Journal of General Virology (1990), **71**, 1065–1073. *Printed in Great Britain*

Nucleotide sequence of the gene encoding the spike glycoprotein of human coronavirus HCV 229E

Thomas Raabe, Barbara Schelle-Prinz and Stuart G. Siddell*

Institute of Virology, University of Würzburg, Versbacher Strasse 7, 8700 Würzburg, F.R.G.

The gene encoding the spike glycoprotein of the human coronavirus HCV 229E has been cloned and sequenced. This analysis predicts an S polypeptide of 1173 amino acids with an M_r of 128600. The polypeptide has 30 potential *N*-glycosylation sites. A number of structural features typical of coronavirus S proteins can be recognized, including a signal sequence, a membrane anchor, heptad repeat structures and a carboxy-terminal cysteine cluster. A detailed, computer-aided comparison with the S proteins of infectious bronchitis virus, feline infectious peritonitis virus, transmissible gastroenteritis virus and murine hepatitis virus, strain JHM is presented. We have also done a Northern blot analysis of viral RNAs in HCV 229E-infected cells using synthetic oligonucleotides. On the basis of this analysis, and by analogy to the replication strategy of other coronaviruses, we are able to propose a model for the organization and expression of the HCV 229E genome.

Introduction

Human coronaviruses (HCV) are a common cause of respiratory disease in man and it has been estimated that they are responsible for up to 20% of common colds (Hierholzer & Tannock, 1988; Isaacs *et al.*, 1983; McIntosh *et al.*, 1974). With a few exceptions, HCVs are difficult to propagate in tissue or organ culture and consequently their biology is relatively poorly understood. Nevertheless, it has been possible to establish that there are two major HCV antigenic groups, represented by HCV 229E and HCV OC43 (Macnaughton, 1981; Pedersen *et al.*, 1978).

The HCV 229E virion consists of the genomic RNA, which if HCV is similar to other coronaviruses will be about 30 kb, a lipid envelope and three major proteins: the nucleocapsid protein, N (M_r of 50K), the membrane glycoprotein, M (M_r of 21K to 25K) and the spike glycoprotein, S (M_r of 186K) (Kemp *et al.*, 1984; Macnaughton & Madge, 1978; Schmidt & Kenny, 1982). Human coronaviruses of the OC43 group possess an additional surface glycoprotein, the haemagglutinin-esterase, HE (M_r of 65K) (Hogue & Brian, 1986).

The HCV replication strategy involves the synthesis of subgenomic RNAs in the cytoplasm of infected cells (Weiss & Leibowitz, 1981). It is assumed that these subgenomic RNAs are synthesized by a process of leader-primed discontinuous transcription as has been described for the murine hepatitis virus (MHV) (Baric *et al.*, 1985; Makino *et al.*, 1986; Shieh *et al.*, 1987). This process involves the recognition of a specific sequence, the so-called 'region of homology', present at the 3′ end of a leader RNA and at each intergenic transcriptional reinitiation site on the antigenomic RNA template (for a review see Lai *et al.*, 1987).

In the case of HCV this process results in a set of six 3′ coterminal subgenomic RNAs (Kamahora *et al.*, 1989; Schreiber *et al.*, 1989). By analogy to other coronaviruses the 5′ unique region in each RNA (i.e. the region not present in the next smallest RNA) should be translated and, at least in the case of the RNAs encoding structural proteins, they should be expressed as a single polypeptide (Spaan *et al.*, 1988).

Recently, the HCV 229E genes encoding the N protein and the M glycoprotein have been cloned and sequenced (Raabe & Siddell, 1989*a*; Schreiber *et al.*, 1989). Also, sequence analysis of the genomic region upstream from the M protein gene has revealed three open reading frames (ORFs) with the potential to encode polypeptides of 15·3K, 10·2K and 9·1K (Raabe & Siddell, 1989*b*). As proteins of this size have not been identified in virions (Schmidt & Kenny, 1982), these genes are thought to encode non-structural components. A similar arrangement of structural and non-structural genes has been shown for a number of other coronaviruses (Spaan *et al.*, 1988).

The spike glycoprotein of coronaviruses forms the characteristic peplomer structures on the surface of the virion. The protein is a large, acylated glycopolypeptide with an M_r, depending upon the virus in question, of

between 170K and 200K (Spaan *et al.*, 1988). Each peplomer consists of a dimer or trimer of S proteins (Cavanagh, 1983) which in the case of MHV and infectious bronchitis virus (IBV), but not feline infectious peritonitis virus (FIPV) or transmissible gastroenteritis virus (TGEV) have been cleaved into two non-identical subunits, the amino-terminal S1 and the carboxy-terminal S2.

In the case of HCV 229E it has been shown that the S protein is the major antigenic determinant in natural infections and has a central role in the induction of the immune response (Macnaughton *et al.*, 1981). Studies on other coronaviruses have shown that the same protein also mediates such essential biological functions as attachment of the virion to the cell surface and the fusion of viral and cellular membranes (de Groot *et al.*, 1989; Sturman & Holmes, 1985).

In the long term, our aim is to define the role of the S protein in the pathogenesis of HCV 229E infections as well as its interaction with the human immune system. As a first step, we present the complete nucleotide sequence of the HCV 229E S gene and compare the predicted amino acid sequence with other recently determined coronavirus S protein sequences. Also, on the basis of analogy to other coronaviruses, recently published sequence data (Raabe & Siddell, 1989*a*, *b*; Schreiber *et al.*, 1989) and a Northern blot analysis of intracellular viral RNA, we propose a model for the organization and expression of the HCV 229E genome.

Methods

Virus and cells. The HCV 229E strain used in these studies was isolated from a volunteer at the MRC Common Cold Unit, Salisbury, U.K. The virus was adapted to tissue culture by passage in C16 cells, a heteroploid cell line of human origin (Phillpotts, 1983). The virus was titrated by limiting dilution and the supernatant from a well with one focus of infection was taken as the primary 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 (Siddell, 1983). Polyadenylated RNA was fractionated by chromatography on poly(U)–Sepharose.

cDNA cloning. Two cDNA libraries were prepared essentially according to the method of Gubler & Hoffman (1983), using either random hexanucleotides or an S gene-specific oligonucleotide (positions 227 to 244, Fig. 1) as first-strand primer. The synthesized ds cDNA was size-fractionated on a Sephacryl S-1000 column, ligated to *Eco*RI linkers and cloned into the Bluescript vector pKS II$^+$ (Stratagene). Recombinant clones were screened 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 Maniatis *et al.* (1982).

Sequence analysis. cDNA was subcloned by digestion with restriction enzymes and ligation into *Sma*I-linearized M13mp19 vector DNA (Messing & Vieira, 1982). The sequence of clone 11B5 was obtained after generation of a series of overlapping deletions using exonuclease III (Henikoff, 1984). Sequencing was done on ds and ss DNA templates using the chain termination method (Sanger *et al.*, 1977) with the M13 universal primer or S gene-specific oligonucleotide primers. The sequences presented were determined completely on both cDNA strands. Sequence data were assembled by the programs of Staden (1982) and analysed by the programs of the University of Wisconsin Computer Genetics Group (Devereux *et al.*, 1984).

Northern blot analysis. Polyadenylated RNA from HCV 229E-infected C16 cells was electrophoresed on 0·9% agarose–formaldehyde gels and transferred onto nitrocellulose membranes using standard procedures (Maniatis *et al.*, 1982). HCV 229E-specific oligonucleotides were synthesized using phosphoramidite chemistry on a Cyclone DNA synthesizer and purified by gel electrophoresis. Oligonucleotides were 5′ end-labelled with [γ-^{32}P]ATP and hybridized using the conditions described by Woods (1984). The oligonucleotides used were 5′ GCAACCACCGGGTATATC 3′ (A), 5′ AACATCAGTCTG-CAATGC 3′ (B), 5′ GAGCCATTACTGTATGTG 3′ (C), 5′ CGAATGGTTTCAGAGCCT 3′ (D), 5′ CAACAGCTGGGTGTT-CAC 3′ (E), 5′ ATACACACTAGTAGTATC 3′ (F) and 5′ TCCCAATTAGCCCAGGTG 3′ (G).

A cDNA probe specific for the HCV 229E N gene was prepared by nick translation of plasmid pSMF1 DNA (Myint *et al.*, 1989) and hybridized under standard conditions (Maniatis *et al.*, 1982).

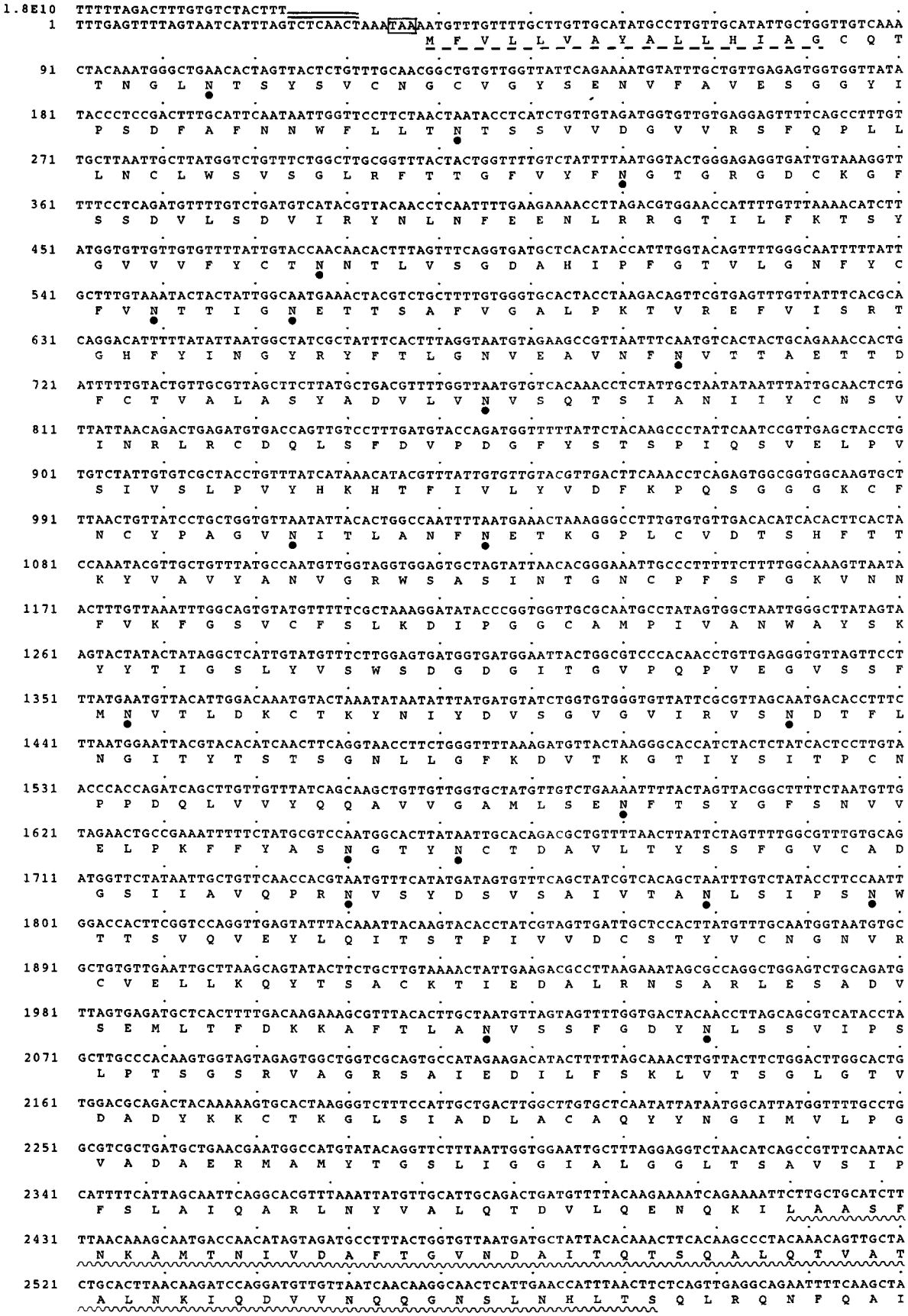
Results

Characterization of HCV 229E-specific cDNA clones

An 18 base oligonucleotide complementary to a sequence near the 5′ end of the HCV 229E N gene was used to screen a randomly primed cDNA library derived from polyadenylated RNA extracted from HCV 229E-infected cells (Raabe & Siddell, 1989*a*). Plasmid 2F7 contained a 4·2 kb cDNA insert which hybridized to all HCV 229E RNAs (data not shown). Sequence analysis of clone 2F7 showed that the insert cDNA extends from a position within the N gene to a position within the S gene (see Fig. 2). An oligonucleotide complementary to the 5′ end of clone 2F7 (using the mRNA orientation, nucleotides 1209 to 1226, Fig. 1) was used to identify clone 11B5 which overlaps 2F7 by 3 kb and extends a further 1 kb in the 5′ direction. Finally, a second series of cDNA clones were synthesized using an oligonucleotide primer based upon sequences derived from the 5′ end of clone 11B5 (nucleotides 227 to 244, Fig. 1). One such clone, 5B5, encompasses the 5′ end of the S gene and extends approximately 2·5 kb in the 5′ direction. Another, 8E10, contains a 250 bp insert and terminates at the 5′ end with a sequence previously identified as the HCV 229E leader RNA (Schreiber *et al.*, 1989). Fig. 2 shows the location of these cDNA clones with respect to the genomic and subgenomic RNAs.

Sequence analysis of the HCV 229E S protein gene

The nucleotide sequence of the HCV 229E S gene together with the predicted amino acid sequence of the S


1.8E10 TTTTTAGACTTTGTGTCTACTTT
1 TTTGAGTTTTAGTAATCATTTAGTCTCAACTAAATAAAATGTTTGTTTTGCTTGTTGCATATGCCTTGTTGCATATTGCTGGTTGTCAAA 90
M F V L L V A Y A L L H I A G C Q T
91 CTACAAATGGGCTGAACACTAGTTACTCTGTTTGCAACGGCTGTGTTGGTTATTCAGAAAATGTATTTGCTGTTGAGAGTGGTGGTTATA 180
T N G L N T S Y S V C N G C V G Y S E N V F A V E S G G Y I
181 TACCCTCCGACTTTGCATTCAATAATTGGTTCCTTCTAACTAATACCTCATCTGTTGTAGATGGTGTTGTGAGGAGTTTTCAGCCTTTGT 270
P S D F A F N N W F L L T N T S S V V D G V V R S F Q P L L
271 TGCTTAATTGCTTATGGTCTGTTTCTGGCTTGCGGTTTACTACTGGTTTTGTCTATTTTAATGGTACTGGGAGAGGTGATTGTAAAGGTT 360
L N C L W S V S G L R F T T G F V Y F N G T G R G D C K G F
361 TTTCCTCAGATGTTTTGTCTGATGTCATACGTTACAACCTCAATTTTGAAGAAAACCTTAGACGTGGAACCATTTTGTTTAAAACATCTT 450
S S D V L S D V I R Y N L N F E E N L R R G T I L F K T S Y
451 ATGGTGTTGTTGTGTTTTATTGTACCAACAACACTTTAGTTTCAGGTGATGCTCACATACCATTTGGTACAGTTTTGGGCAATTTTTATT 540
G V V V F Y C T N N T L V S G D A H I P F G T V L G N F Y C
541 GCTTTGTAAATACTACTATTGGCAATGAAACTACGTCTGCTTTTGTGGGTGCACTACCTAAGACAGTTCGTGAGTTTGTTATTTCACGCA 630
F V N T T I G N E T T S A F V G A L P K T V R E F V I S R T
631 CAGGACATTTTTATATTAATGGCTATCGCTATTTCACTTTAGGTAATGTAGAAGCCGTTAATTTCAATGTCACTACTGCAGAAACCACTG 720
G H F Y I N G Y R Y F T L G N V E A V N F N V T T A E T T D
721 ATTTTTGTACTGTTGCGTTAGCTTCTTATGCTGACGTTTTGGTTAATGTGTCACAAACCTCTATTGCTAATATAATTTATTGCAACTCTG 810
F C T V A L A S Y A D V L V N V S Q T S I A N I I Y C N S V
811 TTATTAACAGACTGAGATGTGACCAGTTGTCCTTTGATGTACCAGATGGTTTTTATTCTACAAGCCCTATTCAATCCGTTGAGCTACCTG 900
I N R L R C D Q L S F D V P D G F Y S T S P I Q S V E L P V
901 TGTCTATTGTGTCGCTACCTGTTTATCATAAACATACGTTTATTGTGTTGTACGTTGACTTCAAACCTCAGAGTGGCGGTGGCAAGTGCT 990
S I V S L P V Y H K H T F I V L Y V D F K P Q S G G G K C F
991 TTAACTGTTATCCTGCTGGTGTTAATATTACACTGGCCAATTTTAATGAAACTAAAGGGCCTTTGTGTGTTGACACATCACACTTCACTA 1080
N C Y P A G V N I T L A N F N E T K G P L C V D T S H F T T
1081 CCAAATACGTTGCTGTTTATGCCAATGTTGGTAGGTGGAGTGCTAGTATTAACACGGGAAATTGCCCTTTTTCTTTTGGCAAAGTTAATA 1170
K Y V A V Y A N V G R W S A S I N T G N C P F S F G K V N N
1171 ACTTTGTTAAATTTGGCAGTGTATGTTTTTCGCTAAAGGATATACCCGGTGGTTGCGCAATGCCTATAGTGGCTAATTGGGCTTATAGTA 1260
F V K F G S V C F S L K D I P G G C A M P I V A N W A Y S K
1261 AGTACTATACTATAGGCTCATTGTATGTTTCTTGGAGTGATGGTGATGGAATTACTGGCGTCCCACAACCTGTTGAGGGTGTTAGTTCCT 1350
Y Y T I G S L Y V S W S D G D G I T G V P Q P V E G V S S F
1351 TTATGAATGTTACATTGGACAAATGTACTAAATATAATATTTATGATGTATCTGGTGTGGGTGTTATTCGCGTTAGCAATGACACCTTTC 1440
M N V T L D K C T K Y N I Y D V S G V G V I R V S N D T F L
1441 TTAATGGAATTACGTACACATCAACTTCAGGTAACCTTCTGGGTTTTAAAGATGTTACTAAGGGCACCATCTACTCTATCACTCCTTGTA 1530
N G I T Y T S T S G N L L G F K D V T K G T I Y S I T P C N
1531 ACCCACCAGATCAGCTTGTTGTTTATCAGCAAGCTGTTGTTGGTGCTATGTTGTCTGAAAATTTTACTAGTTACGGCTTTTCTAATGTTG 1620
P P D Q L V V Y Q Q A V V G A M L S E N F T S Y G F S N V V
1621 TAGAACTGCCGAAATTTTTCTATGCGTCCAATGGCACTTATAATTGCACAGACGCTGTTTTAACTTATTCTAGTTTTGGCGTTTGTGCAG 1710
E L P K F F Y A S N G T Y N C T D A V L T Y S S F G V C A D
1711 ATGGTTCTATAATTGCTGTTCAACCACGTAATGTTTCATATGATAGTGTTTCAGCTATCGTCACAGCTAATTTGTCTATACCTTCCAATT 1800
G S I I A V Q P P R N V S Y D S V S A I V T A N L S I P S N W
1801 GGACCACTTCGGTCCAGGTTGAGTATTTACAAATTACAAGTACACCTATCGTAGTTGATTGCTCCACTTATGTTTGCAATGGTAATGTGC 1890
T T S V Q V E Y L Q I T S T P I V V D C S T Y V C N G N V R
1891 GCTGTGTTGAATTGCTTAAGCAGTATACTTCTGCTTGTAAAACTATTGAAGACGCCTTAAGAAATAGCGCCAGGCTGGAGTCTGCAGATG 1980
C V E L L K Q Y T S A C K T I E D A L R N S A R L E S A D V
1981 TTAGTGAGATGCTCACTTTTGACAAGAAAGCGTTTACACTTGCTAATGTTAGTAGTTTTGGTGACTACAACCTTAGCAGCGTCATACCTA 2070
S E M L T F D K K A F T L A N V S S F G D Y N L S S V I P S
2071 GCTTGCCCACAAGTGGTAGTAGAGTGGCTGGTCGCAGTGCCATAGAAGACATACTTTTTAGCAAACTTGTTACTTCTGGACTTGGCACTG 2160
L P T S G S R V A G R S A I E D I L F S K L V T S G L G T V
2161 TGGACGCAGACTACAAAAAGTGCACTAAGGGTCTTTCCATTGCTGACTTGGCTTGTGCTCAATATTATAATGGCATTATGGTTTTGCCTG 2250
D A D Y K K C T K G L S I A D L A C A Q Y Y N G I M V L P G
2251 GCGTCGCTGATGCTGAACGAATGGCCATGTATACAGGTTCTTTAATTGGTGGAATTGCTTTAGGAGGTCTAACATCAGCCGTTTCAATAC 2340
V A D A E R M A M Y T G S L I G G I A L G G L T S A V S I P
2341 CATTTTCATTAGCAATTCAGGCACGTTTAAATTATGTTGCATTGCAGACTGATGTTTTACAAGAAAATCAGAAAATTCTTGCTGCATCTT 2430
F S L A I Q A R L N Y V A L Q T D V L Q E N Q K I L A A S F
2431 TTAACAAAGCAATGACCAACATAGTAGATGCCTTTACTGGTGTTAATGATGCTATTACACAAACTTCACAAGCCCTACAAACAGTTGCTA 2520
N K A M T N I V D A F T G V N D A I T Q T S Q A L Q T V A T
2521 CTGCACTTAACAAGATCCAGGATGTTGTTAATCAACAAGGCAACTCATTGAACCATTTAACTTCTCAGTTGAGGCAGAATTTTCAAGCTA 2610
A L N K I Q D V V N Q Q G N S L N H L T S Q L R Q N F Q A I

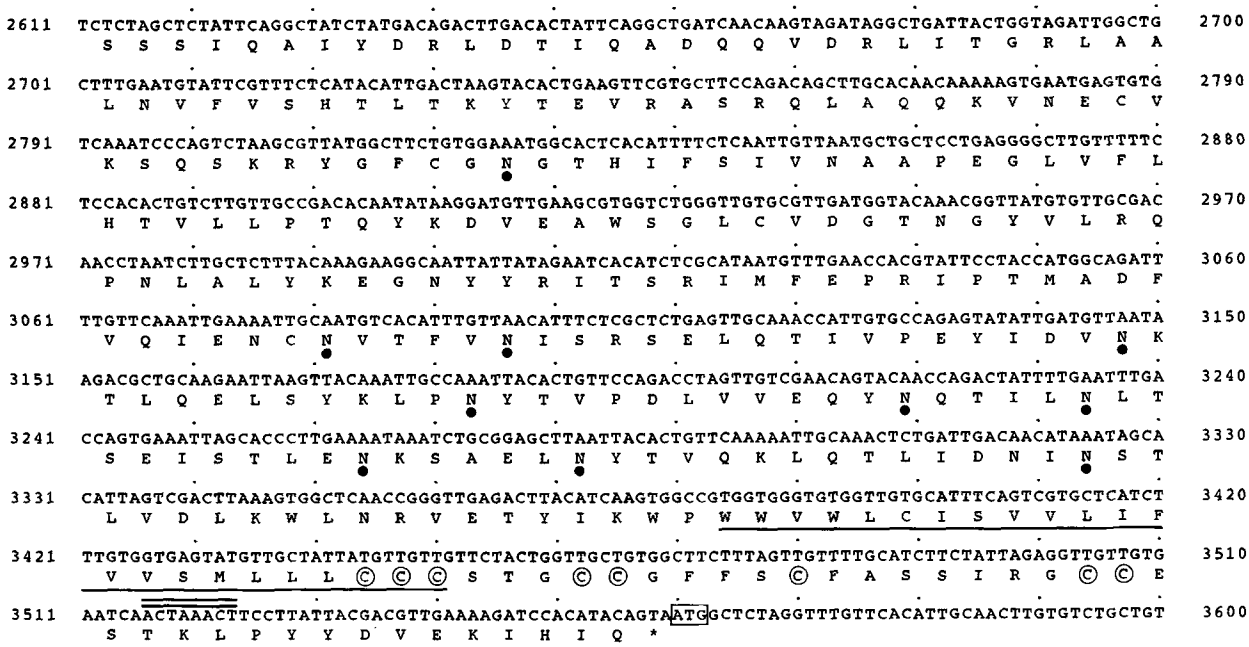


Fig. 1. Nucleotide sequence of the HCV 229E S gene and the predicted amino acid sequence of the S protein precursor. The amino-terminal signal sequence (---), the putative membrane anchor (——) and the heptad repeat region (~~~~) are underlined. Potential *N*-glycosylation sites (●) and the cysteine-rich region (©) are indicated. The region of homology preceding the S gene and the 4a gene are overlined. The positions of the 4a initiation codon and the 5′ upstream ORF termination codon are boxed.

protein are shown in Fig. 1. Immediately upstream of the S gene ORF is a sequence, TCTCAACT, which is similar or identical to sequences found adjacent to the HCV 229E N and M genes, as well as the putative non-structural genes 4a and 5 (see Fig. 2) (Raabe & Siddell, 1989*a*, *b*; Schreiber *et al.*, 1989). This sequence represents the HCV 229E 'region of homology' and the divergence between the clones 5B5 and 8E10 at this point (Fig. 1) confirms this as the site at which fusion of the leader and body sequences of the HCV 229E S protein mRNA has taken place.

The AUG codon which initiates the S protein gene (nucleotide 39, Fig. 1) is in a favoured context (Kozak, 1983) and opens a reading frame of 3519 nucleotides which encodes a polypeptide of 1173 amino acids with an M_r of 128·6K. The predicted S protein polypeptide contains 30 potential *N*-glycosylation sites (NXS or NXT) and the difference in the apparent M_r of the HCV 229E S protein (186K; Schmidt & Kenny, 1982) and the predicted size of the polypeptide suggest that the majority of these sites are used.

At the amino terminus of the polypeptide is a stretch of 14 mainly hydrophobic amino acids followed by two amino acids with small uncharged side-chains, a feature typical of a signal peptidase recognition site (von Heijne, 1984). At the carboxy terminus, between amino acids 1116 and 1138, a second strongly hydrophobic region can be recognized which is believed to serve as the trans-membrane anchor (de Groot *et al.*, 1987*a*). This region is flanked on the amino-terminal side by the sequence KWPWWVWL, which differs by only one amino acid from the sequence KWPWYVWL which is conserved in all coronavirus S protein genes sequenced to date. On the carboxy-terminal side, the membrane anchor region is flanked by an unusually high number of cysteine residues. This feature has also been recognized in other coronavirus S proteins (Rasschaert & Laude, 1987; Schmidt *et al.*, 1987) and it has been proposed that at least some of these residues may be involved in the acylation of the S protein which has been described for MHV (Sturman *et al.*, 1985, van Berlo *et al.*, 1987). In the HCV 229E S protein sequence it is also possible to identify the 'heptad repeat' structures (corresponding to amino acids 794 to 849, Fig. 1) which have been proposed by de Groot *et al.* (1987*a*) and Rasschaert & Laude (1987) to be essential elements in forming the elongated structure of the S protein.

Finally, the predicted sequence of the HCV 229E S protein does not reveal any basic amino acid sequences related to the motifs RRXRR or RRAHR (where X is F, S, H or A) which have been identified as the sites at which MHV and IBV S proteins are proteolytically cleaved to yield the S1 and S2 polypeptides (Spaan *et al.*, 1988). These motifs are also absent in the FIPV and TGEV S proteins, which are apparently not cleaved (Garwes & Reynolds, 1981; Horzinek *et al.*, 1982).

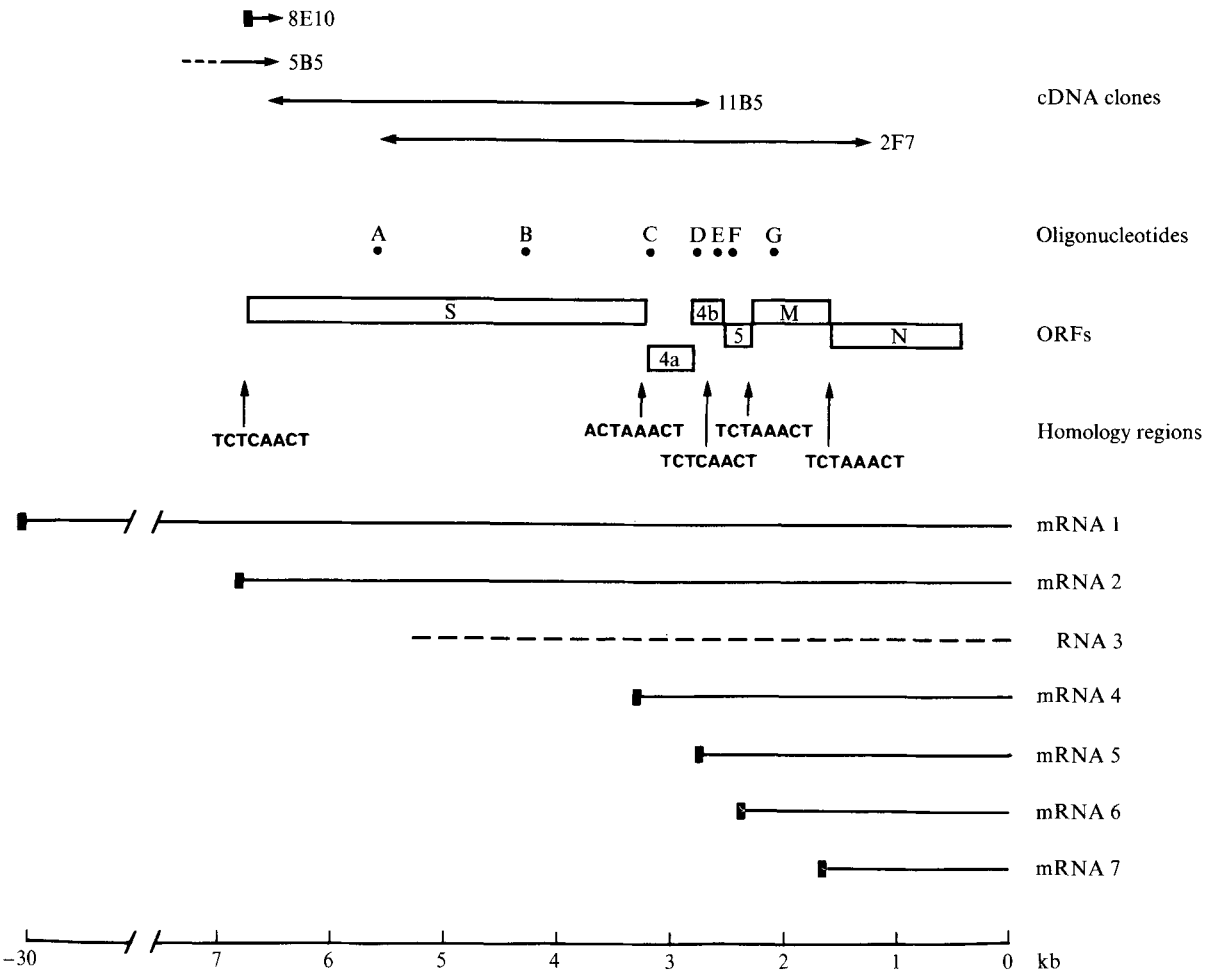


Fig. 2. A proposed model for the organization and expression of the HCV 229E genome. The coding regions for the structural proteins (S, M, N) and the non-structural proteins (4a, 4b, 5) are shown in relation to the genomic and subgenomic RNAs. The black boxes at the 5′ end of the RNAs represent a common leader sequence which has been demonstrated for the S and N mRNAs (this paper; Schreiber *et al.*, 1989). The positions of the oligonucleotides A to G are indicated (●). Also shown are the positions and sequences of the homology regions and the extent of the cDNA clones used in this study.

Genomic organization of HCV 229E

Together with the data presented in this report a continuous sequence of 6·7 kb at the 3′ end of the HCV 229E genome has been determined. Within this sequence the regions encoding the S, M and N proteins have been identified on the basis of the sizes of the ORFs and the characteristics of the predicted polypeptides (Raabe & Siddell, 1989*a*; Schreiber *et al.*, 1989). In addition, three large ORFs which are supposed to encode non-structural proteins have been found between the S and M genes (Raabe & Siddell, 1989*b*). This arrangement of ORFs with respect to the genomic RNA is summarized in Fig. 2.

In order to identify the subgenomic RNAs that code for the S, M, N and non-structural proteins, we have done Northern blot analysis using synthetic oligonucleotides and a cDNA probe (Fig. 3 shows the localization of these probes). The cDNA probe, pSMF1, which encompasses the N protein gene plus the 3′ non-coding region, detects seven virus-specific RNAs which have been numbered 1 to 7 in order of decreasing size (Fig. 3). The M gene-specific oligonucleotide G hybridized to the RNAs 1 to 6, but not RNA 7. The S gene-specific oligonucleotide A (complementary to nucleotides 1209 to 1226, Fig. 1) hybridized only to RNAs 1 and 2. Assuming that the HCV RNAs are arranged as a 3′ coterminal nested set and that the 5′ unique regions are translated, these results lead us to propose that the RNAs 2, 6 and 7 encode the structural proteins S, M and N, respectively.

The oligonucleotide B, as well as all other oligonucleotides except A, hybridized to an RNA which we and others have termed RNA 3. This result was unexpected because the sequences complementary to oligonucleotide B lie within the S gene ORF (nucleotides 2379 to 2396,

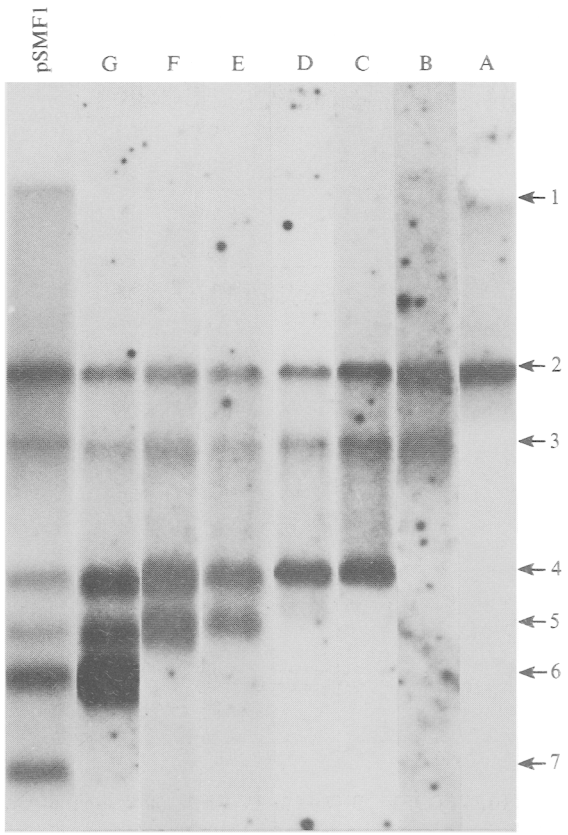


Fig. 3. Northern blot analysis of HCV 229E RNA. The polyadenylated RNA of HCV 229E-infected C16 cells was electrophoresed in formaldehyde–agarose gels and transferred to nitrocellulose membranes. ^{32}P-labelled oligonucleotides A to G (see Methods for the sequence and Fig. 2 for the location) and a cDNA clone corresponding to the HCV 229E N gene and 3′ non-coding region were used as hybridization probes. The HCV 229E-specific RNAs were numbered (1 to 7) according to their decreasing size.

Fig. 1). There are no sequences within the S gene ORF which resemble a 'region of homology' and at the moment we have no reason to suppose that this RNA functions as an mRNA.

We have previously described three ORFs in the region between the S and M genes of HCV 229E (Raabe & Siddell, 1989*b*). However, there are only two viral RNAs (RNA 4 and RNA 5) whose unique regions encompass this area. To assign these ORFs to the RNAs we therefore did hybridizations with oligonucleotides located at the 5′ end of the ORF 4a (oligonucleotide C), at the 5′ end of ORF 4b (oligonucleotide D), at the 3′ end of ORF 4b (oligonucleotide E) and at the 5′ end of ORF 5 (oligonucleotide F) (Fig. 3). As expected, oligonucleotide C hybridized to RNA 4 and oligonucleotide F to RNA 5. Oligonucleotide D which corresponds to the 5′ end of ORF 4b does not hybridize to RNA 5, but a clear positive signal is obtained for RNA 5 using oligonucleotide E. This indicates that the 5′ end of the RNA 5 body extends well into, but not over the complete coding region of ORF 4b. In the light of these data we re-examined the ORF 4b sequence and found a perfect 'region of homology' motif, TCTCAACT, at a position 107 nucleotides downstream from the ORF 4b initiation codon. This indicates, in contrast to our previous suggestion (Raabe & Siddell, 1989*b*), that in functional terms the ORFs 4a and 4b should be assigned to the 'unique' region of RNA 4 and the ORF 5 to the unique region of RNA 5.

On the basis of the available sequence data, analogy to other coronaviruses and the hybridization experiments described here, we propose a model of the organization and expression of the HCV 229E genome as is shown in Fig. 2.

Discussion

As we have described above, inspection of the HCV 229E S protein sequence reveals a number of features which are typical of coronavirus S proteins, for example, the amino-terminal signal sequence, the carboxy-terminal membrane anchor and the carboxy-terminal cysteine cluster. In order to search for further structural features whose conservation may indicate an important functional role, we have made a computer-aided comparison of the HCV 229E S protein sequence with the published S protein sequences of FIPV (de Groot *et al.*, 1987*b*), TGEV (Jacobs *et al.*, 1987), IBV (Binns *et al.*, 1985) and MHV JHM (Schmidt *et al.*, 1987). Firstly, we made an 'optimal' alignment of all sequences using the UWGCG GAP program. This alignment (which is available from the authors upon request) set matches, i.e. identical amino acids, equal to 1·5 and mismatches equal to lower values based upon the evolutionary distance between the amino acids as measured by Dayhoff and normalized by Gribskov & Burgess (1986). These alignments were then displayed using the program GAPSHOW, with a match display threshold of 1·5, i.e. only identical amino acids are displayed. Finally, we marked the positions of potential *N*-glycosylation sites as well as cysteine residues in all sequences. The result is shown in Fig. 4.

A number of important conclusions can be reached. Firstly, it is evident that the similarity of the HCV 229E sequence to the FIPV and TGEV sequences is much greater than to the IBV or MHV sequences. Moreover, as has been previously noted (de Goot *et al.*, 1987*a*), there is more similarity in the carboxy-terminal halves of these proteins than in the amino-terminal halves. These similarities are summarized in Table 1.

Although the amino-terminal halves of the coronavirus S proteins are less well conserved with respect to length and amino acid composition, it is interesting to

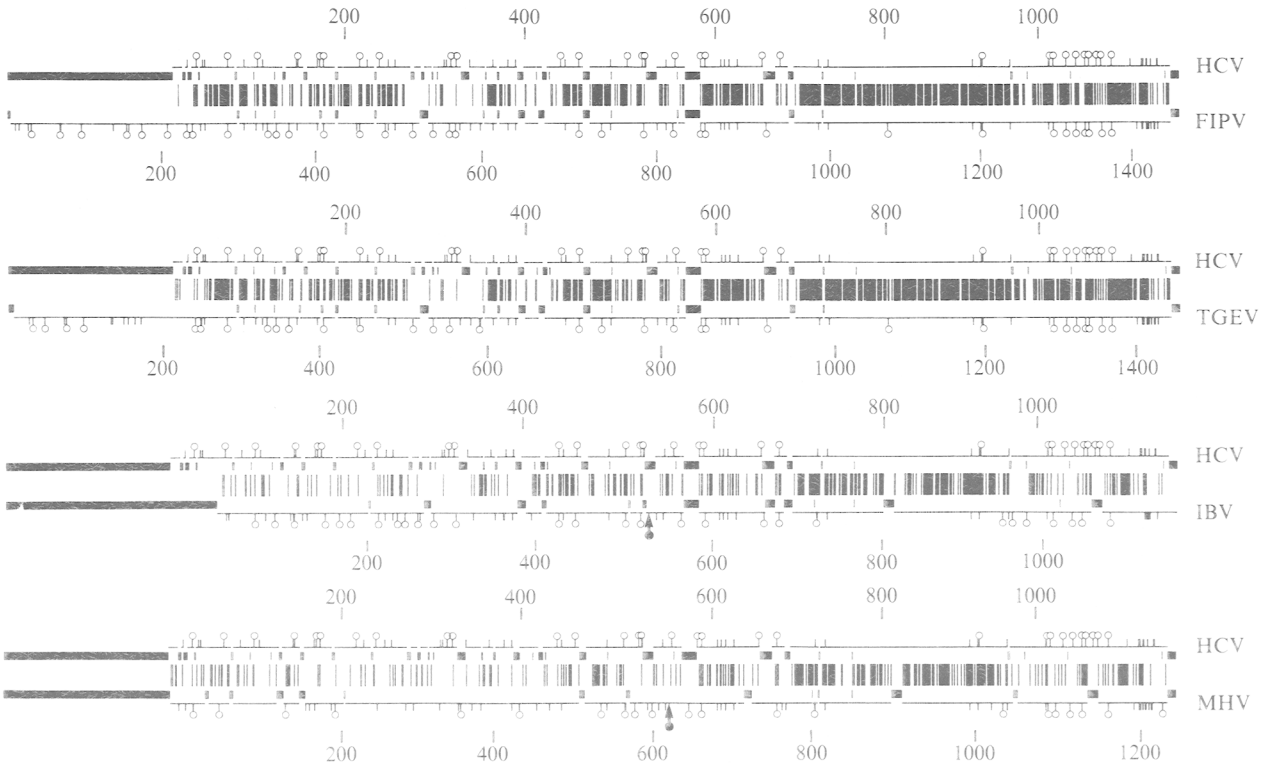


Fig. 4. Structural comparison of the S protein of HCV 229E and the S proteins of FIPV, TGEV, IBV and MHV JHM. The figure shows the positions of identical amino acids after optimal alignment of all sequences (⫴ ⫼). The 'gaps' introduced for alignment are shown as boxes (▬). The positions of potential *N*-glycosylation sites (♀), cysteine residues (|) and the post-translational cleavage sites of the IBV and MHV proteins (▲) are indicated. Details of the UWGCG programs GAP and GAPSHOW are given in the text. The FIPV, TGEV, IBV and MHV JHM S gene sequences were determined by de Groot *et al.* (1987*b*), Jacobs *et al.* (1987), Binns *et al.* (1985) and Schmidt *et al.* (1987).

Table 1. *Sequence comparison* of the S polypeptide of HCV 229E and the S polypeptides of FIPV, TGEV, IBV and MHV JHM*

Viral S protein	Residues	HCV 229E 1–543 Identity (%)	HCV 229E 1–543 Similarity (%)	HCV 229E 544–1173 Identity (%)	HCV 229E 544–1173 Similarity (%)
FIPV	1–786	38†	56	–	–
FIPV	787–1452	–	–	57	75
TGEV	1–781	37	56	–	–
TGEV	782–1447	–	–	57	74
IBV	1–535	18	38	–	–
IBV	536–1163	–	–	36	53
MHV JHM	1–626	16	36	–	–
MHV JHM	627–1235	–	–	32	54

* The sequences were aligned using the UWGCG GAP program.
† The figures given are the percentage amino acid identity or similarity following optimal alignment.

note that the 'optimal' alignment of the HCV S protein sequence to the FIPV or TGEV sequences results in a large amino-terminal gap. Jacobs *et al.* (1987) have reported a striking discontinuity in the levels of amino acid homology within the FIPV and TGEV S proteins. At the amino terminus (nucleotides 1 to 274) the mean homology is 30%, whereas the remaining sequences are 94% homologous. These authors have suggested that this observation could be explained by recombination between coronaviruses and our analysis is consistent with this interpretation.

It is worth noting that although the similarity between the HCV 229E and FIPV S proteins in positions 1 to 543 and 1 to 786, respectively, is only 38%, roughly 50% of the cysteine residues in this region of both sequences are located at the 'same' position. For the corresponding region of the HCV and MHV proteins (positions 1 to 543 and 1 to 580, respectively) only about 17% of the cysteine residues show this relationship.

Within the carboxy-terminal half of all S proteins there is an evident clustering of *N*-glycosylation sites at a position where the polypeptide is thought to emerge to the outside of the lipid bilayer (de Groot *et al.*, 1987*a*). Also, in addition to the carboxy-terminal cysteine cluster, we have now identified a number of cysteine residues that are conserved within the carboxy-terminal half of all S proteins. Striking, for example, are the

residues corresponding to the positions 608, 613, 619, 630, 715, 726, 917, 928 and 967 in the HCV sequence. It is clear that the relevance of features such as these will be fully appreciated only when a three-dimensional image of the S protein becomes available.

The number and sizes of the HCV 229E RNAs identified in our Northern blot analysis are in agreement with previously published results (Schreiber *et al.*, 1989; Weiss & Leibowitz, 1981). By analogy to other coronaviruses and on the basis of new hybridization data, we have now proposed coding assignments for five of these RNAs (Fig. 2). These assignments and the mRNA function of the RNAs need to be confirmed by *in vitro* translation of purified or synthetic RNAs, together with identification of the translation products using HCV protein-specific antibodies. In particular, it will be necessary to determine the coding capacity of RNA 4, which our data suggest has two ORFs in the 5′ unique region, and RNA 5 which appears to have an unusually long 5′ non-coding region. The availability of cDNA clones encompassing these genes will facilitate coupled transcription–translation experiments as have been described for MHV (Budzilowicz & Weiss, 1987). We expect that these studies will show that the replication strategy of HCV 229E closely parallels those of other coronaviruses.

At the moment we are not able to judge the relevance of the RNA 3 species which is detected by our hybridization probes and has been previously identified as a virus-specific RNA by metabolic labelling in the presence of actinomycin D (Schreiber *et al.*, 1989). It is not clear whether the RNA should be considered a putative mRNA or whether it represents, for example, an intracellular defective RNA or even a replicative form component. We hope to be able to resolve this question by sequence analysis of a cDNA corresponding to this RNA.

In addition to the S and M glycoproteins, MHV JHM, BCV and HCV OC43 possess a third glycoprotein, HE, which has both receptor-destroying and receptor-binding activities (M. Pfleiderer & S. Siddell, unpublished; Vlasak *et al.*, 1988). For MHV JHM and BCV, the gene encoding this protein is located immediately upstream of the S protein gene (Parker *et al.*, 1989; Shieh *et al.*, 1989). In the course of these studies, we have sequenced approximately 0·15 kb upstream of the HCV 229E S gene and our analysis revealed an ORF, the deduced amino acid sequence of which displays a high homology with the carboxy terminus of the IBV gene F (polymerase) product (data not shown) (Boursnell *et al.*, 1987). Taken together with the fact that HCV 229E does not have a receptor-binding (haemagglutinating) activity (Hierholzer, 1976), and our Northern blot analysis which did not reveal any additional RNAs between RNA 1 and RNA 2, these data strongly suggest that the HCV 229E genome does not contain a haemagglutinin–esterase gene.

In this paper we have proposed a model for the organization and expression of the HCV 229E genome and presented the predicted amino acid sequence of the spike glycoprotein. These data provide an essential basis to investigate the replication of the virus, as well as the structure, function, immunological and biological properties of the S protein. These studies will undoubtedly be important for our understanding of the pathogenesis and epidemiology of a widespread human infection.

We would like to thank Dr S. Myint for providing the plasmid pSMF1 and Helga Kriesinger for typing the manuscript. This work was supported by Sonderforschungsbereich 165, B1. The sequence data presented in this paper will appear in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under the accession number X16816.

References

BARIC, R. S., STOHLMAN, S. A., RAZAVI, M. K. & LAI, M. M. C. (1985). Characterization of leader-related small RNAs in coronavirus-infected cells: further evidence for leader-primed mechanism of transcription. *Virus Research* **3**, 19–33.

BINNS, M. M., BOURSNELL, M. E. G., CAVANAGH, D., PAPPIN, D. J. C. & BROWN, T. D. K. (1985). Cloning and sequencing of the gene encoding the spike protein of the coronavirus IBV. *Journal of General Virology* **66**, 719–726.

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.

BUDZILOWICZ, C. J. & WEISS, S. R. (1987). *In vitro* synthesis of two polypeptides from a non-structural gene of coronavirus mouse hepatitis virus strain A59. *Virology* **157**, 509–515.

CAVANAGH, D. (1983). Coronavirus IBV: structural characterization of the spike protein. *Journal of General Virology* **64**, 2577–2583.

DE GROOT, R. J., LUYTJES, W., HORZINEK, M. C., VAN DER ZEIJST, B. A. M., SPAAN, W. J. M. & LENSTRA, J. A. (1987*a*). Evidence for a coiled-coil structure in the spike proteins of coronaviruses. *Journal of Molecular Biology* **196**, 963–966.

DE GROOT, R. J., MADURO, J., LENSTRA, J. A., HORZINEK, M. C., VAN DER ZEIJST, B. A. M. & SPAAN, W. J. M. (1987*b*). 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., VAN LEEN, R. W., DALDERUP, M. J. M., VENNEMA, H., HORZINEK, M. C. & SPAAN, W. J. M. (1989). Stably expressed FIPV peplomer protein induces cell fusion and elicits neutralizing antibodies in mice. *Virology* **171**, 493–502.

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

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.

GRIBSKOV, M. & BURGESS, R. R. (1986). Sigma factors from *E. coli*, *B. subtilis*, phage SP01 and phage T4 are homologous proteins. *Nucleic Acids Research* **14**, 6745–6763.

GUBLER, U. & HOFFMAN, B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25**, 263–269.

HENIKOFF, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28**, 351–359.

HIERHOLZER, J. C. (1976). Purification and biophysical properties of human coronavirus 229E. *Virology* **75**, 155–165.

HIERHOLZER, J. C. & TANNOCK, G. A. (1988). *Coronaviridae:* The coronaviruses. In *Viral, Rickettsial and Chlamydial Diseases.* Laboratory diagnosis of infectious diseases. Principles and practices, vol. 2, pp. 451–483. Edited by E. H. Lennette, F. Halonen & F. A. Murphy. New York: Springer-Verlag.

HOGUE, B. G. & BRIAN, D. A. (1986). Structural proteins of human respiratory coronavirus OC43. *Virus Research* **5**, 131–144.

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

ISAACS, D., FLOWERS, D., CLARKE, J. R., VALMAN, B. & MACNAUGHTON, M. R. (1983). Epidemiology of coronavirus respiratory infections. *Archives of Disease in Childhood* **58**, 500–503.

JACOBS, L., DE GROOT, R., VAN DER ZEIJST, B. A. M., HORZINEK, M. C. & SPAAN, W. J. M. (1987). The nucleotide sequence of the peplomer gene of porcine transmissible gastroenteritis virus (TGEV): comparison with the sequence of the peplomer protein of feline infectious peritonitis virus. *Virus Research* **8**, 363–371.

KAMAHORA, T., SOE, L. H. & LAI, M. M. C. (1989). Sequence analysis of nucleocapsid gene and leader RNA of human coronavirus OC43. *Virus Research* **12**, 1–9.

KEMP, M. C., HIERHOLZER, J. C., HARRISON, A. & BURKS, J. S. (1984). Characterization of viral proteins synthesized in 229E-infected cells and effect(s) of inhibition of glycosylation and glycoprotein transport. In *Molecular Biology and Pathogenesis of Coronaviruses*, Advances in Experimental Medicine and Biology, vol. 173, pp. 65–77. Edited by P. J. M. Rottier, B. A. M. van der Zeijst, W. J. M. Spaan & M. C. Horzinek. New York & London: Plenum Press.

KOZAK, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles. *Microbiological Reviews* **47**, 1–45.

LAI, M. M. C., MAKINO, S., SOE, L. H., SHIEH, C.-K., KECK, J. G. & FLEMING, J. O. (1987). Coronavirus: a jumping RNA transcription. *Cold Spring Harbor Symposia on Quantitative Biology* **52**, 359–365.

MCINTOSH, K., CHAD, R. K., KRAUSE, H. E., WASIL, R., MOSEGA, H. E. & MUFSON, M. A. (1974). Coronavirus infection in acute lower respiratory tract disease of infants. *Journal of Infectious Diseases* **130**, 502–507.

MACNAUGHTON, M. R. (1981). Structural and antigenic relationship between human, murine and avian coronaviruses. In *Biochemistry and Biology of Coronaviruses*, Advances in Experimental Medicine and Biology, vol. 142, pp. 19–29. Edited by V. ter Meulen, S. Siddell & H. Wege. New York & London: Plenum Press.

MACNAUGHTON, M. A. & MADGE, M. H. (1978). The genome of human coronavirus strain 229E. *Journal of General Virology* **39**, 497–504.

MACNAUGHTON, M. R. & MADGE, M. H. (1978). The genome of human coronavirus strain 229E. *Journal of General Virology* **39**, 497–504. infected with human coronavirus 229E group viruses. *Infection and Immunity* **31**, 845–849.

MAKINO, S., STOHLMAN, S. A. & LAI, M. M. C. (1986). Leader sequences of murine coronavirus RNA can be freely reassorted: evidence for the role of free leader RNA in transcription. *Proceedings of the National Academy of Sciences, U.S.A.* **83**, 4204–4208.

MANIATIS, T., FRITSCH, E. F. & SAMBROOK, J. (1982). *Molecular Cloning: A Laboratory Manual.* New York: Cold Spring Harbor Laboratory.

MESSING, J. & VIEIRA, J. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. *Gene* **19**, 269–276.

MYINT, S., SIDDELL, S. & TYRRELL, D. (1989). The detection of human coronavirus 229E in nasal washings using RNA:RNA hybridization. *Journal of Medical Virology* **29**, 70–73.

PARKER, M. D., COX, G. J., DEREGT, D., FITZPATRICK, D. R. & BABIUK, L. A. (1989). Cloning and *in vitro* expression of the gene for the E3 haemagglutinin glycoprotein of bovine coronavirus. *Journal of General Virology* **70**, 155–164.

PEDERSEN, N. C., WARD, J. & MENGELING, W. L. (1978). Antigenic relationship of the feline infectious peritonitis virus to coronaviruses of other species. *Archives of Virology* **58**, 45–53.

PHILLPOTTS, R. J. (1983). Clones of MRC-C cells may be superior to the parent line for the culture of 229E-like strains of human respiratory coronavirus. *Journal of Virological Methods* **6**, 267–269.

RAABE, T. & SIDDELL, S. G. (1989*a*). Nucleotide sequence encoding the membrane protein of human coronavirus 229E. *Archives of Virology* **107**, 323–328.

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

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.

SANGER, F., NICKLEN, S. & COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 5463–5467.

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

SCHMIDT, O. W. & KENNY, G. E. (1982). Polypeptides and functions of antigens from human coronaviruses 229E and OC43. *Infection and Immunity* **35**, 515–522.

SCHREIBER, S. S., KAMAHORA, T. & LAI, M. M. C. (1989). Sequence analysis of the nucleocapsid protein gene of human coronavirus 229E. *Virology* **169**, 142–151.

SHIEH, C.-K., SOE, L. H., MAKINO, S., CHANG, M.-F., STOHLMAN, S. A. & LAI, M. M. C. (1987). The 5′ prime end sequence of the murine coronavirus genome: implications for multiple fusion sites in leader-primed transcription. *Virology* **156**, 321–330.

SHIEH, C.-K., SOE, L. H., MAKINO, S., CHANG, M.-F., STOHLMAN, S. A. & LAI, M. M. C. (1987). The 5′ end sequence of the murine coronavirus genome: implications for multiple fusion sites in leader-primed transcription. *Virology* **156**, 321–330.

SIDDELL, S. (1983). Coronavirus JHM: coding assignments of subgenomic mRNAs. *Journal of General Virology* **64**, 113–125.

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 gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Research* **10**, 4731–4751.

STURMAN, L. & HOLMES, K. (1985). The novel glycoproteins of coronaviruses. *Trends in Biochemical Sciences* **10**, 17–20.

STURMAN, L. S., RICARD, C. S. & HOLMES, K. V. (1985). Proteolytic cleavage of the E2 glycoprotein of murine coronavirus: activation of cell-fusing activity of virions by trypsin and separation of two different 90K cleavage fragments. *Journal of Virology* **56**, 904–911.

VAN BERLO, M. F., VAN DEN BRINK, W. J., HORZINEK, M. C. & VAN DER ZEIJST, B. A. M. (1987). Fatty acid acylation of viral proteins in murine hepatitis virus-infected cells. *Archives of Virology* **95**, 123–128.

VLASAK, R., LUYTJES, W., LEIDER, J., SPAAN, W. & PALESE, P. (1988). The E3 protein of bovine coronavirus is a receptor destroying enzyme with acetylesterase activity. *Journal of Virology* **62**, 4686–4690.

VON HEIJNE, G. (1984). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**, 4683–4690.

WEISS, S. R. & LEIBOWITZ, J. L. (1981). Comparison of the RNAs of murine and human coronaviruses. In *Biochemistry and Biology of Coronaviruses*, Advances in Experimental Medicine and Biology, vol. 142, pp. 43–69. Edited by V. ter Meulen, S. Siddell & H. Wege. New York & London: Plenum Press.

WOODS, D. (1984). Oligonucleotide screening of cDNA libraries. *Focus* **6**, 1–3.

(*Received 9 November 1989; Accepted 18 January 1990*)