Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

VIROLOGY 198, 466–476 (1994)

Sequence Analysis of the Porcine Epidemic Diarrhea Virus Genome between the Nucleocapsid and Spike Protein Genes Reveals a Polymorphic ORF

MARIELA DUARTE,* KURT TOBLER,† ANNE BRIDGEN,†,[1] DENIS RASSCHAERT,* MATHIAS ACKERMANN,† AND HUBERT LAUDE*,[2]

*I.N.R.A., Unité de Virologie et Immunologie Moléculaires, CR de Jouy-en-Josas, 78350 Jouy-en-Josas, France; and †Institut für Virologie, Vet. Med. Fak. der Universität Zürich, Winterthurerstrasse 266a, CH-8057 Zürich, Switzerland

Received August 3, 1993; accepted October 6, 1993

In order to investigate the genome organization of the porcine epidemic diarrhea virus (PEDV) further, cDNA clones covering the region between the nucleocapsid and the spike (S) protein genes were independently constructed and sequenced for the two virulent isolates Br1/87 and CV777. Of the three major ORFs identified, two were found to encode the major and minor coronavirus membrane proteins M and sM. A potentially single ORF, designated ORF3 according to the pattern of the viral subgenomic mRNAs, could be identified between the S and sM genes. A striking variability, essentially generated by short deletions clustered in a few loci, was observed in the ORF3 of both isolates. The largest predicted polypeptide of 223 amino acids showed homology with polypeptides potentially encoded by other members of the same genetic subset, including two shorter polypeptides of human respiratory virus HCV 229E and one of transmissible gastroenteritis virus TGEV. This homology suggests that the two HCV ORFs may have originated from a single precursor. The function of these polypeptides is not known, but the predicted products of the PEDV ORF3 and related ORFs share features suggestive of a membrane-associated protein.

INTRODUCTION

Porcine epidemic diarrhea virus (PEDV) causes an acute enteritis in animals of all ages, often fatal for the neonates (Pensaert and Debouck, 1978). Outbreaks of the disease have been reported in many swine-raising countries, notably in Europe and Asia (Debouck, 1982). The causative agent, first described by Pensaert and Debouck in 1978, was provisionally classified as a coronavirus, as a result of its microscopic appearance and polypeptide composition. The major virion components appear to consist of an RNA-binding nucleocapsid protein (N, 58 kDa) and two glycoproteins. The two prominent glycoprotein clusters of 180–200 kDa and 27–32 kDa which can be detected immunologically in PEDV-infected cell lysates (Egberink *et al.*, 1988; Knuchel *et al.*, 1992; Utiger *et al.*, 1993), are thought to correspond to the S and M membrane proteins, respectively. No antigenic relationships to other coronaviruses have been detected using seroneutralization, immunofluorescence, and immunoelectron microscopy. However, some cross-reactivity involving the N and M proteins has been reported in Western immunoblotting with three agents belonging to the same serogroup: feline infectious peritonitis virus (FIPV), transmissible gastroenteritis virus (TGEV), canine coronavirus (CCV), and with the putative mink coronavirus (MCV) (Pensaert *et al.*, 1981; Yaling *et al.*, 1988; Have *et al.*, 1992).

PEDV replicates essentially in the enterocytes covering the small intestine villi, like TGEV (Debouck and Pensaert, 1980; Debouck *et al.*, 1981), yet the two viruses are clearly distinct entities. Strikingly, PEDV is unable to grow in porcine cell cultures permissive to TGEV. A method to propagate the virus in a heterologous, monkey cell line was only first described 5 years ago (Hofmann and Wyler, 1988). This explains in part why the molecular characterization of PEDV is so poorly advanced. Recently, the sequences of the 1700 nucleotides nearest the genome 3′ end and comprising the N gene were determined (Bridgen *et al.*, 1993a); this showed significant homology to the N genes of coronaviruses and thus the status of PEDV as a member of the Coronaviridae was formally established. The present report deals with the cloning and sequencing of a further 1740 nt of the PEDV genome, in which three major ORFs, designated M, sM, and ORF3, could be identified. The results confirm our previous statement that PEDV is more closely related to the human respiratory virus HCV 229E than to TGEV and revealed an unexpected variability of the ORF3 region.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. Z24733.

[1] Current address: Institute of Virology, Church Street, Glasgow G11 5JR, Scotland.

[2] To whom reprint requests should be addressed.

0042-6822/94 $6.00

MATERIALS AND METHODS

Virus and cells

Virus isolates and cell culture have been described elsewhere (Bridgen *et al.*, 1993a; Duarte *et al.*, 1993). The British Br1/87 isolate was analyzed by the Paris group, while the Swiss group studied the Belgian CV777. Propagation of virus in Vero cells essentially followed the protocol of Hofmann and Wyler (1988).

Methods used for the characterization of the Br1/87 isolate

Preparation of RNA and construction of cDNA library. Total RNA from 10^7 PEDV-infected cells was extracted at 48 hr post-infection by the guanidinium thyocianate method and purified by ultracentrifugation through a 5.7 *M* CsCl gradient, as described by Vaquero *et al.* (1982). Polyadenylated RNA was fractionated by chromatography on poly(U) Sepharose (Pharmacia). Five micrograms of the poly(A)$^+$ RNA was used to construct a cDNA expression library in the λ Zap II expression vector (Stratagene). This was done according to the instructions of the manufacturer, except that the RNA was denatured with methylmercuric hydroxide as described for TGEV genomic RNA by Rasschaert *et al.* (1987). The cDNA expression library contained 5.5×10^5 individual recombinants and was amplified 10^4-fold.

Screening of the cDNA library. (i) Immunoscreening: the amplified library was screened with hyperimmune anti-PEDV serum supplied by Station de Pathologie Porcine (CNEVA, Ploufragan). To eliminate background with antibodies to *Escherichia coli,* the antiserum was depleted prior to utilization as follows: a cell lysate (XL1 Blue strain) was adsorbed onto a nitrocellulose membrane which was then suspended in a tank of chloroform vapor for 15 min, cut, and shaken with the antiserum for 1 hr at 37° and overnight at 4°. The immunoscreening was carried out essentially as described by Young *et al.* (1982). (ii) Molecular hybridization: further screening of the library with DNA or oligonucleotides probes was performed using at least 5×10^4 recombinant phages per 150-mm plate. [^{32}P]-dCTP-labeled probes were prepared by random priming of appropriate cDNA clone restriction fragments: a 1150 nt *Pst*I fragment of clone pBE1 (Probe I), a 350 nt *Pst*I/*Bsm*I fragment of clone pBE2 (Probe II) and a 200 nt *Pst*I/*Nar*I fragment of pBE3 (Probe III). A 20-mer oligonucleotide probe (5′-CACTGCACGTGGACCTTTTC-3′) derived from the pBE5 sequence was 5′ end-labeled and hybridized to nitrocellulose filters according to Wood *et al.* (1985). Positive phage clones were detected and plaque-purified two or three times with hybridization carried out at each stage to verify the insert before phage stock of the final positive recombinants were made.

DNA amplification by the polymerase chain reaction (PCR). PCRs were performed according to the instructions of Perkin–Elmer Cetus with *Taq* DNA polymerase (2.5 U, Promega), in a Perkin–Elmer/Cetus DNA thermal cycler for the following purposes: (i) to detect the largest positives clones in each cDNA library screening, following the method described by Hamilton *et al.* (1991): crude positive phage plugs were picked into SM buffer (50 m*M* Tris–HCl, pH 7.5, 100 m*M* NaCl, 8 m*M* $MgSO_4$, 0.01% gelatin) and their sequences amplified using a phage-specific 5′ primer and a PEDV-specific sense primer; (ii) to verify the sequence in a genome region which was found to be divergent among the cDNA clones. The clones p501, 502, and 506 were derived from total RNA by RT/PCR. For RT reaction, 1 μg of RNA denatured as above was used in a 20 μl final volume. After inactivation of the RT for 10 min at 90°, 10 μl of the reaction were used for PCR amplification using oligo 10 (5′-TACTAGACCATT-ATCATTCA-3′) and 1856 (5′-TTGAGGCGAATTATAT-TATG-3′).

DNA sequencing. After excision and rescue (Short *et al.*, 1988), pBluescript SK (−) phagemids containing a PEDV cDNA insert were prepared as large-scale DNA preparations and sequenced. The whole PEDV sequence was determined from both strands of at least two clones by the dideoxynucleotide chain termination method. (i) M13 DNA sequencing: clones pBE1, pBE5, and pBE6 were sonicated, end-repaired with the Klenow fragment of *E. coli,* and subcloned into the desphosphorylated M13 mp18 *Sma*I-cut vector. Subclones containing the random sequences of the PEDV inserts were detected by hybridization and single-stranded templates were prepared from the positive plaques. Sequence data were assembled using Microgenie (Beckman, 1983) and gelassemble option of the University of Wisconsin GCG package (Devereux *et al.*, 1984). (ii) Supercoiled DNA sequencing: this method was used to sequence the extremities of all cDNA inserts using T3 and T7 primers. Clones pBE2 and pBE3 were also partially sequenced using PEDV-specific primers. Most Br1/87 clones were sequenced using the Sequenase 2 kit (U.S. Biochemicals). Alternatively, half of the pBE6 sequences and the PCR fragments were sequenced using a cycle-sequencing protocol (Taq Dye Primer T3, T7, −21 M13 and Taq Dye Deoxy terminator sequencing kits). The sequencing reactions were subsequently loaded on a 373A Applied Biosystems automated DNA sequencer.

Methods used for the characterization of the CV777 isolate

Techniques for virus purification, RNA isolation, cDNA cloning, and sequencing were essentially as de-

scribed in our previous publication (Bridgen *et al.*, 1993a). The following techniques were modified for the work described in this article. The P24/P25 PCR reaction was performed as described previously (Bridgen *et al.*, 1993b) using the Perkin–Elmer Cetus Amplitaq DNA polymerase, but all other PCR reactions were performed using Stratagene *Pfu* polymerase. PCR reactions were performed as recommended by the supplier using the supplied buffer 1. Cycling parameters were generally 1 cycle of 95° 5 min; 45° 1 min, and 74° 5 min followed by 39 cycles of 95° 50 sec, 45° 1 min, and 74° 2.5 min. The following primers derived from the PEDV sequence (except P34, P69, and P75, which have 5′ modifications) were used: P9 (M gene): 5′-TCACCTCATCAACGGGCATA-3′; P10 (sM Gene): 5′-TCATAGACCATTATCATTCA-3′; P34 (N gene): 5′-GAGGATCCTGAAAGCTGACAG-3′; P50 (M gene): 5′-CCTCTACAAGCAATGTACC-3′; P59 (ORF3 gene): 5′-ATTATAAGCATTACCTTCGT-3′; P61 (S gene): 5′-TCCTAGACTTCAACCTTAC-3′; P68 (ORF3 gene): 5′-AACAAAGCCTGCCAATAAG-3′; P69 (M gene): 5′-CTATAAATATGTCTAACGG-3′; P75 (ORF3 gene): 5′-ATGTTATATTACTGTGGTG-3′. Degenerate primers were also used as follows: P24 (N gene): 5′-CTCG-AGCGACCCAGA(C/A)GAC(A/T)CC(G/T)TC-3′; P25 (M gene): 5′-GACTAGTTGGTGGAG(T/A)TTTAA(T/C)CC(A/T)GA-3′; P55 (S gene): 5′-GGA(G/T)AAGGT(G/T)AATGAA(A/G)TG(C/T)GT-3′; P56 (S gene): 5′-CCA(G/T)AC(A/G/T)TACCA(A/T)GGCCA(T/C)TT-3′. All these primers with the exception of P9 and P10, which were kindly provided by the Paris group, were synthesized by Dr Schmidheini (Microsynth). Cloning of the PCR products illustrated in Fig. 1 into the pBluescript KS plasmid (Stratagene) and subcloning of PCR clone restriction fragments were as described previously (Bridgen *et al.*, 1993a). The *Hin*dII and *Xba*I restriction sites illustrated in this figure were used to construct subclones of the cloned 700 bp, M, O, and S PCR products. All clones were sequenced using the universal sequencing primers or the vector SK and KS primers using the T7 sequencing kit II (U.S. Biochemicals). In addition, many of the clones were sequenced with the PEDV-specific primers described above, such that the entire PEDV sequence described was sequenced on both strands of at least two clones. The ORF3 region was sequenced completely on six clones.

Northern blot hybridization

RNA blot hybridizations were carried out by conventional techniques and probed with [^{32}P]-labeled pBE1 insert. Total RNA from PEDV-infected and mock-infected Vero cells was separated in a denaturing formaldehyde gel electrophoresed in 10 m*M* phosphate buffer, pH 7.0 + 2.2 *M* formaldehyde prior to transfer to nylon membrane Hybond-N (Amersham) and hybridized (42°) in 50% formamide, 5× Denhardt, 5× SSC, then washed three times for 30 min each in 0.1× SSC, 0.1% SDS at 55°, before drying. Autoradiography was performed at −70° with intensifying screens.

Sequence analysis and phylogeny

Nucleotide and amino acid sequences analysis were performed with the GCG package. Multiple sequence alignment was generated by the Clustal V package (Higgins and Sharp, 1989). Distance matrix was calculated from the aligned sequences (including positions with gaps) using the "Phylogenetic trees" option of the Clustal package. The phylogenetic tree was then generated submitting the distance matrix to the neighbor-joining method (Saitou and Nei, 1987) of the PHYLIP (Phylogenetic Inference Package program, version 3.4; Felsenstein, 1991). Clustal V and Phylip were both available on the French CITI2/BISANCE network (Dessen *et al.*, 1990). The tree was plotted using the Phylip's Drawtree program.

RESULTS

Derivation and characterization of the PEDV cDNA clones

The Br1/87 cDNA clones were obtained from a cDNA library constructed in a λ expression vector. Immunoscreening of the cDNA library with a polyclonal antiserum allowed us to isolate the clone named pBE1. As we reported earlier, this clone contained N gene-related sequences. Clones pBE2, pBE3, and pBE5 were selected in this order by serial screening of the library with DNA probes consisting of a restriction fragment from the 5′ end of the last clone isolated. Clone pBE6 was obtained by screening the library with a primer complementary to the pBE5 5′ end. Figure 1A shows an ordered map of these cDNA clones together with the strategy used to sequence them.

The CV777 clones illustrated in Fig. 1B were all obtained by PCR using a mixture of degenerate and PEDV sequence-specific primers. The degenerate primers in the N, M, and S genes were based on conserved sequences identified within the corresponding coronavirus genes in our previous publication (Bridgen *et al.*, 1993b). The degenerate primers P24 and P25, in the N and M genes, respectively, have been described in detail elsewhere (Bridgen *et al.*, 1993a). The two degenerate primers in the S gene, P55 and P56, were based on the conserved consensus residues KVNECV and KWPWYVW, respectively. The unique primers were based on previous PEDV sequence data (Bridgen *et al.*, 1993a) or were provided by the Paris group. All amplifications were made on first strand cDNA template which had been primed with the unique or degenerate primers. Amplifications were performed using Strata-

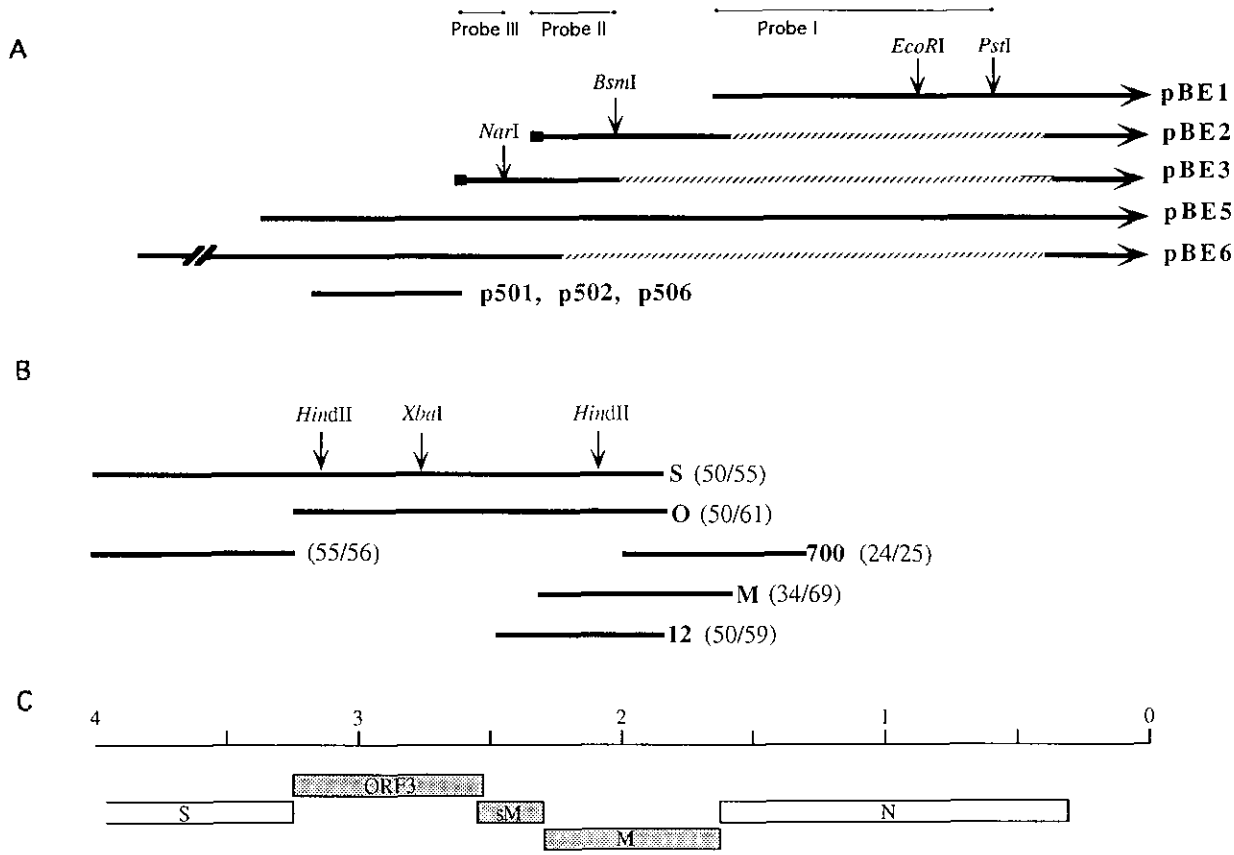

FIG. 1. cDNA clones used for determining the sequence of the PEDV genome downstream of the S gene. (A) Br1/87 clones: the sequenced parts of the indicated clones appear as solid lines. Clones pBE2 and pBE3 end within the PEDV leader sequence (solid box). Clones p501 to p506 are cloned RT/PCR fragments. Clones pBE1, pBE5, and pBE6 were shotgunned. Clones pBE2 and pBE3 were sequenced only at their extremities by using T3, T7, and three PEDV-specific primers. Probes I to III, see Materials and Methods. (B) CV777 clones: the lines represent PCR-generated clones. The primers used comprised a mixture of PEDV-specific primers and degenerate primers based on conserved regions of the coronavirus genome in the M, N, and S genes. Each line represents a set of clones rather than individual clones, of which some were derived from PEDV virion RNA while others were made from PEDV-infected cell mRNAs. Subclones were constructed using the restriction endonuclease sites indicated and clones were sequenced using vector and PEDV sequence-specific primers. (C) Deduced PEDV genomic organization. Newly identified ORFs are shown as shaded boxes.

gene *Pfu* polymerase for the greater accuracy and heat stability of this enzyme as compared with the Perkin–Elmer Cetus Amplitaq we had used previously, and also because the P50/P55 and P55/P56 reactions failed to work with Amplitaq despite using a wide range of different experimental conditions.

Identification of ORFs

The whole sequence of PEDV shown in Fig. 2 was determined on both strands of at least two clones for each of the Br1/87 and CV777 strains. The 1740 nt-long sequence extends upstream from the N gene, the sequence of which we reported recently. The 5′-most ORF (113 nt) was assumed to correspond to the 3′ end of the S gene, based on homology with other coronavirus S genes, on the position with respect to the other ORFs and on the presence of a cysteine cluster typical of the endodomain of the coronavirus S protein (review Spaan *et al.*, 1988). Three major ORFs were identified downstream of this ORF. The first one (nt 116 to 784), designated ORF3, has the coding capacity for a 223 aa-long polypeptide, or possibly 224 aa as discussed below. The next two ORFs have the potential to encode 76 and 226 aa-long polypeptides, homologous to the coronavirus structural proteins sM (Godet *et al.*, 1992) and M (review Spaan *et al.*, 1988), respectively. Alignments of the sM and M sequences revealed a 54 and 57% identity, respectively, for the PEDV-HCV 229E comparison, and 29 and 53%, respectively, for the PEDV-TGEV comparison (sM sequence data: Rasschaert *et al.*, 1987; Raabe and Siddell, 1989; M sequence data: see legend of Fig. 7). The nucleotide sequences of CV777 and Br1/87 were nearly identical: only 4 nt substitutions were noted, 2 in ORFs 3 and 2 in the M gene, all of them leading to amino acid changes (see Fig. 2). The sM gene was strictly conserved between the two isolates. However, as described in the next section, various deletions (I, II, and III in Fig. 2) were observed in ORF3.

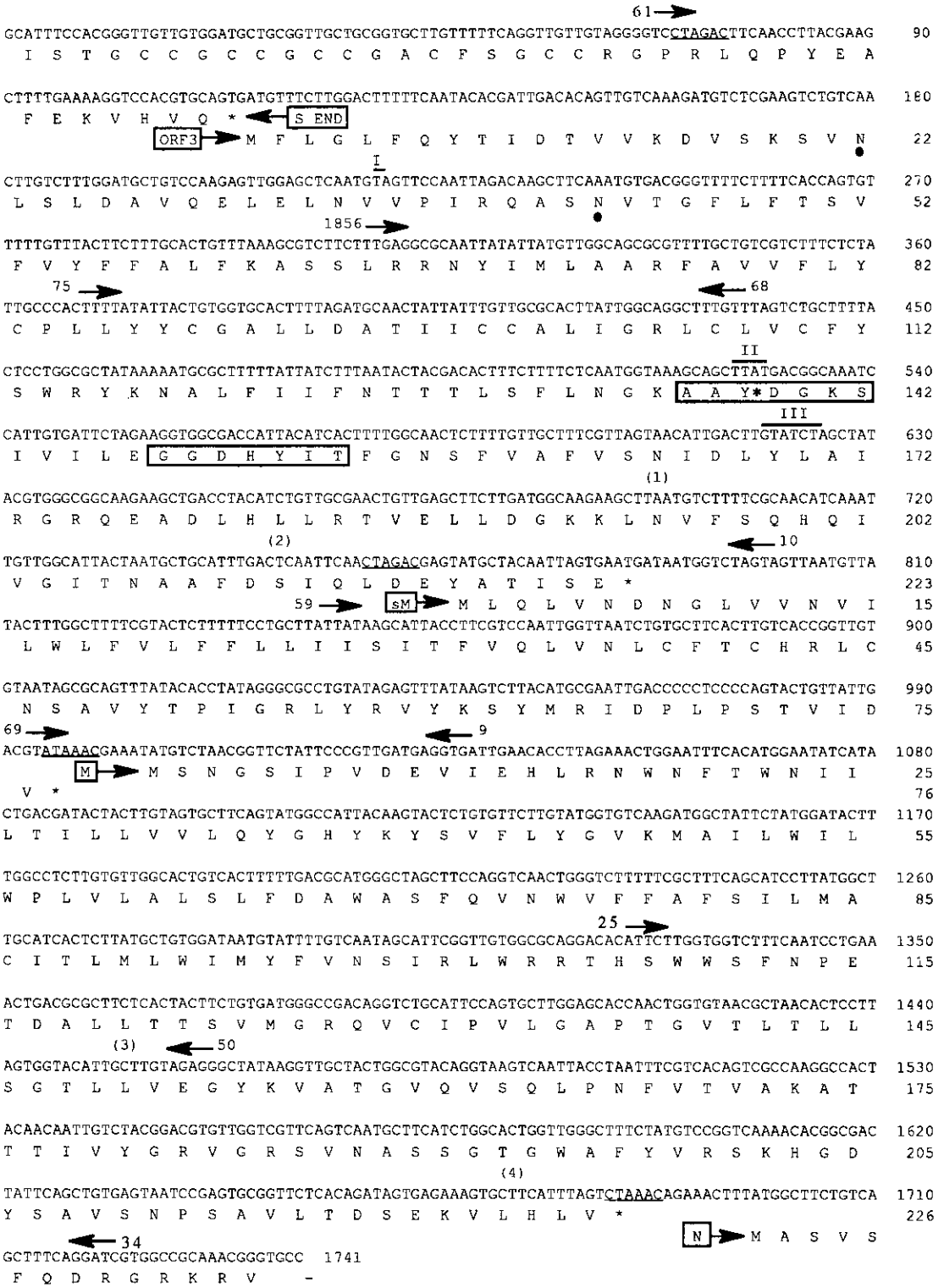

FIG. 2. Nucleotide sequence of the PEDV genome region between the S and N genes. The predicted amino acid sequences of the polypeptides encoded by the major ORFs are shown below the nucleotide sequence in single-letter code. The proposed consensus intergenic sequences are underlined. Differences between the two isolates or between different clones are indicated by arabic numbers above the sequence: (1) T not A in CV777, causing N to Y amino acid change; (2) G not T in pBE3 Br1/87 clone, causing S to A amino acid change; (3) C not T in CV777 and pBE3 Br1/87 clones, causing F to L change; (4) C not T in pBE3 Br1/87 clone causing L to P amino acid change. Variable regions in ORF3 are shown by underlined roman numbers. Arrows above the nt sequence indicate the position of primers. Points indicate potential Asn-linked glysosylation sites. (*) Additional Y amino acid residue predicted in the potential ORF3 224 aa polypeptide (see Results). Two functional motifs are boxed (see Discussion). The sequence data in the EMBL Data Library are available under Accession No. Z24733.

A

cDNA Clones	Variation I	Variation II	Variation III	Predicted polypeptide (aa)
	117-	524-	617-	Var I, Var II, Var III
*07, 018	TGTAGTT	GCTTA..ATGA	TGTATCTAGCT	137 106
pBE5,p501	TGTAGTT	GCTT....TGA	TGTATCTAGCT	143
*020	TGTAGTT	GCTTATTATGA	TGTATCTAGCT	153
*06,017	TG..GTT	GCTTATTATGA	TGTATCTAGCT	208
pBE6,p502,p506	TGTAGTT	GCTTATTATGA	TG.....CT	213
*S	TGTAGTT	GCTTA...TGA	TGTATCTAGCT	**223**
pBE5+pBE6	TGTAGTT	GCTTATTATGA	TGTATCTAGCT	224

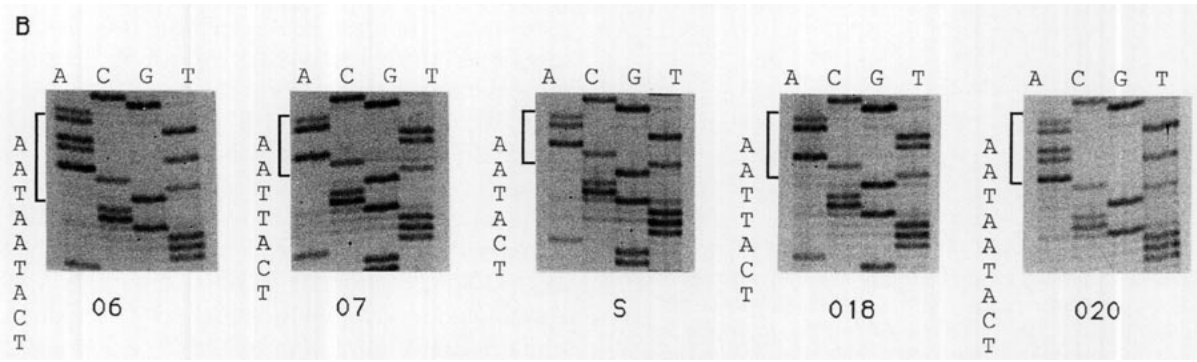

FIG. 3. (A) Nucleotide alignment of the variable regions observed in PEDV ORF3 cDNA clones. CV777 clones are denoted by an asterisk. The nucleotide position of each variable region is indicated on the left. The resulting predicted polypeptides are shown at the right side of the figure; the region homologous to the 223 aa, reference polypeptide (see text) is shown as a solid line. pBE5 + pBE6 stands for hypothetical complete ORF to pBE5 + pBE6. (B) Sequence of five CV777 clones in the variation II region. The autoradiogram shows the antisense sequence, as determined from P10.

Nucleotide variation in the ORF3 clones

A striking feature of our sequence data was the presence within ORF3 of at least 3 regions of variability, consisting of short deletions (2 to 7 nt). These variations are shown in Fig. 3, together with the length of the predicted polypeptides. All these deletions were mutually exclusive. Deletions were present in both strains of the virus sequenced. The largest observed ORF could encode a 223 aa product (sequence displayed in Fig. 1) and this was taken as a reference for the description of the other predicted polypeptides. However, removal of any of the deletions mentioned above would lead to an ORF encoding a polypeptide with one extra amino acid (Tyr between positions 138 and 139 as numbered in Fig. 2) compared to the 223 amino acid reference polypeptide.

Variation I, observed in CV777 only (clones O6 and O17), consists of a 2 nt deletion and results in a 208 aa-long polypeptide, of which the last 192 are homologous to the 223 aa species. Three sizes of deletions were found in the region of variation II; clones O7 and O18 have the coding capacity for 137 and 106 aa polypeptides, which are essentially identical to the amino- and carboxy-sequences of the 223 aa species, respectively. Clone pBE5 encodes a 143 aa polypeptide corresponding to the 223 aa species truncated by the 86 aa nearest to the carboxy-terminus. Clone S encodes the 223 aa reference polypeptide. The 7 nt deletion of the variation III region (pBE6 clone) results in a fused polypeptide 213 aa-long which has only the amino part in common with the 223 aa species. Since part of the variations presented in Fig. 3 have been confirmed by sequencing of cloned RT/PCR fragments encompassing the relevant regions (p501, p502, and p506), they are unlikely to reflect cloning artifacts. Furthermore, direct sequencing of uncloned RT/PCR products revealed the presence of two templates, as expected if deleted and undeleted viral RNAs were present in roughly equivalent proportions in Br1/87-infected cells (data not shown). The six CV777 clones covering ORF3 were derived from both infected cell RNA (S, O17, O18, O20) and from virion RNA (O6, O7), thus excluding the possibility of the deleted RNA purely being a feature of infected cell RNA. Only one cDNA clone (clone O20 derived from CV777) had none of the above mentioned short deletions. However, it turned out to have a 37 nt deletion at the very 5′ end of ORF3

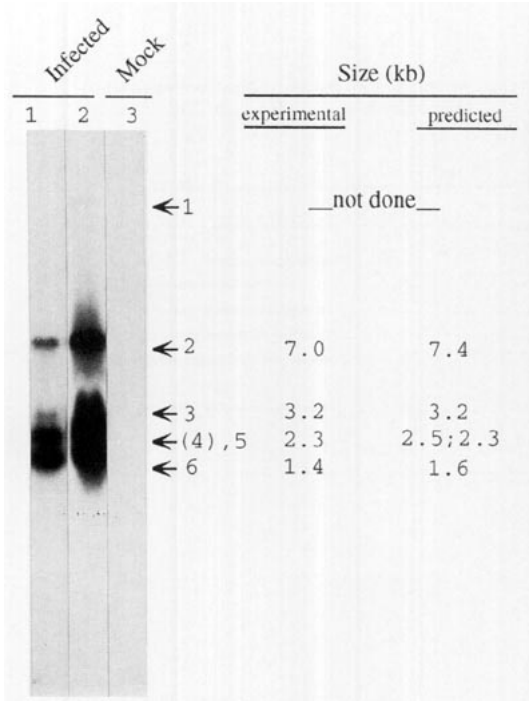

FIG. 4. Northern blot analysis of PEDV intracellular RNAs. (Left) Total RNA was electrophoresed in formaldehyde–agarose gel and transferred to a nylon membrane. A [^{32}P]-labeled cDNA (clone pBE1 insert), corresponding to the N gene and the 3′ non-coding region, was used as the hybridization probe. RNA from mock-infected (lane 3) or PEDV-infected (lanes 1 and 2) cultures. Lane 1, 10 μg, 4-hr exposure. Lane 2, 30 μg, 4-day exposure. (Right) Sizes of PEDV mRNAs: comparison between predicted and experimental values.

(position 102 to 138 in Fig. 2). As a consequence, an ORF 153 aa long is created which is completely homologous with the 223 aa ORF species.

PEDV subgenomic RNAs

Northern blot analysis was performed to see whether the pattern of subgenomic RNAs transcribed in PEDV-infected cells was consistent with those of other coronaviruses. When using a probe covering the genome 3′ end, four major RNA species were detected in addition to the genomic RNA (of >20 kb), as expected for RNAs having the nested set structure typical of the Coronaviridae (Fig. 4). An identical pattern was observed for the two isolates (not shown). Their respective sizes were in good agreement with those predicted from the nucleotide sequence, assuming that the initiation of transcription occurs at the hexameric intergenic sequences indicated in Fig. 2. This would imply that the S, ORF3, M, and N polypeptides are expressed from mRNAs 2, 3, 5, and 6, respectively. The mRNA species potentially encoding the sM polypeptide comigrated with RNA species 5, but could be observed by probing with an sM-specific sequence (not shown). Finally, no RNA species could be detected between the genomic RNA and the subgenomic RNA 2, thus suggesting that PEDV has no additional genes upstream of the S gene except that encoding the polymerase. This finding is in agreement with the lack of hemagglutinase activity associated with PEDV (Utiger, personal communication) since, if this gene is present within a coronaviral genome, it is invariably located between the S and polymerase genes.

DISCUSSION

In this study we have cloned and analyzed the genome region located between the S and N genes of the two PEDV virulent isolates, Br1/87 and CV777. The nucleotide sequences of the two isolates were 99.9% identical, consistent with our previous results for the N gene region (Bridgen *et al.*, 1993a). The 1740 nt-long sequenced stretch revealed two ORFs encoding the coronavirus transmembrane proteins M and sM, and a large, potentially single ORF, which we have designated ORF3 on the basis of the mRNA pattern seen in infected cells. The deduced structural characteristics of the membrane protein M and of the minor structural protein sM are very similar to those of the other coronaviruses. The N-terminal domain of the PEDV M protein is predicted (i) to have no signal sequence such as is found in TGEV (Laude *et al.*, 1987); (ii) to extend 17 amino acids from the virion envelope, i.e., one of the shortest ectodomains together with that of the HCV 229E M protein (16 residues) (Jouvenne *et al.*, 1990); and (iii) to have one N-linked glycosylation site (Asn at position 3, Fig. 2). The predicted features of the M apoprotein are consistent with experimental findings. For example, the MW of the PEDV M apoprotein was predicted to be 25 kDa, similar to the reported M_r (Utiger *et al.*, 1993), and recent glycosylation inhibition experiments have confirmed the N-linked nature of the M protein glycosylation (Utiger, personal communication). Comparison of the predicted M and sM proteins of PEDV with those of other coronaviruses yielded similar results to those of the N protein, with higher homologies to the TGEV, FIPV, and CCV proteins than to those of murine hepatitis virus (MHV), bovine coronavirus (BCV), and infectious bronchitis virus (IBV).

In addition to the genes for structural proteins, coronavirus genomes contain a number of ORFs thought to encode non-structural proteins, which are currently at best poorly characterized (review Lai, 1990). The number and position of such ORFs differ among the coronaviruses, generally two of them being located between the S and sM genes. Thus the presence of a potentially single ORF leaving little non-coding space between S and sM appears to be a distinctive feature of the PEDV genome. With the exception of certain coronavirus strains having altered ORFs, as is discussed later, such a situation has only been encountered for HCV OC43. In the latter case, however, a

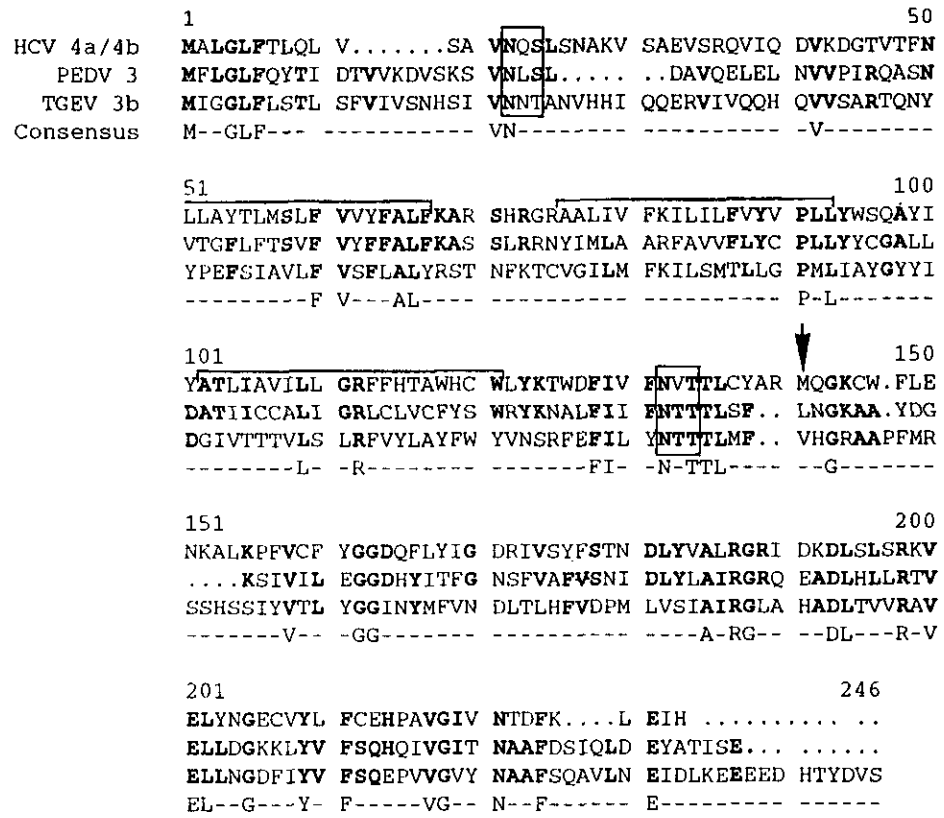

FIG. 5. Alignment of the predicted amino acid sequence of PEDV ORF3 with the ORFs 4a + 4b of HCV 229E and ORF3b of TGEV. The HCV 229E ORF4a and ORF4b polypeptides are presented in tandem, without spacing. The arrowhead indicates the first Met residue of the HCV ORF4b polypeptide. Identical amino acid residues between PEDV and at least one of the other two sequences are in boldface. Potential N-glycosylation sites conserved in the 3 sequences are boxed. The three large, overall markedly hydrophobic regions predicted in the three sequences, which are shown as the shaded region in Fig. 6, are overlined. The alignment was constructed using the UWGCG Pileup and Pretty programs and was adjusted manually. Sequence data for TGEV and HCV are according to Kapke *et al.* (1988) and Raabe and Siddell (1989).

short (11 residues), vestigial ORF, which is followed by a 47 nt stretch homologous to the leader sequence, is found downstream of the S gene. Since this may reflect a recent recombinational event (Mounir and Talbot, 1992) it remains to be discovered whether other OC43 isolates show the same organization.

We would, however, also like to suggest from our results for PEDV that the HCV 229E genome also potentially encodes, or at one time encoded, a single ORF between the S and sM genes. Pairwise alignments performed in order to see whether counterparts of PEDV ORF3 were present in the corresponding genome region of other coronaviruses yielded positive results for four viruses: (i) The ORF3b (244 aa) of TGEV (Kapke *et al.*, 1988), ORF3-1 of porcine respiratory coronavirus PRCV (Rasschaert *et al.*, 1990), and ORF3b of CCV (Horsburgh *et al.*, 1982) showed a 30% identity. (ii) The chimaeric polypeptide ORF4, representing ORFs 4a (133 aa) and 4b (88 aa) of HCV 229E fused tail-to-head, showed an overall 35% identity (Raabe and Siddell, 1989). This analysis also revealed a substantial, previously unnoticed identity (28%) between the HCV ORF4a/b and TGEV ORF3b. A multialignment of the above sequences is shown in Fig. 5 in which 42 residues fully conserved between the three viruses have been identified. In addition, the aligned sequences share several features. First, as is also illustrated in Fig. 6, several long hydrophobic stretches are present at homologous positions in the amino terminal half of these proteins. Second, two Asn-linked glycosylation motifs are conserved, one of which is located upstream of the first hydrophobic stretch. Such characteristics are suggestive of a membrane protein having at least one transmembrane segment. The first hydrophobic stretch, comprising 17 consecutive, nonpolar residues, could well have this role. It was also noted that six Cys residues are clustered within the downstream hydrophobic stretch of PEDV ORF3. The presence of a cluster of Cys residues near the carboxy-end of the anchor region and serving as palmitoylation sites is not unfrequent in transmembrane proteins (Schmidt, 1982). However, no such residues are present in the corresponding domains of TGEV and HCV. Finally, an interesting possibility is that the ORF3 product might represent an additional virion envelope protein which has remained unrecognized since it is very similar in size to the major structural protein M. This, however, is unlikely since virus particles were found

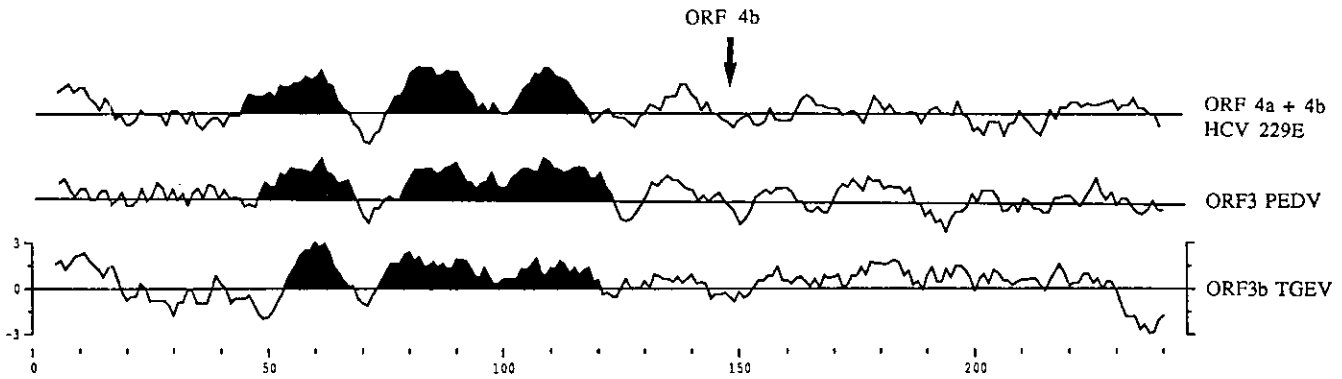

FIG. 6. Compared hydropathy of the HCV, PEDV, and TGEV polypeptides shown in Fig. 5. The curves are the average of a residue specific hydrophobicity index over a window of nine residues. Peaks extending upward represent hydrophobic regions; the shaded areas represent potential membrane-associated regions. The analysis was performed using the UWGCG PepPlot program.

defective for the full-length ORF3 RNA and since no homolog of this protein is present in MHV, IBV or BCV.

Two possible functional motifs were observed in the PEDV ORF3 sequence (see Fig. 2): (i) An arginase signature (at residues GGDHYIT) which is partially conserved in the TGEV and HCV 229E proteins; (ii) A possible ATP/GTP binding site, present in the putative 224 aa polypeptide at residues AAY(Y)DGKS; however, this motif occurs in the variation II region and is not conserved in TGEV and HCV proteins. As mentioned under Results, an important variability was observed among the cDNA clones spanning ORF3. Due to the presence of various deletions within the ORF, most of them encoded a truncated product. These deletions could well imply that the PEDV ORF3 product is not required for replication in cell culture. We have no good explanation for the clustering of the deletion sites to a few specific loci. Variable region I occurs at a region of extensive secondary structure in the viral RNA, but this does not apply to the other regions. Deletions within genes are not unknown in coronaviruses, however. Our findings are reminiscent of the apparent flexibility (deletion, premature stop codon, or lack of start codon) emerging from sequence data of different strains of MHV (ORF5, Yokomori and Lai, 1991; ORF4, Weiss *et al.*, 1993.), TGEV (ORF3, Rasschaert *et al.*, 1987; Britton *et al.*, 1993; Wesley *et al.*, 1990) and HCV 229E (ORF4, Jouvenne *et al.*, 1992).

One interesting feature of the variations observed in the PEDV ORF 3 was that, although deletions were sometimes present at the same location in the two PEDV strains studied, the nature of the deletion was never the same in the two strains. This favors the view that CV777 and Br1/87 represent two distinct isolates of PEDV which originated from the same virus population but which later evolved independently. These two isolates, although in many respects virtually identical, differ in the markedly higher efficiency of replication in cell culture of CV777 compared to Br1/87 (data not shown). One possible explanation of this could be that the latter isolate possesses no full-length ORF3 (capable of encoding a 223 or 224 aa polypeptide). One question which certainly requires investigation is whether the ORF3 product is expressed as a 223 or 224 aa polypeptide in non-cell culture-adapted isolates of PEDV and whether such isolates have internal deletions in the ORF.

Based on the size and distribution of the plus sense RNAs synthesized in infected cells, each of the ORFs presently identified in the PEDV genome is likely to be expressed from a functionally monocistronic mRNA. The initiation of transcription of the subgenomic RNAs is likely to occur at the "unique" hexameric motif present upstream of each ORF, similar to that found in other coronaviruses (review Lai, 1990). This appeared to be the case for the mRNA 4, which encodes the sM gene, since a 40 nt stretch 50% homologous to the TGEV leader sequence was found immediately upstream of the motif CTAGAC at the 5′ end of clone pBE3 (not shown). Similarly, a 30 nt, leader-like sequence is present immediately upstream of the motif CTAAAC at the 5′ end of clone pBE2, which is likely to represent a cDNA copy of mRNA5, encoding the M protein. This indicates that the proposed motif is functional, thus implying that the conserved intergenic sequence is more flexible than previously thought. This finding is consistent with that of a recent report, which shows that, in a stable context, each nt of the MHV conserved motif can be mutated without affecting the efficiency of transcription (Joo *et al.*, 1992).

The findings presented in this study further strengthen the view that PEDV belongs to the same subgroup, tentatively designated genetic subset, as HCV 229E and the antigenically clustered viruses TGEV, PRCV, FIPV, feline enteric coronavirus and CCV. This subset of the coronavirus genus has been proposed to acknowledge the fact that the above viruses, although exhibiting few if any subset-specific epitopes, share several distinguishing genetic traits (Duarte *et al.*, 1993). This proposal is also supported by

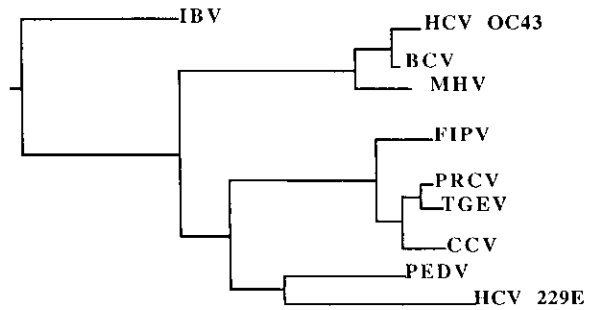

FIG. 7. Neighbor-joining distance matrix tree based on the M amino acid sequences of nine coronaviruses. The regions used for the analysis were the C terminal last: (i) 192 aa of the M protein of HCV 229E, TGEV, PRCV, FIPV, CCV, and PEDV (Jouvenne *et al.*, 1990; Laude *et al.*, 1987; Rasschaert *et al.*, 1990; Venemma *et al.*, 1991; Horsburgh *et al.*, 1992; this paper); (ii) 189 aa of the M protein of BCV and OC43 (Lapps *et al.*, 1987; Mounir and Talbot, 1992); (iii) 186 aa of the M protein of MHV-JHM (Pfleiderer *et al.*, 1986); (iv) 188 aa of the M protein of IBV (Boursnell *et al.*, 1984). This method gives an unrooted tree, which means that the direction of evolution through the tree is not shown. To infer the root we assume a degree of constancy in the molecular clock, and place the root, half way along the longest branch. The final tree was made by hand, maintaining the distances of the unrooted tree. The horizontal branch lengths are indicative of the evolutionary distances. The vertical lines are only for clarity.

the phylogenetic analysis of the conserved region of the M protein (Fig. 7) and of the sM protein (not shown). It is also apparent from this figure that PEDV is more closely related to the human respiratory virus 229E than to the other porcine enteric virus, TGEV. Indeed, pairwise alignment of the amino acid sequences deduced for each of the three newly identified ORFs of PEDV (M, sM and ORF3) with their counterparts revealed a higher identity with HCV 229E than with TGEV, as already reported for the N protein (Bridgen *et al.*, 1993a). The PEDV sM protein, in particular, is far more homologous to that of HCV 229E than to that the TGEV sM protein. In addition, neither PEDV nor HCV 229E have an ORF like the TGEV ORF3a or an ORF downstream of the N gene. Thus, it is likely that these two viruses diverged later than TGEV (and its related viruses) from their common ancestor. However, since intergenomic recombination is a recognized feature of coronavirus evolution (review Lai, 1990), it remains to be investigated whether such a conclusion would hold true for the remaining part of the PEDV genome.

Finally, the evolutionary relationship demonstrated between PEDV, HCV 229E and TGEV raises stimulating questions with respect to the virus tropism. First, it might be worth examining whether PEDV multiplication could also take place in the respiratory tract or whether it is fully restricted to the intestine, where the major histopathological changes are located. Second, TGEV and HCV 229E recognize aminopeptidase N (APN) as a major membrane receptor (Delmas *et al.*, 1992; Yeager *et al.*, 1992) and this in a strictly species-specific manner (Delmas *et al.*, in preparation). This leads to the speculation that PEDV, despite the lack of permissiveness of the APN-expressing porcine cell lines, may gain entry into the enterocytes by utilization of the same molecule.

ACKNOWLEDGMENTS

M.D. and H.L. thank the following people from their laboratory: J. Gelfi for her expert assistance in propagating the virus, P. Lambert and N. Woloszyn for their help in nucleotide sequencing, and F. Lefèvre for advice in using the PHYLIP package. M. Duarte (C.I.V.E.T., Pando, Uruguay), Ph.D. student, held a fellowship from the French Government. K.T., A.B., and M.A. thank A. Utiger, M. Rosskopf and M. Schwyzer for advice during the course of this work. K. Tobler was supported by Grant 012.91.7 from the Swiss Federal Veterinary Services.

REFERENCES

BOURSNELL, M. E. G., BROWN, T. D. K., and BINNS, M. M. (1984). Sequence of the membrane protein gene from avian coronavirus IBV. *Virus Res.* **1,** 303–313.

BRIDGEN, A., DUARTE, M., TOBLER, K., LAUDE, H., and ACKERMANN, M. (1993a). Sequence determination of the nucleocapsid protein gene of the porcine epidemic diarrhoea virus confirms that this virus is a coronavirus related to human coronavirus 229E and porcine transmissible gastroenteritis virus. *J. Gen. Virol.* **74,** 1795–1804.

BRIDGEN, A., TOBLER, K., and ACKERMANN, M. (1993b). Identification of coronaviral conserved sequences and application to viral genome amplification. *In* "Coronaviruses: Molecular Biology and Virus–Host Interactions" (H. Laude and J. V. Vautherot, Eds.). Plenum, New York, in press.

BRITTON, P., KOTTIER, S., CHEN, C.-M., POCKOK, D. H., SALMON, H., and AYNAUD, J. M. (1993). The use of PCR genome mapping for the characterisation of TGEV strains. *In* "Coronaviruses: Molecular Biology and Virus–Host Interactions" (H. Laude and J. V. Vautherot, Eds.). Plenum, New York, in press.

DEBOUCK, P., and PENSAERT, M. (1980). Experimental infection of pigs with a new porcine enteric coronavirus, CV 777. *Am. J. Vet. Res.* **41,** 219–223.

DEBOUCK, P., PENSAERT, M., and COUSSEMENT, W. (1981). The pathogenesis of an enteric infection in pigs, experimentally induced by the coronavirus-like agent, CV 777. *Vet. Microbiol.* **6,** 157–167.

DEBOUCK, P. (1982). Prevalence of the porcine epidemic diarrhea (PED) virus in the pig population of different countries. Proc. 7th Int. Congr. Pig Vet. Soc., Mexico City.

DELMAS, B., GELFI, J., L'HARIDON, R., VOGEL, L. K., SJOSTROM, H., NOREN, O., and LAUDE, H. (1992). Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV. *Nature* **357,** 417–420.

DESSEN, P., FONDRAT, C., VALENCIEN, C., and MUGNIER, C. (1990). BISANCE A french service for access to biomolecular sequence databases. *Comput. Appl. Biosci.* **6,** 355–356.

DEVEREUX, J., HAEBERLI, P., and SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12,** 387–395.

DUARTE, M., GELFI, J., LAMBERT, P., RASSCHAERT, D., and LAUDE, H. (1993). Genomic organization of porcine epidemic diarrhoea virus (PEDV). *In* "Coronaviruses: Molecular Biology and Virus–Host Interactions" (H. Laude and J. V. Vautherot, Eds.). Plenum, New York, in press.

EGBERINK, H. F., EDERVEEN, J., CALLEBAUT, P., and HORZINEK, M. C. (1988). Characterization of the structural proteins of porcine epi-

zootic diarrhea virus, strain CV 777. *Am. J. Vet. Res.* **49,** 1320–1324.

FELSENSTEIN, J. (1991). "Phylip: Phylogenetic Inference Package, Version 3.4." University of Washington, Seattle, WA.

GODET, M., L'HARIDON, R., VAUTHEROT, J. F., and LAUDE, H. (1992). TGEV coronavirus ORF4 encodes a membrane protein that is incorporated into virions. *Virology* **188,** 666–675.

HAMILTON, B., PALAZZOLO, M., and MEYEROWITZ, M. (1991). Rapid isolation of long cDNA clones from existing libraries. *Nucleic Acids Res.* **18,** 1951.

HAVE, P., MOVING, V., SVANSSON, V., UTTENTHAL, A., and BLOCH, B. (1992). Coronavirus infection in mink (*Mustela vison*). Serological evidence of infection with a coronavirus related to transmissible gastroenteritis virus and porcine epidemic diarrhea virus. *Vet. Microbiol.* **31,** 1–10.

HIGGINS, D., and SHARP, P. (1989). Fast and sensitive multiple sequence alignments on a micro computer. *Comput. Appl. Biosci.* **5,** 151–153.

HOFMANN, M., and WYLER, R. (1988). Propagation of the virus of porcine epidemic diarrhea in cell culture. *J. Clin. Microbiol.* **26,** 2235–2239.

HORSBURGH, B. C., BRIERLEY, I., and BROWN, T. D. K. (1992). Analysis of a 9.6 kb sequence from the 3' end of canine coronavirus genomic RNA. *J. Gen. Virol.* **73,** 2849–2862.

JOO, M., and MAKINO, S. (1992). Mutagenic analysis of the coronavirus intergenic consensus sequence. *J. Virol.* **66,** 6330–6337.

JOUVENNE, P., RICHARDSON, C. D., SCHREIBER, S. S., LAI, M. M., and TALBOT, P. J. (1990). Sequence analysis of the membrane protein gene of human coronavirus 229E. *Virology* **174,** 608–612.

JOUVENNE, P., MOUNIR, S., STEWARD, J., RICHARDSON, C. D., and TALBOT, P. (1992). Sequence analysis of human coronavirus 229E mRNA4 and 5: Evidence for polymorphism and homology with myelin basis protein, *Virus Res.* **22,** 125–141.

KAPKE, P. A., TUNG, F. Y. T., and BRIAN, D. A. (1988). Nucleotide sequence between the peplomer and matrix protein genes of the porcine transmissible gastroenteritis coronavirus identifies three largeopen reading frames. *Virus Genes* **2,** 293–294.

KNUCHEL, M., ACKERMANN, M., MULLER, H., and KIHM, U. (1992). An ELISA for detection of antibodies against porcine epidemic diarrhoea virus (PEDV) based on the specific solubility of the viral surface glycoprotein. *Vet. Microbiol.* **32,** 117–134.

LAI, M. M. C. (1990). Coronavirus: Organization, replication and expression of genome. *Annu. Rev. Microbiol.* **44,** 303–333.

LAPPS, W., HOGUE, B. G., and BRIAN, D. A. (1987). Sequence analysis of the bovine coronavirus nucleocapsid and matrix protein genes. *Virology* **157,** 47–57.

LAUDE, H., RASSCHAERT, D., and HUET, J. C. (1987). Sequence and N-terminal processing of the transmembrane proteine E1 of the coronavirus transmissible gastroenteritis virus. *J. Gen. Virol.* **68,** 1687–1693.

MOUNIR, S., and TALBOT, P. J. (1992). Sequence analysis of the membrane protein gene of human coronavirus OC43 and evidence for O-glycosylation. *J. Gen. Virol.* **73,** 2731–2736.

PENSAERT, M. B., and DEBOUCK, P. (1978). A new coronavirus-like particle associated with diarrhea in swine. *Arch. Virol.* **58,** 243–247.

PENSAERT, M. B., DEBOUCK, P., and REYNOLDS, D. J. (1981). An immunoelectron microscopic and immunofluorescent study on the antigenic relationship between the coronavirus-like agent, CV 777, and several coronaviruses. *Arch. Virol.* **68,** 45–52.

PFLEIDERER, M., SKINNER, M. A., and SIDDELL, S. G. (1986). Coronavirus MHV-JHM: nucleotide sequence of the mRNA that encodes the membrane protein. *Nucleic Acids Res.* **14,** 6338.

RAABE, T., and SIDDELL, S. (1989). Nucleotide sequence of the human coronavirus HCV 229E mRNA 4 and mRNA 5 unique regions. *Nucleic Acids Res.* **17,** 6387.

RASSCHAERT, D., GELFI, J., and LAUDE, H. (1987). Enteric coronavirus TGEV: partial sequence of the genomic RNA, its organization and expression. *Biochimie* **69,** 591–600.

RASSCHAERT, D., DUARTE, M., and LAUDE, H. (1990). Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. *J. Gen. Virol.* **71,** 2599–2607.

SAITOU, N., and NEI, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.

SAMBROOK, J., FRITSCH, E. F., and MANIATIS, T. (1989). "Molecular Cloning: A Laboratory Manual," 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

SCHMIDT, M. F. (1982). Acylation of viral spike glycoproteins: A feature of enveloped RNA viruses. *Virology* **116,** 327–338.

SHORT, J., FERNANDEZ, J., SORGE, J., and HUSE, W. (1988). 1ZAP: a bacteriophage 1 expression vector with *in vivo* excision properties. *Nucleic Acids Res.* **16,** 7583–7599.

SPAAN, W., CAVANAGH, D., and HORZINEK, M. C. (1988). Coronaviruses: structure and genome expression. *J. Gen. Virol.* **69,** 2939–2952.

UTIGER, A., ROSSKOPF, M., GUSCETTI, F., and ACKERMANN, M. (1993). Preliminary characterization of a monoclonal antibody specific for a viral 27 kD glycoprotein family synthesized in porcine epidemic diarrhoea virus infected cells. *In* "Coronaviruses: Molecular Biology and Virus–Host Interactions" (H. Laude and J. V. Vautherot, Eds.). Plenum, New York, in press.

VAQUERO, C., SANCEAU, J., CATINOT, L., ANDREU, G., FALCOFF, E., and FALCOFF, R. (1982). Translation of mRNA from phytohemagglutinin-stimulated human lymphocytes: Characterization of interferon mRNAs. *J. Interf. Res.* **2,** 217–218.

VENNEMA, H., DE GROOT, R., HARBOUR, D., HORZINEK, M., and SPAAN, W. (1991). Primary structure of the membrane and nucleocapsid protein genes of feline infectious peritonitis virus and immunogenicity of recombinant vaccinia virus in kittens. *Virology* **131,** 327–335.

WESLEY, R., WOODS, R., and CHEUNG, A. (1990). Genetic basis for the pathogenesis of transmissible gastroenteritis virus. *J. Virol.* **64,** 4761–4766.

WEISS, S. R., ZOLTICK, P. W., and LEIBOWITZ, J. L. (1993). The ns4 gene of mouse hepatitis virus (MHV), strain A59 contains two ORFs and thus differs from ns4 of the JHM and S strains. *Arch. Virol.* **129,** 301–309.

WOOD, W., GITSCHIER, J., LASKY, L., and LAWN, R. (1985). Base composition-independent hybridization in tetramethylammonium chloride: A method for oligonucleotide screening of highly complex gene libraries. *Proc. Natl. Acad. Sci. USA* **82,** 1585–1588.

YALING, Z., EDERVEEN, J., EGBERINK, H., and PENSAERT, M. (1988). Porcine epidemic diarrhoea virus (CV777) and feline infectious peritonitis virus (FIPV) are antigenically related. *Arch. Virol.* **102,** 63–71.

YEAGER, C. L., ASHMUN, R. A., WILLIAMS, R. K., CARDELLICHIO, C. B., SHAPIRO, L. H., LOOK, A. T., and HOLMES, K. V. (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357,** 420–423.

YOKOMORI, K., and LAI, M. C. C. (1991). Mouse hepatitis virus S RNA sequence reveals that nonstructural proteins ns4 and ns5a are not essential for murine coronaviruses replication. *J. Virol.* **65,** 5605–5608.

YOUNG, R., BLOOM, B., GROSSKINSKY, C., IVANYI, J., THOMAS, D., and DAVIS, R. (1985). Dissection of *Mycobacterium tuberculosis* antigens using recombinant DNA. *Proc. Natl. Acad. Sci. USA* **82,** 2583–2587.