SEQUENCE ANALYSIS OF CCV AND ITS RELATIONSHIP TO FIPV, TGEV AND PRCV

Brian C. Horsburgh and T. David K. Brown

Division of Virology
Department of Pathology
University of Cambridge
Tennis Court Road
Cambridge CB2 1QP
UK

INTRODUCTION

Canine coronavirus (CCV), a causative agent of enteritis in newborn dogs, was first identified in 1971[1]. The disease is characterized by infection of the absorptive epithelium of the villi and the onset of diarrhoea followed by villus atrophy[2]. The CCV virion is known to contain at least three protein species viz. the 204 kDa spike glycoprotein, S, the 32 kDa membrane glycoprotein, M, and the 50 kDa nucleocapsid protein, N[3].

CCV belongs to one of the major antigenic groups of coronaviruses[4,5] and is serologically related to feline infectious peritonitis virus (FIPV), feline enteric coronavirus (FECV), transmissible gastroenteritis virus (TGEV), and porcine respiratory coronavirus (PRCV)[6]. 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[7,8] and swine[9]. Likewise TGEV can also infect the other member species[10,11], and FIPV can infect swine[12]. This close relationship indicates that the viruses may have a common ancestor[13,6].

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[14,15,16,17,18,19]. A comparison of the 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[20,21,22]. Despite this high degree of homology 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 organization.

The close relationship between TGEV, PRCV, FIPV and CCV has considerable epidemiological implications. Clearly, more information is needed in order to better understand the taxonomic relationship of this antigenic group of viruses. CCV is the least characterized virus from this antigenic group; it was hoped that cloning, sequencing and subsequent analyses would help to illuminate the taxonomic relationship of this family of viruses.

Coronaviruses, Edited by H. Laude and J.F. Vautherot
Plenum Press, New York, 1994

METHODS

The materials and methods are described in detail by Horsburgh *et al.*[23]. Briefly, virus was pelleted from the supernatant of CCV-Insavc-1 infected A72 cells, and the pellet homogenized in guanidinium isothiocyanate solution. The mixture was layered onto a CsCl pad, and the viral RNA pelleted by centrifugation. The RNA was dissolved in TE, oligo d(T) selected, and the concentration determined (A260). A cDNA library was produced from approximately 25µg poly $(A)^+$ tailed RNA using oligo d(T) and random pentanucleotides as primers[24]. The cDNA was blunted ended using T4 DNA polymerase and ligated into the SmaI site of pUC119. Portions of the ligation mixture were transformed into E.coli strain TG-1 and clones identified by colour selection. Positive colonies were probed with radio-labelled, randomly primed CCV cDNA. 'Minipreps' of plasmid DNA were prepared from colonies which gave the strongest signals. A number of the colonies contained inserts of 1.8 kb or over and were selected for further analysis.

PCR-amplified fragments were obtained using cDNA:RNA heteroduplexes as templates and oligonucleotides 7 (5' GTT GCA ATT GCG GCC GCA CAG TTA TTA TTG TTG) and 8 (5' CCC ATT GGC AAC GCG GCC GCT GTC ACC AAA ATT GGC), each of which was modified to contain a NotI site, as primers. Amplification of the cDNA was performed as described by Sambrook *et al.*[25] using Taq polymerase. The generated fragment was cleaved with NotI, gel purifed and ligated into the NotI site of pKL1. (pKL1 is a pUC based vector with a modified polylinker and was a gift from Dr. K Law, University of Cambridge).

The nucleotide sequence of the overlapping cDNA clones at the 3'-end of the genome (BH5 - 10) was determined using the M13/dideoxynucleotide method[26]. Briefly, insert DNA was excised from vector sequences, self ligated and sonicated. The sonicated DNA fragments were end-repaired with the Klenow fragment of E.coli DNA polymerase and T4 DNA polymerase prior to size selection on a 1.2% agarose gel. Fragments in the size range 300-500 bps were purified and subcloned into SmaI digested, phosphatased M13mp8, from which single stranded DNA was prepared and used as template for sequencing reactions.

The nucleotide sequence data obtained was analyzed using the programs of Staden[27] on a VAX 8350 computer.

RESULTS

To clone the 3'-end of the CCV genome we prepared a cDNA library from CCV genomic RNA. Inserts from recombinant clones of approximately 1.8kb or greater in size were selected for further analysis. In order to map the clones, we took advantage of the suspected nucleotide sequence homology between the genomes of CCV and TGEV[28]. Double stranded sequencing of recombinant plasmid DNA, using the pUC forward and reverse primers, revealed approximately, 150 nts of sequence at each end of the viral insert. Comparison of these sequences with published TGEV sequences disclosed 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 five clones (BH5, 7, 8, 9 and 10) were identified which spanned in total, some 8.5 kb at the 3'-end (see Fig. 1). A region at the 3'-end for which large clones were not present in the library was prepared by PCR amplification (BH6; Fig. 1). Briefly, partial sequence information obtained from the 3'-end of BH7 and the 5'-end of BH5, allowed the design of primers 7 and 8, each of which was modified to contain a NotI site. PCR-amplified fragments were obtained using cDNA:RNA heteroduplexes as template and oligonucleotides 7 and 8 as primers. The generated DNA fragment was cloned into the NotI site of pKL1 and transformed into E.coli strain TG-1. A positive clone, pBH6 was selected for analysis. The relationships between putative overlapping clones were confirmed by Southern hybridization.

The consensus sequence determined from the six overlapping cDNA clones, 9580 bps long (exclusive of the poly (A) tail), is presented elsewhere[23]; EMBL/Genbank accession number D13096). This consensus sequence was analyzed using the SAP programs of Staden[27]. Analysis revealed the presence of 10 open reading frames (ORFs) over 50 amino acids in length (see Fig. 2). Pairwise alignment of these ORFs with their likely counterparts from other members of this coronavirus group disclosed very high

levels of identity and indicated that the CCV structural proteins, S, M and N are encoded by ORFs 2, 5 and 6 respectively.

With respect to subgenomic mRNA synthesis, it is known that the 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.*[5]). Indeed, analysis of the CCV sequence revealed that this sequence was present upstream of all the

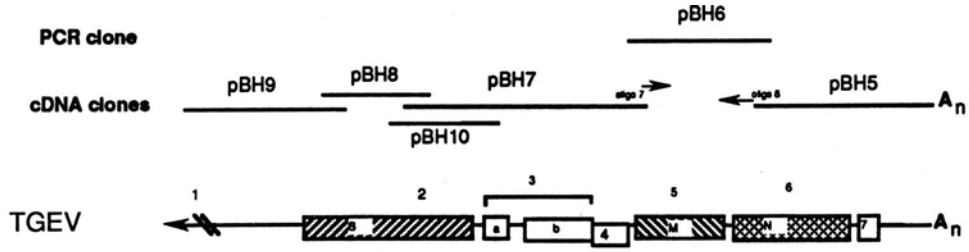

Fig. 1
Alignment of CCV cDNA clones with respect to the TGEV genome using partial sequence information.

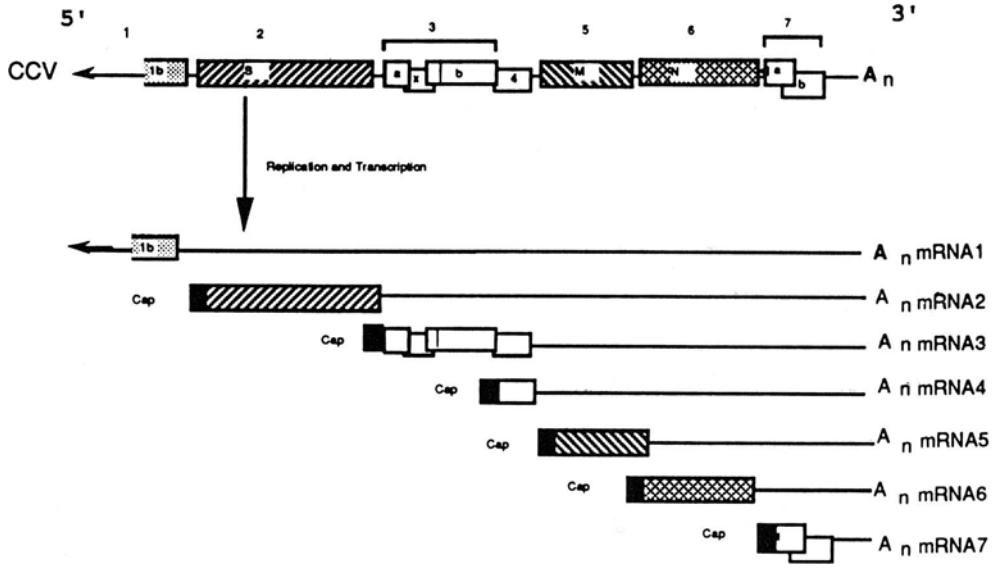

Fig. 2
Gene and subgenomic message organization predicted from the sequence data. The vertical line in ORF 3b represents a stop codon and black boxes represent leader sequences. Numbers represent ORFs encoded by that message.

ORFs with the exception of 1b, 3x and 7b. As ORF 1 is incomplete (see below), an additional CTAAAC sequence is presumably located at the 5'-end of the genomic RNA. When we analyzed intracellular RNAs produced during CCV infection of canine A72 cells: seven species of mRNA were observed[23]. 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. 2. This organization is very similar to that

observed by Groot *et al.*[29] for FIPV except that CCV has an extra message (mRNA4) which has not yet been detected in FIPV-infected cells. The ORFs encoded by each mRNA are described below.

ORF 1 is incomplete, it has no ATG start codon, consists of 168 amino acids and terminates in a UGA stop codon at position 510. A comparison of this ORF with TGEV strain FS772/70 shows 99.2% similarity to 1b and 47% and 52.7% identity to gene 1b of IBV and MHV respectively[30,31,32]. Thus this ORF represents the 3'-end of the putative polymerase encoding region of genome mRNA1.

ORF 2, located immediately downstream of the polymerase gene would be translated from the 9.1 kb subgenomic message 2. This ORF, 4356 nts long representing 1452 amino acids with a calculated MW of 160 kDa, is preceded by the potential RNA polymerase-leader complex binding site, CTAAAC, 32 bps upstream of the translation initiation site. Comparison of this ORF with sequences held in the EMBL database reveals remarkably high identity to the FIPV spike glycoprotein coding sequence (91.2%) and to a lesser degree, the porcine S genes (79% TGEV; 75% PRCV), identifying this ORF as the CCV S gene.

The CCV S protein shares characteristic features of a type one membrane protein viz. a putative signal sequence and transmembrane domain. There are also 30 potential N-glycosylation sites; glycosylation probably accounts for the higher apparent MW of the S protein found in the virion[3].

There are four ORFs distal of the S gene coding sequence which are likely to be encoded by mRNAs 3 and 4 (Fig. 2). Three of these have high identity to their porcine counterparts, TGEV and PRCV, and have been named 3a (8.6 kDa; 83.5%/48.8%), 3b (28.4 kDa; 92.7%/92.6%) and 4 (9.3 kDa; 88.4%/88.4%). The fourth ORF; which to date has not been detected in this family of viruses could potentially encode a 71 amino acid protein with a predicted MW of 10 kDa and overlaps ORFs 3a and 3b (see Fig. 2). This ORF has been designated 3x. The CCV 3b ORF was expected to encode a 28 kDa protein like its TGEV counterpart[33], however, this strain of CCV has an internal termination codon, TAA, 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 IBV 3c ORF shares similar features with the gene 4 polypeptide of CCV, such as a hydrophobic core preceded by aspartate or glutamate residues and followed by conserved cysteine and proline residues. It is possible that the CCV gene 4 polypeptide, like IBV 3c, is found in the virion envelope.

Message 4, as predicted from our sequence data, could only encode ORF 4, as the supposed signal for transcription, CTAAAC, is found 43 nts upstream of the predicted ORF 4 start codon. A counterpart for this CCV message has not yet been detected in FIPV-infected cells[29].

Messenger RNA species 5 and 6 encode ORFs 5 and 6 which have 84%/88% and 77%/90% identity to the M and N coding sequences of FIPV and TGEV respectively. Translation of poly (A) selected CCV intracellular RNA in the rabbit reticulocyte lysate system produced products of the expected size of M (25 kDa) and N (45 kDa) when analyzed by SDS-PAGE (data not shown).

ORFs 7a and 7b are likely to be encoded on a single RNA species (mRNA7) since smaller messages were not seen on Northern blots (data not shown), nor is another message predicted from the sequence data. Furthermore, an equivalent RNA in FIPV is thought to be bicistronic[34] 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 the TGEV and PRCV revealed that the 7a ORF of the porcine coronaviruses has undergone a deletion of 69 nts and furthermore, they have no counterpart to ORF 7b. Nevertheless, the CCV structural ORFs, with the exception of S, have higher identities to TGEV than to FIPV.

The CCV sequence contains the octameric sequence, GGAAGAGC, at the 3'-end of the genome, upstream of the poly (A) site, which is conserved in all coronavirus sequences to date.

DISCUSSION

In this study approximately 9.6 kb from the 3' genomic end of CCV strain Insavc-1 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 the other antigenically related coronaviruses, namely TGEV, PRCV and FIPV. The respective genetic organizations are shown in Fig. 3.

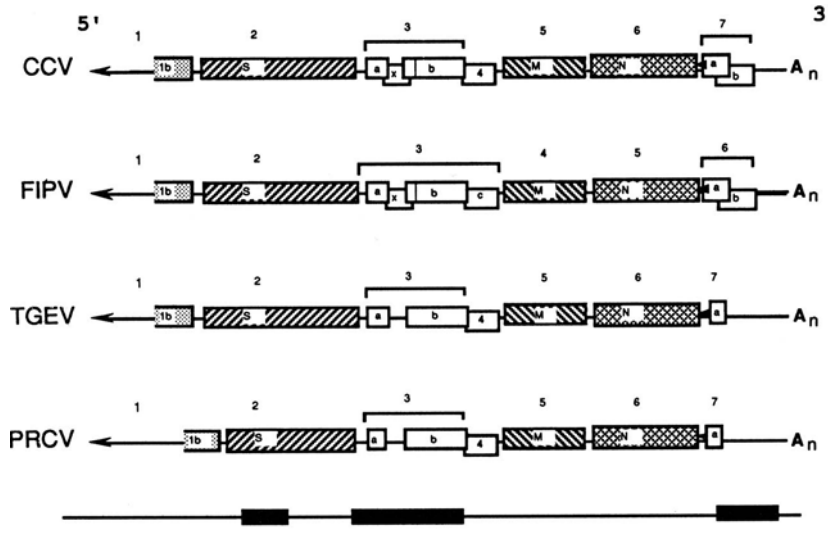

Fig. 3
Genomic organization of CCV, FIPV, TGEV and PRCV. The black boxes represent putative 'hot spot' regions.

Fig. 3
Genomic organization of CCV, FIPV, TGEV and PRCV. The black boxes represent putative 'hot spot' regions.

From antigenic data and cross infectivity studies, the viruses within this group have been termed 'host range mutants'[13]. This close relationship is emphasized by our analyses of the CCV sequence data. The CCV spike 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 reflect changes in cell tropism, genetic drift and selection from the host's immune system. Similarly, interspecies comparison of the other structural proteins M and N, again revealed very high levels of identity (see results).

The genome organizations of both CCV and FIPV are virtually identical (Fig. 3), implying that CCV is more related to the feline coronaviruses than to the porcine coronaviruses. Nevertheless, the four viruses have strikingly similar genetic organizations. However, it would appear that there are 3 'hot spot' regions where recombination, insertions or usually deletions occur at a higher frequency relative to the surrounding sequences. These are within S, between S and M, and downstream of N (Fig. 3). The dynamism of the coronavirus genome is well documented[35,36] 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 different template.

The PRCV S gene has acquired a deletion which may be responsible for the change in virus tropism. The polymorphism of S found in MHV strains with differing passage histories is mainly due to deletions in that gene which can be up to 159 amino acids. This can have 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 the capacity to produce a non-fatal demyelinating disease in mice[37].

Polymorphism is also observed in the second 'hot spot' region, between the S and M genes (Fig. 3). The CCV ORFs which lie between the spike and membrane genes like the other ORFs so far analyzed have high identities to their porcine counterparts and presumably perform similar functions. 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 40bp to over 200bp.

Alignment of the ORF 3a amino acid sequences reveals that in addition, variation is found at the ends of these coding sequences. Moreover, another deletion in the PRCV genome results in the loss of a functional 3a protein.

A previously undetected ORF, 3x was identified which could potentially encode a 10 kDa polypeptide. However codon usage and base preference programs of Staden[27] suggest that this ORF is not a coding sequence. Furthermore, the proximal ATG is in a poor context for translation initiation[38], and the only other ATG is found at the very 3'-end of the coding sequence. Therefore it is very unlikely that this ORF will be expressed in this strain of CCV; it may represent an evolutionarily redundant sequence which is no longer required by the virus or may contain signals important for translation of downstream ORFs. Analysis of TGEV genomic sequence in this region revealed a counterpart for this canine pseudogene; 92 nucleotides have however been deleted. This deletion also results in a frame-shift in the sequence which explains why this ORF has not hitherto been noticed. In addition to the likely non-functionality of ORF 3x, it is also unlikely that ORF 3b is expressed in this strain of CCV as there is an internal termination codon (TAA; represented by a vertical bar in Fig.3) some 93 nts downstream of the first ATG and subsequent ATG codons are in poor contexts for ribosome binding[38]. *In vitro* transcription and translation of this ORF did not yield any discernible products on SDS-polyacrylamide gels (data not shown). Further evidence that this region may function as a recombinational 'hot spot' comes from a study by Cavanagh *et al.*[39]. In it they reported an IBV strain (Port/322/85) which appears to have arisen as a result of recombination between the M and S genes from two other strains of IBV.

The third "hot spot region" is found downstream of the N gene (Fig. 3). The porcine coronaviruses have a 69 nt deletion in ORF 7a whilst ORF 7b is not present[34]. This phenomenon is not unique to the coronaviruses from this antigenic group. Deletions of up to 170 nts are found downstream of the N gene in some strains of IBV[40]. It is interesting to note that the CCV ORF 7b has 57% identity to FIPV ORF 6b. This ORF is the least conserved between the two viruses.

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 differ genetically. Clearly, there is a need to clone and sequence other strains in order to build a consensus picture of the CCV genome.

REFERENCES

1. Binn, L.N., Lazar, E.C., Keenan, K.P., Huxsoll, D.L., Marchwicki, R.M. & Scott, F.W. Proc.U.S. Animal Health Assoc. 78:359-366 (1974).
2. Keenan, K.P., Jervis, H.R., Marchwicki, R.H. & Binn, L.N. Amer. J. Vet. Res. 37: 247-256 (1976).
3. Garwes, D.J. & Reynolds, D.J. J. gen. Virol. 52: 153-157 (1981).
4. Siddell, S., Wege, H. & ter Meulen, V.. J. gen. Virol. 64: 761-776 (1983).
5. Spaan, W., Cavanagh, D. & Horzinek, M.C. J. gen. Virol. 69: 2939-2952 (1988).
6. Sanchez, C.M., Jimenez, G., Laviada, M.D., Correa, I., Sune, C., Bullido, M.J., Gebaues, F., Smerdou, C., Callebaut, P., Escribano, J.M. & Enjuanes, L. Virol. 174: 410-417 (1990).
7. Barlough, J.E., Stoddart, C.A., Sorresso, G.P., Jacobson, R.H. & Scott, F.W. Lab. Animal Sci. 34: 592-597 (1984).
8. Stoddart, C.A., Barlough, J.E., Baldwin, C.A. & Scott, F.W. Res. Vet. Sci. 45: 383-388 (1988).
9. Woods, R.D. & Wesley, R.D. Amer. J. Vet. Res. 47: 1239-1242 (1986).
10. Norman, J.O., McClurkin, A.W. & Stark, S.L. Can. J. Comp. Med. 34: 115-117 (1970).
11. Woods, R.D. & Pedersen, N.C. Vet. Microbiol. 4: 11-16 (1979).
12. Woods, R.D., Cheville, N.F. & Gallagher, J.E. Amer. J. Vet. Res. 42: 1163-1169 (1981).
13. Horzinek, M.C., Lutz, H. & Pedersen, N. Infect. and Immun. 37: 1148-1155 (1982).
14. Groot, R.J. de, Maduro, J., Lenstra, J.A., Horzinek, M.C., Zeijst, B.A.M. van der & Spaan, W.J.M. J. gen. Virol. 68: 2639-2646 (1987).
15. Rasschaert, D. & Laude, H. J. gen. Virol. 68: 1883-1890 (1987).

16. Britton, P., Carmenes, R.S., Page, K.W., Garwes, D.J. & Parra, F. Mol. Microbiol. 2: 89-99 (1988).
17. Britton, P., Carmenes, R.S., Page, K.W. & Garwes, D.J. Mol. Microbiol. 2: 497-505 (1988).
18. Rasschaert, D., Duarte, M. & Laude, H. J. gen. Virol. 71: 2599-2607 (1990).
19. Vennema, H., Groot, R.J. de., Harbour, D.A., Horzinek, M.C. & Spaan, W.J.M. Virol. 181: 327-335 (1991).
20. Skinner, M.A. & Siddell, S.G. Nucl. Acids Res. 11: 5045-5054 (1983).
21. Armstrong, J., Neimann, H., Smeekens, S., Rottier, P. & Warren, G. Nature 308: 751-752 (1984).
22. Schmidt, I., Skinner, M. & Siddell, S.G. J. gen. Virol. 68: 47-56 (1987).
23. Horsburgh, B.C., Brierley, I. & Brown, T.D.K. J. gen. Virol. 73: 2849-2862 (1992).
24. Gubler, U. & Hoffman, B.J. Gene 25: 263-269 (1983).
25. Sambrook, J., Fritsch, E.F. & Maniatis, T. Molecular cloning: a laboratory manual 2nd edn., Cold Spring Harbour Laboratory, New York 1989).
26. Bankier A.T., Weston, K.W. & Barrell, B.G. Methods in Enzymology 155: 51-93 (1987).
27. Staden, R. Nucl. Acids Res. 14: 217-231 (1986).
28. Shockley, L.J., Kapke, P.A., Lapps, W., Brian, D.A., Potgeiter, L.N.D. & Woods, R. J. Clin. Microbiol. 25: 1591-1596 (1987).
29. Groot, R.J. de, Haar, R.J. ter, Horzinek, M.C. & Zeijst, B.A.M van der . J. gen. Virol. 68: 995-1002 (1987).
30. Boursnell, M.E.G., Brown, T.D.K., Foulds, I.J., Green, P.F., Tomley, F.M. & Binns, M.M. J. gen. Virol. 68: 57-77 (1987).
31. Bredenbeek, P.J., Pachuk, C.J., Noten, A.F.H., Charite, J., Lutyes, W., Weiss, S.R. & Spaan, W.J.M. Nucl. Acids Res. 18: 1825-1832 (1990).
32. Britton, P, & Page, K.W. Virus Res. 18: 71-80 (1990).
33 Jacobs, L., Zeijst, B.A.M. van der & Horzinek, M.C. J. Virol. 57: 1010-1015 (1986).
34. Groot, R.J. de, Andeweg, A.C., Horzinek, M.C. & Spaan, W.J.M. Virol. 167: 370-376 (1988).
35. Keck, J.G., Matsushima, G.K., Makino, S., Fleming, J.O., Vannier, D.M., Stohlman, S.A. & Lai, M.M.C. J. Virol. 62: 1810-1813 (1988).
36. Kusters, J.G., Jager, E.J. & Zeijst, B.A.M. van der. Nucl. Acids Res 17: 6726-6729 (1989).
37. Parker, S.E., Gallagher, T.M. & Buchmeier, M.J. Virol. 173: 664-673 (1989).
38. Kozak, M. Adv.Virus Res. 31: 229-292 (1986).
39. Cavanagh, D., Davis, P, Cook, J. & Li, D. Adv. Exptl. Med. Biol. 276: 369-372 (1990).
40. Collisson, E.W., Williams, A.K., Haar, R.V., Li, W. & Sneed, L.W. Adv. Exptl. Med. Biol. 276: 373-377 (1990).