VIROLOGY 195, 680–691 (1993)

Nucleotide Sequence of the Human Coronavirus 229E RNA Polymerase Locus

J. HEROLD,[1] T. RAABE,[2] B. SCHELLE-PRINZ, AND S. G. SIDDELL

Institute of Virology, University of Würzburg, Versbacher Str. 7, 8700 Würzburg, Germany

Received February 11, 1993; accepted April 12, 1993

The nucleotide sequence of the human coronavirus 229E (HCV 229E) RNA polymerase gene and the 5′ region of the genome has been determined. The polymerase gene is comprised of two large open reading frames, ORF1a and ORF1b, that contain 4086 and 2687 codons, respectively. ORF1b overlaps ORF1a by 43 bases in the (−1) reading frame. The *in vitro* translation of SP6 transcripts which include HCV 229E sequences encompassing the ORF1a/ORF1b junction show that expression of ORF1b can be mediated by ribosomal frame-shifting. The predicted translation products of ORF1a (454,200 molecular weight) and ORF1a/1b (754,200 molecular weight) have been compared to the predicted RNA polymerase gene products of infectious bronchitis virus (IBV) and murine hepatitis virus (MHV) and conserved structural features and putative functional domains have been identified. This analysis completes the nucleotide sequence of the HCV 229E genome. © 1993 Academic Press, Inc.

INTRODUCTION

The human coronaviruses (HCV) are the cause of upper respiratory illness and it has been estimated that up to 20% of common colds are caused by HCV (McIntosh *et al.*, 1974; Isaacs *et al.*, 1983; Macnaughton *et al.*, 1983). Although the symptoms of HCV-related colds are generally mild and the duration of illness is short, the economic consequences of HCV infection are significant (Hierholzer and Tannock, 1988). Also, the possible association of HCV infection with more severe respiratory tract illness in children (Matsumoto and Kawano, 1992) or as a precipitant of asthmatic exacerbations (Pattemore *et al.*, 1992) needs to be further investigated.

It has been established that there are two major antigenic groups of HCV, represented by the prototypes HCV 229E and HCV OC43. The major structural components of HCV 229E and HCV OC43 virions have been identified and there is some limited information on the synthesis of viral RNA and proteins in the infected cell (Schmidt and Kenny, 1982; Schmidt, 1984; Kemp *et al.*, 1984; Hogue and Brian, 1986; Schreiber *et al.*, 1989; Raabe *et al.*, 1990; Arpin and Talbot, 1990).

The HCV 229E genome is a positive-strand RNA with an estimated size of 6×10^6 (Macnaughton and Madge, 1978). To date, the nucleotide sequence of approximately 7 kilobases extending from the 3′ end of the genome has been determined. This region encodes the nucleocapsid protein, N (Schreiber *et al.*, 1989; Myint *et al.*, 1990), the membrane glycoprotein, M (Raabe and Siddell, 1989b; Jouvenne *et al.*, 1990) and the surface glycoprotein, S (Raabe *et al.*, 1990). Additionally, there are three small open reading frames (ORFs), ORF4a, ORF4b, and ORF5, located between the S and the M protein genes (Raabe and Siddell, 1989a; Jouvenne *et al.*, 1992). It seems likely that the putative HCV 229E ORF5 gene product is a virion structural protein (Liu and Inglis 1991; Godet *et al.*, 1992) but the function of the putative ORF4a and ORF4b gene products is unknown.

In coronavirus-infected cells, the viral genes are expressed from the genomic and subgenomic mRNAs. The subgenomic mRNAs form a 3′ coterminal set and are synthesized by a process of discontinuous transcription (for a recent review see Lai, 1990). In the case of HCV 229E, seven positive-strand RNA species (numbered 1 to 7 in order of decreasing size) have been identified in the infected cell. The translation products of the S, ORF4a and 4b, ORF5, M, and N protein genes have been provisionally assigned to RNA 2, 4, 5, 6, and 7, respectively (Raabe *et al.*, 1990; Schreiber *et al.*, 1989), although messenger RNA function has been confirmed only for RNA 7 (Myint *et al.*, 1990). It is not yet clear whether RNA 3 should be considered as a putative mRNA (Raabe *et al.*, 1990).

The remainder of the HCV 229E genome encompasses the unique region of RNA 1, i.e., the genomic RNA. This region, which is referred to as gene 1, the RNA polymerase gene or the RNA polymerase locus, has been entirely sequenced for IBV and MHV (Boursnell *et al.*, 1987; Bredenbeek *et al.*, 1990; Lee *et al.*, 1991) and is comprised of two large ORFs, ORF1a and

[1] To whom reprint requests should be addressed.

[2] Current address: Institute of Zoology, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland.

0042-6822/93 $5.00
Copyright © 1993 by Academic Press, Inc.
All rights of reproduction in any form reserved.

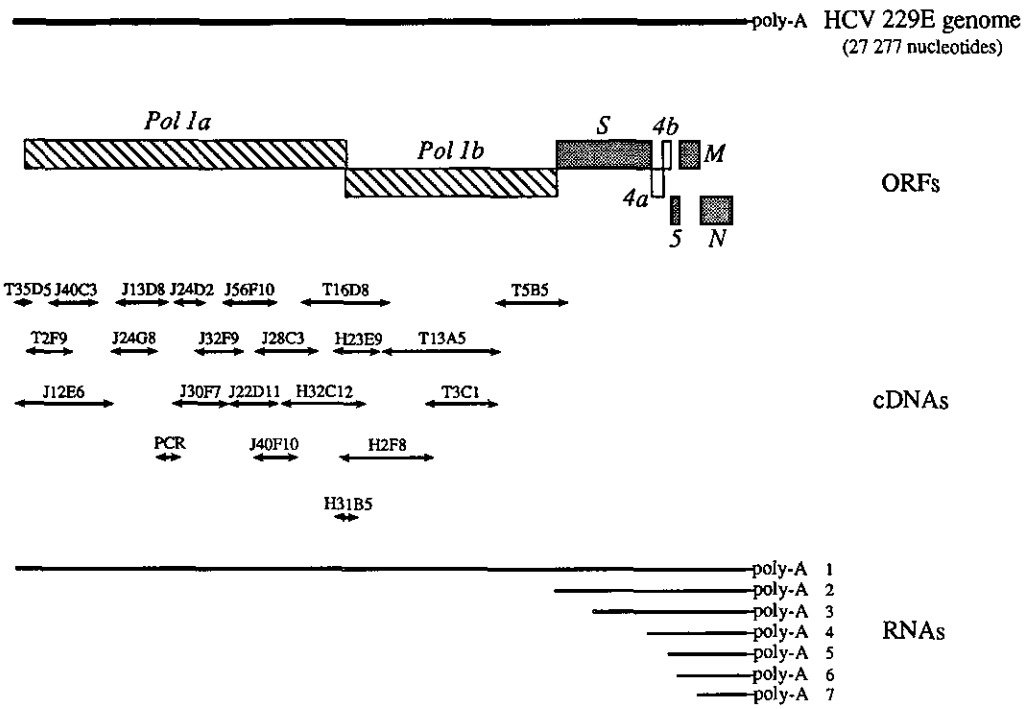

FIG. 1. Organization of the HCV 229E genome and the position of cDNA clones encompassing gene 1. The major ORFs are represented as boxes in the 0, −1, and −2 reading frames. The known and putative structural genes (S, 5, M, and N) are shaded and gene 1 (Pol1a and Pol1b) is cross-hatched. The size and position of the cDNA clones and PCR amplification products used to determine the HCV 229E gene 1 sequence are shown. The relationship of the intracellular poly(A) RNA species to the genome is also illustrated.

ORF1b, which overlap by 40–80 bases. The upstream ORF1a potentially encodes a polypeptide of 450,000 to 500,000 molecular weight. The downstream ORF1b potentially encodes a polypeptide of 300,000 molecular weight. The downstream ORF1b is expressed, however, as a fusion protein together with the ORF1a gene product by a mechanism involving (−1) ribosome slippage (Brierley *et al.*, 1987). This ribosomal frameshift is mediated by a "slippery sequence" and pseudoknot structure located in a region of the genome encompassing the overlap of ORF1a and ORF1b (Brierley *et al.*, 1989, 1991, 1992; Bredenbeek *et al.*, 1990; Lee *et al.*, 1991).

In the case of IBV and MHV, the ORF1b regions are relatively conserved whereas the ORF1a regions have diverged, in particular toward their 5′ ends. It is evident that these two large ORFs must encode a number of different functions. First, there are functions related to RNA replication. Complementation analysis of MHV ts mutants with a RNA minus phenotype has shown that there are at least five distinct viral functions related to RNA synthesis (Leibowitz *et al.*, 1982; Schaad *et al.*, 1990). Analysis of these mutants by genetic recombination allows the different functions to be located and ordered within the gene 1 locus (Keck *et al.*, 1987; Baric *et al.*, 1990). Also, both IBV and MHV contain in their ORF1b sequence motifs characteristic of RNA polymerases, helicase and metal binding proteins (Gorbalenya *et al.*, 1989; Bredenbeek *et al.*, 1990).

Second, there is genetic and biochemical evidence that the MHV gene 1 contains viral encoded proteases. The complementation frequencies of MHV ts mutants are indicative of intergenic, rather than intragenic complementation (Leibowitz *et al.*, 1982) and an autoproteolytic activity has been mapped to the middle of the MHV ORF1a (Baker *et al.*, 1989). Motifs characteristic of both papain-like and picornavirus 3C-like cysteine proteases have also been identified in ORF1a of MHV and IBV (Gorbalenya *et al.*, 1989; Lee *et al.*, 1991).

Finally, the large size of the coronavirus gene 1 region (approximately 20 kilobases) suggests that it may encode many, as yet unidentified, functions. One obvious candidate would be a methyltransferase activity necessary for the generation of capped viral RNA in the cytoplasm of infected cells. Other functions may be related to the conserved "membrane protein," "cysteine-rich," and "X" domains which have been identified in gene 1 of IBV and MHV (Gorbalenya *et al.*, 1989; Lee *et al.*, 1991).

In this paper we report the nucleotide sequence of the human coronavirus 229E gene 1 and the 5′ region of the genome. This analysis completes the nucleotide sequence of HCV 229E. Furthermore, we provide evidence that, in common with IBV and MHV, HCV 229E ORF1b expression is mediated by (−1) ribosomal frame-shifting. The identification of structural and putative functional motifs in the predicted HCV gene 1 product and a comparison of their organization in the

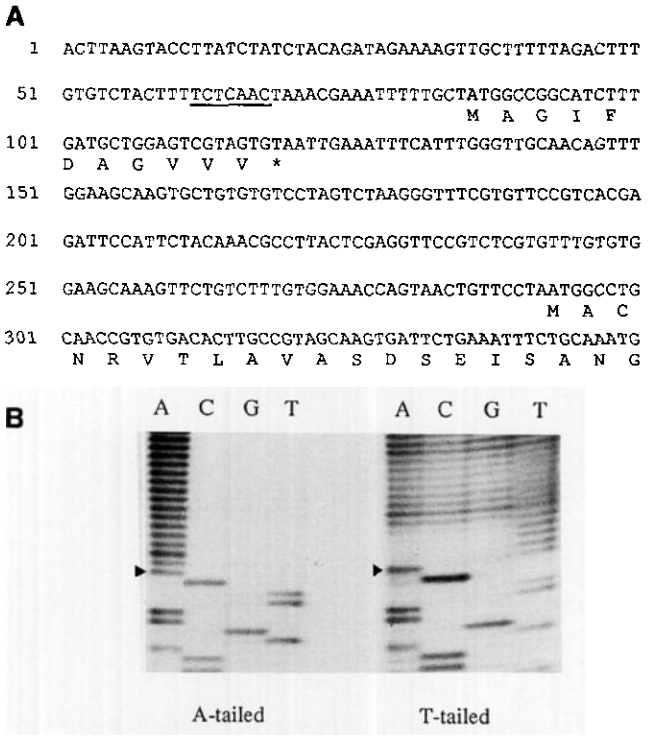

FIG. 2. Consensus sequence of cDNA clones representing the 5' region of the HCV 229E genome. (A) The consensus sequence. The intergenic motif, UCUCAAC, is underlined. ORF1a is initiated with AUG at position 293 and the conserved 5' "minicistron" is indicated. (B) Sequence analysis of the 5' RACE products. The sequence of the A-tailed and T-tailed 5' RACE products derived from the HCV leader RNA were determined as described in Methods. The 5' terminal A nucleotide is indicated.

gene 1 proteins of HCV 229E, IBV, and MHV is also presented.

MATERIALS AND METHODS

Virus and cells

The HCV 229E isolate used in these studies, the methods of virus propagation in C16 cells, and the isolation of cytoplasmic, poly(A)-containing RNA from HCV 229E-infected cells have been described (Raabe *et al.*, 1990).

cDNA cloning

cDNA synthesis was done by the method of Gubler and Hoffman (1983) using random hexanucleotides or the HCV 229E S gene specific oligonucleotide 1 (Raabe *et al.*, 1990) as reverse transcription primers. The synthesized double-stranded cDNA was size-fractionated on a Sephacryl S-1000 column, cloned into pBluescript II KS^+ and transformed into competent *Escherichia coli* TG-1 cells. Recombinant clones were screened by colony hybridization with HCV 229E-specific, ^{32}P-labeled oligonucleotides or HCV 229E-specific cDNAs. Standard recombinant DNA procedures were done as described by Sambrook *et al.* (1989) and colony hybridizations were done as described by Woods (1984).

PCR

PCR was done using a GeneAmp/RNA PCR Kit according to the manufacturers procedures (Perkin–Elmer Cetus, Überlingen, Germany). The biotinylated oligonucleotide 2 was used as upstream primer and oligonucleotide 3 was used as downstream and reverse transcription primer. The resulting cDNA strands were separated using streptavidin-coupled magnetic beads, according to the manufacturers protocol (Dynal, Hamburg, Germany) and the nucleotide sequence of both strands was determined.

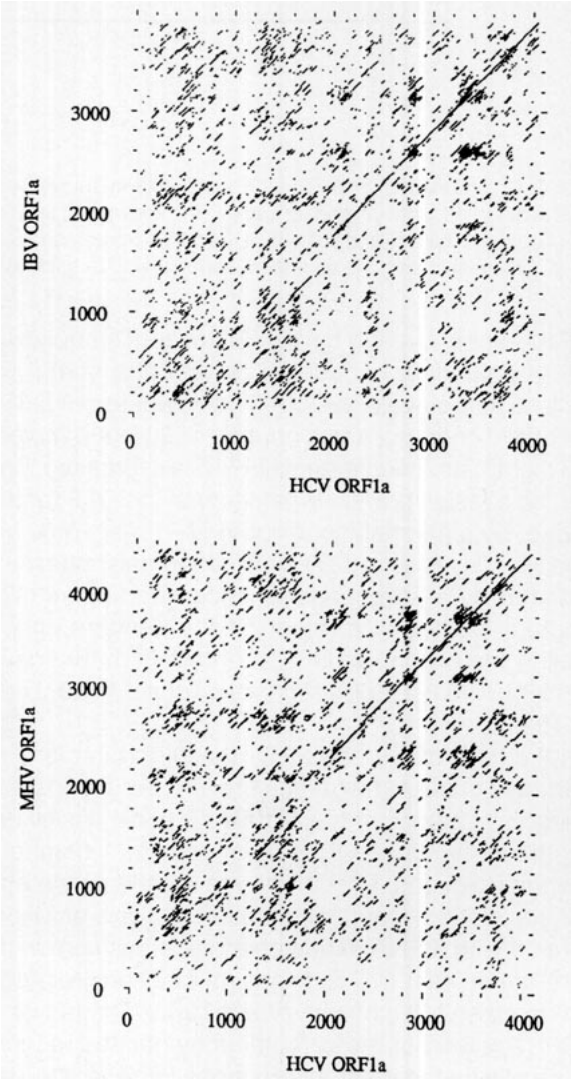

FIG. 3. Dot matrix comparisons of the predicted amino acid sequences of the ORF1a proteins of HCV 229E, IBV and MHV. Comparisons of the HCV and IBV proteins (upper panel) and the HCV and MHV proteins (lower panel) were generated using the GCG program COMPARE (window, 100; stringency, 30; default comparison table) and displayed with the program DOTPLOT.

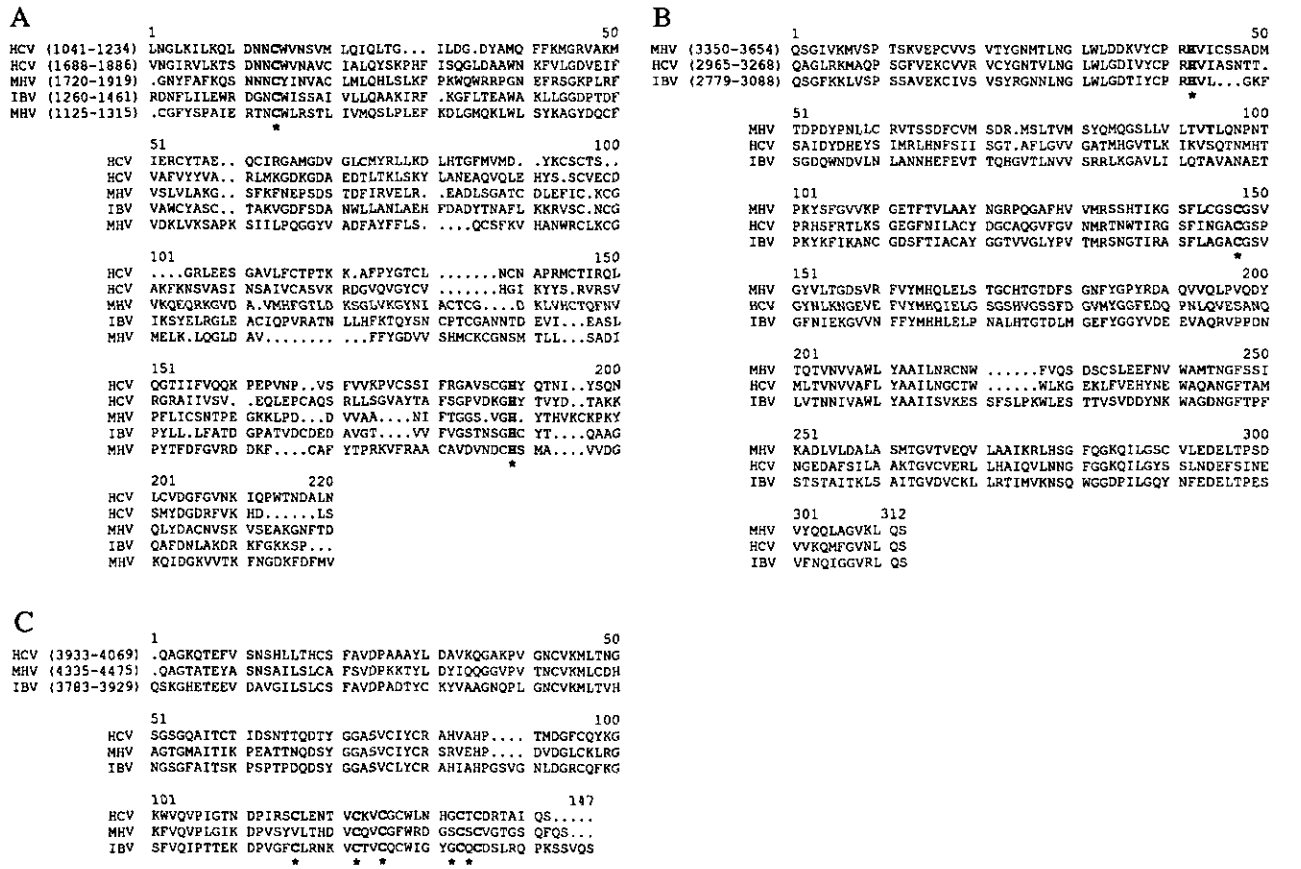

FIG. 4. Putative functional domains of the HCV 229E ORF1a translation product. The amino acid sequences of ORF1a of HCV 229E, MHV, and IBV were aligned using the UWGCG program PileUp (default settings) and the structgappep.cmp (A) or pam250.cmp (B and C) comparison tables. (A) The papain-like protease motifs; (B) the 3C-like protease motif: (C) The growth factor/receptor-like motif. In (A) and (B) the catalytic residues proposed by Lee *et al.* (1991) are shown in bold type and marked with an asterisk. In (C) the putative disulphide bond residues proposed by Gorbalenya *et al.* (1989) are similarly highlighted. The numbering of the aligned sequences is for reference only.

DNA sequencing

Sequencing was done on single-strand and double-strand templates using the chain termination method and M13, T7, T3, and HCV 229E-specific sequencing primers. To generate cDNA sequencing templates, overlapping deletions were introduced by unidirectional exonuclease III digestion (Henikoff, 1984). Both strands of all cDNAs and the PCR product were sequenced. Sequence data was assembled by the program of Staden (1982) and analysed by the programs of the Genetics Computer Group, Inc. (Devereux *et al.*, 1984).

5′ RACE

Sequences at the 5′ end of the HCV 229E leader RNA were determined by a "rapid amplification of cDNA ends" method (Frohmann *et al.*, 1988). A ^{32}P-labeled oligonucleotide, 4, complementary to a region of the HCV 229E leader RNA (Schreiber *et al.*, 1989), was used as primer for the reverse transcription of cytoplasmic, poly(A)-containing RNA from HCV-infected cells. Reverse transcription was done with Superscript RNase H$^-$ reverse transcriptase (Gibco, Eggenstein, Germany) using the manufacturers protocol. The largest product was purified by gel electrophoresis and tailed with dATP or dTTP using terminal transferase. The tailed cDNAs were then amplified in separate 3 primer PCRs (A-tailed product: Oligonucleotides 4 and 5 and biotinylated oligonucleotide 2; T-tailed product: Oligonucleotides 4 and 6 and biotinylated oligonucleotide 2). The amplifications were done using AmpliTaq DNA polymerase (Perkin–Elmer Cetus) by heating the reaction to 94°, followed by 3 cycles of denaturation (94°, 1 min), annealing (45°, 1 min), and extension (72°, 5 sec), 30 cycles of denaturation (94°, 1 min), annealing (51°, 1 min) and extension (72°, 5 sec) and a final extension step of 72° for 10 min. The cDNA strands were separated using streptavidin coupled magnetic beads and the biotinylated strand was sequenced using primer 4.

Oligonucleotides

Oligonucleotides were synthesized using phosphoamidite chemistry on a Cyclone DNA synthesizer and purified by gel electrophoresis. The 5′ biotinylated

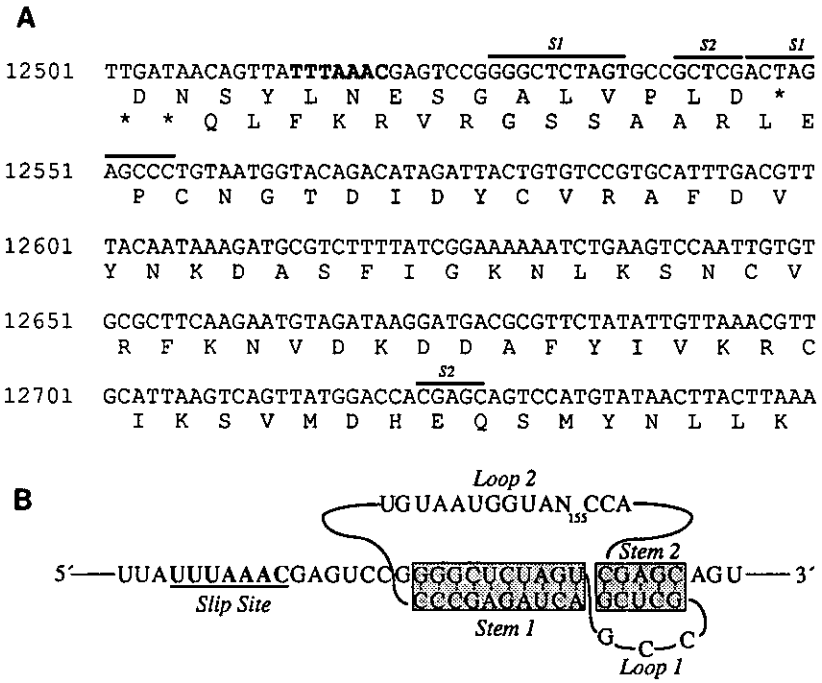

FIG. 5. Analysis of HCV RNA-mediated ribosomal frame-shifting. (A) The consensus sequence of cDNA clones in the region of the ORF1a/ORF1b overlap. The ends of the ORF1a and ORF1b sequences are indicated. The putative slippage site, TTTAAAC, is shown in bold type and the complementary sequences which we propose to form the S1 and S2 stems are overlined. (B) A proposed model of the HCV 229E pseudoknot structure at the ORF1a-ORF1b junction. (C) The structure of plasmids pFS and pΔFS. The DNA structure of pFS and pΔFS is schematically shown together with the position of the HCV 229E ORF1a/ORF1b overlap. The size of the SP6 run off transcription products and the translation products predicted in the event of ORF1a termination or (−1) ribosomal frame-shifting are shown. (D) *In vitro* translation products of pFS and pΔFS mRNA. Lane M, molecular weight markers (CFA626, Amersham Buchler, Braunschweig, Germany); lane 1, no RNA; lane 2, pΔFS/*Bam*HI RNA; lane 3, pFS/*Afl*II RNA; lane 4, pFS/*Bst*EII RNA.

oligonucleotide was purchased (MWG-Biotech, Ebersberg, Germany). The oligonucleotides used for cDNA synthesis, PCR, and 5′ RACE were

1. 5′-CAT CTA CAA CAG ATG AGG-3′
2. 5′-BIOTIN GCC TAT GAA AGT GCT GTT GTT AAT GG-3′
3. 5′-TTA GAT TTA AGA ACA GCC TGT GAC GC-3′
4. 5′-GTA GAC ACA AAG TCT AAA AAG C-3′
5. 5′-GCC TAT GAA AGT GCT GTT GTT AAT GGT_{18}-3′
6. 5′-GCC TAT GAA AGT GCT GTT GTT AAT GGA_{18}-3′.

Construction of plasmids pFS and pΔFS

A 1264 base pair *Nde*I–*Hpa*I fragment of clone T16D8 (corresponding to bases 12,293–13,557 in the HCV genome, see Fig. 1), or a 427 base pair *Nde*I–*Sau*96I fragment of clone T16D8 (corresponding to bases 12,293–12,720) was treated with the Klenow fragment of DNA polymerase and exchanged with the small (230 base pair) *Eco*RV fragment of pSP65-GUS (Prüfer *et al.*, 1992). The clones containing the HCV DNA fragments in the correct orientation (pFS and pΔFS, respectively) were identified by restriction enzyme analysis and the constructions were verified by sequencing.

In vitro transcription and translation

Plasmid DNA was linearized with *Afl*II or *Bst*EII (pFS) or *Bam*HI (pΔFS) and transcribed with SP6 RNA polymerase as described by Melton *et al.* (1984). The *in vitro* synthesized, capped RNAs were translated in a rabbit reticulocyte lysate in the presence of [^{35}S]-methionine and the products were analyzed on 10% polyacrylamide–SDS gels as described previously (Siddell, 1983). The radioactivity incorporated into the translation products was determined using a PhosphorImager Model 400E (Molecular Dynamics, Sunnyvale, USA).

RESULTS

Molecular cloning and nucleotide sequence of the HCV 229E gene 1

The HCV 229E gene 1 was cloned in a series of 21 cDNA clones. One region, corresponding to bases 5781–5934 in the genomic RNA, was not represented in the three cDNA libraries screened and it was sequenced by PCR techniques. More than 85% of the sequence presented was determined on two or more

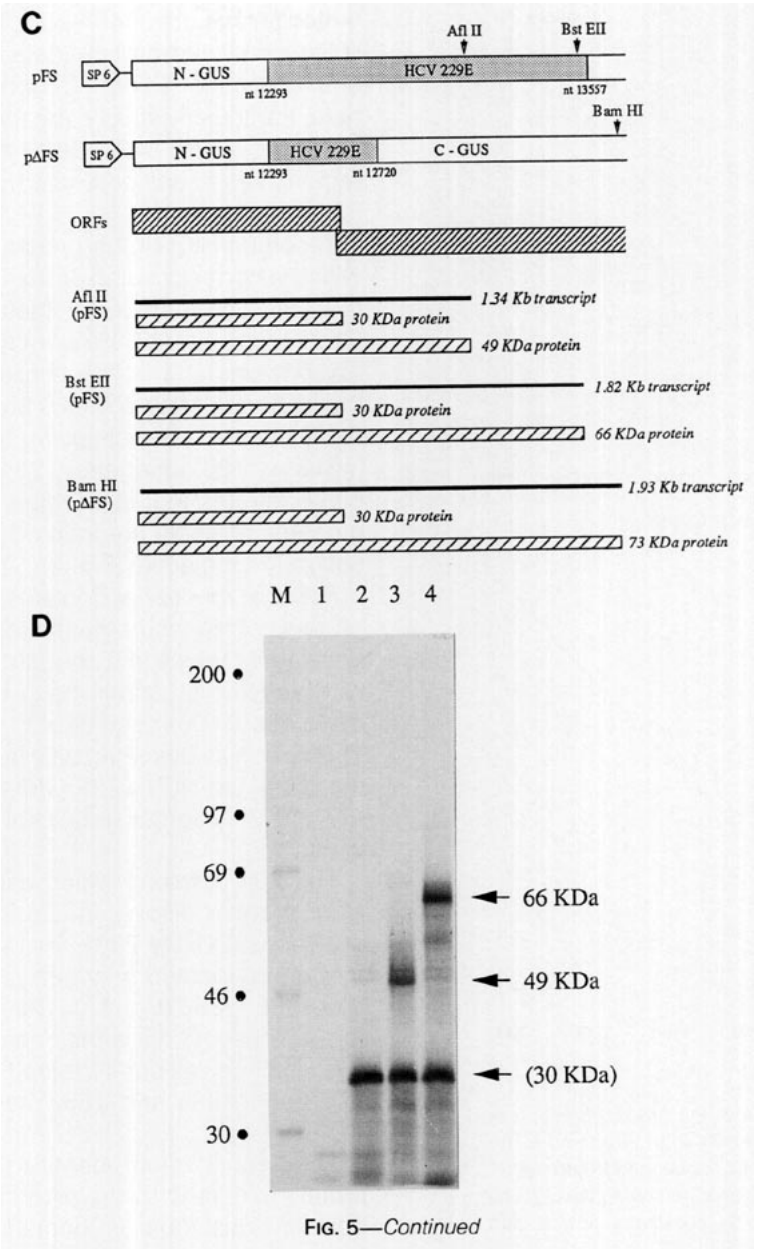

FIG. 5—*Continued*

cDNA clones. The sequence encompassing the "frameshift region" (see below) was obtained on five independent cDNA clones from two cDNA libraries (Fig. 1). The length of the consensus cDNA sequence is 20,774 nucleotides, extending from base 1 at the 5′ end of the genome to base 20,774, which corresponds to the 68th codon of the HCV S protein gene. Within this sequence are two large ORFs. ORF1a is initiated with an AUG at base 293 and contains 4086 codons. ORF1b, which is initiated at base 12,508 with CAG contains 2687 codons and overlaps ORF1a by 43 bases in the (−1) reading frame. The nucleotide sequence of the HCV 229E gene 1 has been deposited with the EMBL/GenBank/DDBJ nucleotide sequence Data Libraries and is available under accession number X69721.

5′ Region of the genome

The consensus sequence of cDNAs which encompass the region of the HCV 229E genome preceding the ORF1a initiation codon was deduced from the sequence of cDNA clones J12E6 and T35D5 together with 5′ RACE clones produced from poly(A)-containing RNA (Fig. 2). The validity of the HCV 229E sequence, therefore, depends upon the assumption that the mRNA leader sequence is equivalent to the 5′ end of the genomic RNA. This equivalence has been demonstrated for MHV (Shieh *et al.*, 1987).

The genomic sequence begins with an adenine. At position 62–68 the sequence UCUCAAC is found. This or a closely related sequence is located adjacent to the 5′ end of all HCV 229E genes and represents the so-

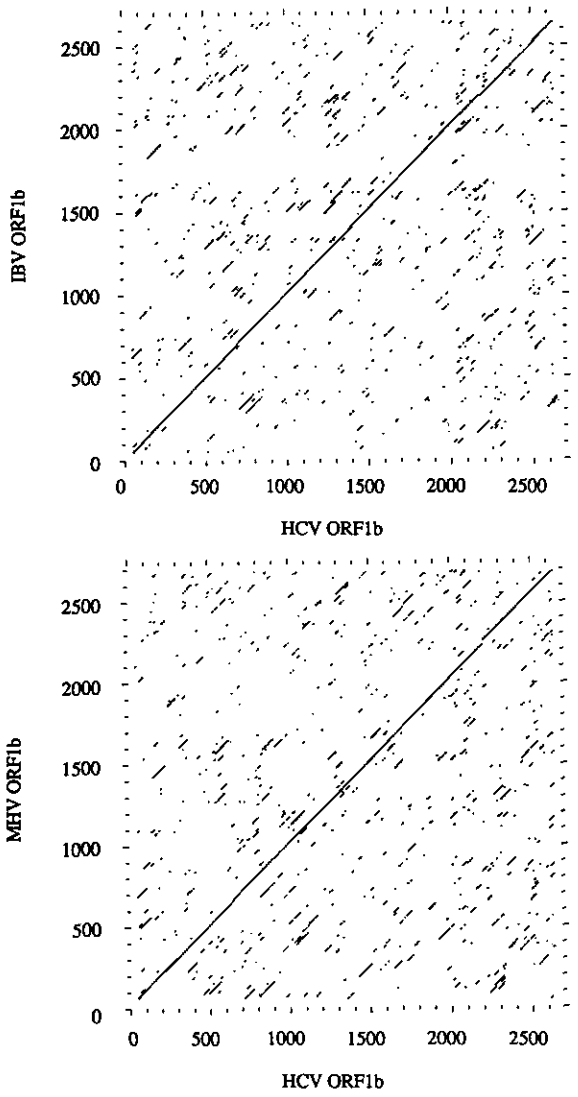

FIG. 6. Dot matrix comparisons of the predicted amino acid sequences of the ORF1b proteins of HCV 229E, IBV, and MHV. Comparisons of the HCV and IBV proteins (upper panel) and the HCV and MHV proteins (lower panel) were generated using the GCG program COMPARE (window, 100; stringency, 30; default comparison table) and displayed with the program DOTPLOT.

called "intergenic consensus" sequence which is believed to have a pivotal role in the discontinuous transcription of coronavirus mRNAs (Joo and Makino, 1992). At position 86 a short ORF of 12 codons is initiated with AUG. This ORF would be unremarkable except that similar ORFs are conserved in the genomes of IBV and MHV (Boursnell *et al.*, 1987; Soe *et al.*, 1987). It might be speculated that this ORF has a role, for example, in the regulation of the initiation of protein synthesis from genomic RNA.

ORF1a

Structural features. ORF1a has the potential to encode a protein of 4085 amino acids with a predicted molecular weight of 454,200. The hydrophilicity profile of the predicted protein (data not shown) shows several regions in which hydrophobic residues predominate. Particularly striking are the regions encompassed by amino acids 2720–2890 and 3270–3510. These regions represent potential membrane spanning domains.

A comparison of the predicted HCV ORF1a protein with the corresponding proteins of IBV and MHV using the GCG program GAP (default settings) indicates 51.2% similarity (27.3% identity) between the HCV and IBV proteins and 51.4% similarity (28.0% identity) for the HCV and MHV proteins after optimal alignment. A more detailed analysis using the COMPARE program (window, 100; stringency, 30; default comparison table) illustrates that in all three proteins the regions of greatest similarity are located in the carboxy-terminal half of the molecule (Fig. 3).

Putative functional domains. The predicted ORF1a proteins of IBV and MHV have been analyzed in detail and motifs which are thought to represent domains with specific functions have been identified (Gorbalenya *et al.*, 1989; Lee *et al.*, 1991; Bredenbeek *et al.*, 1990). This analysis can be extended by comparison of the predicted HCV 229E ORF1a protein with those of IBV and MHV and the results of this analysis are shown in Fig. 4.

The first domains which can be recognized in the HCV protein, display motifs indicative of papain-like proteases. In HCV 229E, as in MHV, two such motifs are found, located between amino acids 1041–1234 and 1688–1886 (Fig. 4A). The most characteristic feature of these motifs is the conserved putative catalytic Cys and His residues located at positions 1054 and 1701 and 1305 and 1663, respectively, in the HCV protein.

The second motif identified in the predicted HCV protein is related to the picornavirus 3C-like protease domain. This motif is located between amino acids 2965–3265 (Fig. 4B). It should be noted that the features which distinguish the coronavirus 3C-like motif from other 3C-like protease motifs (a Gly → Tyr substitution in the vicinity of the proposed catalytic Cys residue and the absence of a conserved Asp/Glu as a third catalytic site residue) are maintained in the predicted HCV protein.

The third HCV ORF1a motif which has been identified is a cysteine-rich domain located between amino acids 3933–4069 (Fig. 4C). This motif has been recognized in the MHV and IBV genomes and is related to motifs found in growth factors and their receptors.

The frame-shifting region

By analogy to IBV and MHV it seems likely that expression of the HCV ORF1b is mediated by a (−1) ribosomal frame-shifting event during translation of the genomic RNA in the region of the ORF1a/ORF1b overlap.

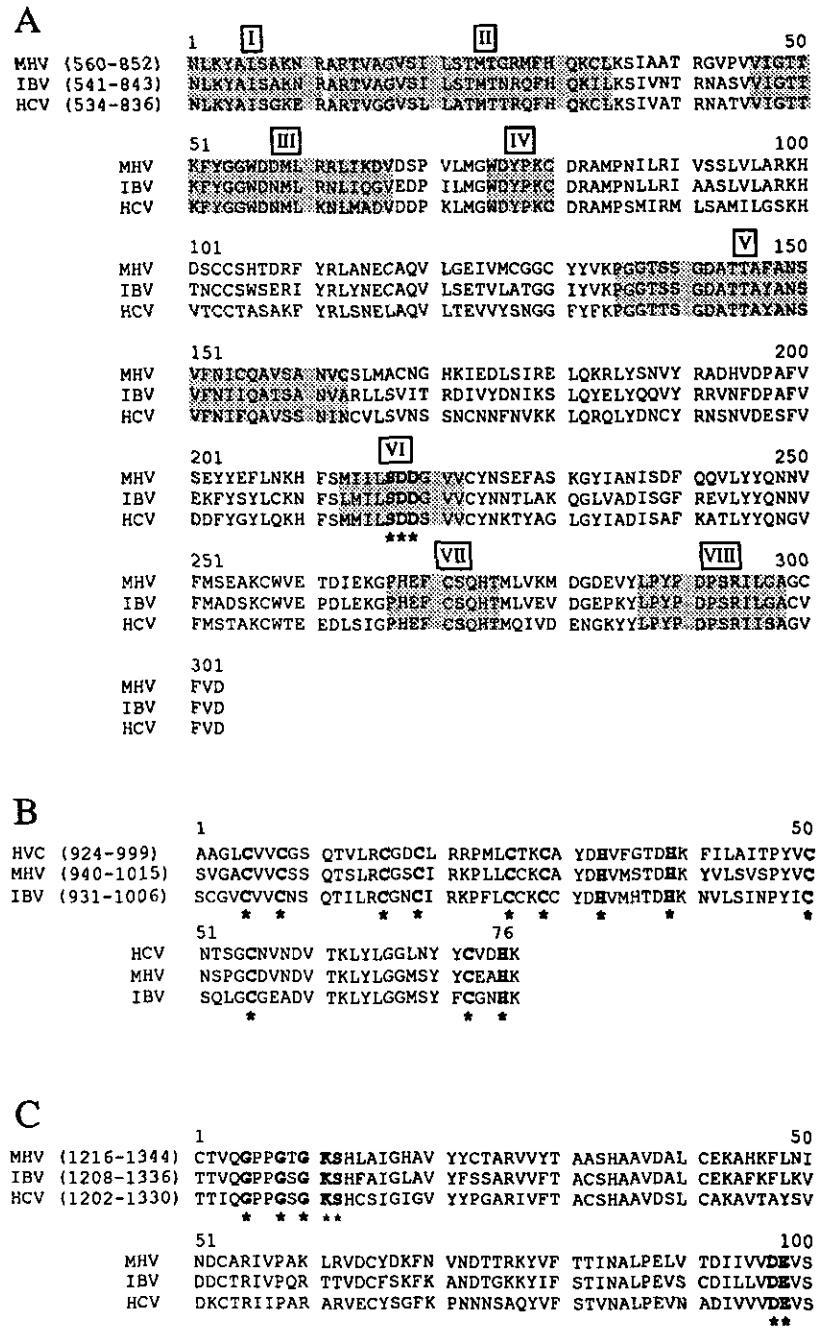

FIG. 7. Putative functional domains of the HCV 229E ORF1b translation product. The amino acid sequences of ORF1b of HCV 229E, MHV, and IBV were aligned using the UWGCG program PileUp (default settings) and the pam250.cmp comparison table. (A) The RNA polymerase domain, (B) the metal binding domain, (C) the helicase domain. In (A) the characteristic GDD (SDD) motif is highlighted and the polymerase domains I to VIII (Koonin, 1991) are shaded. In figure (B) the conserved Cys/His residues which may be involved in metal ion ligation (Lee *et al.*, 1991) are shown in bold type and marked with an asterisk. In (C) the characteristic "A" and "B" sites (Gorbalenya and Koonin, 1989) are shown and conserved residues are similarly highlighted. The numbering of the aligned sequence is for reference only.

The consensus sequence of the cDNA clones which encompass this region is shown in Fig. 5A. The sequence UUUAAAC which is found at position 12,514–12,520 in the HCV sequence, 27 bases upstream of the ORF1a termination codon, is identical to the slippage site of IBV (Brierley *et al.*, 1992) and the putative slippage site of MHV-JHM (Lee *et al.*, 1991) and MHV-A59 (Bredenbeek *et al.*, 1990). The HCV sequences 3′ of this site can also folded into a tertiary RNA structure, the pseudoknot, which is the second element required for efficient frame shifting (Brierley *et al.*, 1989). Figure 5B illustrates a pseudoknot structure in which the stem S1 is formed by base pairing of nucleotides 12,528–12,537 and 12,546–12,555 and the stem S2 is formed by base pairing of nucleotides 12,541–12,545 and 12,723–12,727 (overlined in Fig. 5A). This structure

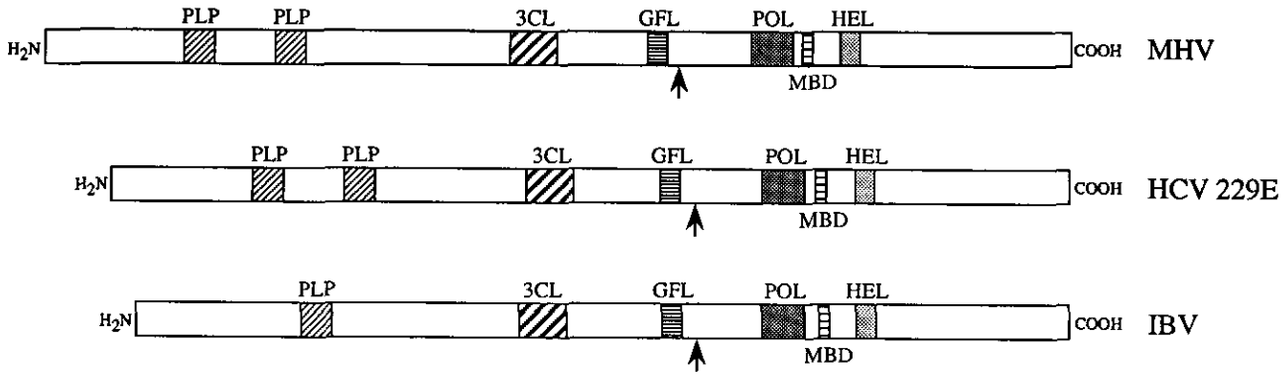

FIG. 8. Position of the putative functional domains in the gene 1 translation products of HCV 229E, MHV, and IBV. The figure is drawn to scale although the boundaries of the motifs cannot be defined precisely. PLP, papain-like protease; 3CL, 3C-like protease; GFL, growth factor/receptor-like; POL, polymerase module; MBD, metal binding domain; HEL, helicase (NTP-binding) domain. The arrow indicates the position of the ORF1a/ORF1b junction.

would necessitate an L1 loop of 3 bases and an L2 loop of 168 bases, values that exceed the minimum required length (Brierley *et al.*, 1991). Clearly, further experimental evidence will be needed to confirm or refute this model.

To confirm that the HCV 229E ORF1a/ORF1b overlap region is able to mediate (−1) ribosomal frameshifting we constructed two plasmids for the *in vitro* transcription of mRNA (Fig. 5C). Plasmid pFS contains the putative frame-shifting region (nucleotides 12,293–13,557) flanked by and in frame with DNA encoding the amino- and carboxy-terminal regions of the *E. coli* β-glucuronidase (GUS) protein. Plasmid pΔFS was identical, except that the HCV 229E sequences extended only from position 12,293 to 12,720, i.e., did not include the pentanucleotide sequence CGAGC which is complementary to the sequence GCUCG which we propose to be in the S2 stem of the pseudoknot structure. The plasmids were linearized with *Afl*II or *Bst*EII (pFS) or *Bam*HI (pΔFS) and capped SP6 run-off transcripts were synthesized *in vitro*. The transcripts were translated in rabbit reticulocyte lysate and the results are shown in Fig. 5D.

The pFS/*Afl*II transcript directed the synthesis of 34,000 and 49,000 molecular weight proteins. The pFS/*Bst*EII transcript directed the synthesis of 34,000 and 66,000 molecular weight proteins and the pΔFS/*Bam*HI transcript directed the synthesis of a 34,000 molecular weight protein. By reference to Fig. 5C, it can be seen that these are the results expected if the HCV sequence of pFS mediates (−1) ribosomal frameshifting and the proposed pentanucleotide base-pairing interaction is necessary to produce a functional frame-shifting element. A quantitative PhosphorImager analysis of the data shown in Fig. 5D indicates that in the pFS transcripts frame-shifting occurs at a frequency between 18 and 30%.

Careful analysis of the translation products directed by the pΔFS/*Bam*HI transcript reveals a protein of 73,000 molecular weight, which would be expected if (−1) frame-shifting has occurred. The amount of protein synthesized represents a frame-shifting frequency of < 1%. We believe this could be explained by a less stable S2 stem formed between the GCUCG pentanucleotide at position 12,541–12,545 and the sequence CGUGC located at nucleotides 12,586–12,590. Further studies are required to confirm this interpretation. We have also noted that in all translations the N-GUS-HCV ORF1a product, predicted to have a molecular weight of 30.700, has a slower electrophoretic mobility than expected. At the moment we have no explanation for this anomaly.

ORF1b

Structural features. ORF1b has the potential to encode a protein of 2686 amino acids with a molecular weight of 300,300. If, however, (−1) ribosomal frameshifting takes place at the slippage site in ORF1a (see above), the ORF1a/ORF1b fusion protein has a potential molecular weight of 754,200. The hydrophilicity profile of the predicted ORF1b translation product (data not shown) shows both hydrophilic and hydrophobic regions but none are indicative of extensive membrane spanning regions. A comparison of the predicted HCV ORF1b protein with the corresponding proteins of IBV and MHV using the GAP program (default settings) indicates 69.7% similarity (53.8% identity) for the HCV and IBV proteins and 70.5% similarity (54.2% identity) for the HCV and MHV proteins after optimal alignment. This high degree of similarity is essentially uniform over the entire length of all three proteins, as is evident in the dot matrix comparisons shown in Fig. 6 (program COMPARE, window, 100; stringency, 30; default comparison table).

Putative functional domains. As with ORF1a, the HCV ORF1b gene product can be compared with the ORF1b proteins of MHV and IBV and putative func-

tional motifs can be identified. The first such motif is the RNA polymerase element located between amino acids 534 and 836 (Fig. 7A). The HCV motif aligns well with the MHV and IBV motifs and can be divided into eight distinct regions recognized by Koonin (1991) as characteristic of a wide variety of putative RNA polymerases. The alteration of the RNA polymerase "core" sequence Glu-Asp-Asp to Ser-Asp-Asp is maintained in the HCV ORF1b protein.

The second motif recognized in the HCV protein is related to the "finger" domain characteristic of numerous DNA and RNA binding proteins. This motif located between amino acids 924–999 in the HCV protein, consists of a defined sequence of Cys and His residues. As for the homologous region of the MHV protein, not all of the residues which were originally proposed to be involved in the IBV ORF1b metal binding domain are conserved in the HCV sequence (Fig. 7B) (Gorbalenya *et al.,* 1989).

The third motif identified in the predicted HCV protein is the purine NTP binding sequence pattern which is thought to be a feature of duplex unwinding (i.e., helicase) activities (Gorbalenya and Koonin, 1989). This motif is located in the HCV ORF1b protein at position 1202–1330 and is highly conserved in comparison to the same motif in the MHV and IBV proteins (Fig. 7C).

In addition to the sequence similarities in the RNA polymerase genes of HCV, IBV and MHV, recent analysis of arterivirus and torovirus RNA polymerase genes (Snijder *et al.,* 1990; Kuo *et al.,* 1991; Den Boon *et al.,* 1991) have revealed evolutionary links between arteri-, toro-, and coronaviruses. The polymerase core motif, the finger domain and the NTP binding sequence pattern described above are found, for example, in the polymerase genes of equine arteritis virus and Berne virus. Also, a conserved domain located at the carboxy-terminus of coronavirus, arterivirus and torovirus ORF1b proteins has been recognized, but a function has not yet been proposed (Snijder *et al.,* 1990).

DISCUSSION

Coronaviruses have been traditionally divided into four antigenic groups (Holmes, 1990). HCV 229E belongs to group 1, together with transmissible gastroenteritis virus (TGEV), canine coronavirus (CCV), feline infectious peritonitis virus (FIPV) and feline enteric coronavirus (FECV) (see, however, Sanchez *et al.,* 1990). Thus the nucleotide sequence of a group 1 (HCV229E), a group 2 (MHV), and a group 3 (IBV) coronavirus is now available. HCV 229E is also the first human coronavirus to be entirely sequenced and we hope that many questions concerning the biology and pathogenesis of these viruses can now be investigated more easily.

The HCV 229E gene 1 is comparable in size and organization to gene 1 of IBV and MHV. The predicted gene product displays a number of structural features and putative functional domains (Fig. 8). These include functions related to RNA synthesis (POL, MBD, and HEL) in the ORF1b gene product and proteolytic activities (PLP and 3CL) in the ORF1a gene product. The experiments of ourselves and others (Brierley *et al.,* 1987; Lee *et al.,* 1991; Bredenbeek *et al.,* 1990) show that expression of these functions can be regulated via the mechanism of ribosomal frame-shifting. At the same time, we predict that *in vivo* they are also likely to be coordinated by the activation or inactivation of one set of functions (RNA synthesis) by the other (proteases). Clearly, it will be a difficult task to unravel these complex interactions. However, the availability of a complete set of cDNAs encompassing the HCV polymerase gene serves as a useful starting point.

First, the cDNAs can be used to generate a collection of immunological reagents which facilitate the analysis of polymerase gene expression in HCV-infected cells. Without such reagents it will be very difficult to identify and characterize the low amounts of gene 1 products which can be expected. In this respect, an important step forward has also been the recent identification of the cellular receptor for HCV 229E as aminopeptidase N (Yeager *et al.,* 1992). This finding may allow the development of better cell culture systems for the biochemical analysis of HCV replication.

Second, the HCV polymerase cDNAs together with recently developed vaccinia virus vectors (Merchlinsky and Moss, 1992) should make it possible to (over) express the HCV 229E polymerase gene in eucaryotic cells. This will also facilitate the analysis of polymerase gene expression and more importantly provide an opportunity to investigate the function of polymerase gene products via reverse genetics. Experiments toward these goals are in progress.

ACKNOWLEDGMENTS

We thank Atiye Toksoy for technical help and D. Prüfer for the plasmid pSP65-GUS. This work was financed by the DFG (SFB 165-B1) and the Bundesministerium für Forschung und Technologie (BMFT 01 KI 8838/0).

REFERENCES

ARPIN, N., and TALBOT, P. J. (1990). Molecular characterization of the 229E strain of human coronavirus. *Adv. Exp. Med. Biol.* **276,** 73–80.

BAKER, S. C., SHIEH, C. K., SOE, L. H., CHANG, M. F., VANNIER, D. M., and LAI, M. M. C. (1989). Identification of a domain required for autoproteolytic cleavage of murine coronavirus gene A polyprotein. *J. Virol.* **63,** 3693–3699.

BARIC, R. S., FU, K., SCHAAD, M. C., and STOHLMAN, S. A. (1990). Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. *Virology* **177,** 646–656.

BOURSNELL, M. E., BROWN, T. D. K., FOULDS, I. J., GREEN, P. F., TOMLEY, F. M., and BINNS, M. M. (1987). Completion of the sequence

of the genome of the coronavirus avian infectious bronchitis virus. *J. Gen. Virol.* **68,** 57–67.

BREDENBEEK, P. J., PACHUK, C. J., NOTEN, A. F., CHARITE, J., LUYTJES, W., WEISS, S. R., and SPAAN, W. J. (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.

BRIERLEY, I., BOURSNELL, M. E., BINNS, M. M., BILIMORIA, B., BLOK, V. C., BROWN, T. D. K., and INGLIS, S. C. (1987). An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. *EMBO J.* **6,** 3779–3785.

BRIERLEY, I., DIGARD, P., and INGLIS, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell* **57,** 537–547.

BRIERLEY, I., JENNER, A. J., and INGLIS, S. C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.* **227,** 463–479.

BRIERLEY, I., ROLLEY, N. J., JENNER, A. J., and INGLIS, S. C. (1991). Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.* **220,** 889–902.

DEN BOON, J. A., SNIJDER, E. J., CHIRNSIDE, E. D., DE VRIES, A. A. F., HORZINEK, M. C., and SPAAN, W. J. M. (1991). Equine arteritis virus is not a togavirus but belongs to the coronavirus-like superfamily. *J. Virol.* **65,** 2910–2920.

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

FROHMAN, M. A., DUSH, M. K., and GAIL, M. R. (1988). Rapid production of full-length cDNAs from rare transcripts: Amplification using single gene specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85,** 8998–9002.

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

GORBALENYA, A. E., and KOONIN, E. V. (1989). Viral proteins contain the purine NTP-binding sequence pattern. *Nucleic Acids Res.* **17,** 8413–8440.

GORBALENYA, A. E., KOONIN, E. V., DONCHENKO, A. P., and BLINOV, V. M. (1989). Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. *Nucleic Acids Res.* **17,** 4847–4861.

GUBLER, U., and HOFFMANN, B. J. (1983). A simple method 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., and TANNOCK, G. A. (1988). Coronaviridae: The coronaviruses. *In* "Viral, Rickettsial and Chlamydial Diseases: Laboratory Diagnosis of Infectious Diseases. Principles and practices" (E. H. Lennette, F. Halonen, and F. A. Murphy, Eds.), Vol. 2, pp 451–483. Springer Verlag, New York.

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

HOLMES, K. V. (1990). Coronaviridae and their replication. *In* "Virology," 2nd edition. Edited by (B. N. Fields, D. M. Knipe *et al.*, Eds.), 2nd ed., pp 841–856, Raven Press: New York.

ISAACS, D., FLOWERS, D., CLARKE, J. R., VALMAN, H. B., and MACNAUGHTON, M. R. (1983). Epidemiology of coronavirus respiratory infections. *Arch. Dis. Child.* **58,** 500–503.

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

JOUVENNE, P., MOUNIR, S., STEWART, J. N., RICHARDSON, C. D., and TALBOT, P. J. (1992). Sequence analysis of human coronavirus 229E mRNAs 4 and 5: Evidence for polymorphism and homology with myelin basic protein. *Virus Res.* **22,** 125–141.

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

KECK, J. G., STOHLMAN, S. A., SOE, L. H., MAKINO, S., and LAI, M. M. C. (1987). Multiple recombination sites at the 5'-end of murine coronavirus RNA. *Virology* **156,** 331–341.

KEMP, M. C., HIERHOLZER, J. C., HARRISON, A., and BURKS, J. S. (1984). Characterization of viral proteins synthesized in 229E infected cells and effect(s) of inhibition of glycosylation and glycoprotein transport. *Adv. Exp. Med. Biol.* **173,** 65–77.

KOONIN, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive stranded viruses. *J. Gen. Virol.* **72,** 2197–2206.

KUO, L., HARTY, J. T., ERICKSON, L., PALMER, G. A., and PLAGEMANN, P. G. W. (1991). A nested set of eight RNAs is formed in macrophages infected with lactate dehydrogenase-elevating virus. *J. Virol.* **65,** 5118–5123.

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

LEE, H. J., SHIEH, C. K., GORBALENYA, A. E., KOONIN, E. V., LA, M. N., TULER, J., BAGDZHADZHYAN, A., and LAI, M. M. C. (1991). The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. *Virology* **180,** 567–582.

LEIBOWITZ, J. L., DEVRIES, J. R., and HASPEL, M. D. (1982). Genetic analysis of murine hepatitis virus strain JHM. *J. Virol.* **42,** 1080–1087.

LIU, D. X., and INGLIS, S. C. (1991). Association of the infectious bronchitis virus 3c protein with the virion envelope. *Virology* **185,** 911–917.

MACNAUGHTON, M. R., and MADGE, M. H. (1978). The genome of human coronavirus strain 229E. *J. Gen. Virol.* **39,** 497–504.

MACNAUGHTON, M. R., FLOWERS, D., and ISAACS, D. (1983). Diagnosis of human coronavirus infections in children using enzyme-linked immunosorbent assay. *J. Med. Virol.* **11,** 319–325.

MATSUMOTO, I., and KAWANA, R. (1992). Virological surveillance of acute respiratory tract illnesses of children in Morioka, Japan. III. Human respiratory coronavirus. *Kansenshogaku Zasshi* **66,** 319–326.

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

MELTON, D. A., KRIEG, P. A., REBAGLIATI, M. R., MANIATIS, T., ZINN, K., and GREEN, M. R. (1984). Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promotor. *Nucleic Acids Res.* **12,** 7035–7056.

MERCHLINSKY, M., and MOSS, B. (1992). Introduction of foreign DNA into the vaccinia virus genome by *in vitro* ligation: Recombination-independent selectable cloning vectors. *Virology* **190,** 522–526.

MYINT, S., HARMSEN, D., RAABE, T., and SIDDELL, S. G. (1990). Characterization of a nucleic acid probe for the diagnosis of human coronavirus 229E infections. *J. Med. Virol.* **31,** 165–172.

PATTEMORE, P. K., JOHNSTON, S. L., and BARDIN, P. G. (1992). Viruses as precipitants of asthma symptoms. I. Epidemiology. *Clin. Exp. Allergy* **22,** 325–336.

PRÜFER, D., TACKE, E., SCHMITZ, J., KULL, B., KAUFMANN, A., and ROHDE, W. (1992). Ribosomal frameshifting in plants: A novel signal directs the −1 ribosomal frameshift in the synthesis of the putative viral replicase of potato leafroll luteovirus. *EMBO J.* **11,** 1111–1117.

RAABE, T., and SIDDELL, S. (1989a). Nucleotide sequence of the hu-

man coronavirus HCV 229E mRNA 4 and mRNA 5 unique regions. *Nucleic Acids Res.* **17,** 6387.

RAABE, T., and SIDDELL, S. G. (1989b). Nucleotide sequence of the gene encoding the membrane protein of human coronavirus 229E. *Arch. Virol.* **107,** 323–328.

RAABE, T., SCHELLE-PRINZ, B., and SIDDELL, S. G. (1990). Nucleotide sequence of the gene encoding the spike glycoprotein of human coronavirus HCV 229E. *J. Gen. Virol.* **71,** 1065–1073.

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

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

SCHAAD, M. C., STOHLMAN, S. A., EGBERT, J., LUM, K., FU, K., WEI, T., JR., and BARIC, R. S. (1990). Genetics of mouse hepatitis virus transcription: Identification of cistrons which may function in positive and negative strand RNA synthesis. *Virology* **177,** 634–645.

SCHMIDT, O. W. (1984). Antigenic characterization of human coronaviruses 229E and OC43 by enzyme-linked immunosorbent assay. *J. Clin. Microbiol.* **20,** 175–180.

SCHMIDT, O. W., and KENNY, G. E. (1982). Polypeptides and functions of antigens from human coronavirus 229E and OC43. *Infect. Immun.* **35,** 515–522.

SCHREIBER, S. S., KAMAHORA, T., and 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., and 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. *J. Gen. Virol.* **64,** 113–125.

SNIJDER, E. J., DEN BOON, J. A., BREDENBEEK, P. J., HORZINEK, M. C., RIJNBRAND, R., and SPAAN, W. J. M. (1990). The carboxy-terminal part of the putative Berne virus polymerase is expressed by ribosomal frameshifting and contains sequence motifs which indicate that toro- and coronaviruses are evolutionarily related. *Nucleic Acids Res.* **18,** 4535–4542.

SOE, L. H., SHIEH, C. K., BAKER, S. C., CHANG, M. F., and LAI, M. M. C. (1987). Sequence and translation of the murine coronavirus 5′-end genomic RNA reveals the N-terminal structure of the putative RNA polymerase. *J. Virol.* **61,** 3968–3976.

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.

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

YEAGER, C. L., ASHMUN, R. A., WILLIAMS, R. K., CARDELLICHIO, C. B., SHAPIRO, L. H., LOOK, A. T., and HOLMES, K. V. (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357,** 420–422.