulenberg *et al*., 1993). The GAV polymerase core contains homologies in the eight domains described for (+)-RNA viruses and all absolutely conserved amino acids in Supergroup I viruses are preserved (Koonin, 1991). The most striking feature is the presence of the SDD, rather than GDD, motif in the polymerase core which is unique to nidoviruses.

ORF1b also contains 3 MIB or 'zinc-finger' motifs defined by conservation of potentially active Cys and His residues and the positioning of aromatic residues similarly to TFIIIA-like fingers (Fig. 4B) (Gorbalenya *et al*., 1989). The potential structure formed by the GAV MIB domain conforms with the general model proposed for numerous RNA and DNA binding proteins (data not shown) (Gorbalenya *et al*., 1989). However, the regions linking the binding domains are generally a little longer and bare little resemblance to those of corona- or toroviruses (Lee *et al*., 1991; Herold *et al*., 1993; Snijder *et al*., 1990*a*).

The ORF1b helicase domain contains the characteristic purine NTP-binding motifs A (GppGtGKT/S) and B (DE) believed to be associated with dsRNA duplex unwinding during RNA replication (Hodgman, 1988; Gorbalenya & Koonin, 1989) (Fig. 4B). At the terminal position of motif A, GAV possesses a Thr residue characteristic of toro- and arteriviruses rather than a Ser residue conserved among coronaviruses (Lee *et al*., 1991; Snijder *et al*., 1990*a*; den Boon *et al*., 1991; Godeny *et al*., 1993).

In addition to the three domains with predicted functions, ORF1b contains a region (aa 2086–2199) homologous to the

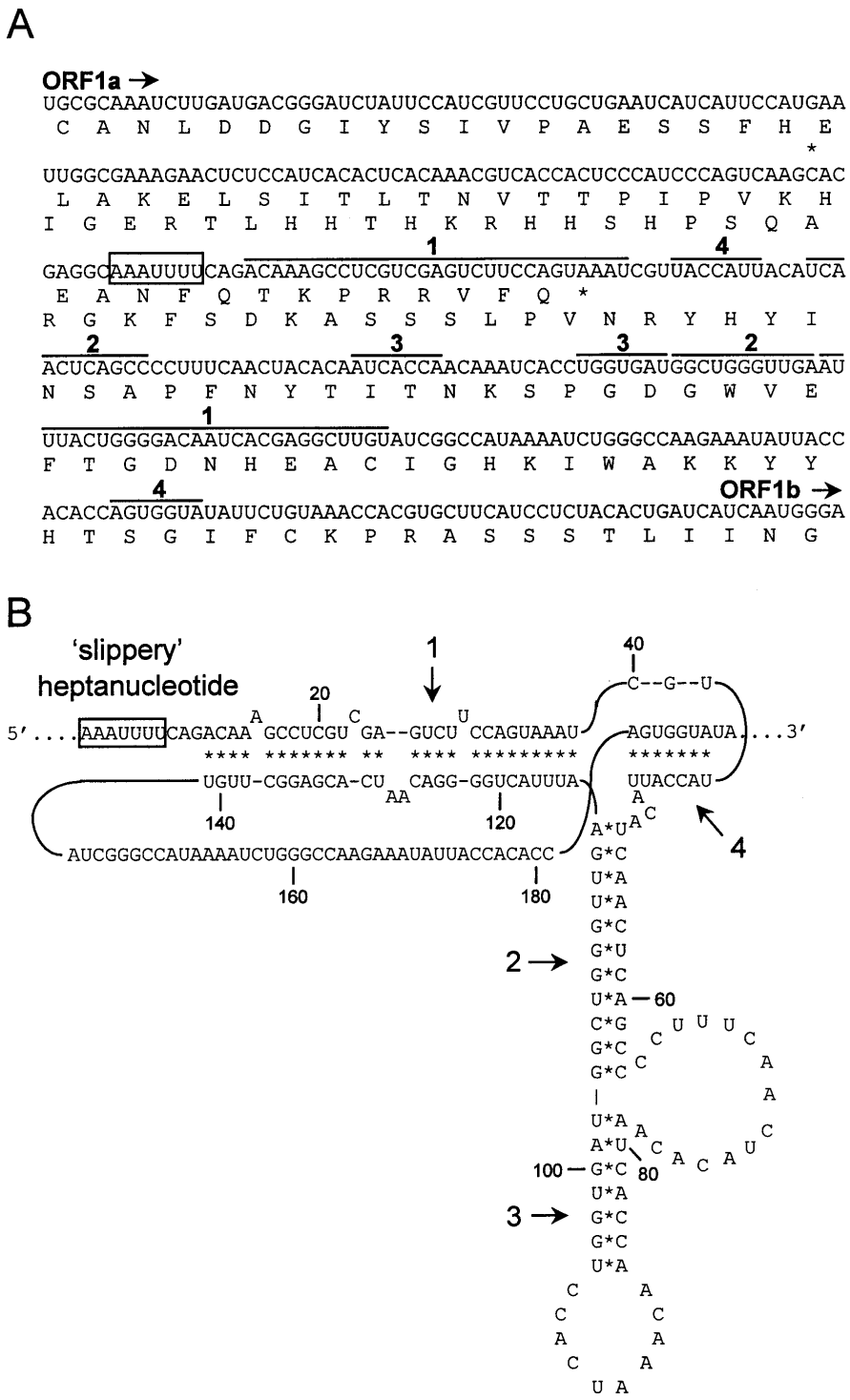

Fig. 7. (A) RNA and deduced amino acid sequence in the region where ORF1a and ORF1b of GAV overlap by 33 codons. The putative 'slippery' heptanucleotide AAAUUUU of the −1 frameshift site is boxed and sequences potentially forming four helices of an RNA pseudoknot are overlined and numbered. (B) RNA folding structure predicted in the GAV ORF1a/1b overlap using MFOLD (Zuker, 1989). A 131 nt stem–loop containing two ssRNA bulges is predicted 3 nt downstream of the AAAUUUU sequence (boxed). A 7 nt sequence with potential to base pair with a ssRNA bulge (helix 4) to form a complex pseudoknot occurs 40 nt downstream of the stem–loop (helices 1–3).

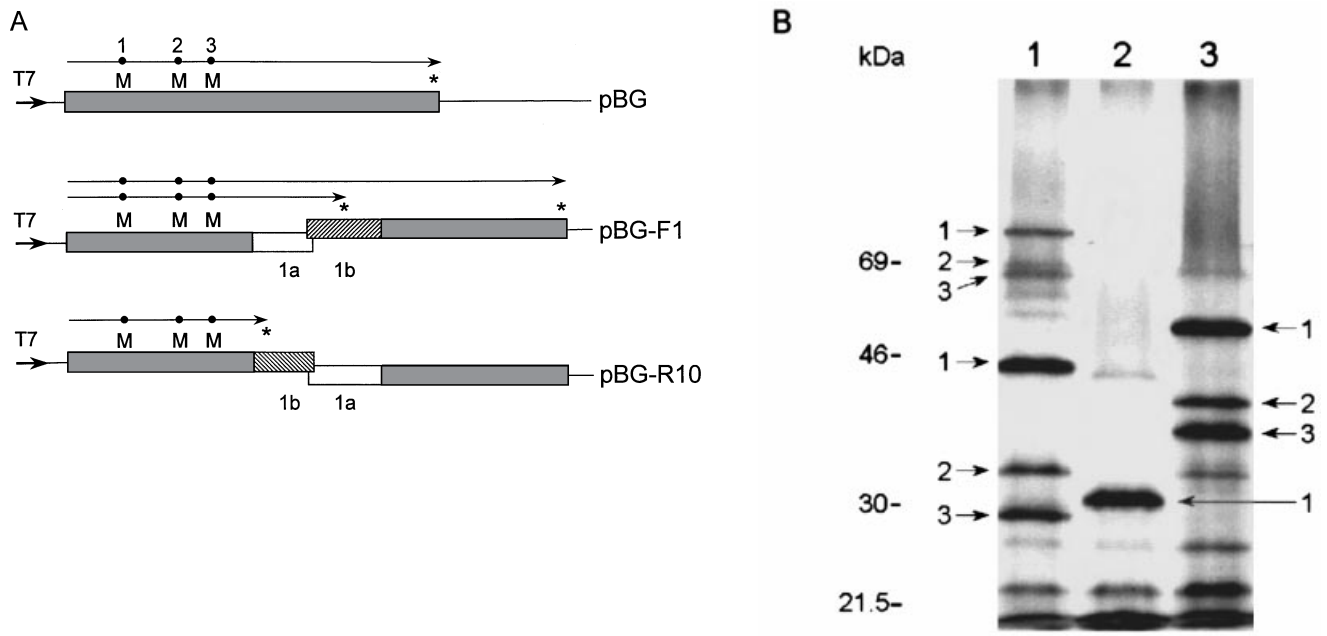

Fig. 8. (A) Strategy for the insertion of a 693 nt sequence containing the GAV ORF1a/1b overlap into a plasmid (pBG) containing the BEFV G protein gene downstream of a T7 RNA promoter. The relative positions of three potential internal translation initiation (AUG) sites (●) and the termination sites (*) in each construct are indicated. (B) SDS–PAGE and fluorography of [^{35}S]methionine-labelled proteins expressed in a coupled T7 transcription/translation system from plasmids pBG-F1 (lane 1), pBG-R10 (lane 2) and pBG (lane 3). Polypeptides initiated at putative internal AUG sites at positions 1, 2 and 3 are indicated.

Motif 1 domain (Fig. 4B) identified in corona- and toroviruses, but not arteriviruses (de Vries *et al.*, 1997). Sequence identity, however, is restricted primarily to conserved Cys and His residues in the C-terminal half of Motif 1. The near C terminus of GAV ORF1b (aa 2528–2600) also displays limited homology to the C-terminal region of Motif 3 of corona- and toroviruses (Snijder *et al.*, 1990*a*; den Boon *et al.*, 1991; Godeny *et al.*, 1993; de Vries *et al.*, 1997)

The phylogenetic relationship of GAV to nidoviruses was determined for the polymerase and helicase domains described in Fig. 4(B). Previous analyses of coronavirus and torovirus polymerases indicate they are distinct from other (+)-RNA virus polymerases (Koonin, 1991). Phylogenetic comparisons based on Clustal W alignments and the neighbour-joining or maximum parsimony methods generated similar phenograms which placed the GAV polymerase on a branch distinct from the corona-, toro- and arteriviruses, which clustered into the expected lineages (Godeny *et al.*, 1993) (Fig. 6A). Indeed, GAV was almost as distantly related to nidoviruses as the potyvirus TEV included as a representative of Supergroup I virus polymerases (Koonin, 1991). Phylogenetic comparisons of the helicase clustered corona-, toro- and arteriviruses, although less tightly, into the same general relationships (Fig. 6B). In this case, however, GAV was placed on a branch marginally more closely related to coronaviruses than was the torovirus ETV-Berne.

Overlapping ORFs and putative −1 ribosomal frameshift site

The C terminus of ORF1a of GAV overlaps the N terminus of ORF1b by 33 codons (Fig. 7A). In nidoviruses, similar ORF1a/1b overlaps contain the conserved 'slippery' sequence G/UUUAAAC and 'H-type' pseudoknots that mediate a −1 ribosomal frameshift leading to the generation of an ORF1ab polyprotein (Boursnell *et al.*, 1987; Brierley *et al.*, 1989; Snijder *et al.*, 1990*a*; Bredenbeek *et al.*, 1990; Lee *et al.*, 1991; den Boon *et al.*, 1991; Godeny *et al.*, 1993; Meulenberg *et al.*, 1993). Although this motif is not present in the GAV overlap, we identified another potential 'slippery' sequence (AAAUUUU) found to facilitate −1 ribosomal frameshifts in the translation of retrovirus RNAs (Sonigo *et al.*, 1986; Jacks *et al.*, 1988; ten Dam *et al.*, 1990; Brierley, 1995; Brierley *et al.*, 1992). By the mechanism proposed by Jacks *et al.* (1988), slippage would be predicted at the Asn and Phe codons HEANFQTK of ORF1a resulting in the read-through sequence HEANF/SDK at the ORF1a/1b junction. Translational read-through at this position would generate an ORF1ab polyprotein comprising 6673 aa with a deduced molecular mass of 758812 Da.

MFOLD and PLOTFOLD analysis of the GAV ORF1a/1b overlap predicted a highly stable, 131 nt RNA stem–loop (ΔG = −46·1 kcal/mol) 3 nt downstream of the AAAUUUU sequence (Fig. 7B). Visual scanning also identified a 7 nt region

(5′ UACCAUU 3′) in an ssRNA bulge of the stem–loop with potential to base pair with a sequence 40 nt downstream of the stem base. The complex structure predicted for the putative pseudoknot is characteristic of the −1 ribosomal frameshift sites employed by some retroviruses (Jacks *et al.*, 1988; Brierley *et al.*, 1989).

Function analysis of the −1 ribosomal frameshift site

Functional activity of the GAV −1 ribosomal frameshift site was examined *in vitro* by expression in a coupled transcription/translation system. A 693 nt sequence encompassing the ORF1a/1b overlap was inserted into the BEFV G protein gene (Walker *et al.*, 1992) such that ORF1a and ORF1b were in-frame with the N- and C-terminal portions of the G ORF, respectively (Fig. 8 A). SDS–PAGE identified three major polypeptides (52·5, 41 and 37·5 kDa) expressed from the BEFV G-gene plasmid (Fig. 8 B, lane 3). The size of these proteins is consistent with their translation from the first three in-frame AUG codons downstream of the G protein initiation codon. The reason for the translational bypass of the N-terminal AUG is not known. Consistent with this pattern of initiation, the plasmid with the GAV frameshift site in the reverse orientation translated only a truncated N-terminal portion (30·5 kDa) of the G protein (Fig. 8 B, lane 2) whilst that with the in-frame insertion translated three major (46, 33·5 and 29·5 kDa) and three minor larger polypeptides (81, 69·5 and 66 kDa) (Fig. 8 B, lane 1). The sizes of the major polypeptides suggest they terminated at the GAV ORF1a stop codon. The sizes of the larger proteins indicate they were translated by read-through of the ORF1a/1b frameshift site, linking the N- and C-terminal ORFs of the BEFV G protein. Radioactivity in the terminated and extended polypeptides was quantified using a phosphorimager and normalized for the number of Met residues. This indicated that translational read-through at the GAV ORF1a/1b frameshift site occurred at 23–25% efficiency.

Discussion

In this paper, we describe a 20089 nt sequence from the 5′ end of the (+)-ssRNA genome of GAV, a yellow head-like virus of *Penaeus monodon* prawns (Spann *et al.*, 1997). The sequence was derived from cDNA clones generated from a cellular ~22 kb dsRNA likely to represent a replicative intermediate of GAV genomic RNA. The region comprises a single gene containing two long overlapping ORFs with motifs which indicate a functional equivalence to the ORF1a and ORF1ab polyproteins of nidoviruses (Lai & Cavanagh, 1997; Snijder & Spaan, 1995; de Vries *et al.*, 1997). GAV ORF1a (4060 aa) is similar in size to coronaviruses (3951–4488 aa) and ORF1b (2646 aa) is also closer in size to coronaviruses (2652–2733 aa) than to the torovirus ETV-Berne (2291 aa).

ORF1a of GAV contains a 3C-like Cys protease motif likely to be involved in processing the large (759 kDa) ORF1ab polyprotein into functional units. Interestingly, BLAST searches gave the closest match with the NIa proteases of TEV and other plant potyviruses that have a QG/S cleavage specificity similar to coronavirus 3CLPs (Carrington & Dougherty, 1987). Although the GAV 3CLP contains a Cys as its 'catalytic nucleophile', alignment of other residues important to substrate binding and cleavage suggests a structural similarity to arterivirus chymotrypsin-like Ser proteases (Snijder *et al.*, 1996). However, the GAV 3CLP is distantly related and cleavage sites in the ORF1ab polyprotein will need to be identified to determine whether its site specificity resembles that of the equivalent coronavirus or arterivirus proteases. Its unique sequence, however, may provide definitive evidence of their common ancestry.

The ORF1b coding region of GAV contains functional motifs associated with RNA synthesis, including a polymerase, a metal ion binding domain and a helicase. It also contains distantly related homologues of the Motif 1 and 3 domains that have undefined functions and are present in toro- and coronaviruses but not arteriviruses (Lai & Cavanagh, 1997; Snijder & Spaan, 1995; de Vries *et al.*, 1997). Although ordered similarly, the spatial distribution of the ORF1b domains is distinct from corona- and toroviruses and no significant similarity exists in the intervening sequences. Pairwise comparisons of the ORF1b sequence using the GAP and COMPARE programs indicate that coronaviruses and the torovirus ETV-Berne are more closely related to each other than to GAV. Moreover, phylogenetic analyses using the polymerase and helicase domains suggest that corona-, toro- and arteriviruses have diverged from GAV to similar extents. Nevertheless, its ancestral relationship to nidoviruses is clearly established by the presence of the 'SDD' polymerase motif unique to these (+)-ssRNA viruses (Koonin, 1991; Snijder & Spaan, 1995). Considering that GAV replicates in penaeid shrimp, the low levels of homology in the conserved functional components of the replicase polyproteins of the vertebrate nidoviruses are not unexpected. Indeed, as marine invertebrates were abundant prior to the evolution of higher plants and animals, GAV may be a relict of an ancient progenitor nidovirus.

A prerequisite for translation of nidovirus ORF1b coding regions is the presence of a −1 ribosomal frameshift site at the ORF1a/1b overlap (Spaan *et al.*, 1988; Lai, 1990; Lai & Cavanagh, 1997; de Vries *et al.*, 1997). In GAV, the ORF1a/1b overlap contains a recognized 'slippery' heptanucleotide, AAAUUUU, and a stable but complex RNA pseudoknot 3 nt downstream. Although unlike the G/UUUAAAC heptanucleotides and 'H-type' pseudoknots characteristic of the −1 frameshift elements of vertebrate nidoviruses (Brierley *et al.*, 1989; Pleij & Bosch, 1989; ten Dam *et al.*, 1990; Snijder *et al.*, 1990*a*; Bredenbeek *et al.*, 1990; Lee *et al.*, 1991), the GAV frameshift site promotes translational read-through of the

ORF1a/1b overlap at an efficiency (23–25 %) comparable to these viruses (Brierley *et al.*, 1989, 1992). Interestingly, a frameshift site similar to that of GAV, utilizing an AAAUUUU sequence and complex pseudoknot formed by base-pairing of a downstream sequence with a bulge region of a long stem–loop, facilitates a translational frameshift at the *gag/pol* junction of the retrovirus Rous sarcoma virus (Jacks *et al.*, 1988; Brierley *et al.*, 1989). Although diverse heptanucleotides and accompanying RNA folding structures can promote −1 ribosomal frameshifting, those used by vertebrate nidoviruses and GAV are among the most efficient yet identified (Brierley *et al.*, 1992). It is apparent that the requirement for efficient translation of the GAV ORF1ab replicase polyprotein is fulfilled by a frameshift site quite distinct from those of vertebrate nidoviruses.

Sequencing of 5′-RACE clones suggests that the 5′ end of the GAV genome initiates at a 5′ AC sequence 68 nt upstream of the putative ORF1a start codon. The short length of the putative 5′ untranslated sequence contrasts with the relatively long sequences present in the torovirus ETV-Berne (> 700 nt) (Snijder *et al.*, 1991) and in the coronaviruses IBV, HCV and MHV (209–528 nt) (Boursnell *et al.*, 1987; Shieh *et al.*, 1987; Herold *et al.*, 1993; Bonilla *et al.*, 1994). For coronaviruses, the 5′ end of their long intracellular mRNA 1 is equivalent to that of the genomic (+)-ssRNA packaged in virions. Although our data suggest that this is also the case with GAV, we have identified subgenomic (sg) mRNAs that initiate at common 5′ AC positions in highly conserved regions of intergenic sequences downstream of ORF1ab (J. A. Cowley and others, unpublished). Thus, we cannot as yet discount the possibility that the RNA 5′ end identified by 5′-RACE is a product of transcription initiation at a similar internal sequence upstream of ORF1a. Further characterization of the 5′ termini of the putative genomic and sgmRNAs will be required to determine whether GAV utilizes a transcription mechanism like that suggested for the torovirus ETV-Berne (Snijder *et al.*, 1990*b*) or a leader-fusion mechanism like that used by coronaviruses (Lai *et al.*, 1984; Spaan *et al.*, 1988; Lai, 1990) and arteriviruses (de Vries *et al.*, 1990).

We have recently shown that regions of ORF1b coding sequences of GAV from Australia and YHV from Thailand share 81–85 % amino acid and 87–96 % nucleotide identity, suggesting that they represent geographical topotypes (Cowley *et al.*, 1999). The viruses are morphologically indistinguishable. Their tubular helical nucleocapsids (16–18 nm diameter) vary considerably in length (166–435 nm) and their rod-shaped virions (35–50 × 150–200 nm) possess ~ 11 nm surface projections and are quite flexible, often forming pleomorphic bent or doughnut-shaped structures. Although longer, there is a conspicuous resemblance to the rod forms of the highly pleomorphic torovirions (Weiss *et al.*, 1983) but not to the spherical coronavirions (Lai & Cavanagh, 1997). However, in common with these viruses, GAV and YHV virions mature by nucleocapsid budding primarily at endoplasmic reticulum and Golgi membranes, often leading to the formation of virion arrays in cytoplasmic vesicles (Boonyaratpalin *et al.*, 1993; Chantanachookin *et al.*, 1993; Spann *et al.*, 1995, 1997).

Based on its rod-shaped morphology and data indicating an ssRNA genome, YHV was suggested to be a rhabdovirus or coronavirus (Wongteerasupaya *et al.*, 1995). Evidence that YHV contains four major structural proteins (170, 135, 67 and 22 kDa) and that its ~ 22 kb genomic ssRNA may have (−) polarity led to its provisional classification as a rhabdovirus (Nadala *et al.*, 1997; Loh *et al.*, 1997). More recent strand-specific hybridization analyses, however, suggest that YHV possesses a (+)-ssRNA genome (Tang & Lightner, 1999). The molecular data presented here for the Australian yellow head-like virus GAV unequivocally support a relationship, albeit distant, to the (+)-ssRNA nidoviruses. This evidence includes: (i) cloning of viral RNA from an extremely large (~ 22 kb) intracellular dsRNA likely to represent a genomic replicative intermediate; (ii) an ORF1a polyprotein containing a 3C-like Cys protease; (iii) an ORF1b coding sequence with replicase functions including an 'SDD' polymerase, a metal ion binding domain, a helicase domain and other motifs conserved in corona- and toroviruses; and (iv) an efficient −1 ribosomal frameshift site at the ORF1a/1b overlap that will facilitate translation of a 759 kDa ORF1ab polyprotein. As such, the penaeid shrimp virus GAV and the closely related YHV (Cowley *et al.*, 1999) appear to be new members of the *Nidovirales* and the first to be isolated from an invertebrate. Additional data on the nature and organization of the structural genes and on the replication strategy employed by these viruses will be needed to resolve the issue of their taxonomic classification within the Order. As GAV and YHV replicate primarily in the prawn lymphoid (or 'Oka') organ, and as a defining clinical feature of YHV infection is a yellow (or ochre; meaning pale brownish-yellow colour) discoloration of the prawn cephalothorax, we suggest the name 'Okavirus' to describe these viruses.

We thank Gold Coast Marine Aquaculture for supplying healthy prawns and Dr Ross Tellam for useful discussions on protein structure analysis.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Bonilla, P. J., Gorbalenya, A. E. & Weiss, S. R. (1994). Mouse hepatitis virus strain A59 RNA polymerase gene ORF 1a: heterogeneity among MHV strains. *Virology* **198**, 736–740.

Boonyaratpalin, S., Supamattaya, K., Kasornchandra, J., Direkbusaracom, S., Aekpanithanpong, U. & Chantanachookin, C. (1993). Non-occluded baculo-like virus, the causative agent of yellow-head disease in the black tiger shrimp (*Penaeus monodon*). *Fish Pathology* **28**, 103–109.

Boursnell, M. E. G., Brown, T. D. K., Foulds, I. J., Green, P. F., Tomley, F. M. & Binns, M. M. (1987). Completion of the sequence of the coronavirus avian infectious bronchitis virus. *Journal of General Virology* **68**, 57–77.

Bredenbeek, P. J., Pachuk, C. J., Noten, A. F. H., Charite, J., Luytjes, W., Weiss, S. R. & Spaan, W. J. M. (1990). The primary structure and expression of the second open reading frame of the polymerase gene of the coronavirus MHV-A59; a highly conserved polymerase is expressed by an efficient ribosomal frame shifting mechanism. *Nucleic Acids Research* **18**, 1825–1832.

Brierley, I. (1995). Ribosomal frameshifting on viral RNAs. *Journal of General Virology* **76**, 1885–1892.

Brierley, I., Digard, P. & Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell* **57**, 537–547.

Brierley, I., Jenner, A. J. & Inglis, S. C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *Journal of Molecular Biology* **227**, 463–479.

Carrington, J. C. & Dougherty, W. G. (1987). Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. *Journal of Virology* **61**, 2540–2548.

Chantanachookin, C., Boonyaratpalin, S., Kasornchandra, J., Sataporn, D., Ekpanithanpong, U., Supamataya, K., Riurairatana, S. & Flegel, T. W. (1993). Histology and ultrastructure reveal a new granulosis-like virus in *Penaeus monodon* affected by yellow-head disease. *Diseases of Aquatic Organisms* **17**, 145–157.

Cowley, J. A., Dimmock, C. M., Wongteerasupaya, C., Boonsaeng, V., Panyim, S. & Walker, P. J. (1999). Yellow head virus from Thailand and gill-associated virus from Australia are closely related but distinct prawn viruses. *Diseases of Aquatic Organisms* **36**, 153–157.

Cowley, J. A., Dimmock, C. M., Spann, K. M. & Walker, P. J. (2000). Detection of Australian gill-associated virus (GAV) and lymphoid organ virus (LOV) of *Penaeus monodon* by RT-nested PCR. *Diseases of Aquatic Organisms* (in press).

den Boon, J. A., Snijder, E. J., Chirnside, E. D., de Vries, A. A. F., Horzinek, M. C. & Spaan, W. J. M. (1991). Equine arterivirus is not a togavirus but belongs to the coronavirus-like superfamily. *Journal of Virology* **65**, 2910–2920.

de Vries, A. A. F., Chirnside, E. D., Bredenbeek, P. J, Gravestein, L. A., Horzinek, M. C. & Spaan, W. J. M. (1990). All subgenomic mRNAs of equine arteritis virus contain a common leader sequence. *Nucleic Acids Research* **18**, 3241–3247.

de Vries, A. A. F., Horzinek, M. C., Rottier, P. J. M. & de Groot, R. J. (1997). The genome organization of the Nidovirales: similarities and differences between arteri-, toro-, and coronaviruses. *Seminars in Virology* **8**, 33–47.

Dumas, J. B., Edwards, M., Delort, J. & Mallet, J. (1991). Oligodeoxynucleotide ligation of single-stranded cDNAs: a new tool for cloning 5′ ends of mRNAs and for constructing cDNA libraries by *in vitro* amplification. *Nucleic Acids Research* **19**, 5227–5232.

Eleouet, J.-F., Rasschaert, D., Lambert, P., Levy, L., Verde, P. & Laude, H. (1995). Complete sequence (20 kilobases) of the polyprotein-encoding gene 1 of transmissible gastroenteritis virus. *Virology* **206**, 817–822.

Felsenstein, J. (1993). PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle, WA, USA.

Froussard, P. (1992). A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Research* **20**, 2900.

Godeny, E. K., Chen, L., Kumar, S. N., Methven, S. L., Koonin, E. V. & Brinton, M. A. (1993). Complete genomic sequence and phylogenetic analysis of the lactate dehydrogenase elevating virus (LDV). *Virology* **194**, 585–596.

Gorbalenya, A. E. & Koonin, E. V. (1989). Viral proteins containing the purine NTP-binding sequence pattern. *Nucleic Acids Research* **17**, 8413–8440.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. (1989). Coronavirus genome: predictive functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. *Nucleic Acids Research* **17**, 4847–4861.

Herold, J., Raabe, T., Schelle-Prinz, B. & Siddell, S. G. (1993). Nucleotide sequence of the human coronavirus 229E RNA polymerase locus. *Virology* **195**, 680–691.

Hodgman, T. C. (1988). A new superfamily of replicative proteins. *Nature* **333**, 22–23.

Jacks, T., Madhani, H. D., Masiarz, F. R. & Varmus, H. E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus *gag–pol* region. *Cell* **55**, 447–458.

Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *Journal of General Virology* **72**, 2197–2206.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292.

Lai, M. M. C. (1990). Coronavirus: organization, replication and expression of genome. *Annual Review of Microbiology* **44**, 303–333.

Lai, M. M. C. & Cavanagh, D. (1997). The molecular biology of coronaviruses. *Advances in Virus Research* **48**, 1–100.

Lai, M. M. C., Baric, R. S., Brayton, P. R. & Stohlman, S. A. (1984). Characterization of leader RNA sequences on the virion and mRNAs of mouse hepatitis virus, a cytoplasmic RNA virus. *Proceedings of the National Academy of Sciences, USA* **81**, 3626–3630.

Lee, H.-J., Shieh, C.-K., Gorbalenya, A. E., Koonin, E. V., La Monica, N., Tuler, J., Bagdzhadzhyan, A. & Lai, M. M. C. (1991). The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. *Virology* **180**, 567–582.

Limsuwan, C. (1991). *Handbook for Cultivation of Black Tiger Prawns*. Bangkok: Tansetakit Co. Ltd.

Loh, P. C., Tapay, L. M., Lu, Y. & Nadala, E. C. B., Jr (1997). Viral pathogens of the penaeid shrimp. *Advances in Virus Research* **48**, 263–312.

Meulenberg, J. J., Hulst, M. M., de Meijer, E. J., Moonen, P. L. J. M., den Besten, A., de Kluyver, E. P., Wensvoort, G. & Moormann, R. J. M. (1993). Lelystad virus, the causative agent of porcine epidemic abortion and respiratory syndrome (PEARS), is related to LDV and EAV. *Virology* **192**, 62–72.

Nadala, E. C. B., Tapay, L. M. & Loh, P. C. (1997). Yellow-head virus: a rhabdovirus-like pathogen of penaeid shrimp. *Diseases of Aquatic Organisms* **31**, 141–146.

Patel, P. H. & Preston, B. D. (1994). Marked infidelity of human immunodeficiency virus type 1 reverse transcriptase at RNA and DNA template ends. *Proceedings of the National Academy of Sciences, USA* **91**, 549–553.

Peliska, J. A. & Benkovic, S. J. (1992). Mechanism of DNA strand transfer reactions catalysed by HIV-1 reverse transcriptase. *Science* **258**, 1112–1118.

Pleij, C. W. A. & Bosch, L. (1989). RNA pseudoknots: structure, detection, and prediction. *Methods in Enzymology* **180**, 289–303.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Shieh, C. K., Soe, L. H., Makino, S., Chang, M. F., Stohlman, S. A. & Lai, M. M. C. (1987). The 5′-end sequence of the murine coronavirus genome: implications for multiple fusion sites in leader-primed transcription. *Virology* **156**, 321–330.

Snijder, E. J. & Spaan, W. J. M. (1995). The coronaviruslike superfamily. In *The Coronaviridae*, pp. 239–255. Edited by S. G. Siddell. New York: Plenum Press.

Snijder, E. J., den Boon, J. A., Bredenbeek, P. J., Horzinek, M. C., Rijnbrand, R. & Spaan, W. J. M. (1990*a*). The carboxyl-terminal part of the putative Berne virus polymerase is expressed by ribosomal frameshifting and contains sequence motifs which indicate that toro- and coronaviruses are evolutionarily related. *Nucleic Acids Research* **18**, 4535–4542.

Snijder, E. J., Horzinek, M. C. & Spaan, W. J. M. (1990*b*). A 3′-coterminal nested set of independently transcribed mRNAs is generated during Berne virus replication. *Journal of Virology* **64**, 331–338.

Snijder, E. J., den Boon, J. A., Horzinek, M. C. & Spaan, W. J. M. (1991). Characterization of defective interfering Berne virus RNAs. *Journal of General Virology* **72**, 1635–1643.

Snijder, E. J., Wassenaar, A. L. M., van Dinten, L. C., Spaan, W. J. M. & Gorbalenya, A. E. (1996). The arterivirus Nsp4 protease is the prototype of a novel group of chymotrypsin-like enzymes, the 3C-like serine proteases. *Journal of Biological Chemistry* **271**, 4864–4871.

Sonigo, P., Barker, C., Hunter, E. & Wain-Hobson, S. (1986). Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. *Cell* **45**, 375–385.

Spaan, W., Cavanagh, D. & Horzinek, M. C. (1988). Coronaviruses: structure and genome expression. *Journal of General Virology* **69**, 2939–2952.

Spann, K. M., Vickers, J. E. & Lester, R. J. G. (1995). Lymphoid organ virus of *Penaeus monodon* from Australia. *Diseases of Aquatic Organisms* **23**, 127–134.

Spann, K. M., Cowley, J. A., Walker, P. J. & Lester, R. J. G. (1997). Gill-associated virus (GAV), a yellow head-like virus from *Penaeus monodon* cultured in Australia. *Diseases of Aquatic Organisms* **31**, 169–179.

Tang, K. F.-J. & Lightner, D. V. (1999). A yellow head virus gene probe: application to *in situ* hybridization and determination of its nucleotide sequence. *Diseases of Aquatic Organisms* **35**, 165–173.

ten Dam, E. B., Pleij, C. W. A. & Bosch, L. (1990). RNA pseudoknots: translational frameshifting and readthrough on viral RNAs. *Virus Genes* **4**, 121–136.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Walker, P. J., Byrne, K. A., Riding, G. A., Cowley, J. A., Wang, Y. & McWilliam, S. M. (1992). The genome of bovine ephemeral fever rhabdovirus contains two related glycoprotein genes. *Virology* **191**, 49–61.

Walker, P. J., Wang, Y., Cowley, J. A., McWilliam, S. M. & Prehaud, C. J. N. (1994). Structural and antigenic analysis of the nucleoprotein of bovine ephemeral fever rhabdovirus. *Journal of General Virology* **75**, 1889–1899.

Weiss, M., Steck, F. & Horzinek, M. C. (1983). Purification and partial characterization of a new enveloped RNA virus (Berne virus). *Journal of General Virology* **64**, 1849–1858.

Wongteerasupaya, C., Sriurairatana, S., Vickers, J. E., Akrajamorn, A., Boonsaeng, V., Panyim, S., Tassanakajon, A., Withyachumnarnjul, B. & Flegel, T. W. (1995). Yellow-head virus of *Penaeus monodon* is an RNA virus. *Diseases of Aquatic Organisms* **22**, 45–50.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48–52.

Received 21 October 1999; Accepted 3 February 2000