Virus Genes 23:2, 137–144, 2001

Completion of the Porcine Epidemic Diarrhoea Coronavirus (PEDV) Genome Sequence

ROLF KOCHERHANS,[1] ANNE BRIDGEN,[2] MATHIAS ACKERMANN[1] & KURT TOBLER[1*]

[1]*Virologisches Institut der Veterinär-Medizinischen Fakultät, Universität Zürich, Winterthurerstrasse 266a, CH-8057 Zürich*

[2]*Division of Virology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G11 5JR*

Received March 07, 2001; Accepted April 10, 2001

Abstract. The sequence of the replicase gene of porcine epidemic diarrhoea virus (PEDV) has been determined. This completes the sequence of the entire genome of strain CV777, which was found to be 28,033 nucleotides (nt) in length (excluding the poly A-tail). A cloning strategy, which involves primers based on conserved regions in the predicted ORF1 products from other coronaviruses whose genome sequence has been determined, was used to amplify the equivalent, but as yet unknown, sequence of PEDV. Primary sequences derived from these products were used to design additional primers resulting in the amplification and sequencing of the entire ORF1 of PEDV. Analysis of the nucleotide sequences revealed a small open reading frame (ORF) located near the 5′ end (no 99–137), and two large, slightly overlapping ORFs, ORF1a (nt 297–12650) and ORF1b (nt 12605–20641). The ORF1a and ORF1b sequences overlapped at a potential ribosomal frame shift site. The amino acid sequence analysis suggested the presence of several functional motifs within the putative ORF1 protein. By analogy to other coronavirus replicase gene products, three protease and one growth factor-like motif were seen in ORF1a, and one polymerase domain, one metal ion-binding domain, and one helicase motif could be assigned within ORF1b. Comparative amino acid sequence alignments revealed that PEDV is most closely related to human coronavirus (HCoV)-229E and transmissible gastroenteritis virus (TGEV) and less related to murine hepatitis virus (MHV) and infectious bronchitis virus (IBV). These results thus confirm and extend the findings from sequence analysis of the structural genes of PEDV.

Key words: porcine epidemic diarrhoea virus, coronavirus, ORF1, replicase gene

Introduction

Porcine epidemic diarrhoea virus (PEDV) is a causative agent for diarrhoea in pigs, particularly in neonates. The disease has been recognised for approximately thirty years, but the causative virus was only first described in 1978 [1], while another ten years elapsed before a method was developed for propagation of the virus in cell culture [2]. During this time, outbreaks of the disease were reported from numerous European countries as well as Korea, China and Japan. The epidemiology and pathogenesis of the disease have been well described by Pensaert [3]. The biological behaviour, electron microscopic appearance and polypeptide structure of PEDV resulted in its provisional classification as a coronavirus [2,4,5].

Coronaviruses belong to the taxonomic order of *Nidovirales* and contain a single stranded RNA genome of positive polarity, which is approximately thirty kilobases in length. The genes encoding the structural proteins are located at the 3′ end of the genome. An astonishing two-thirds of the genome consist of the replicase gene, which is located at the

*Author for all correspondence: E-mail: kurtt@vetvir.unizh.ch
GenBank Accession#: AF353511

5′ end of the genome. The replicase proteins are encoded by ORF1a and ORF1b. These two long, slightly overlapping ORFs are connected by a ribosomal frame shift site in all coronaviruses sequenced to date. This regulates the ratio of the two polypeptides encoded by ORF1a and the read-through product ORF1ab. About 70–80% of the translation products are terminated at the end of ORF1a, and 20–30% continue to the end of ORF1b. The polypeptides are post-translationally processed by viral encoded proteases [reviewed by 6]. These proteases are encoded within ORF1a; the polymerase- and the helicase-function are encoded by ORF1b.

We have previously completed the sequencing of the nucleocapsid- (N), membrane- (M), small membrane- (E), ORF3 and spike- (S) genes of the PEDV strain CV777 [7–9]. The alignment of the deduced amino acid sequences indicated that PEDV occupies an interesting intermediate position between the two well-characterized members of the group I coronaviruses, transmissible gastroenteritis virus (TGEV) and human coronavirus (HCoV)-229E. In this study, we have continued to determine and analyse nucleotide sequences of PEDV. To our knowledge, only two group I coronaviruses have been sequenced completely, HCoV-229E and TGEV [10,11]. In addition, two strains of mouse hepatitis virus (MHV), JHM and A59 belonging to the group II coronaviruses, and infectious bronchitis virus (IBV) have been completely sequenced [12–15]. Therefore, the sequence presented in this paper is the sixth sequence of a coronavirus covering the entire genome.

Materials and Methods

Growth of Virus and Preparation of Viral RNA

Growth of cell adapted PEDV strain CV777 was performed essentially as has been described elsewhere [2,8], except that virus-infected cells were harvested at approximately 18 h post infection. Cells were freeze-thawed three times and cell debris removed by low speed centrifugation. Virus was pelleted by centrifugation for 2 h at 22,000 rpm and 4°C in a SW28 rotor of a Beckman centrifuge. Virus pellets prepared from two 175 cm^2 flasks were pooled and resuspended in 1 ml Trizol™ (Gibco-BRL), and RNA was prepared as recommended by the manufacturer.

cDNA Synthesis and PCR Amplification of Viral Sequences

In order to obtain the first partial PEDV specific sequences, the predicted amino acid sequences of the HCoV-229E and TGEV polymerase ORFs were aligned and homologous regions identified. The homologous regions were used to design degenerate primers [9] that were used for RT-PCR amplifications. These initial amplicons were cloned and sequenced [9]. Later, a mixture of up to six anti-genome sense primers based on PEDV specific sequences or the degenerate primers and random hexamer primer (purchased from Schmidheini AG; Balgach, Switzerland) was used for first strand cDNA synthesis. RNA prepared from two 175 cm^2 flasks of virus-infected cells was denatured for 10 min at 65°C and first strand cDNA was performed in a 20 µl total reaction volume using SuperscriptII™ (GibcoBRL; Basel, Switzerland) according to the manufacture's protocol. This was modified to create the longer reverse transcription products by including a denaturation step at 95°C for 5 min following the first 1 h incubation at 42°C, followed by the addition of 1 µl SuperscriptII™ and a second prolongation step of 1 h at 42°C. Template RNA was digested by adding 1 µl RNaseH (GibcoBRL; Basel, Switzerland) to the reaction mix and incubating at 37°C for 20 min.

PCR amplification was performed as described elsewhere. In brief, *Pfu* DNA polymerase (Stratagene; Basel, Switzerland) was used for the amplifications, which were performed on a DNA Engine (MJ Research) machine. PCR fragments were subsequently cloned into pBluescript®II KS+ or pUC19 vectors using standard procedures. The nucleotide sequence was determined on these cDNA clones. Direct sequencing was performed on a RT-PCR product (see Fig. 1B), which was cleaned through an agarose gel. The contigs of the sequence determinations were constructed using SeqMan (DNA*, Lasergene, Madison WI, USA). We previously reported the determination of the PEDV leader sequence on the mRNA encoding the N-gene [16]. This sequence was used for the primer design in order to amplify the 5′ end of the genome. The leader

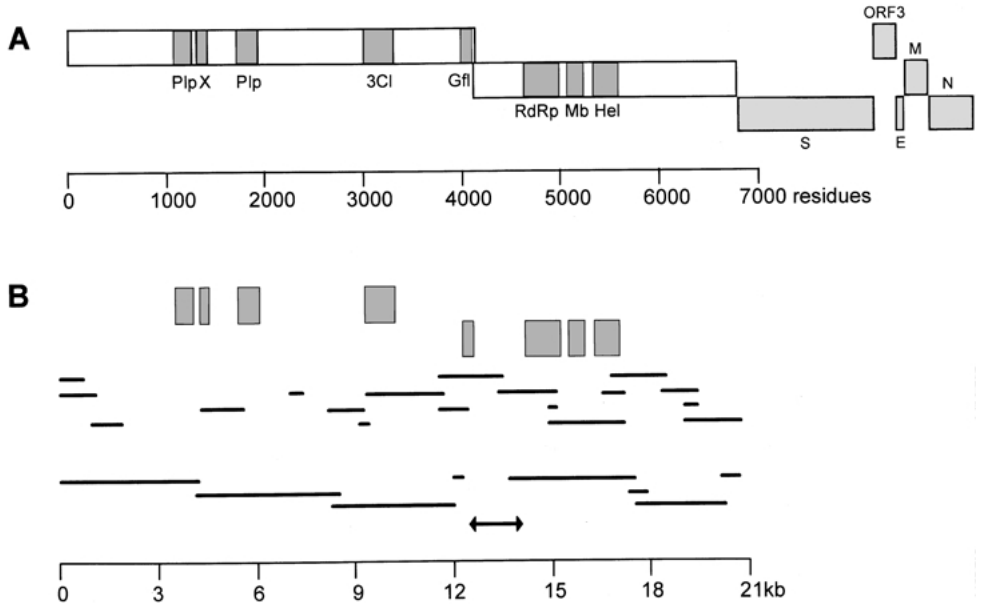

Fig. 1. Schematic presentation of the PEDV genome and map of cDNA clones. (A) The open reading frames (ORFs) are represented as boxes. The following domains located in ORF1ab are shown: papain-like proteinase (Plp), X domain (X), poliovirus 3C-like proteinase (3C1), growth factor-like domain (Gfl), RNA-dependent RNA polymerase (RdRp), metal ion-binding-domain (Mb), and helicase (Hel). An amino acid scale is shown at the bottom. (B) The cDNA clones and the RT-PCR product used for the determination of the nucleotide sequence presented in this paper are shown as located on the genomic RNA of PEDV. The upper part of the figure shows the initial RT-PCR products amplified with degenerate primers designed from conserved coronavirus sequences. The lower part shows RT-PCR products amplified with primers based on PEDV sequences of initial cDNA clones. Clones are depicted as lines, the RT-PCR product which was sequenced directly as a two-headed arrow. A nucleotide scale is shown at the bottom.

sequence was used for the *in silico* construction of the genomic RNA sequence, which is available on GenBank database (Accession Number AF353511).

Sequence Analysis

Virus sequences covering replicase genes were obtained from the GenEMBL sequence database. The files with the accession numbers X69721, Z34093, AF029248, and M95169 for HCoV-229E, TGEV (Purdue 115), MHV-A59, and IBV (Beaudette) respectively were used.

The deduced amino acid sequences were compared as indicated in the text using PILEUP and GAP (GCG Package version 10.0; Madison, WI, USA). The files generated by PILEUP were used in DISTANCES (GCG Package version 10.0; Madison, WI, USA) to determine the Kimura protein sequence distances, which were subsequently used for the construction of unrooted dendrogram using TreeGen on the CBRG server (http://cbrg.inf.ethz.ch/)

Results and Discussion

Cloning Strategy

The cloning approach we used previously to clone the PEDV M and N genes involved designing primers based on conserved regions of the coronavirus M and N genes to amplify the equivalent to the unknown PEDV sequence. In this study, we employed this technique to clone parts of the ORF1 of PEDV. Such a method is useful for viruses which do not grow to high titre, avoids lengthy screening of clones and could potentially be applied to the cloning of any group I coronavirus. However, the large size of ORF1 and the paucity of sequence data from other coronaviruses made this an ambitious objective. A number of conserved functional domains were

identified in the predicted ORF1 products, but these domains are mainly located in the ORF1b region and leave large regions of the ORF1a product with no known function and only a low level of sequence conservation between different coronavirus genomes. In order to clone and determine the sequences for the PEDV ORF1, the predicted amino acid sequences of the HCoV-229E and TGEV ORF1 were aligned and homologous regions identified. The HCoV-229E and TGEV ORFs were sufficiently closely related to allow complete alignment of the predicted expression products. In contrast, the MHV and IBV sequences were much more divergent, and could only be aligned with the group I sequences in some of the conserved regions. Degenerate primers were designed from regions conserved between the HCoV-229E and TGEV and, where possible, MHV and IBV ORF1. These primers were used both to prime reverse transcription and for the PCR amplifications. Sequence data derived from these PCR products allowed us to design sequence-specific primers which were then used to amplify the entire ORF1 (see Fig. 1B).

Analysis of the Nucleotide Sequence, Prediction of ORFs

Numerous small cDNA clones, five large cDNA clones and one RT-PCR product covering the 5′ two-thirds of the PEDV genome were used to determine the nucleotide sequence of the PEDV ORF1 (Fig. 1). This analysis completes the nucleotide sequence of PEDV, and thereby the sixth entire sequence determined from a coronavirus genome [10–13,15]. The genome of PEDV (CV777) excluding the poly A-tail is 28033 nt in length.

Analysis of the newly determined nucleotide sequence revealed a pattern of ORFs typical of coronaviruses. A small ORF with the potential to code for a 12-amino acid peptide was found at the 5′ end of the genome from nucleotide position 99–137. Such small ORFs (uORFs) are present in all coronaviruses sequenced so far. The uORFs of HCoV-229E [17] and IBV [15] are found to be eleven codons in length, while that of MHV is eight codons long [18,19]. That of TGEV can only encode a three-amino acid peptide [20]. Two long ORFs of 12354 and 8037 nt, which overlap by 46 nt, covered most of the newly determined sequence. By analogy to published coronavirus sequences [15,17,20], the ORFs were designated ORF1a and ORF1b. The predicted ORF1a of FEDV extended from nucleotide 297 to 12650. This resulted in a 4117-codon ORF. The overlapping ORF1b starting at nucleotide 12605 and ending at nucleotide 20641 had the capacity to code for 2678 amino acids.

It has been proposed for coronaviruses and other members of the order *Nidovirales* [21] that the nucleotide sequences in the overlapping regions of ORF1a and ORF1b are able to fold into a pseudoknot tertiary structure [22,23]. This region allows the ribosome shifting of the reading frame during translation of the ORF1a and subsequently continues the translation in ORF1b. The function of these RNA structures as ribosomal frame shift sites was demonstrated for the analogous sequences of IBV [24] and HCoV-229E [25]. It seems likely that the translation of the PEDV ORF1b is mediated by such a ribosomal frame shifting. The nucleotide sequences of PEDV, HCoV-229E, and TGEV covering the ribosomal frame shift site are more conserved to each other than to MHV-A59 or IBV. In order to identify the sequence which could be involved in the formation of the tertiary structure, the nucleotide sequences covering the end of ORF1a and the beginning of ORF1b from HCoV-229E [25] and TGEV [20] were aligned with the corresponding sequence of PEDV. Fig. 2A shows the predicted frame shift region of PEDV based on this comparison. The so-called slippery site (UUUAAAC) at which frame shifting occurs is identical in all coronaviruses sequenced so far. The stems and loops required to provide the tertiary structure of the frame shift regions of TGEV and HCoV-229E were compared and Fig. 2B shows the predicted tertiary structure required for the frame shift of PEDV based on this comparison.

Amino Acid Sequence Comparison

Pairwise comparison of the deduced amino acid sequences (using GAP) revealed that ORF1b of PEDV is more conserved than ORF1a to corresponding sequences of other coronaviruses. The percentage of similarities and identities is shown in Table 1. The putative protein sequence of ORF1a was most similar to the sequence of ORF1a of HCoV-229E (59.4%) and less similar to the corresponding ORF1a of TGEV (52.1%), MHV-A59 (39.5%) and IBV (38.7%). The same relationship, but at a higher level of similarity, was true for the deduced amino

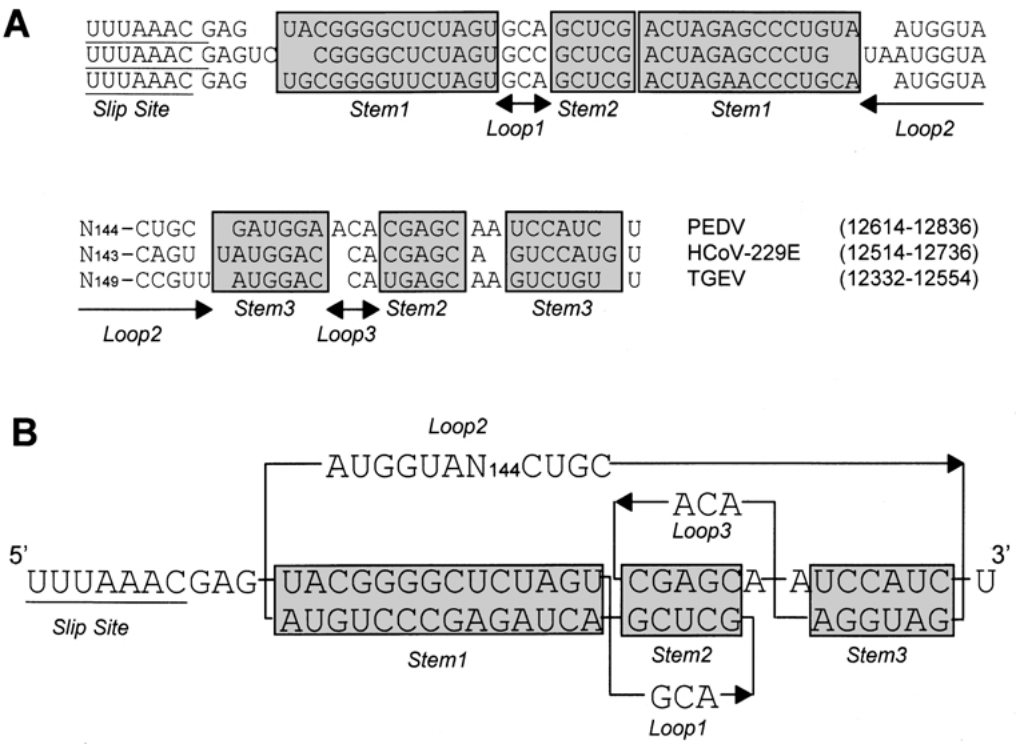

Fig. 2. The putative ribosomal frame shift region. (A) Nucleotide sequence covering the pseudoknot structures of HCoV-229E and TGEV aligned to the corresponding sequence from PEDV using DNA*. The putative slippery sites are underlined, the stems are boxed and the loops are indicated by two headed arrows. (B) The putative tertiary structure of PEDV is predicted from the sequence alignment from the upper part of the figure.

acid sequence of the predicted PEDV ORF1b. It was most similar to the amino acid sequence of HCoV-229E ORF1b and TGEV ORF1b (83.2% and 80.3%, respectively). The similarity to the ORF1b from MHV-A59 and IBV was around 64%. The deduced amino acid sequences of ORF1a and ORF1b from PEDV were aligned with the corresponding sequences of HCoV-229E, TGEV, MHV-A59, and IBV using PILEUP. The degrees of amino acid homologies are graphically presented as dendrograms (Fig. 3A,B).

The multiple sequence alignments revealed several putative functional domains common to coronavirus sequences [23,26] located on the deduced amino acid sequence of ORF1ab of PEDV. Some of these had been used to design the primers for the RT-PCR amplification. In the ORF1a region the following motifs were observed. Two motifs indicative of papain-like proteases (Plp) were present at amino acid positions 1077–1266 and 1716–1917. The Plp motif is found twice in the replicase genes of HCoV-229E, TGEV and MHV, but only once in that of IBV. In this respect, PEDV resembles HCoV-229E, TGEV and MHV rather than IBV. A highly conserved region (X-domain) was found between the two Plp motifs. Despite this motif being present in all coronavirus sequences, its function is not yet known. A picornavirus 3C-like (3C1) protease domain is located between amino acids 2998 and 3299 of the PEDV ORF1a. All corona- and arteriviruses encode this motif, which is the main protease for the coronavirus mediated processing of the polyproteins. Three markedly hydrophobic domains conserved among coronaviruses are found in ORF1a. The first is located after the second Plp motif and the others flank the 3Cl motif. Finally, a growth factor-like (Gfl) domain was located close to the end of ORF1a (amino acid position 3965–4000). In the ORF1b region, three structural protein motifs could be recognized, which all play a role in viral replication. A sub-sequence at amino acid position 4636–4939 containing the characteristic tripeptide

SDD (or GDD in most RNA viruses) [26] is probably the active site for the RNA dependent RNA polymerase. A metal ion-binding domain covering amino acids 5027–5103 and a helicase motif at amino acid positions 5309–5624 were also observed in the PEDV ORF1b product. Alignments of the deduced amino acid sequences of the 3Cl protease and the polymerase motif from five different coronaviruses are shown in Fig. 4A and 4B, respectively. The findings concerning conserved domains are summarised in Fig. 1A.

A deletion of about 180 amino acids located between the X-domain and the second Plp motif in the putative ORF1a sequence of TGEV compared to that of HCoV-229E was reported by Eleouet et al. [20]. This additional sequence was present in the PEDV ORF1a product. The alignment (using GAP) of the HCoV-229E and PEDV amino acid sequences revealed 42.5% similarity and 31.5% identity in this region.

Table 1. Percentage of similarity and identity between ORF 1a and 1b of Coronavirus sequences. The sequences were aligned GAP (GCG Package) using a blossum62 weight matrix and default settings (Gap Weight 8, Length Weight 2)

Similarity	HCoV-299E	TGEV	MHV-A59	IBV
1a				
PEDV	59.4	52.1	39.4	38.7
HCoV-229E		53.1	37.9	38.0
TGEV			37.6	39.6
MHV-A59				38.8
1b				
PEDV	83.2	80.3	64.5	64.3
HCoV-229E		79.9	63.4	63.6
TGEV			64.9	64.6
MHV-A59				65.7
Identity	**HCoV-299E**	**TGEV**	**MHV-A59**	**IBV**
1a				
PEDV	50.1	42.9	30.9	29.8
HCoV-229E		43.2	28.9	28.9
TGEV			28.1	30.2
MHV-A59				30.0
1b				
PEDV	77.8	73.3	55.8	55.4
HCoV-229E		72.7	54.6	54.5
TGEV			55.5	54.9
MHV-A59				57.1

Conclusion

Earlier sequence analysis of PEDV based on the structural protein sequences has shown that PEDV is most closely related to HCoV-229E and TGEV [7–9,27], less related to MHV-A59, and least related to IBV. However, it was not possible to determine the relative similarities of HCoV-229E, TGEV and PEDV. In this study, the similarities and identities of the amino acid sequence alignments based on ORF1a and ORF1b show clearly that PEDV is most closely related to HCoV-229E and, moreover, that

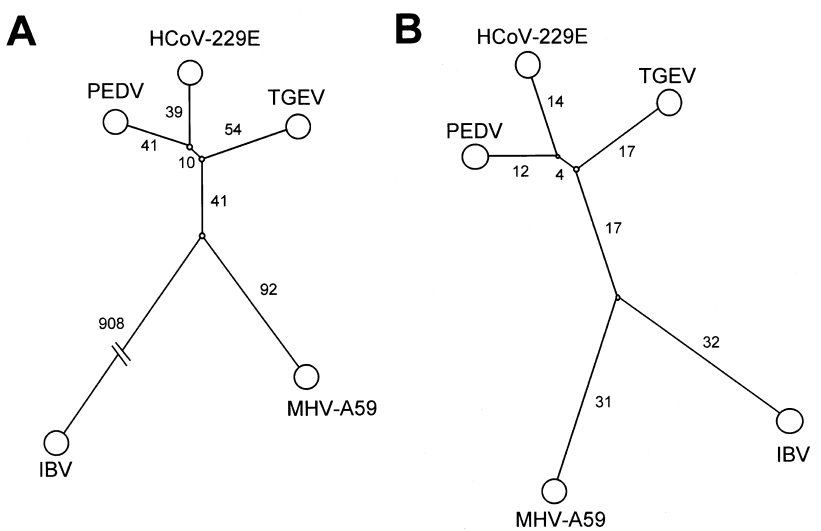

Fig. 3. Phylogenetic trees generated using GenTree (CBRG server). Unrooted dendrogram showing the Kimura's distances of (A) ORF1a and (B) ORF1b.

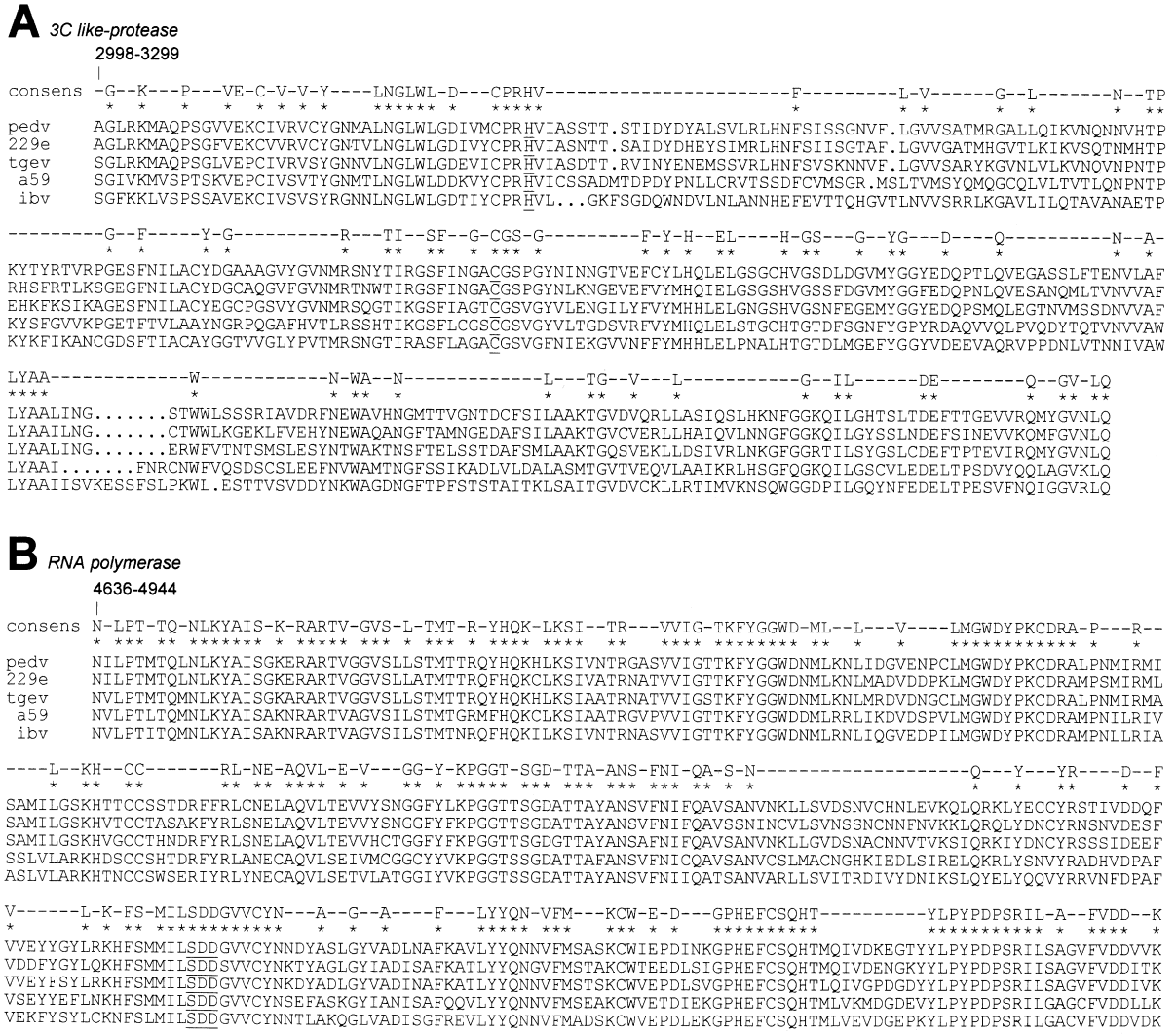

Fig. 4. Amino acid alignments of five coronavirus sequences covering the 3C like protease (A) and the RNA dependent RNA polymerase-motif (B). The consensus sequence of residues conserved in all sequences is shown on the top of the alignments and marked by asterisks. The conserved SDD motif in polymerases is underlined. The locations of the aligned sequences relative to the start of the PEDV fusion protein are shown for both motifs.

HCoV-229E is more similar in sequence to PEDV than it is to TGEV.

In addition to the sequence analysis, the presented work offers various possibilities for future research on coronaviruses. Functional analysis and processing of the as yet uncharacterised PEDV ORF1 is now possible. Recently, Almazan et al. and Yount et al. achieved the generation of infectious TGEV from cDNA [28,29] and Thiel et al. suceeded in generating full length cDNA clones of HCoV-229E and IBV in a recombinant vaccinia virus system [30]. The sequence and the cDNA clones covering the entire genome of PEDV would allow the development of a mini-genome system to study viral replication or the generation of an assembled, infectious cDNA clone. Bearing in mind the close relationship of PEDV and HCoV-229E, the latter approach could be used to exchange functional parts of these viruses to gain new insights into the biology of these viruses. Furthermore, the

generation of a PEDV infectious clone could allow the use of PEDV as a vaccine against TGEV.

Acknowledgements

The authors thank Christa Meyer for excellent technical assistance. These studies were supported by the Swiss National Science Foundation, grant #31-43503.95.

References

1. Pensaert M.B. and Debouck P., Arch Virol *58*, 243–247, 1978.
2. Hofmann M. and Wyler R., J Clin Microbiol *26*, 2235–2239, 1988.
3. Pensaert M.B., *Porcine Epidemic Diarrhea Virus Virus Infections of Porcines*. Elsevier, pp. 167–176, 1989.
4. Egberink H.F., Ederveen J., Callebaut P., and Horzinek M.C., Am J Vet Res *49*, 1320–1324, 1988.
5. Utiger A., Tobler K., Bridgen A., and Ackermann M., Virus Genes *10*, 137–148, 1995.
6. Ziebuhr J., Snijder E.J., and Gorbalenya A.E., J Gen Virol *81(4)*, 853–879, 2000.
7. Duarte M., Tobler K., Bridgen A., Rasschaert D., Ackermann M., and Laude H., Virology *198*, 466–476, 1994.
8. Bridgen A., Duarte M., Tobler K., Laude H., and Ackermann M., J Gen Virol *74*, 1795–1804, 1993.
9. Bridgen A., Kocherhans R., Tobler K., Carvajal A., and Ackermann M., Adv Exp Med Biol *440*, 781–786, 1998.
10. Herold J., Raabe T., and Siddell S., Arch Virol Suppl *7*, 63–74, 1993.
11. Eleouet J.F., Rasschaert D., Lambert P., Levy L., Vende P., and Laude H., Adv Exp Med Biol *380*, 459–461, 1995.
12. Lee H.-J., Shieh C.-K., Gorbalenaya A.E., Koonin E.V., La Monica N., Tuler J., Bagdazhadzhyan A., and Lai M.M., Virology *180*, 567–582, 1991.
13. Bonilla P.J., Gorbalenya A.E., and Weiss S.R., Virology *198*, 736–740, 1994.
14. Bredenbeek P.J., Pachuk C.J., Noten A.F., Charite J., Luytjes W., Weiss S.R., and Spaan W.J., Nucleic Acids Res *18*, 1825–1832, 1990.
15. Boursnell M.E., Brown T.D., Foulds I.J., Green P.F., Tomley F.M., and Binns M.M., J Gen Virol *68*, 57–77, 1987.
16. Tobler K., and Ackermann M., Adv Exp Med Biol *380*, 541–542, 1995.
17. Herold J., Raabe T., Schelle-Prinz B., and Siddell S.G., Virology *195*, 680–691, 1993.
18. Pachuk C.J., Bredenbeek P.J., Zoltick P.W., Spaan W.J., and Weiss S.R., Virology *171*, 141–148, 1989.
19. Soe L.H., Shieh C.K., Baker S.C., Chang M.F., and Lai M.M., J Virol *61*, 3968–3976, 1987.
20. Eleouet J.F., Rasschaert D., Lambert P., Levy L., Vende P., and Laude H., Virology *206*, 817–822, 1995.
21. Cavanagh D., Arch Virol *142*, 629–633, 1995.
22. Brierley I., Boursnell M.E., Binns M.M., Bilimoria B., Brown V.C., Blok T.D., and Inglis S.C., EMBO J *6*, 3779–3785, 1987.
23. de Vries A.A.F., Horzinek M.C., Rottier P.J.M., and de Groot R.J., Seminars in Virology *8*, 33–47, 1997.
24. Brierley I., Digard P., and Inglis S.C., Cell *57*, 537–547, 1989.
25. Herold J. and Siddell S.G., Nucleic Acids Res *21*, 5838–5842, 1993.
26. Gorbalenya A.E., Koonin E.V., Donchenko A.P., and Blinov V.M., Nucleic Acids Res *17*, 4847–4861, 1989.
27. Duarte M., and Laude H., J Gen Virol *75*, 1195–1200, 1989.
28. Almazan F., Gonzalez J.M., Penzes Z., Izeta A., Calvo E., Plana-Duran J., and Enjuanes L., Proc Natl Acad Sci USA *97*, 5516–5521, 2000.
29. Yount B., Curtis K.M., and Baric R.S., J Virol *74*, 10600–10611, 2000.
30. Thiel V., Casais T., Cavanagh D., and Britton P., A reverse genetic system for coronaviruses. European Congress of Virology, Glasgow, 2000.