Journal of General Virology (1992), **73**, 2849–2862. *Printed in Great Britain*

Analysis of a 9·6 kb sequence from the 3′ end of canine coronavirus genomic RNA

Brian C. Horsburgh, Ian Brierley and T. David K. Brown*

Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, U.K.

We have analysed the organization of the 3′ end of the genomic RNA of canine coronavirus (CCV), a virus which has a close antigenic relationship to transmissible gastroenteritis virus (TGEV), porcine respiratory coronavirus (PRCV) and feline infectious peritonitis virus (FIPV). Genomic RNA isolated from CCV strain Insavc-1-infected A72 cells was used to generate a cDNA library. Overlapping clones, spanning approximately 9·6 kb [from the 3′ end of the polymerase gene, 1b, to the poly(A) tail] were identified. Sequencing and subsequent analyses revealed 10 open reading frames (ORFs). Three of these code for the major coronavirus structural polypeptides S, M and N; a fourth codes for a small membrane protein, SM, a putative homologue of the IBV structural polypeptide 3c, and five code for polypeptides, designated 1b, 3a, 4, 7a and 7b, homologous to putative non-structural polypeptides encoded in the TGEV or FIPV genomes. An extra ORF which had not hitherto been identified in this antigenic group of coronaviruses was designated 3x. Pairwise alignment of these ORFs with their counterparts in TGEV, PRCV and FIPV revealed high levels of identity and highlighted the close relationship between the members of this group of viruses.

Introduction

Canine coronavirus (CCV), a causative agent of enteritis in neonatal dogs, was first identified in 1971 (Binn *et al.*, 1974). The disease is characterized by infection of the absorptive epithelium of the villi and the onset of diarrhoea followed by villus atrophy (Keenan *et al.*, 1976). CCV belongs to the *Coronaviridae*, a family of enveloped viruses possessing a ssRNA genome of positive polarity. In infected cells, a set of 3′-coterminal subgenomic RNAs are produced and, as a result, the 5′ end of each mRNA contains unique sequence information not present on smaller RNAs in the nested set. Only this unique region of each mRNA is translated (reviewed by Spaan *et al.*, 1988), thus the mRNAs are, in principle, functionally monocistronic. Nevertheless, some mRNAs contain two or more coding regions within the unique sequence and thus may be functionally bi- or tricistronic (Brierley *et al.*, 1987; Liu *et al.*, 1991; Liu & Inglis, 1992). The CCV virion is known to contain at least four protein species: the 204K spike glycoprotein, S; the 32K membrane glycoprotein, M; the 9·2K small membrane protein, SM; and the 50K nucleocapsid protein, N (Garwes & Reynolds, 1981; Godet *et al.*, 1992).

CCV belongs to one of the major antigenic groups of coronaviruses (Siddell *et al.*, 1983; Spaan *et al.*, 1988) and is serologically related to feline infectious peritonitis virus (FIPV), feline enteric coronavirus (FECV), transmissible gastroenteritis virus (TGEV) and porcine respiratory coronavirus (PRCV) (Sanchez *et al.*, 1990). These viruses have been distinguished mainly by their host species of origin. It has been reported, however, that some strains of CCV can also infect cats (Barlough *et al.*, 1984; Stoddart *et al.*, 1988) and swine without causing any apparent disease (Woods & Wesley, 1986). Likewise, TGEV can also infect other species (Woods & Pedersen 1979; Norman *et al.*, 1970) and FIPV can infect swine (Woods *et al.*, 1981). This close relationship indicates that the viruses may have a common ancestor (Horzinek *et al.*, 1982; Sanchez *et al.*, 1990).

Molecular analysis has helped to elucidate some of the aspects of this phylogenetic relationship and some of the mechanisms involved in pathogenesis. TGEV, PRCV and FIPV have been characterized in some detail and the genes encoding the structural proteins have been cloned and sequenced (de Groot *et al.*, 1987; Vennema *et al.*, 1991; Britton *et al.*, 1988*a*, *b*; Rasschaert & Laude, 1987; Rasschaert *et al.*, 1990). A comparison of the available

The nucleotide sequence data reported in this paper have been submitted to GenBank and EMBL and assigned the accession number D13096.

FIPV amino acid sequences with the corresponding sequences of TGEV and PRCV has revealed that the structural genes are very closely related. For S the identities were 81·6% (TGEV) and 76% (PRCV), for M 84·4% and 85·9%, and for N 77% and 75·6%, respectively. This contrasts greatly with the relationship to murine hepatitis virus (MHV), a prototypic coronavirus from another antigenic group, where the identities for these polypeptides are 24%, 30% and 27%, respectively (Schmidt *et al.*, 1987; Skinner & Siddell, 1983; Armstrong *et al.*, 1984). Despite this high degree of similarity amongst the structural proteins of these three viruses there are, nevertheless, differences at the 3′ end of their viral genomes and in their subgenomic message organisation.

CCV is the least characterized virus from this antigenic group. Here we report the cloning and sequencing of 9·6 kb from the 3′ end of the RNA of the avirulent CCV strain Insavc-1, subgenomic message analysis and comparison to available TGEV, PRCV and FIPV sequence data which illuminate the evolutionary relationship of this family of viruses. The presented sequence, which includes all of the CCV coding information except for the polymerase region, represents the first report of cloning and sequencing of a canine coronavirus.

Methods

Virus and cells. Canine A72 cells and CCV strain Insavc-1 were obtained from Dr W. Baxendale (Intervet UK, Houghton, U.K.). A72 cells were grown in Gibco's Wellcome formula, a modified Eagle's medium supplemented with 10% foetal calf serum (FCS) containing penicillin (100 units/ml) and streptomycin (100 μg/ml) (MEM). Flasks (175 cm^2) of A72 cells were washed with PBS and infected with CCV at an m.o.i. of 0·1 in 10 ml MEM. Virus adsorption was allowed to proceed for 60 min at 37 °C and the inoculum was then replaced by MEM–10% FCS.

Preparation of CCV genomic and messenger RNAs. CCV genomic RNA was prepared as follows. At 48 h post-infection (p.i.) the culture supernatant was harvested, chilled to 4 °C and the cell debris removed by low-speed centrifugation (3000 *g* for 15 min). Virus was pelleted from the supernatant at 53000 *g* for 2 h (Beckman type 19 rotor) and the pellet homogenized in 6 M-guanidinium isothiocyanate, 0·5% *N*-lauroyl sarcosinate, 5 mM-sodium citrate. The mixture was layered onto a 5·7 M-CsCl pad and viral RNA pelleted by centrifugation (108000 *g* for 12 h at 18 °C). The RNA was dissolved in 10 mM-Tris–HCl, 0·1 mM-EDTA (TE) containing 0·1% SDS and stored at −70 °C. Samples were analysed on a 1% Tris–borate–EDTA agarose gel containing 0·1% SDS. A single species of high M_r RNA was identified with the characteristic mobility of coronavirus genomic RNA.

Subgenomic RNAs were prepared in a similar manner. Briefly, at 36 h p.i. the infected cells were chilled to 4 °C, washed three times with ice cold PBS then pelleted at 3000 *g* for 10 min. The cell pellet was homogenized in 6 M-guanidinium isothiocyanate, 0·5% *N*-lauroyl sarcosinate, 5 mM-sodium citrate then treated as described above.

Cloning of CCV genomic RNA

(i) *cDNA cloning.* A cDNA library from CCV genomic RNA was prepared by reverse transcription after priming with oligo(dT) and random pentanucleotides using the instructions and contents of the Boehringer Mannheim Biochemica cDNA synthesis kit. The resulting cDNA was blunt-ended using T4 DNA polymerase and ligated into the *Sma*I site of pUC119. Portions of the ligation mixture were transformed into *Escherichia coli* strain TG-1 and clones were identified by colour selection. Inserts of viral origin were confirmed by colony hybridization using cDNA prepared by random priming of CCV RNA as a probe. CCV-derived recombinant clones were analysed by restriction enzyme digestion and those containing inserts of 1·8 kb or greater in size were retained for further study.

(ii) *Polymerase chain reaction (PCR).* PCR-amplified fragments were obtained using cDNA:RNA heteroduplexes as template and oligonucleotides 7 and 8 (each of which contains a *Not*I site; Fig. 1) as primers. *Taq* DNA polymerase (Promega) was used to amplify the region of interest according to the recommendations of Sambrook *et al.* (1989) and 25 cycles (95 °C, 1 min; 60 °C, 1 min; and 72 °C, 2 min) were performed in a Techne PHC-1 machine. The generated DNA fragment was cleaved with *Not*I, gel-purified, ligated into the *Not*I site of pKL1 and transformed into *E. coli* strain TG-1. (pKL1 is a pUC-based vector with a modified polylinker and was a gift from Dr K. Law, University of Cambridge, U.K.)

Sequencing

(i) *M13 DNA sequencing.* DNA sequencing was performed by Sanger's dideoxynucleotide chain termination method as described by Bankier *et al.* (1987). Briefly, insert DNA was excised from vector sequences, self-ligated and sonicated in a cup-horn sonicator (Heat Systems, Ultrasonics). The sonicated DNA fragments were end-repaired with the Klenow fragment of *E. coli* DNA polymerase I and T4 DNA polymerase prior to size selection on a 1·2% agarose gel. Fragments in the size range 300 to 500 bp were purified and cloned into *Sma*I-digested, phosphatase-treated M13mp8. Shotgun sequence data were assembled using the SAP programs of Staden (1982) on a VAX 8350 and microVAX 3100 (Digital Equipment Corporation).

(ii) *Supercoiled DNA sequencing.* DNA templates were prepared as described by Lim & Pène (1988). CsCl-purified plasmid DNA (3 μg) was denatured with 0·15 M-NaOH and 0·15mM-EDTA for 30 min at 37 °C, then centrifuged through a Sepharose CL6-B column equilibrated in TE. Sequencing reactions were carried out on the eluate as described using the pUC forward and reverse primers.

(iii) *RNA sequencing.* Primer (50 pmol) was annealed to either 1 μg genomic RNA or 10 μg total infected cell RNA at room temperature for 15 min. Sequencing reactions were performed as described by Fichot & Girad (1990).

Northern blot hybridization. Total RNA extracted from CCV-infected cells was denatured for 15 min at 56 °C in 50% deionized formamide, 2·2 M-formaldehyde and 0·5 mM-EDTA. The samples were cooled on ice after the addition of loading buffer containing 0·5% SDS, 0·025% bromophenol blue and 25% glycerol. The samples were electrophoresed overnight in a horizontal submerged gel containing 1·1 M-formaldehyde and 0·8% agarose. RNA was blotted from the gel to a nitrocellulose filter (Schleicher and Schuell). Prehybridization was carried out in 5 × SSC (1 × SSC is 150m mM-sodium chloride and 15 mM-sodium citrate), 10 × Denhardt's solution (1 × Denhardt's solution is 0·02% polyvinylpyrrolidone, 0·02% Ficoll and 0·02% bovine serum albumin), 100 μg/ml sonicated salmon sperm DNA and 0·1% SDS for 2 h at 65 °C. Hybridization was carried out at 65 °C overnight after addition of a ^{32}P-radiolabelled DNA probe prepared by random priming the CCV-specific insert purified from pBH5 (Sambrook *et al.*, 1989). Following hybridization, the filter was washed twice at 65 °C with 2 × SSC, then washed three times at 42 °C with 0·2 × SSC, prior to exposure to X-ray film.

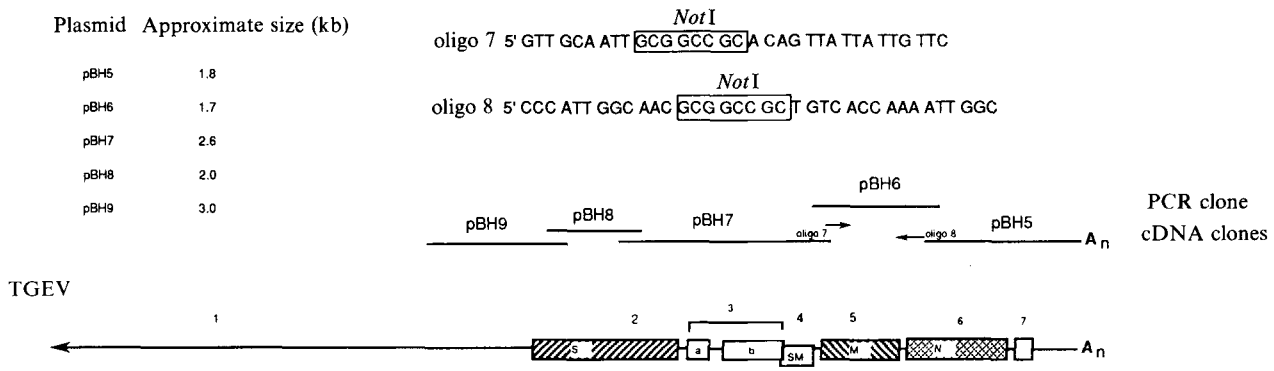


Fig. 1. Alignment of CCV cDNA clones with respect to the TGEV genome using partial sequence information. Oligonucleotides 7 and 8 were used as primers in a PCR reaction to obtain clone pBH6. Overlaps were confirmed by Southern blotting.

Southern blotting and other cloning procedures. These were carried out according to the protocols of Sambrook *et al.* (1989). Enzymes were used according to the manufacturers' specifications (Boehringer Mannheim and New England Biolabs).

Results

Generation and mapping of CCV clones

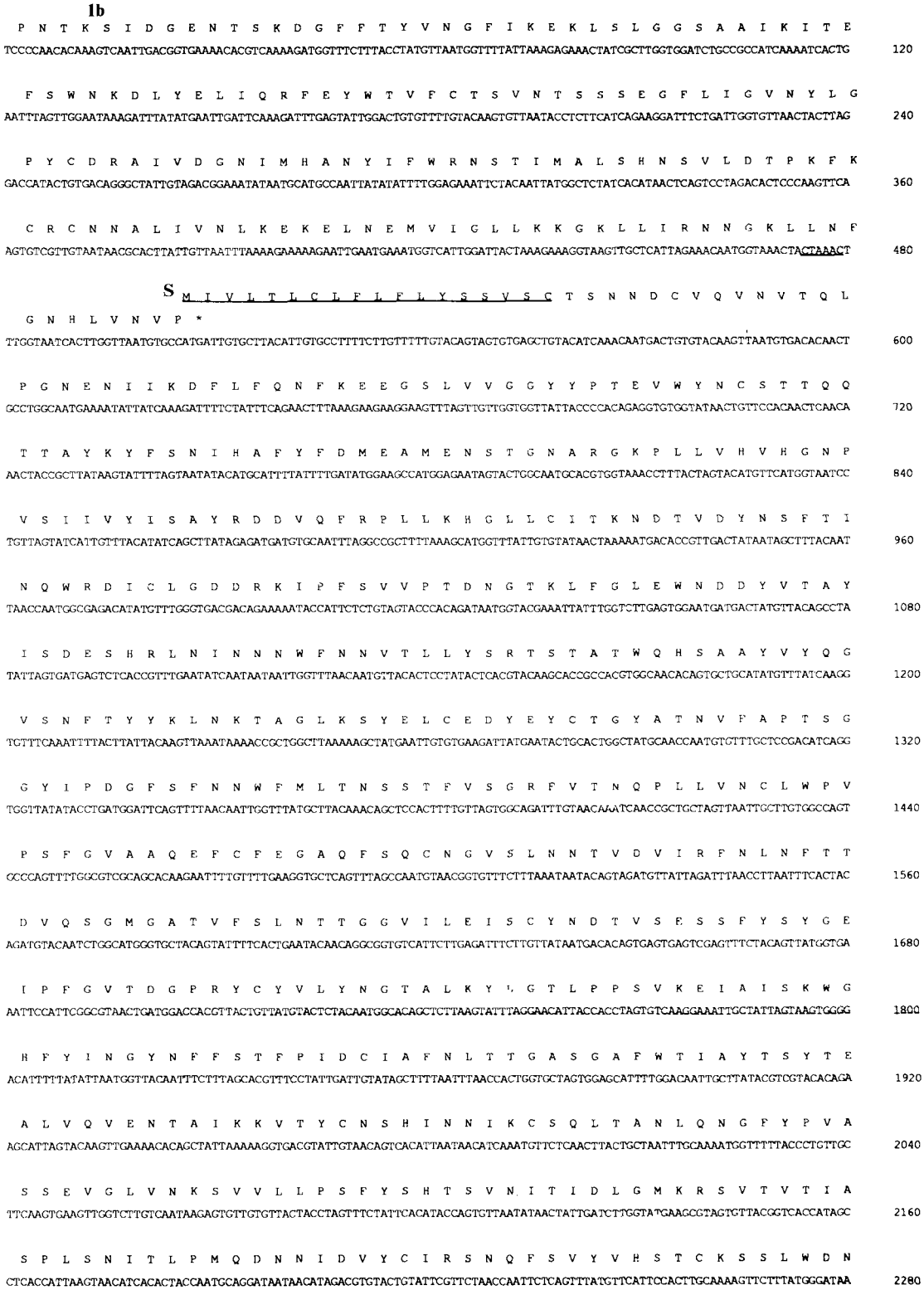
To clone the 3′ end of the CCV genome we prepared a cDNA library from CCV genomic RNA. Inserts from recombinant clones of 1·8 kb or greater were selected for further analyses. In order to map the clones, we took advantage of the suspected nucleotide sequence homology between the genomes of CCV and TGEV. Partial sequencing of recombinant clones revealed identity in excess of 95%. This permitted initial alignment of the CCV clones with respect to the TGEV genome. This approach proved fruitful in that four clones were identified which spanned some 8·5 kb at the 3′ end (Fig. 1). A region at the 3′ end for which large clones were not represented in the library was prepared by PCR amplification (BH6; Fig. 1). The relationships between putative overlapping clones were confirmed by Southern hybridization. Therefore, partial sequencing and Southern blotting identified five overlapping clones which covered approximately 9·6 kb from the 3′ end of the CCV genome.

Shotgun DNA sequencing and sequence analyses

The inserts from the plasmids detailed in Fig. 1 were sequenced using the shotgun methods of Bankier *et al.* (1987). The consensus nucleotide sequence of 9624 bp presented in Fig. 2 was analysed using the SAP programs of Staden (1982). Analysis revealed the presence of 10 open reading frames (ORFs) (Fig. 3). Pairwise alignment of these ORFs with their likely counterparts from other members of this coronavirus group disclosed very high levels of identity (Table 1) and indicated that the CCV structural proteins S, M and N are encoded by ORFs 2, 4 and 5, respectively. Each of the 10 ORFs is described in more detail below.

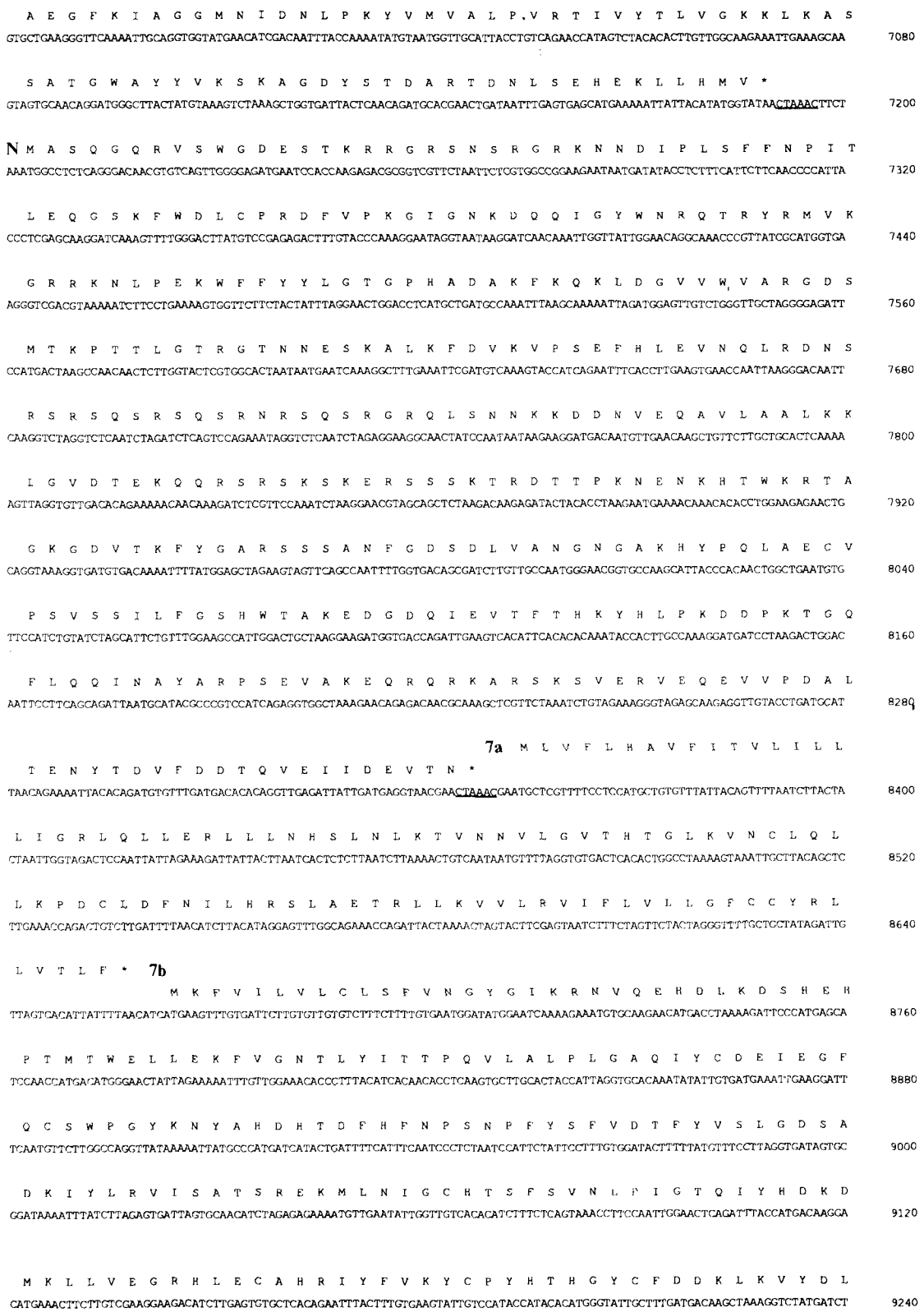
With respect to subgenomic mRNA synthesis, it is known that the minimal conserved signal for transcription in this coronavirus group, CTAAAC, is identical in TGEV, PRCV and FIPV and is therefore likely to be conserved in CCV (as reviewed by Spaan *et al.*, 1988). Indeed, analysis of the CCV sequence revealed that this sequence was present upstream of all the ORFs with the exception of the first and last. As ORF 1 is incomplete (see below), an additional CTAAAC sequence is presumably located at the 5′ end of the genomic RNA. When we analysed intracellular RNAs produced during CCV infection of canine A72 cells, eight species of RNA were observed (Fig. 4); the species observed between species 5 and 6 could not be accounted for in terms of the 3′-coterminal nested arrangement of coronavirus subgenomic RNAs and the observed positions of consensus transcription initiation signals. Taking into account the predicted size of each mRNA and the known location of the CTAAAC sequences, we predict a subgenomic message organization as depicted in Fig. 3. The ORFs encoded by each mRNA are described below.

The numbering of CCV RNAs used here is based on that currently employed by workers studying the Purdue-115 and FS772/70 strains of TGEV. The RNA organization of CCV strain Insavc-1 is most closely related to that described for these TGEV strains. This numbering scheme is not, however, applicable in a straightforward fashion to all members of the antigenic group. In the case of the Miller strain of TGEV, an RNA originally designated 4b (Wesley *et al.*, 1989) may be involved in the expression of ORF 3b and ORF 4; no additional RNA was detected between this RNA and the RNA coding for the membrane protein (RNA 5), but the


1b
P N T K S I D G E N T S K D G F F T Y V N G F I K E K L S L G G S A A I K I T E
TCCCCAACACAAAGTCAATTGACGGTGAAAACACGTCAAAAGATGGTTTCTTTACCTATGTTAATGGTTTTATTAAAGAGAAACTATCGCTTGGTGGATCTGCCGCCATCAAAATCACTG 120
F S W N K D L Y E L I Q R F E Y W T V F C T S V N T S S S E G F L I G V N Y L G
AATTTAGTTGGAATAAAGATTTATATGAATTGATTCAAAGATTTGAGTATTGGACTGTGTTTTGTACAAGTGTTAATACCTCTTCATCAGAAGGATTTCTGATTGGTGTTAACTACTTAG 240
P Y C D R A I V D G N I M H A N Y I F W R N S T I M A L S H N S V L D T P K F K
GACCATACTGTGACAGGGCTATTGTAGACGGAAATATAATGCATGCCAATTATATATTTTGGAGAAATTCTACAATTATGGCTCTATCACATAACTCAGTCCTAGACACTCCCAAGTTCA 360
C R C N N A L I V N L K E K E L N E M V I G L L K K G K L L I R N N G K L L N F
AGTGTCGTTGTAATAACGCACTTATTGTTAATTTAAAAGAAAAAGAATTGAATGAAATGGTCATTGGATTACTAAAGAAAGGTAAGTTGCTCATTAGAAACAATGGTAAACTACTAAACT 480
S M I V L T L C L F L F L Y S S V S C T S N N D C V Q V N V T Q L
G N H L V N V P *
TTGGTAATCACTTGGTTAATGTGCCATGATTGTGCTTACATTGTGCCTTTTCTTGTTTTTGTACAGTAGTGTGAGCTGTACATCAAACAATGACTGTGTACAAGTTAATGTGACACAACT 600
P G N E N I I K D F L F Q N F K E E G S L V V G G Y Y P T E V W Y N C S T T Q Q
GCCTGGCAATGAAAATATTATCAAAGATTTTCTATTTCAGAACTTTAAAGAAGAAGGAAGTTTAGTTGTTGGTGGTTATTACCCCACAGAGGTGTGGTATAACTGTTCCAGAACTCAACA 720
T T A Y K Y F S N I H A F Y F D M E A M E N S T G N A R G K P L L V H V H G N P
AACTACCGCTTATAAGTATTTTAGTAATATACATGCATTTTATTTTGATATGGAAGCCATGGAGAATAGTACTGGCAATGCACGTGGTAAACCTTTACTAGTACATGTTCATGGTAATCC 840
V S I I V Y I S A Y R D D V Q F R P L L K H G L L C I T K N D T V D Y N S F T I
TGTTAGTATCATTGTTTACATATCAGCTTATAGAGATGATGTGCAATTTAGGCCGCTTTTAAAGCATGGTTTATTGTGTATAACTAAAAATGACACCGTTGACTATAATAGCTTTACAAT 960
N Q W R D I C L G D D R K I P F S V V P T D N G T K L F G L E W N D D Y V T A Y
TAACCAATGGCGAGACATATGTTTGGGTGACGACAGAAAAATACCATTCTCTGTAGTACCCACAGATAATGGTACGAAATTATTTGGTCTTGAGTGGAATGATGACTATGTTACAGCCTA 1080
I S D E S H R L N I N N N W F N N V T L L Y S R T S T A T W Q H S A A Y V Y Q G
TATTAGTGATGAGTCTCACCGTTTGAATATCAATAATAATTGGTTTAACAATGTTACACTCCTATACTCACGTACAAGCACCGCCACGTGGCAACACAGTGCTGCATATGTTTATCAAGG 1200
V S N F T Y Y K L N K T A G L K S Y E L C E D Y E Y C T G Y A T N V F A P T S G
TGTTTCAAATTTTACTTATTACAAGTTAAATAAAACCGCTGGCTTAAAAAGCTATGAATTGTGTGAAGATTATGAATACTGCACTGGCTATGCAACCAATGTGTTTGCTCCGACATCAGG 1320
G Y I P D G F S F N N W F M L T N S S T F V S G R F V T N Q P L L V N C L W P V
TGGTTATATACCTGATGGATTCAGTTTTAACAATTGGTTTATGCTTACAAACAGCTCCACTTTGTTAGTGGCAGATTTGTAACAAATCAACCGCTGCTAGTTAATTGCTTGTGGCCAGT 1440
P S F G V A A Q E F C F E G A Q F S Q C N G V S L N N T V D V I R F N L N F T T
GCCCAGTTTTGGCGTCGCAGCACAAGAATTTTGTTTTGAAGGTGCTCAGTTTAGCCAATGTAACGGTGTTTCTTTAAATAATACAGTAGATGTTATTAGATTTAACCTTAATTTCACTAC 1560
D V Q S G M G A T V F S L N T T G G V I L E I S C Y N D T V S E S S F Y S Y G E
AGATGTACAATCTGGCATGGGTGCTACAGTATTTTCACTGAATACAACAGGCGGTGTCATTCTTGAGATTTCTTGTTATAATGACACAGTGAGTGAGTCGAGTTTCTACAGTTATGGTGA 1680
I P F G V T D G P R Y C Y V L Y N G T A L K Y L G T L P P S V K E I A I S K W G
AATTCCATTCGGCGTAACTGATGGACCACGTTACTGTTATGTACTCTACAATGGCACAGCTCTTAAGTATTTAGGAACATTACCACCTAGTGTCAAGGAAATTGCTATTAGTAAGTGGGG 1800
H F Y I N G Y N F F S T F P I D C I A F N L T T G A S G A F W T I A Y T S Y T E
ACATTTTTATATTAATGGTTACAATTTCTTTAGCACGTTTCCTATTGATTGTATAGCTTTTAATTTAACCACTGGTGCTAGTGGAGCATTTTGGACAATTGCTTATACGTCGTACACAGA 1920
A L V Q V E N T A I K K V T Y C N S H I N N I K C S Q L T A N L Q N G F Y P V A
AGCATTAGTACAAGTTGAAAACACAGCTATTAAAAAGGTGACGTATTGTAACAGTCACATTAATAACATCAAATGTTCTCAACTTACTGCTAATTTGCAAAATGGTTTTTACCCTGTTGC 2040
S S E V G L V N K S V V L L P S F Y S H T S V N I T I D L G M K R S V T V T I A
TTCAAGTGAAGTTGGTCTTGTCAATAAGAGTGTTGTGTTACTACCTAGTTTCTATTCACATACCAGTGTTAATATAACTATTGATCTTGGTATGAAGCGTAGTGTTACGGTCACCATAGC 2160
S P L S N I T L P M Q D N N I D V Y C I R S N Q F S V Y Y V H S T C K S S L W D N
CTCACCATTAAGTAACATCACACTACCAATGCAGGATAATAACATAGACGTGTACTGTATTCGTTCTAACCAATTCTCAGTTTATGTTCATTCCACTTGCAAAAGTTCTTTATGGGATAA 2280

```
N F N S A C T D V L D A T A V I K T G T C P F S F D K L N N Y L T F N K F C L S
CAATTTTAATTCAGCATGTACCGACGTTTAGACGCCACAGCTGTTATAAAAACTGGTACTTGTCCTTTCTCATTTGATAAATTGAATAATTACTTAACTTTAACAAGTTCTGTTTGTC  2400

L N P V G A N C K L D V A A R T R T N E Q V F G S L Y V I Y E E G D N I V G V P
GTTGAATCCCGTTGGTGCCAACTGTAAGTTAGATGTTGCCGCCCGTACAAGAACCAATGAGCAGGTTTTTGGAAGTTTATATGTAATATATGAAGAAGGAGACAACATAGTGGGTGTACC  2520

S D N S G L H D L S V L H L D S C T D Y N I Y G R T G V G I I R K T N S T L L S
GTCTGATAATAGTGGTTTGCACGATTTGTCAGTGTTGCACTTAGACTCTTGTACAGATTACAATATATATGGTAGAACTGGTGTTGGTATTATTAGAAAAACTAACAGCACACTACTTAG  2640

G L Y Y T S L S G D L L G F K N V S D G V V Y S V T P C D V S A Q A A V I D G A
TGGCTTATATTACACATCACTATCAGGTGATTTGTTAGGTTTTAAAAATGTTAGTGATGGTGTTGTCTACTCTGTAACGCCATGTGATGTAAGTGCACAAGCTGCTGTTATTGATGGTGC  2760

I V G A M T S I N S E L L G L T H W T T T P N F Y Y Y S I Y N Y T N V M N R G T
CATAGTTGGAGCTATGACTTCCATTAATAGTGAACTGTTAGGTCTAACTCATTGGACAACAACACCTAATTTTTATTACTACTCCATATATAATTATACAAATGTGATGAATCGTGGCAC  2880

A I D N D I D C E P I I T Y S N I G V C K N G A L V F I N V T H S D G D V Q P I
GGCAATTGATAATGATATTGATTGTGAACCTATCATAACATATTCTAATATAGGTGTTTGTAAAAATGGAGCTTTGGTTTTTATTAACGTCACACATTCTGATGGAGACGTTCAACCAAT  3000

S T G N V T I P T N F T I S V Q V E Y I Q V Y T T P V S I D C A R Y V C N G N P
TAGCACCGGTAATGTCACGATACCCACAAATTTTACTATATCTGTGCAAGTCGAATATATTCAGGTTTACACTACACCAGTTTCAATAGACTGTGCAAGATACGTTTGCAATGGTAACCC  3120

R C N K L L T Q Y V S A C Q T I E Q A L A M G A R L E N M E I D S M L F V S E N
AAGATGCAATAAGTTATTAACACAATACGTTTCTGCATGTCAAACTATTGAGCAAGCGCTTGCAATGGGTGCCAGACTTGAAAACATGGAGATTGATTCCATGTTATTTGTTTCGGAAAA  3240

A L K L A S V E A F N S T E N L D P I Y K E W P N I G G S W L G G L K D I L P S
TGCCCTTAAATTGGCATCTGTTGAAGCATTCAATAGTACGGAAAATTTAGACCCTATTTATAAAGAATGGCCTAACATTGGTGGTTCTTGGCTAGGAGGTTTAAAAGATATATTGCCATC  3360

H N S K R K Y R S A I E D L L F D K V V T S G L G T V D E D Y K R S A G G Y D I
TCATAATAGCAAACGTAAGTACCGCTCGGCTATAGAAGACTTGCTTTTGATAAGGTTGTAACATCTGGCTTAGGTACAGTTGACGAAGATTACAAACGTTCTGCAGGTGGTTATGACAT  3480

A D L V C A R Y Y N G I M V L P G V A N D D K M T M Y T A S L T G G I T L G A L
AGCTGACTTAGTGTGTGCACGATATTACAATGGCATCATGGTGCTACCTGGTGTAGCTAATGATGACAAGATGACTATGTACACTGCATCTCTTACAGGTGGTATAACATTAGGTGCACT  3600

S G G A V A I P F A V A V Q A R L N Y V A L Q T D V L N K N Q Q I L A N A F N Q
TAGTGGTGGCGCAGTGGCTATACCTTTTGCAGTAGCAGTTCAGGCTAGACTTAATTATGTTGCTCTACAAACTGATGTATTGAACAAAAACCAACAAATCTTGGCTAATGCTTTCAATCA  3720

A I G N I T Q A F G K V N D A I H Q T S K G L A T V A K A L A K V Q D V V N T Q
AGCTATTGGTAACATTACACAGGCATTTGGTAAGGTTAATGACGCTATACATCAAACATCAAAAGGTCTTGCTACTGTTGCTAAAGCATTGGCAAAGGTGCAAGATGTTGTTAACACGCA  3840

G Q A L S H L T V Q L Q N N F Q A I S S S I S D I Y N R L D E L S A D A Q V D R
AGGTCAAGCTTTAAGCCACCTAACAGTACAATTGCAAAACAATTTTCAAGCCATTAGCAGTTCTATTAGTGACATTTATAACAGGCTTGATGAATTGAGTGCTGATGCACAAGTTGACAG  3960

L I T G R L T A L N A F V S Q T L T R Q A E V R A S R Q L A K D K V N E C V R S
GCTGATTACAGGACGACTTACAGCACTTAATGCATTTGTGTCTCAGACTTTAACCAGACAAGCAGAGGTTAGGGCTAGTAGACAACTTGCTAAAGACAAGGTTAATGAATGCGTTAGGTC  4080

Q S Q R F G F C G N G T H L F S L A N A A P N G M I F F H T V L L P T A Y E T V
TCAATCCCAGAGATTTGGATTCTGTGGTAATGGTACACATTTGTTTTCACTTGCAAATGCGGCACCAAATGGCATGATTTTCTTTCACACAGTGCTATTACCAACAGCTTATGAAACTGT  4200

T A W S G I C A S D G S R T F G L V V E D V Q L T L F R N L D E K F Y L T P R T
GACGGCCTGGTCAGGTATTTGTGCGTCAGATGGCAGTCGCACTTTTGGACTTGTTGTTGAGGATGTCCAGCTGACGCTATTTCGCAATTTAGATGAAAAATTTTATTTGACGCCCAGAAC  4320

M Y Q P R V A T S S D F V Q I E G C D V L F V N G T V I E L P S I I P D Y I D I
TATGTATCAGCCCAGAGTTGCAACTAGTTCTGATTTTGTTCAAATAGAAGGCTGTGATGTGTTGTTTGTTAATGGAACTGTAATTGAATTGCCTAGTATCATACCTGACTATATCGATAT  4440

N Q T V Q D I L E N F R P N W T V P E L P L D I F H A T Y L N L T G E I N D L E
TAATCAAACTGTTCAGGACATATTAGAAAATTTCAGACCAAATTGGACTGTACCCGAGTTGCCACTTGACATTTTTCATGCAACCTACTTAAACCTGACTGGTGAAATTAATGACTTAGA  4560

F R S E K L H N T T V E L A I L I D N I N N T L V N L E W L N R I E T Y V K W P
ATTTAGGTCAGAAAAGTTACATAACACCACAGTAGAACTTGCTATTCTCATTGATAATATTAATAACACATTAGTCAATCTTGAATGGCTCAACAGAATTGAAACTTATGTAAAATGGCC  4680
```

```
 W Y V W L L I G L V V I F C I P I L L F C C C S T G C C G C I G C L G S C C H S
TTGGTATGTTTGGCTACTAATTGGATTAGTAGTAATATTCTGCATACCCATATTGCTATTTTGTTGTTGTAGTACTGGTTGTTGTGGATGTATCGGGTGTTTAGGAAGCTGTTGTCATTC  4800

 I C S R G Q F E S Y E P I E K V H V H *
CATATGTAGTAGAGGCCAATTTGAAAGTTATGAACCTATTGAAAAAGTTCATGTTCACTGAATTCAAAATGTTAAGTCTACTATTTTAATTACACCCTTGGCAACACAAGTGATATAAAG

GTGGTGTCGTAATTCATACCAGTCAATTTTAGCATTAATAAAACACACTTCTATGGCTGGTAATACCGGTTATATATAATGTGTTTTTAATTTTTAAGAACTAAACTTATGAGTCATTAC  5040

 3a M D I V K S I D T S V D A V L D E F D C A Y F A V T L K V E F K T G K Q L
AGGTCTTGTATGGACATTGTCAAATCTATTGACACATCCGTAGACGCTGTACTTGACGAATTTGATTGCGCATACTTTGCTGTAACTCTTAAAGTAGAGTTCAAGACTGGTAAGCAACTT  5160

 V C I G F G D T L L E A K D K A Y A K L G L S I I E E V N S H T V V *
                                3x M L N L V S L L L K K S I V I Q L F D I T V Y K
GTGTGTATAGGTTTTGGTGATACACTTTTAGAGGCTAAGGACAAAGCATATGCTAAACTTGGTCTCTCTATTATTGAAGAAGTCAATAGTCATACAGTTGTTTGATATTACTGTTTATAA  5280

 F K A K F W Y K L P F E T R L R I I K H T K P K A L S A T K Q V K R D Y R K T A
GTTTAAGGCCAAATTTTGGTACAAATTACCTTTTGAAACTAGACTTCGTATCATTAAACACACAAAACCTAAAGCATTAAGTGCTACAAAACAAGTAAAGAGAGATTATAGAAAAACTGC  5400

                 3b M I G G L F L N T L S F V I V S N H V I V N N T A N V H H T Q * D
 I L N S M R K *
CATTCTAAATTCCATGAGAAAATGATTGGTGGACTTTTTCTTAACACTCTGAGTTTTGTAATTGTTAGCAACCATGTCATTGTTAACAATACAGCAAATGTGCATCACACACAATAAGAC  5520

 H V I V Q Q H Q F V S A R T Q N Y Y P E F S I A V L F V S F L A L Y R S T N F K
CATGTTATAGTACAACAACATCAGTTTGTTAGTGCTAGAACACAAAATTACTACCCGGAGTTCAGCATTGCTGTACTCTTTGTATCCTTTCTAGCTTTGTACCGTAGTACAAACTTTAAG  5640

 T C V G I L M F K I V S M T L I G P M L I A F G Y Y I D G I V T T I V L A L R F
ACGTGTGTCGGTATCTTAATGTTTAAGATTGTATCAATGACACTTATAGGACCTATGCTTATAGCATTTGGTTACTACATTGATGGCATTGTTACAACAATTGTCTTAGCTTTAAGATTT  5760

 I Y V S Y F W Y V N N R F E F I L Y N T T T L M F V H G R A A P F M R S S H S S
ATTTACGTATCATATTTCTGGTATGTTAATAATAGATTTGAATTCATTTTATACAATACGACGACACTCATGTTTGTACATGGCAGAGCTGCACCGTTTATGAGAAGTTCTCACAGCTCT  5880

 I Y V T L Y G G I N Y M F V N D L T L H F V D P M L V S I A T R G L A H A D L T
ATTTATGTCACATTGTACGGTGGCATAAATTATATGTTTGTGAATGACCTCACGTTGCATTTTGTAGACCCTATGCTTGTAAGCATAGCAACACGTGGCTTAGCTCATGCTGATCTAACT  6000

 V V R A V E L L N G D F I Y V F S Q E P V V G V Y N A A F S Q A V L N E I D L K
GTTGTTAGAGCAGTTGAACTTCTCAATGGTGATTTTATTTATGTATTTTCACAGGAGCCCGTAGTCGGTGTTTACAATGCAGCCTTTTCTCAGGCGGTTCTAAACGAAATTGACTTAAAA  6120

 E E E E D H I Y D V P S G I D C H R *
                  4 M T F P R A L T V I D D N G M V I S I I F W F L L I I I L I L F S
GAAGAAGAAGAAGACCATATCTATGACGTTCCCTCGGGCATTGACTGTCATAGATGACAATGGAATGGTCATTAGTATCATTTTCTGGTTCCTGTTGATAATTATATTGATATTATTTTC  6240

 I A L L N I I K L C M V C C N L G R T V I I V P A R H A Y D A Y K N F M Q I R A
AATAGCATTGCTAAATATAATTAAGCTATGCATGGTATGTTGCAATTTAGGAAGAACAGTTATTATTGTTCCAGCTCGACATGCCTATGATGCCTATAAGAATTTTATGCAAATTAGAGC  6360

 Y N P D E A L L V *
                    M M K K I L F L L A C A I A C V Y G E R Y C A M T E S S
ATACAACCCTGATGAAGCACTCCTTGTTTGAACTAAACAAAATGAAGAAAATTTTGTTTTTACTAGCGTGTGCAATTGCATGCGTCTATGGAGAACGCTATTGTGCCATGACTGAAAGTT  6480

 T S C R N S T A G N C A S C F E T G D L I W H L A N W N F S W S V I L I I F I T
CTACGTCATGTCGTAATAGCACGGCTGGCAACTGTGCTTCATGCTTCGAAACAGGTGATCTTATTTGGCATCTTGCAAACTGGAACTTCAGCTGGTCTGTAATATTGATCATTTTTATAA  6600

 V L Q Y G R P Q F S W F V C G I K M L I M W L L W P I V L A L T I F N A Y L E Y
CAGTGTTACAATATGGAAGACCTCAATTTAGCTGGTTCGTGTGTGGCATTAAAATGCTTATTATGTGGCTGTTATGGCCCATTGTTTTAGCTCTTACGATTTTTAATGCATACCTGGAAT  6720

 R V S R Y V M F G F S V A G A T V T F I L W I M Y F V R S I Q L Y R R T K S W W
ACCGAGTTTCCAGATATGTAATGTTCGGCTTTAGTGTTGCAGGTGCAACTGTTACATTTATACTTTGGATTATGTATTTTGTTAGATCCATTCAGTTATACAGAAGGACTAAGTCTTGGT  6840

 S F N P E T S A I L C V S A L G R S Y V L P L E G V P T G V T L T L L S G N L C
GGTCTTTCAACCCTGAAACTAGCGCAATTCTTTGCGTTAGTGCGTTAGGAAGAAGCTATGTGCTTCCTCTTGAAGGTGTGCCAACTGGTGTCACTCTAACATTGCTTTCAGGGAATTTGT  6960
```


A E G F K I A G G M N I D N L P K Y V M V A L P V R T I V Y T L V G K K L K A S
GTGCTGAAGGGTTCAAAATTGCAGGTGGTATGAACATCGACAATTTACCAAAATATGTAATGGTTGCATTACCTGTCAGAACCATAGTCTACACACTTGTTGGCAAGAAATTGAAAGCAA 7080
S A T G W A Y Y V K S K A G D Y S T D A R T D N L S E H E K L L H M V *
GTAGTGCAACAGGATGGGCTTACTATGTAAAGTCTAAAGCTGGTGATTACTCAACAGATGCACGAACTGATAATTTGAGTGAGCATGAAAAATTATTACATATGGTATAACTAAACTTCT 7200
N M A S Q G Q R V S W G D E S T K R R G R S N S R G R K N N D I P L S F F N P I T
AAATGGCCTCTCAGGGACAACGTGTCAGTTGGGGAGATGAATCCACCAAGAGACGCGGTCGTTCTAATTCTCGTGGCCGGAAGAATAATGATATACCTCTTTCATTCTTCAACCCCATTA 7320
L E Q G S K F W D L C P R D F V P K G I G N K D Q Q I G Y W N R Q T R Y R M V K
CCCTCGAGCAAGGATCAAAGTTTTGGGACTTATGTCCGAGAGACTTTGTACCCAAAGGAATAGGTAATAAGGATCAACAAATTGGTTATTGGAACAGGCAAACCCGTTATCGCATGGTGA 7440
G R R K N L P E K W F F Y Y L G T G P H A D A K F K Q K L D G V V W V A R G D S
AGGGTCGACGTAAAAATCTTCCTGAAAAGTGGTTCTTCTACTATTTAGGAACTGGACCTCATGCTGATGCCAAATTTAAGCAAAAATTAGATGGAGTTGTCTGGGTTGCTAGGGGAGATT 7560
M T K P T T L G T R G T N N E S K A L K F D V K V P S E F H L E V N Q L R D N S
CCATGACTAAGCCAACAACTCTTGGTACTCGTGGCACTAATAATGAATCAAAGGCTTTGAAATTCGATGTCAAAGTACCATCAGAATTTCACCTTGAAGTGAACCAATTAAGGGACAATT 7680
R S R S Q S R S Q S R N R S Q S R G R Q L S N N K K D D N V E Q A V L A A L K K
CAAGGTCTAGGTCTCAATCTAGATCTCAGTCCAGAAATAGGTCTCAATCTAGAGGAAGGCAACTATCCAATAATAAGAAGGATGACAATGTTGAACAAGCTGTTCTTGCTGCACTCAAAA 7800
L G V D T E K Q Q R S R S K S K E R S S S K T R D T T P K N E N K H T W K R T A
AGTTAGGTGTTGACACAGAAAAACAACAAAGATCTCGTTCCAAATCTAAGGAACGTAGCAGCTCTAAGACAAGAGATACTACACCTAAGAATGAAAACAAACACACCTGGAAGAGAACTG 7920
G K G D V T K F Y G A R S S S A N F G D S D L V A N G N G A K H Y P Q L A E C V
CAGGTAAAGGTGATGTGACAAAATTTTATGGAGCTAGAAGTAGTTCAGCCAATTTTGGTGACAGCGATCTTGTTGCCAATGGGAACGGTGCCAAGCATTACCCACAACTGGCTGAATGTG 8040
P S V S S I L F G S H W T A K E D G D Q I E V T F T H K Y H L P K D D P K T G Q
TTCCATCTGTATCTAGCATTCTGTTTGGAAGCCATTGGACTGCTAAGGAAGATGGTGACCAGATTGAAGTCACATTCACACACAAATACCACTTGCCAAAGGATGATCCTAAGACTGGAC 8160
F L Q Q I N A Y A R P S E V A K E Q R Q R K A R S K S V E R V E Q E V V P D A L
AATTCCTTCAGCAGATTAATGCATACGCCCGTCCATCAGAGGTGGCTAAAGAACAGAGACAACGCAAAGCTCGTTCTAAATCTGTAGAAAGGGTAGAGCAAGAGGTTGTACCTGATGCAT 8280
7a M L V F L H A V F I T V L I L L
T E N Y T D V F D D T Q V E I I D E V T N *
TAACAGAAAATTACACAGATGTGTTTGATGACACACAGGTTGAGATTATTGATGAGGTAACGAACTAAACGAATGCTCGTTTTCCTCCATGCTGTGTTTATTACAGTTTTAATCTTACTA 8400
L I G R L Q L L E R L L L N H S L N L K T V N N V L G V T H T G L K V N C L Q L
CTAATTGGTAGACTCCAATTATTAGAAAGATTATTACTTAATCACTCTCTTAATCTTAAAACTGTCAATAATGTTTTAGGTGTGACTCACACTGGCCTAAAAGTAAATTGCTTACAGCTC 8520
L K P D C L D F N I L H R S L A E T R L L K V V L R V I F L V L L G F C C Y R L
TTGAAACCAGACTGTCTTGATTTTAACATCTTACATAGGAGTTTGGCAGAAACCAGATTACTAAAAGTAGTACTTCGAGTAATCTTTCTAGTTCTACTAGGGTTTTGCTGCTATAGATTG 8640
L V T L F * 7b
M K F V I L V L C L S F V N G Y G I K R N V Q E H D L K D S H E H
TTAGTCACATTATTTTAACATCATGAAGTTTGTGATTCTTGTGTTGTGTCTTTCTTTTGTGAATGGATATGGAATCAAAAGAAATGTGCAAGAACATGACCTAAAAGATTCCCATGAGCA 8760
P T M T W E L L E K F V G N T L Y I T T P Q V L A L P L G A Q I Y C D E I E G F
TCCAACCATGACATGGGAACTATTAGAAAAATTTGTTGGAAACACCCTTTACATCACAACACCTCAAGTGCTTGCACTACCATTAGGTGCACAAATATATTGTGATGAAATTGAAGGATT 8880
Q C S W P G Y K N Y A H D H T D F H F N P S N P F Y S F V D T F Y V S L G D S A
TCAATGTTCTTGGCCAGGTTATAAAAATTATGCCCATGATCATACTGATTTTCATTTCAATCCCTCTAATCCATTCTATTCCTTTGTGGATACTTTTTATGTTTCCTTAGGTGATAGTGC 9000
D K I Y L R V I S A T S R E K M L N I G C H T S F S V N L F I G T Q I Y H D K D
GGATAAAATTTATCTTAGAGTGATTAGTGCAACATCTAGAGAGAAAATGTTGAATATTGGTTGTCACACATCTTTCTCAGTAAACCTTCCAATTGGAACTCAGATTTACCATGACAAGGA 9120
M K L L V E G R H L E C A H R I Y F V K Y C P Y H T H G Y C F D D K L K V Y D L
CATGAAACTTCTTGTCGAAGGAAGACATCTTGAGTGTGCTCACAGAATTTACTTTGTGAAGTATTGTCCATACCATACACATGGGTATTGCTTTGATGACAAGCTAAAGGTCTATGATCT 9240

```
K R V K S R K D F E K I S Q Y Q K S E L *
GAAGCGTGTCAAAAGCAGGAAGGATTTTGAGAAAATCAGCCAATATCAGAAAAGTGAGTTGTAAGGCCACCCGATGTTTAAAATGGTTTTTCCGAGGAATTACTGGTCATCGCGCTGTCT   9360

ACTCTTGTACAGAATGGTAAGCCAAGTGTCAATAGGAGGTACAAGCAACCTATTGCATATTAGGAAGTTTAGATTTGATTTGGCAATGCTAGATTTAGTAATTTAGAGAAGTTTAAAGAG   9480

TCCGTATGACGAGCCAACAATGCGAAGCCTAACGTCTGGATCTAGTGATTGTTTAAAATGTAAAATTGTTTGAAAATTTTCCTTTTGATAGTGATTCACCAAAAAAAAAAAAAAAAAAAAA   9600

AAAAAAAAAAAAAAAAAA   9620
```

Fig. 2. Sequence of the extreme 3′ 9624 nucleotides of the CCV genome and the deduced amino acid sequences encoded by the ORFs. The consensus intergenic sequences are underlined. The octanucleotide sequence conserved in the 3′ non-coding region of all coronaviruses is shown in bold. The predicted ORFs are translated into the single-letter amino acid code. The putative signal peptides for S and M proteins are underlined.

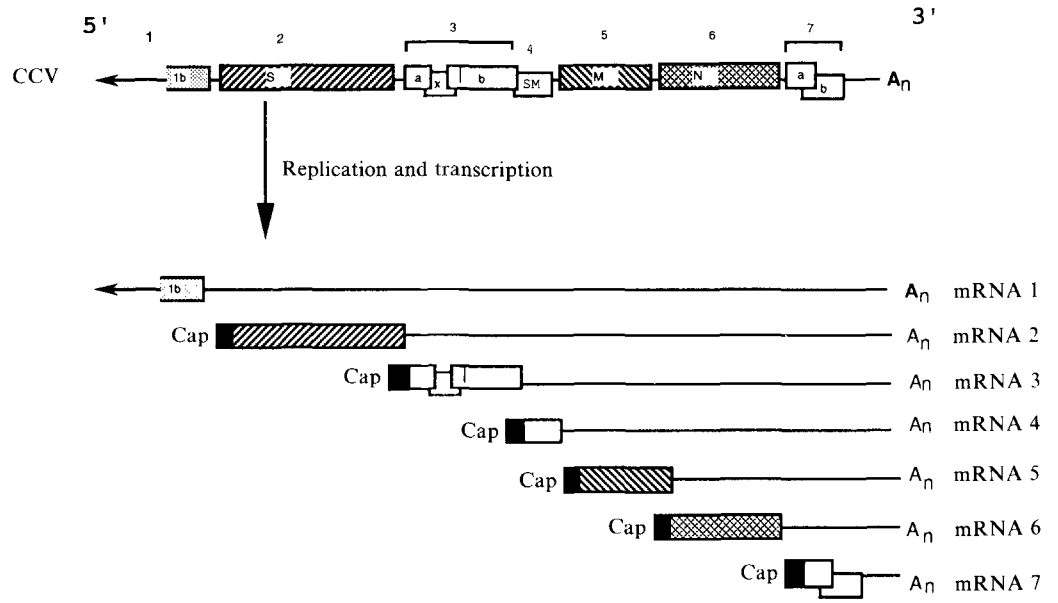


Fig. 3. Gene and subgenomic message organization predicted from the sequence data and Northern blot analyses cited in Results. Genes are designated according to the recommendations of the coronavirus study group (Cavanagh *et al.*, 1990*a*, *b*). ORFs are represented by boxes. The vertical line in ORF 3b represents a stop codon and the black boxes represent leader sequences. Numbers represent ORFs encoded by that message.

Table 1. *Pairwise sequence homology between CCV and FIPV, TGEV, PRCV and MHV ORFs*

CCV* ORF	$M_r \times 10^{-3}$	No. of amino acids	Pairwise identity (%) FIPV	TGEV	PRCV	MHV
1b	NK†	168‡		95·2	96·4	52·7
2 (S)	160	1452	91·1	79	74·7	23·4
3a	8·6	71	NK	83·5	48·8	
3b	28·4§	251	NK	92·7	92·6	
4 (SM)	9·3	82	NK	88·4	88·4	
5 (M)	29·5	262	83·7	88·3	86·3	30·3
6 (N)	43·4	401	76·4	89·6	86·9	27·3
7a	11·5	101	78·4	68·5	68	
7b	29·4	213	57			

* ORF 3x is not included (see Results).
† NK, Not known.
‡ Incomplete.
§ Disregarding terminator, otherwise $M_r = 4000$.

possibility that such an RNA may be synthesized at a low level must be considered because a CTAAAC signal was observed. Similarly, in the case of FIPV strain 79-1146 no RNA has been detected between RNA 3 and the membrane polypeptide RNA (de Groot *et al.*, 1987) but the possibility of an equivalent of the TGEV RNA 4 has been alluded to (de Groot, 1989). Thus, the numbering conventions employed do not deal adequately with the variations in expression strategy observed in this region of genome within this group of closely related viruses.

ORFs encoded by mRNAs 1 and 2

ORF 1 is incomplete, has no AUG start codon, encodes 168 amino acids and terminates in a UGA stop codon at position 510 (Fig. 2). A comparison of this ORF

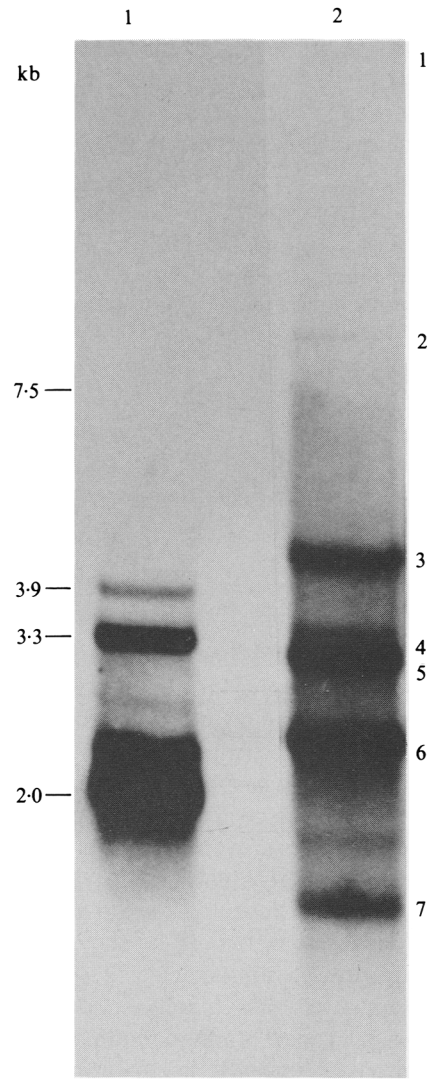


Fig. 4. Northern blot analysis of IBV Beaudette mRNA size markers (lane 1) and CCV strain Insavc-1 (lane 2) mRNAs. Unlabelled intracellular RNAs were separated by formaldehyde gel electrophoresis. The RNAs were transferred to a membrane filter and hybridized with radiolabelled inserts IBV-N and pBH5, respectively. IBV-N, a PCR product of the IBV Beaudette N gene was kindly supplied by Dr David Cavanagh. No CTAAAC motif which could give rise to a messenger species was found between the CTAAAC motifs associated with the messenger species 6 and 7.

with TGEV strain FS772/70 shows 99·2% similarity to 1b and 47 and 52·7% identity to genes 1b of avian infectious bronchitis virus (IBV) and MHV, respectively (Britton & Page, 1990; Boursnell *et al.*, 1987; Bredenbeek *et al.*, 1990). Thus, this ORF represents the 3′ end of the putative polymerase-encoding region of genome mRNA 1.

ORF 2 located immediately downstream of the polymerase gene would be translated from the 9·1 kb subgenomic message 2. This ORF is 4356 nucleotides long representing 1452 amino acids with a calculated M_r of 160K. Comparison of this ORF with sequences held in the EMBL database reveals remarkably high identity to the FIPV spike glycoprotein-encoding sequences (91·1%) and, to a lesser degree, the porcine virus S genes (Table 1), indicating that this is the CCV S gene. In some strains of MHV, the haemagglutinin–esterase glycoprotein gene (HE) is found downstream of the polymerase gene (Luytjes *et al.*, 1988) but it is clear that CCV, like TGEV and IBV, encodes only the polymerase gene upstream of the S gene (Britton & Page, 1990; Boursnell *et al.*, 1987).

The CCV S protein shows features characteristic of a type I membrane protein, i.e. a putative signal sequence (Von Heijne, 1986; positions 506 to 563; Fig. 2) and transmembrane domain (positions 4682 to 4742; Fig. 2). There are also 30 potential *N*-glycosylation sites which probably account for the increased size of the S protein found in the virion (Garwes & Reynolds, 1981).

ORFS encoded by mRNAs 3 and 4

There are four ORFs distal to the S gene coding sequence which are likely to be encoded by messages 3 and 4 (Fig. 3). Three of these have close similarity to their porcine virus counterparts and have been named 3a (8·6K), 3b (28·4K) and 4 (9·3K) (Table 1). The fourth ORF, which to date has not been detected in this group of viruses, could potentially encode a 71 amino acid protein with a predicted M_r of 10K and overlaps ORFs 3a and 3b (Fig. 2 and 3). This ORF has been designated 3x. The CCV 3b ORF was expected to encode a 28K protein like its TGEV counterpart (Jacobs *et al.*, 1986). However, this strain of CCV has acquired a termination codon, UAA (at position 5515; Fig. 2), which would result in a truncated polypeptide of only 33 amino acids. Direct sequencing of the viral genomic and mRNAs has confirmed the authenticity of this stop codon (data not shown). The CCV 4 ORF encodes a small membrane protein that is related to the 3c product of IBV (Fig. 5).

Message 4, as predicted from our sequence data, was detected in Northern blots (see Fig. 4). This message could only express ORF 4, as the proposed signal for transcription, CTAAAC, is found 43 nucleotides upstream of the predicted ORF 4 start codon. This arrangement is found in a number of strains of TGEV.

ORFs encoded by mRNAs 5, 6 and 7

Messenger RNA species 5 and 6 encode ORFs which resemble the coding sequences for the other coronavirus structural proteins, M and N, respectively (Table 1). Translation of poly(A)-selected CCV intracellular RNA in the rabbit reticulocyte lysate system produced pro-

```
IBV-Beaudette  MMNLLNKSLEENGSFLTALYIIVGFLALYLLGRALQAFVQAADACCLFWYTWVVI
CCV-Insavc-1   MTFPRALTVIDDNGMVISIIFWFLLIIILILFSIALLNIIKLCMVCCNLGRTVIIV
TGEV-Miller    MTFPRALTVIDDNGMVISIIFWFLLIIILILLSIALLNIIKLCMVCCNLGRTVIIV
MHV-JHM          MFNLFLTDTVWYVGQIIFIVAVCLMVTIIVVAFLASIKRCIQLCGLCNTLLLS
BCV-Mebus      MFMADAYFADTVWYVGQIIFIVAICLLVIIVVVAFLATFKLCIQLCGMCNTLGLS
                         *                                   *    *

                                    Hydrophobic region

IBV-Beaudette  PGAKGTAFVYKYTYGRKLNNPELEAVIVNEFPKNGWNNKNPANFQDAQRDKLYS
CCV-Insavc-1   PARHAYDAYKNFMQIRAYNPDEALLV
TGEV-Miller    PVQHAYDAYKNFMRIKAYNPDGALLV
MHV-JHM        PSIYLYNRSKQLYKYYNEEVRPPPLEVDDNIIQTL
BCV-Mebus      PSIYVFNRGRGFYEFYNDVKPPVLDVDDV
               *
```

Fig. 5. Alignment of the putative small membrane protein amino acid sequences from five different strains of coronaviruses. The hydrophobic core is shown in bold. Asterisks represent conserved features.

ducts of the sizes expected for M and N when analysed by SDS–PAGE (data not shown). ORFs 7a and 7b are likely to be encoded on a single RNA species (mRNA 7) since smaller messages were not seen on Northern blots, nor is another message predicted from the sequence data. Furthermore, an equivalent RNA in FIPV is thought to be bicistronic (de Groot *et al.*, 1988) and the levels of identity between the 7a and 7b ORFs of CCV and the 6a and 6b ORFs of FIPV are 78·4% and 57% respectively. Alignment of this region of CCV with the related regions of TGEV and PRCV reveals that the 7a ORF of the porcine coronaviruses has undergone a deletion of 69 nucleotides and furthermore they have no counterpart to ORF 7b. Nevertheless, the CCV structural protein ORFs, with the exception of S, have higher identities to TGEV than to FIPV ORFs.

Discussion

In this study approximately 9·6 kb of the 3′ end of the CCV strain Insavc-1 genome was cloned and sequenced. This region is likely to include all of the viral genes excluding the polymerase gene for which only the 3′-terminal 168 amino acids have been determined. Therefore, a substantial part of the virus' genetic information was available for comparison with other antigenically related coronaviruses, namely TGEV, PRCV and FIPV. The deduced sequence and genetic organization of CCV are shown in Fig. 2 and 3, respectively.

From antigenic data and cross-infectivity studies, the viruses within this group have been termed 'host range mutants' (Horzinek *et al.*, 1982). This close evolutionary relationship is emphasized by our analyses of the CCV sequence data. The CCV spike protein is closely related to the other spikes and has the features typical of coronavirus peplomer glycoproteins. Any variation in the sequence of this protein within the group presumably reflects changes in cell tropism, drift as a result of polymerase errors and selection by the host's immune system. Similarly, interspecies comparison of the other structural proteins, M and N, revealed very high levels of identity (Table 1). Alignment of the M gene product amino acid sequences revealed that any variation was primarily found on what would be the exposed amino terminus of the protein (amino acids 22 to 44; Fig. 2), i.e. between the putative signal sequence (Von Heijne, 1986) and the first transmembrane domain. However, the single potential *N*-glycosylation site and the three cysteine residues are conserved. These cysteine residues are probably important in forming interchain disulphide bridges, as M of HCV-229E has been shown to form oligomers under non-reducing conditions (Arpin & Talbot, 1990). The variation in this region is again probably a result of selection pressure from the host's immune system. Interestingly, alignment of the N gene amino acid sequences indicated that FIPV N has diverged to a greater extent than those of both CCV and TGEV (Fig. 6). This is unusual as N proteins are normally highly conserved; alignment of N gene amino acid sequences from five isolates of MHV showed at least 90% identity (Masters *et al.*, 1990). Nevertheless, variation was mainly clustered in two regions of the N molecule, between positions 204 and 210, and 352 and 359 (Fig. 6). It has been proposed that these two loci represent spacers, which have little sequence specificity but connect conserved domains of the molecule involved in interaction with the RNA genome (Masters *et al.*, 1990).

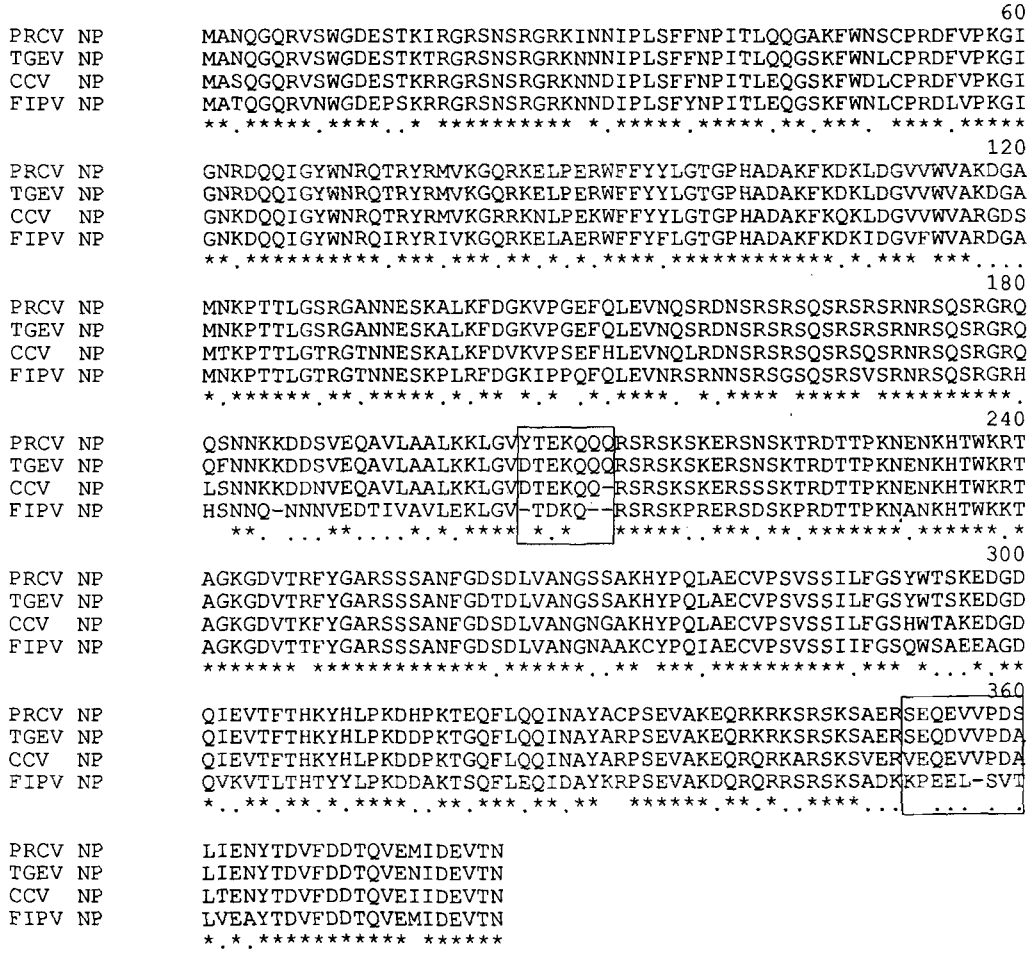


Fig. 6. Alignment of the nucleocapsid protein amino acid sequences from PRCV strain 86/137004 (Britton *et al.*, 1991), TGEV strain FS772/70 (Britton & Page, 1990) CCV strain Insavc-1 (this paper) and FIPV strain 79-1146 (Vennema *et al.*, 1991) using the CLUSTAL program (Higgins & Sharp, 1989). The asterisks below the sequences show identical amino acids in all four viruses and a dot is used if there has been a conservative substitution. The minus signs represent deletions. The boxed areas represent putative spacer regions (Masters *et al.*, 1990).

The ORFs that lie between the S and M genes have, like the other ORFs so far analysed, a high degree of identity to their porcine virus counterparts (Table 1) and presumably perform similar functions. A previously undetected ORF, 3x, was identified which could potentially encode a 10K polypeptide. However, codon usage and base preference programs of Staden (1982) suggest that this ORF does not encode a functional viral protein. Furthermore, the proximal AUG is in a poor context for translation initiation (Kozak, 1986) and the only other AUG is found at the very 3′ end of the coding sequence. Therefore, it is very unlikely that this ORF is expressed in this strain of CCV and it probably represents an evolutionarily redundant sequence which is no longer required by the virus. Analysis of TGEV genomic sequence in this region revealed a counterpart for this canine virus pseudogene; 92 nucleotides have, however, been deleted. This deletion also results in a frameshift in the sequence which explains why this ORF has not hitherto been noticed (Fig. 7). In addition to the likely non-functionality of ORF 3x, it is also unlikely that ORF 3b is expressed in this strain of CCV. Although a transcription signal, CTAAAC, is present upstream of ORF 3b (Fig. 2, position 5213), we were unable to detect an mRNA of the predicted size on Northern blots. Even if low-level transcription occurs from this site, it is unlikely that ORF 3b is expressed as there is a termination codon (UAA) some 93 nucleotides downstream of the first AUG and subsequent AUG codons are in poor contexts for ribosome binding (Kozak, 1986). In fact, *in vitro* transcription and translation of this ORF did not yield any discernible products by SDS–PAGE analysis (data not shown). Alignment of ORF 4 amino acid sequences disclosed features in common with the

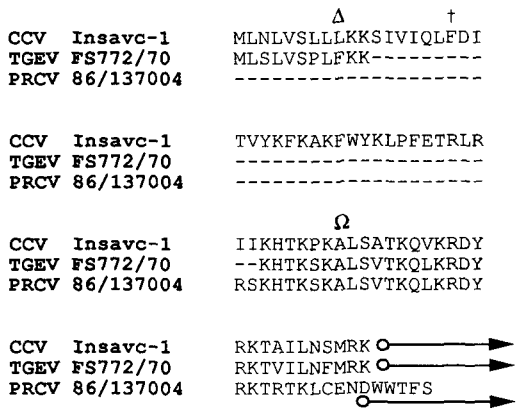


Fig. 7. Alignment of the TGEV and PRCV amino acid sequences with the CCV pseudogene 3x, which overlaps ORFs 3a and 3b. Circles with arrows represent the start of ORF 3b. The end of the 3a ORFs in CCV, TGEV and PRCV are indicated by the symbols †, Δ and Ω, respectively. Deletions are represented by minus signs. The intervening sequence between 3a and 3b ranges from 60 to over 200 nucleotides between these strains, indicating that deletions may occur in this region at a higher frequency than in the surrounding sequence.

IBV 3c protein (Fig. 5). This protein is found in the viral envelope (Smith *et al.*, 1990) and it has been suggested that it may be translated from IBV mRNA 3 by a cap-independent mechanism (Liu, 1991). However, CCV strain Insavc-1 expresses a message species, mRNA 4, appropriate for conventional cap-dependent expression of ORF 4. This RNA is difficult to detect since it is very similar in size to the abundant M message and is present in low abundance, possibly as the result of a suboptimal RNA transcriptional leader binding site, CTAAAC, which is found 43 bp upstream of the AUG start codon of ORF 4.

The degree of variability in the lengths of the non-coding sequences that lie upstream and downstream of ORF 3a in members of this antigenic group is striking. The lengths of these sequences range from 40 bp to over 200 bp (Fig. 7). Alignment of the ORF 3a amino acid sequences reveals that, in addition, variation is found at the ends of these coding sequences. Perhaps the non-coding regions proximal and distal to 3a are 'hot spot regions' where recombination, insertions or usually deletions can occur at a higher frequency relative to the surrounding sequences. The dynamism of the genome is well documented in coronaviruses (Keck *et al.*, 1988; Kusters *et al.*, 1989) and may be related to the propensity of the replicase complex to fall off its template and then to reinitiate RNA replication on the same or a different template. It would appear that there are three regions where deletions can occur at a higher frequency: within S, between S and M, and downstream of N. The polymorphism of S found in MHV strains with differing passage histories is mainly due to deletions in that gene which can lead to deletions of up to 159 amino acids. Consequently, this has an effect on pathogenicity, as deletions in the MHV-4 S coding sequence apparently result in a loss of ability to induce fatal encephalitis and the acquisition of a non-fatal demyelinating disease in mice (Parker *et al.*, 1989). Polymorphism has also been observed in the S gene and in the region between the S and M genes for different strains of TGEV and the respiratory tract mutant, PRCV (Wesley *et al.*, 1990; Rasschaert *et al.*, 1990). In fact, an IBV strain (Port/322/85) has been reported which appears to have arisen as a result of recombination between the M and S genes from two other strains of IBV (Cavanagh *et al.*, 1990*b*). The third 'hot spot region' is found downstream of the N gene. The porcine coronaviruses have a 69 nucleotide deletion in ORF 7a and ORF 7b is not present (de Groot *et al.*, 1988). This phenomenon is not unique to the coronaviruses from this antigenic group. Deletions of up to 170 nucleotides are found downstream of the N gene in some strains of IBV (Collisson *et al.*, 1990).

CCV ORF 7b has 57% identity to FIPV 6b. This ORF is the least conserved between the two viruses. Whether the protein produced from this ORF plays an important role in the immune-mediated disease seen in felines remains to be seen as all the viruses from this antigenic group can infect cats but only FIPV will produce this disease.

In conclusion, sequencing and subsequent analyses stress the very close relationship CCV has to the other viruses within its antigenic group. We must, however, be careful when generalizing about the CCV sequence data from this limited information. Coronavirus genomes are dynamic, subject to recombination, insertion and deletion, and as a consequence strains may show significant genetic differences. Clearly, there is a need to clone and sequence other strains in order to build a consensus picture of the CCV genome.

We wish to thank Dr W. Baxendale for supplying cells and virus stocks. This work was supported by a research grant from Intervet BV, Boxmeer, The Netherlands.

References

ARMSTRONG, J., NIEMANN, H., SMEEKENS, S., ROTTIER, P. & WARREN, G. (1984). Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. *Nature, London* **308**, 751–752.

ARPIN, N. & TALBOT, P. J. (1990). Molecular characterization of the 229E strain of human coronavirus. *Advances in Experimental Medicine and Biology* **276**, 73–80.

BANKIER, A. T., WESTON, K. W. & BARRELL, B. G. (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods in Enzymology* **155**, 51–93.

BARLOUGH, J. E., STODDART, C. A., SORRESSO, G. P., JACOBSON, R. H. & SCOTT, F. W. (1984). Experimental inoculation of cats with canine coronavirus and subsequent challenge with feline infectious peritonitis virus. *Laboratory Animal Science* **34**, 592–597.

BINN, L. N., LAZAR, E. C., KEENAN, K. P., HUXSOLL, D. L., MARCHWICKI, R. H. & SCOTT, F. W. (1974). Recovery and characterization of a coronavirus from military dogs with diarrhea. *Proceedings of the Annual Meeting of the U.S. Animal Health Association* **78**, 359–366.

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 genome 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., BOURSNELL, M. E. G., BINNS, M. M., BILIMORIA, B., BLOK, V. C., BROWN, T. D. K. & INGLIS, S. C. (1987). An efficient ribosomal frame-shifting signal in the polymerase encoding region of the coronavirus IBV. *EMBO Journal* **6**, 3779–3784.

BRITTON, P. & PAGE, K. W. (1990). Sequence of the S gene from a virulent British field isolate of transmissible gastroenteritis virus. *Virus Research* **18**, 71–80.

BRITTON, P., CARMENES, R. S., PAGE, K. W., GARWES, D. J. & PARRA, F. (1988*a*). Sequence of the nucleoprotein from a virulent British field isolate of transmissible gastroenteritis virus and its expression in *Saccharomyces cerevisiae*. *Molecular Microbiology* **2**, 89–99.

BRITTON, P., CARMENES, R. S., PAGE, K. W. & GARWES, D. J. (1988*b*). The integral membrane protein from a virulent isolate of transmissible gastroenteritis virus: molecular characterization, sequence and expression in *Escherichia coli*. *Molecular Microbiology* **2**, 497–505.

BRITTON, P., MAWDITT, K. L. & PAGE, K. W. (1991). The cloning and sequencing of the virion protein genes from a British isolate of porcine respiratory coronavirus: comparison with transmissible gastroenteritis virus genes. *Virus Research* **21**, 181–198.

CAVANAGH, D., BRIAN, D. A., ENJUANES, L., HOLMES, K. V., LAI, M. M., LAUDE, H., SIDDELL, S. G., SPAAN, W. J. M., TAGUCHI, F. & TALBOT, P. J. (1990*a*). Recommendations of the coronavirus study group for the nomenclature of the structural proteins, mRNAs and genes of coronaviruses. *Virology* **176**, 306–307.

CAVANAGH, D., DAVIS, P., COOK, J. & LI, D. (1990*b*). Molecular basis of the variation exhibited by avian infectious bronchitis virus (IBV). *Advances in Experimental Medicine and Biology* **276**, 369–372.

COLLISSON, E. W., WILLIAMS, A. K., HAAR, R. V., LI, W. & SNEED, L. W. (1990). Sequence comparisons of the 3′ end of the genomes of five strains of avian infectious bronchitis virus. *Advances in Experimental Medicine and Biology* **276**, 373–377.

DE GROOT, R. J. (1989). *A molecular study of feline infection peritonitis virus*. Ph.D. thesis, University of Utrecht.

DE GROOT, R. J., MADURO, J., LENSTRA, J. A., HORZINEK, M. C., VAN DER ZEIJST, B. A. M. & SPAAN, W. J. M. (1987). cDNA cloning and sequence analysis of the gene encoding the peplomer protein of feline infectious peritonitis virus. *Journal of General Virology* **68**, 2639–2646.

DE GROOT, R. J., ANDEWEG, A. C., HORZINEK, M. C. & SPAAN, W. J. M. (1988). Sequence analysis of the 3′ end of the feline coronavirus FIPV 79-1146 genome: comparison with the genome of porcine coronavirus TGEV reveals large insertions. *Virology* **167**, 370–376.

FICHOT, O. & GIRAD, M. (1990). An improved method for sequencing RNA templates. *Nucleic Acids Research* **18**, 6162.

GARWES, D. J. & REYNOLDS, D. J. (1981). The polypeptide structure of canine coronavirus and its relationship to porcine transmissible gastroenteritis virus. *Journal of General Virology* **52**, 153–157.

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

HIGGINS, D. G. & SHARP, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* **5**, 151–153.

HORZINEK, M. C., LUTZ, H. & PEDERSEN, N. (1982). Antigenic relationships among homologous structural polypeptides of porcine, feline and canine coronaviruses. *Infection and Immunity* **37**, 1148–1155.

JACOBS, L., VAN DER ZEIJST, B. A. M. & HORZINEK, M. C. (1986). Characterization and translation of transmissible gastroenteritis mRNAs. *Journal of Virology* **57**, 1010–1015.

KECK, J. G., MATSUSHIMA, G. K., MAKINO, S., FLEMING, J. O., VANNIER, J. O., STOHLMAN, S. A. & LAI, M. M. (1988). *In vivo* RNA–RNA recombination of coronavirus in mouse brain. *Journal of Virology* **64**, 6270–6273.

KEENAN, K. P., JERVIS, H. R., MARCHWICKI, R. H. & BINN, L. N. (1976). Intestinal infection of neonatal dogs with canine coronavirus 1-71: studies by virologic, histologic, histochemical and immunofluorescent techniques. *American Journal of Veterinary Research* **37:3**, 247–256.

KOZAK, M. (1986). Regulation of protein synthesis in virus-infected animal cells. *Advances in Virus Research* **31**, 229–292.

KUSTERS, J. G., JAGER, E. J. & VAN DER ZEIJST, B. A. M. (1989). Sequence evidence for *in vivo* RNA recombination in avian coronavirus IBV. *Nucleic Acids Research* **17**, 6726–6730.

LIM, H. M. & PÈNE, J. J. (1988). Optimal conditions for supercoil DNA sequencing with the *Escherichia coli* DNA polymerase I large fragment. *Gene Analysis Techniques* **5**, 32–39.

LIU, D. X. (1991). *Translational analysis of the coronavirus infectious bronchitis virus messenger RNAs*. Ph.D. thesis, University of Cambridge.

LIU, D. X. & INGLIS, S. C. (1992). Identification of two new polypeptides encoded by mRNA5 of the coronavirus infectious bronchitis virus. *Virology* **186**, 342–347.

LIU, D. X., CAVANAGH, D., GREEN, I. J. & INGLIS, S. C. (1991). A polycistronic mRNA specified by the coronavirus infectious bronchitis virus. *Virology* **184**, 531–544.

LUYTJES, W., BREDENBEEK, P., NOTEN, A., HORZINEK, M. C. & SPAAN, W. J. M. (1988). Sequence of the mouse hepatitis virus A59 mRNA2: indications for RNA recombination between coronaviruses and influenza C virus. *Virology* **166**, 415–422.

MASTERS, P. S., PARKER, M. M., RICARD, C. S., DUCHALA, C., FRANA, M. F., HOLMES, K. V. & STURMAN, L. S. (1990). Structure and function studies of the nucleocapsid protein of mouse hepatitis virus. *Advances in Experimental Medicine and Biology* **276**, 239–246.

NORMAN, J. O., MCCLURKIN, A. W. & STARK, S. L. (1970). Transmissible gastroenteritis (TGE) of swine: canine serum antibodies against an associated virus. *Canadian Journal of Comparative Medicine* **34**, 115–117.

PARKER, S. E., GALLAGHER, T. M. & BUCHMEIER, M. J. (1989). Sequence analysis reveals extensive polymorphism and evidence of deletions within the E2 glycoprotein gene of several strains of murine hepatitis virus. *Virology* **173**, 664–673.

RASSCHAERT, D. & LAUDE, H. (1987). The predicted primary structure of the peplomer protein E2 of the porcine coronavirus transmissible gastroenteritis virus. *Journal of General Virology* **68**, 1883–1890.

RASSCHAERT, D., DUARTE, M. & LAUDE, H. (1990). Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. *Journal of General Virology* **71**, 2599–2607.

SAMBROOK, J., FRITSCH, E. F. & MANIATIS, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. New York: Cold Spring Harbor Laboratory.

SANCHEZ, C. M., JIMENEZ, G., LAVIADA, M. D., CORREA, I., SUNE, C., BULLIDO, M. J., GEBAUER, F., SMERDOU, C., CALLEBAUT, P., ESCRIBANO, J. M. & ENJUANES, L. (1990). Antigenic homology among coronaviruses related to transmissible gastroenteritis virus. *Virology* **174**, 410–417.

SCHMIDT, I., SKINNER, M. & SIDDELL, S. (1987). Nucleotide sequence of the gene encoding the surface projection glycoprotein of the coronavirus MHV-JHM. *Journal of General Virology* **68**, 47–56.

SIDDELL, S., WEGE, H. & TER MEULEN, V. (1983). The biology of coronaviruses. *Journal of General Virology* **64**, 761–776.

SKINNER, M. A. & SIDDELL, S. G. (1983). Coronavirus JHM: nucleotide sequence of the mRNA that encodes the nucleocapsid protein. *Nucleic Acids Research* **11**, 573–578.

SMITH, A. R., BOURSNELL, M. E. G., BINNS, M. M., BROWN, T. D. K. & INGLIS, S. C. (1990). Identification of a new membrane-associated polypeptide specified by the coronavirus infectious bronchitis virus. *Journal of General Virology* **71**, 3–11.

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

STADEN, R. (1982). Automation of the computer handling of the gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Research* **10**, 4731–4751.

STODDART, C. A., BARLOUGH, J. E., BALDWIN, C. A. & SCOTT, F. W. (1988). Attempted immunization of cats against feline infectious peritonitis using canine coronavirus. *Research in Veterinary Science* **45**, 383–388.

VENNEMA, H., DE GROOT, R. J., HARBOUR, D. A., HORZINEK, M. C. & SPAAN, W. J. M. (1991). Primary structure of the membrane and nucleocapsid protein genes of feline infectious peritonitis virus and immunogenicity of recombinant vaccinia viruses in kittens. *Virology* **181**, 327–335.

VON HEIJNE, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**, 4683–4690.

WESLEY, R. D., CHEUNG, A. K., MICHAEL, D. D. & WOODS, R. D. (1989). Nucleotide sequence of coronavirus TGEV genomic RNA: evidence for 3 mRNA species between the peplomer and matrix protein genes. *Virus Research* **13**, 87–100.

WESLEY, R. D., WOODS, R. D. & CHEUNG, A. K. (1990). Genetic basis for the pathogenesis of transmissible gastroenteritis virus. *Journal of Virology* **64**, 4761–4766.

WOODS, R. D. & PEDERSEN, N. C. (1979). Cross-protection studies between feline infectious peritonitis virus and porcine transmissible gastroenteritis viruses. *Veterinary Microbiology* **4**, 11–16.

WOODS, R. D. & WESLEY, R. D. (1986). Immune response in sows given transmissible gastroenteritis virus or canine coronavirus. *American Journal of Veterinary Research* **47**, 1239–1242.

WOODS, R. D., CHEVILLE, N. F. & GALLAGHER, J. E. (1981). Lesions in the small intestine of new-born pigs inoculated with porcine, feline and canine coronaviruses. *American Journal of Veterinary Research* **42**, 1163–1169.

(Received 30 March 1992; Accepted 7 July 1992)