Arch Virol (1993) [Suppl] 7: 63–74

Molecular analysis of the human coronavirus (strain 229E) genome

J. Herold, T. Raabe, and **S. Siddell**

Institute of Virology, University of Würzburg, Würzburg, Federal Republic of Germany

Summary. The nucleotide sequence of the human coronavirus strain 229E (HCV 229E) has been determined. This article describes the organization of the virus genome, the predicted viral gene products and the mechanisms which regulate viral gene expression. This information provides a basis to investigate the biology and pathogenesis of HCV.

Introduction

Human coronaviruses are one of the causative agents of the common cold. HCV infections are generally mild, last only a few days and are seldom associated with severe symptoms such as headache, fever or diarrhea. Nevertheless, the economic consequence of respiratory disease caused by HCVs is significant [10]. HCVs can be divided into 2 major serological groups, represented by the prototypes, HCV 229E and HCV OC43. The two groups are about equally responsible for human respiratory infections [11]. Both virus types are difficult to isolate and both grow poorly in tissue culture. HCV 229E can, however, be adapted to grow in human lung fibroblasts, but even so, there is only a limited amount of information on the protein components of the virion, viral RNA and protein synthesis, the regulation of viral gene expression and the function of the viral gene products [1, 12, 21].

As a basis for our studies on the biology and pathogenesis of HCV 229E, we have undertaken a sequence analysis of the viral genome. With this information it is possible to

1. describe the organization of the HCV 229E genome
2. identify structural features which may play a role in the regulation of viral gene expression
3. predict the entire complement of viral gene products and
4. make predictions concerning the possible structure-function relationships of the viral proteins.

Materials and methods

Virus

The HCV 229E isolate used in these studies was obtained from a volunteer at the MRC Common Cold Unit, Salisbury, U.K. The virus was adapted to culture in C16 cells [16], titrated to limiting dilution and the supernatant from a well with one focus of infection was used to prepare a virus stock. C16 cells were infected with HCV 229E at an m.o.i. of 3, incubated at 33°C and cytoplasmic RNA was isolated 48 h.p.i. using standard procedures. Poly-A RNA was obtained by poly-U-Sepharose chromatography.

cDNA cloning

cDNA libraries were prepared essentially by the method of Gubler and Hoffman [9] using random hexanucleotide or virus specific oligonucleotide primers. The cDNA was size fractionated and cloned into the Bluescript vector pKS II^+ (Stratagene). Recombinant clones were identified by colony hybridization with HCV 229E specific oligonucleotides. Plasmid purification, agarose gel electrophoresis, colony hybridizations and standard recombinant DNA procedures were done as described by Sambrook et al. [20].

PCR amplification

PCR was performed using a GeneAmp/RNA PCR kit (Perkin Elmer Cetus) according to the manufacture's procedures using biotinylated and non-biotinylated, HCV specific primers. The resulting cDNA strands were separated with streptavidin coupled magnetic beads (Dynal) and the nucleotide sequence of both strands was determined.

Sequence analysis

Sequencing was done on double and single strand DNA templates (cDNA and PCR amplification products) using the chain termination method and M13, T7, T3 and HCV 229E specific primers. To generate sequencing templates, cDNAs were subcloned by restriction enzyme digestion and overlapping deletions were introduced by exonuclease III. Sequence data were assembled by the programmes of Staden [23] and analysed by the programmes of the University of Wisconsin Genetics Computer Group [8].

Results

The genomic RNA

The genomic RNA of HCV 229E is comprised of 27,277 nucleotides and a 3′ poly-A tract of not less than 50 residues. By analogy to murine hepatitis virus (MHV), it is assumed that the 5′ end of the genome is

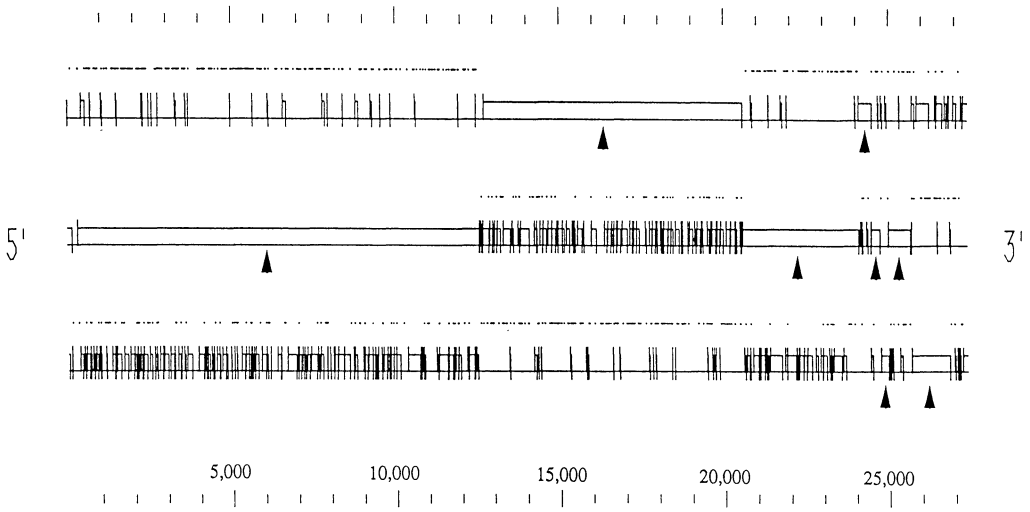

Fig. 1. Open reading frame and codon usage analysis of the HCV 229E genomic RNA. The analysis was performed using the UWGCG programmes FRAMES and CODONPREFERENCE

linked to a "cap" structure but this has not been directly demonstrated for HCV. Computer assisted analysis of the HCV 229E genome sequence reveals 8 non-redundant open reading frames (nrORFs) of more than 50 codons. There are also smaller nrORFs at the 5′ and 3′ ends of the genome and many redundant open reading frames (rORFs) located within the coding regions of the larger nrORFS. If a codon frequency table is deduced on the basis of the 8 nrORFs and applied to the entire genome (Fig. 1) there is no indication that the small nrORFs or the rORFs are expressed.

This analysis would not, however, be sensitive enough to predict viral gene products generated, for example, by mechanisms such as RNA editing or frameshift mutation. Also, it does not exclude the translation of an 11 codon nrORF located at the 5′ end of the genome, in frame with and preceding the first large nrORF. Indeed, this small ORF is conserved in the genomes of HCV 229E, MHV and IBV [3, 13] and it may be speculated to have a role in the regulation of the initiation of protein synthesis from genomic RNA.

By analogy to the MHV and IBV genomes and their gene products, it is possible to assign the products of the HCV 229E genome to the 8 nrORFs described above. In some cases, these assignments can be strengthened by comparing the properties of predicted gene products with those of the known HCV proteins. This assignment is shown in Fig. 2.

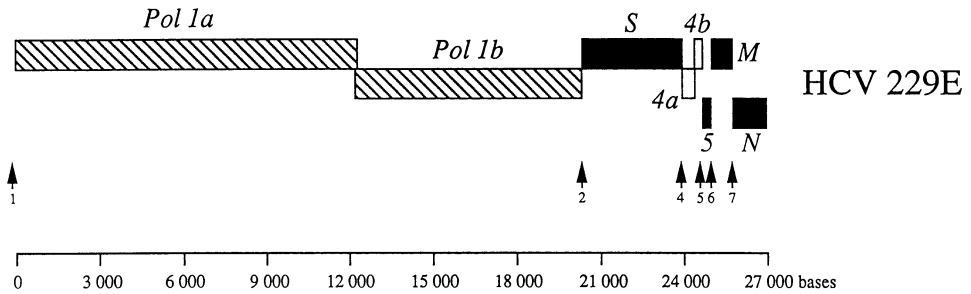

Fig. 2. Organization of the HCV 229E genome. The diagram is drawn to scale and the non-redundant ORFs are drawn in the correct reading frames. The structural protein genes are shown as black boxes, the non-structural protein genes as open (unknown function) or hatched (RNA polymerase) boxes

The genes of HCV 229E

The genes of HCV 229E are listed in Table 1 together with the predicted sizes of their assigned gene products.

The nucleocapsid gene

The nucleocapsid protein gene lies at the 3′ end of the genome. It encodes a polypeptide of M_r 43,500 which is in agreement with the apparent molecular weight of the HCV 229E N protein in SDS-PAGE [14]. In common with other coronavirus nucleocapsid proteins, the HCV

Table 1. The gene products of HCV 229E

Gene	Bases	Codons	Protein	Molecular mass
5′ UTR	293	–	–	–
ORF 1a	12,258	4,086	[polymerase]	454,200
ORF 1ab	20,277	6,759	[polymerase]	754,200
S	3,522	1,174	surface	128,600
ORF 4a	402	134	unknown	15,300
ORF 4b	267	89	unknown	10,200
ORF 5	234	78	[small membrane]	9,100
M	678	226	membrane	26,000
N	1,170	390	nucleocapsid	43,500
3′ UTR	422 +poly A	–	–	–

229E N protein is a serine-rich, basic protein (net charge +16 at neutral pH) and the protein is most probably phosphorylated. The distribution of basic and acidic residues is compatible with a three domain structure, as proposed for the MHV N protein by Parker et al. [15].

The membrane glycoprotein gene

The membrane glycoprotein gene is located adjacent to the N protein gene and encodes a polypeptide of 225 amino acids with an M_r of 26,000. The HCV M protein has several features which are characteristic of a coronavirus membrane protein. First, there are 3 potential N-linked glycosylation sites, one of which is near the amino terminus. It has been shown that the HCV 229E M protein is N-glycosylated [12]. Second, the polypeptide displays three internal hydrophobic domains within the amino terminal half and a relatively hydrophilic carboxy terminus. Third, the polypeptide is slightly basic with a net charge of +4 at neutral pH. These data suggest that the membrane topology of the HCV 229E M protein is very similar to that proposed by Rottier et al. [19] for the MHV M protein.

The surface glycoprotein gene

The HCV 229E surface glycoprotein gene encodes a polypeptide of 1,173 amino acids with an M_r of 128,600. The polypeptide has 30 potential N-glycosylation sites. The difference in the predicted M_r of the S protein and its apparent molecular weight in SDS-PAGE (180,000; [21]) suggests that the majority of these sites are used. A number of structural features typical of coronavirus S proteins can be recognized in the HCV 229E S protein gene product. These include an amino terminal signal sequence, a carboxy terminal membrane anchor, heptad repeat structures and a carboxy terminal cysteine cluster. In contrast to the S proteins of MHV and IBV, the HCV 229E S protein does not contain a basic region with the motif RRXRR or RRAHR (where X is F, S, H or A) which have been identified as the sites at which the MHV and IBV S proteins are proteolytically cleaved. Apparently, the HCV S protein is not post-translationally cleaved.

A detailed, computer-assisted comparison of the HCV 229E S protein with the published S protein sequences of other coronaviruses has also revealed a number of highly conserved cysteine residues (excluding those within the cysteine cluster) which are probably important in determining the three-dimensional structure of the protein. Clearly, however, the

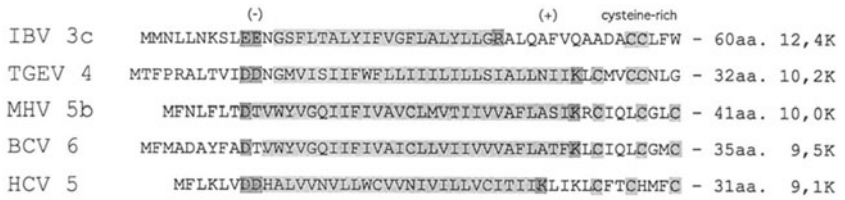

Fig. 3. Sequence similarity of different coronavirus SM proteins. The protein sequences are taken from the literature and have been aligned to emphasize the amino proximal hydrophobic domain, the flanking charged residues and the cysteine cluster

relevance of these features will only become apparent when a detailed structural analysis of the HCV 229E S protein is made.

The ORF 5 gene

On the basis of its structural similarity to the SM proteins of IBV and TGEV (Fig. 3), the HCV ORF 5 gene product is most probably a structural protein of the virus. The predicted M_r of the HCV SM protein is 9,100. To date, no protein of this size has been identified in HCV 229E virions, however, this may be due to the low amount of protein which is incorporated into the virus particle. In addition to its similarity to other coronavirus SM proteins, there is a striking resemblance with the M2 protein of influenza virus. This protein has recently been shown to have an associated ion channel activity selective for monovalent ions [17] but the possible relevance of this activity for the coronavirus infection process is unknown.

The coronavirus non-structural protein genes fall into two categories. On the one hand, a single, large gene located at the 5′ end of the genome encodes the viral RNA polymerase, whilst, a variable number of small genes encoding proteins of as yet unknown function are interspersed between the N, M and S protein structural genes. In the case of HCV 229E this pattern is repeated.

The RNA polymerase gene

The RNA polymerase gene is comprised of two overlapping ORFs of 4,086 and 6,759 codons respectively. Together these ORFs have the potential to encode polypeptides with a total M_r of 754,200. As described below, there is evidence from studies on MHV and IBV that the coronavirus polymerase gene is polycistronic and its expression appears to involve mechanisms such as ribosomal frame shifting and autoproteolytic

processing. The size of the gene leads one to suspect that this region of the genome may also encode functions which are not related to viral RNA synthesis.

A computer-assisted analysis of the HCV 229E polymerase gene reveals a number of sequence motifs which have been associated with RNA replicative functions (an RNA polymerase module, a helicase motif and a metal binding domain) as well as protease motifs (two papain-like and one 3C-like motif) which may encode activities involved in the post-translational proteolytic processing of the RNA polymerase (and perhaps other) gene products. These motifs have also been reported for the MHV and IBV RNA polymerase genes [13]. Additionally, a number of highly conserved regions can be identified in the HCV 229E, MHV and IBV polymerase genes, although it is not yet possible to ascribe them to any particular function.

The ORF 4a and ORF 4b gene

The remaining two HCV 229E nrORFs are ORF 4a and ORF 4b. The proteins encoded by these genes have predicted M_rs of 15,300 and 10,200 respectively. To date, these proteins have not been identified in the infected cell or in virions and they are provisionally considered to be non-structural proteins of unknown function.

Gene regulation

Transcriptional regulation

The expression of coronavirus genes is mediated by a set of 3′ coterminal subgenomic mRNAs. In the case of HCV 229E, 5 subgenomic mRNAs have been identified [18]. With the exception of the smallest mRNA each of these subgenomic RNAs is structurally polycistronic but, in general, they are believed to be functionally monocistronic. This principle has been established, at least for MHV and IBV. As one mRNA encodes only one gene product, the amount of each subgenomic mRNA will largely determine the amount of the individual proteins synthesized.

The coronavirus mRNAs are synthesized in non-equimolar amounts but in a constant ratio throughout the infection cycle. Studies on MHV [22] have shown that the generation of coronavirus mRNAs involves a process of discontinuous transcription. A specific sequence, related to motif UCUAAAC, is found at those positions in the genome which define the 5′ end of the unique region of each mRNA, i.e. the region not

Table 2. The nucleotide sequence of HCV 229E intergenic regions

Region	Sequence
5' UTR / Polymerase	G U C U A C U U U U **U C U C A A C** U A A A C G A A A [N_{214}] A U G
Polymerase / S	A U C A U U U A G **U C U C A A C** U A A A U A A A A U G
S / ORF 4	U G U G A A U C A **A C U A A A C** U U C C U U U U A [N_{30}] A U G
ORF 4 / ORF 5	U U U C U U A U U **U C U C A A C** U A A C G A C U U [N_{146}] A U G
ORF 5 / M	U U A U U G A U U **U C U A A A C** U A A A C G A C A A U G
M / N	U U C A U U U U U **U C U A A A C** U G A A C G A A A A G A U G

found in the next smallest RNA. This motif is thought to have an important function in mRNA synthesis, however, its precise role is still uncertain. On the one hand, it may be involved in determining the frequency of leader primed initiation on genome length negative strands or it could also act as a termination signal for negative strand RNA synthesis. At the present time, it is not possible to present a definitive model for the genesis of coronavirus subgenomic mRNAs. Because of the exceedingly low amounts of viral RNA in the earliest stages of infection, it will be difficult to distinguish between alternative models. In Table 2, the sequence of the HCV 229E intergenic regions which precede each of the subgenomic mRNA body sequences are shown.

Translational regulation

The basic premise of coronavirus translation is that only the information encoded in the 5′ unique region of each mRNA is expressed as protein. (Fig. 4). For the majority of HCV 229E mRNAs, the 5′ unique region encompasses only a single ORF and it can be assumed that these ORFs are functionally monocistronic. There are 2 mRNAs (mRNA 1 and mRNA 4) in which the 5′ unique region contains more than 1 ORF, and in these cases translational strategies appear to regulate gene expression.

mRNA 1

The HCV 229E RNA polymerase gene (or perhaps more correctly, RNA polymerase locus) is comprised of two large overlapping ORFs,

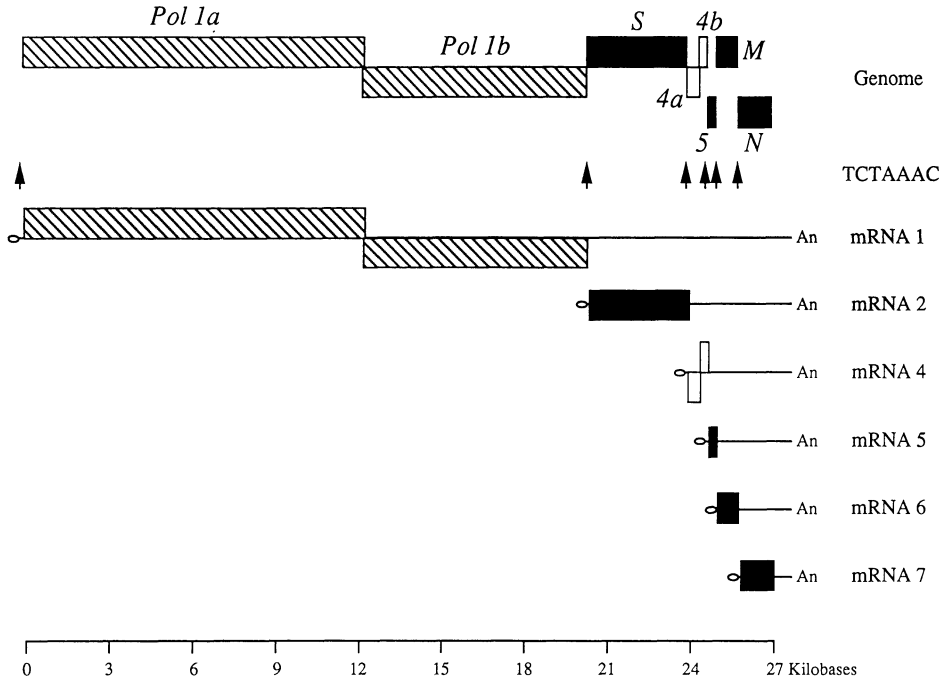

Fig. 4. Expression of the HCV 229E genome. The genomic organization and the 3′coterminal set of subgenomic mRNAs are illustrated. The structural and non-structural protein genes are shown as in Fig. 1

ORF 1a and ORF 1b. There is now substantial evidence that ORF 1b is expressed by a (−1) ribosomal frame shifting event mediated by a specific structure, the RNA pseudoknot, which is located at the ORF 1a/ORF 1b junction. This element has been identified for IBV, MHV and HCV 229E and its function in vitro and in vivo has been demonstrated by transcription/translation experiments [4, 5, 13]. In the absence of any further processing, the HCV 229E polymerase gene could encode an ORF 1a product (454,200 M_r) or an ORF 1a/b product (754,200 ORF M_r).

In addition to the flexibility offered by ribosomal frame shifting, there is both genetic and biochemical evidence which suggests that the coronavirus polymerase gene contains 5 or 6 complementation groups [2]. Almost certainly, the functionally separate gene products are generated by post-translational proteolytic processing. This conclusion is supported by the presence within the polymerase locus of protease motifs characteristic of both papain (×2) and 3C-like proteases. These protease "domains" are located in ORF 1a. Recently, the first experiment on proteolytic processing of the MHV gene products in vivo have been reported, [6, 7] but it is too early to propose a processing pathway.

mRNA 4

The second HCV 229E mRNA which contains more than one nrORF in its 5′ unique region is mRNA 4. At the present time it is premature to speculate on the mechanisms used to express the downstream ORF (ORF 4b) because neither gene product has been identified in vivo or in vitro.

Conclusions

The molecular analysis of the HCV 229E genome reported here forms the basis for a detailed study of the biology and pathogenesis of this infectious agent. Should these studies reveal that HCV 229E infection is relatively benign, it may also open up the possibility of using this agent as a live virus vaccine for respiratory pathogens.

Acknowledgements

We thank Barbara Schelle-Prinz and Atiye Toksoy for excellent technical help and Andrea Feyrer for typing the manuscript.

References

1. Arpin N, Talbot JP (1990) Molecular characterization of the 229E strain of human coronavirus. In: Cavanagh D, Brown TDK (eds) Coronaviruses and their diseases. Advances in experimental biology and medicine, vol 276. Plenum Press, New York, pp 73–80
2. Baric RS, Fu KS, Schaad MC, Stohlman SA (1990) Establishing a genetic recombination map for MHV-A59 complementation groups. Virology 177: 646–656
3. Boursnell MEG, Brown TDK, Foulds IJ, Green PF, Tomley FM, Binns MM (1987) Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. J Gen Virol 68: 57–77
4. Breedenbeek PJ, Pachuk CJ, Noten AFH, Charite A, Luytjes W, Weiss SR, Spaan WJM (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 frameshifting mechanism. Nucleic Acids Res 18: 1825–1832
5. Brierley I, Digard P, Inglis SC (1989) Characterization of an efficient ribosomal frameshifting sequence: requirement for an RNA pseudoknot. Cell 57: 537–547
6. Denison MR, Zoltick PW, Hughes SA, Giangreco B, Olsen AL, Perlman S, Leibowitz JL, Weiss SR (1992) Intracellular processing of the N-terminal ORF 1a

proteins of the coronavirus MHV requires multiple proteolytic events. Virology 189: 274–284
7. Denison MR, Zoltick PW, Leibowitz JL, Pachuk CJ, Weiss SR (1991) Identification of polypeptides encoded in open reading frame 1b of the putative polymerase gene of the murine coronavirus mouse hepatitis virus A59. J Virol 65: 3076–3082
8. Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395
9. Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25: 263–269
10. Hierholzer JC, Tannock GA (1988) Coronaviridae: the coronaviruses. In: Lenette EH, Halonen P, Murphy FA (eds) Viral, rickettsial, and chlamydial diseases. Laboratory diagnosis of infectious disease-principles and practice, vol 2. Springer, Berlin Heidelberg New York Tokyo, pp 451–483
11. Isaacs D, Flowers D, Clarke JR, Valman B, Macnaughton MR (1983) Epidemiology of coronavirus respiratory infections. Arch Dis Child 38: 500–503
12. Kemp MC, Hierholzer JC, Harrison A, Burks JS (1984) Characterization of viral proteins synthesized in 229E infected cells and effect(s) of inhibition of glycosylation and glycoprotein transport. In: Rottier PJM, van dere Zeijst BAM, Spaan WJM, Horzinek MC (eds) Molecular biology and pathogenesis of coronavirus. Advances in experimental biology and medicine, vol 173. Plenum Press, New York, pp 65–79
13. Lee H-J, Shieh C-K, Gorbalenya AE, Koonin EV, LaMonica N, Tuler J, Bagdzhadzhyan A, Lai MMC (1991) The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. Virology 180: 567–582
14. Myint S, Harmsen D, Raabe T, Siddell S (1990) Characterization of a nucleic acid probe for the diagnosis of human coronavirus 229E infections. J Med Virol 31: 165–172
15. Parker MM, Masters PS (1990) Sequence comparison of the N genes of five strains of mouse hepatitis virus suggest a three domain structure for the nucleocapsid protein. Virology 179: 463–468
16. Phillpots JR (1983) Clones of MRC-C cells may be superior to the parent line for the culture of 229E-like strains of human respiratory coronaviruses. J Virol Methods 6: 267–269
17. Pinto LH, Holsinger LJ, Lamb RA (1992) Influenza virus M2 protein has ion channel activity. Cell 69: 517–528
18. Raabe T, Schelle-Prinz B, Siddell SG (1990) Nucleotide sequence of the gene encoding the spike glycoprotein of human coronavirus HCV 229E. J Gen Virol 71: 1065–1073
19. Rottier PJM, Welling GW, Welling-Wester S, Niesters HGM, Lenstra JA, van der Zeijst BAM (1986) Predicted membrane topology of the coronavirus protein E1. Biochemistry 25: 1335–1339
20. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor
21. Schmidt OW, Kenny GE (1982) Polypeptides and functions of antigens from human coronaviruses 229E and OC43. Infect Immun 35: 515–522
22. Spaan W, Rottier P, Smeekens S, van der Zeijst BAM, Delius H, Armstrong J, Skinner M, Siddell SG (1983) Coronavirus mRNA synthesis involves fusion of noncontigious sequences. EMBO J 2: 1839–1844

23. Staden R (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res 10: 4731–4751

Authors' address: Dr. S.G. Siddell, Institute of Virology, University of Würzburg, Versbacher Strasse 7, D-97078 Würzburg, Federal Republic of Germany.