VIROLOGY 169, 142–151 (1989)

Sequence Analysis of the Nucleocapsid Protein Gene of Human Coronavirus 229E

STEVEN S. SCHREIBER,[1] TOSHIO KAMAHORA,[2] AND MICHAEL M. C. LAI

Departments of Neurology and Microbiology, University of Southern California, School of Medicine, 2025 Zonal Avenue, Los Angeles, California 90033

Received August 24, 1988; accepted November 2, 1988

Human coronaviruses are important human pathogens and have also been implicated in multiple sclerosis. To further understand the molecular biology of human coronavirus 229E (HCV-229E), molecular cloning and sequence analysis of the viral RNA have been initiated. Following established protocols, the 3′-terminal 1732 nucleotides of the genome were sequenced. A large open reading frame encodes a 389 amino acid protein of 43,366 Da, which is presumably the nucleocapsid protein. The predicted protein is similar in size, chemical properties, and amino acid sequence to the nucleocapsid proteins of other coronaviruses. This is especially evident when the sequence is compared with that of the antigenically related porcine transmissible gastroenteritis virus (TGEV), with which a region of 46% amino acid sequence homology was found. Hydropathy profiles revealed the existence of several conserved domains which could have functional significance. An intergenic consensus sequence precedes the 5′-end of the proposed nucleocapsid protein gene. The consensus sequence is present in other coronaviruses and has been proposed as the site of binding of the leader sequence for mRNA transcriptional start. This region was also examined by primer extension analysis of mRNAs, which identified a 60-nucleotide leader sequence. The 3′-noncoding region of the genome contains an 11-nucleotide sequence, which is relatively conserved throughout the Coronavirus family and lends support to the theory that this region is important for the replication of negative-strand RNA.

INTRODUCTION

Human coronavirus 229E (HCV-229E) belongs to one of two major antigenic groups of human coronaviruses (MacNaughton, 1981). It shares antigenic relationships with other coronaviruses, such as porcine transmissible gastroenteritis virus (TGEV), feline infectious peritonitis virus (FIPV), and canine coronavirus (CCV). The other well-characterized human coronavirus, HCV-OC43, is in a separate antigenic group which includes mouse hepatitis virus (MHV) and bovine coronavirus (BCV). Both human coronaviruses are mainly respiratory pathogens and have been estimated to cause up to 25% of common colds (McIntosh *et al.*, 1974; Wege *et al.*, 1982). They have also been implicated in gastrointestinal diseases (Resta *et al.*, 1985). Furthermore, the isolation of coronaviruses bearing an antigenic relationship to HCV-OC43 from the central nervous system of two patients with multiple sclerosis has suggested a possible etiologic relationship between human coronaviruses and multiple sclerosis (Burks *et al.*, 1980). This possibility is supported by the observation that neurotropic strains of MHV cause demyelination in the central nervous system of rodents (Weiner and Stohlman, 1978). Thus, human coronaviruses are important human pathogens.

The structural and biochemical properties of several coronaviruses, particularly MHV and avian infectious peritonitis virus (IBV), have been well characterized (Lai *et al.*, 1987; Boursnell *et al.*, 1987). The virion contains a single-stranded, positive-sense RNA molecule (molecular weight $6–8 \times 10^6$ Da) (Lai and Stohlman, 1978) associated in a helical conformation with nucleocapsid proteins (N). The viral nucleocapsid is enclosed by an envelope, in which are embedded at least two types of viral proteins, the peplomer (E2) and matrix (E1) glycoproteins. Coronavirus RNA replication occurs in the cytoplasm of infected cells and is mediated by a virus-encoded RNA-dependent RNA polymerase (Brayton *et al.*, 1982). The virus-specific mRNA in infected cells comprises a genomic-sized RNA plus six subgenomic mRNA species. These mRNAs are arranged in a nested-set structure, which is characterized by RNAs having common 3′-termini but extending for varying lengths in the 5′ direction (Lai *et al.*, 1981). Only the 5′-proximal regions of each mRNA are translated (Rottier *et al.*, 1981). A unique feature of the structure of coronavirus is the existence, at the 5′-end of each mRNA, of an identical leader sequence. This sequence is derived from the 5′-end of the genomic RNA and is of approximately 70 nucleotides in length (Lai *et al.*, 1983, 1984). Recent evidence has supported a role for the leader sequence in mediating a novel type of discontinuous transcription of genomic RNA (Baric *et al.*, 1985; Makino *et al.*, 1986; Shieh *et al.*, 1987).

In contrast to other coronaviruses, the molecular biology of human coronaviruses is relatively poorly understood. The genomic RNA of both HCV-229E and HCV-OC43 has a molecular weight of approximately 6

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. J04419.

[1] To whom requests for reprints should be addressed.

[2] Present address: Department of Virology, Tottori University, School of Medicine, Yonago 683, Japan.

0042-6822/89 $3.00

$\times 10^6$ Da (Hierholzer *et al.*, 1981). The six subgenomic RNA species appear to have lower molecular weights than those of the corresponding MHV RNAs (Weiss and Leibowitz, 1981). The structure of these mRNAs is not yet known. Analysis of purified HCV-229E virions has revealed three major polypeptides: a glycosylated protein with a molecular weight of 180 kDa, a phosphorylated nucleocapsid protein of 50 kDa, and a family of polypeptides with molecular weights of 25, 23, and 21 kDa (Kemp *et al.*, 1984). In addition, several minor nonstructural polypeptides of 107, 92, and 39 kDa have been identified (Kemp *et al.*, 1984). The functions of these proteins have not yet been characterized.

To further understand the molecular biology of HCV-229E, we have initiated molecular cloning and sequence analysis of HCV-229E RNA. In this paper we report the sequence analysis of the gene encoding the nucleocapsid protein of HCV-229E. In addition, the mRNA leader sequence was also identified. The results are compared with sequences of other coronaviruses including MHV, BCV, IBV, and TGEV.

MATERIALS AND METHODS

Virus and cells

HCV-229E (obtained from Dr. J. Fleming, University of Southern California) was propagated at low multiplicities of infection in human fetal lung cells L132 (Kennedy and Johnson-Lussenberg, 1975/1976), using Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum.

Virus purification and preparation of virion RNA

Following a virus adsorption period of 1 hr at 37°, HCV-229E-infected L132 monolayers were incubated at 37° for 24 to 48 hr, at which time the cell culture fluid was harvested. Viruses were precipitated from 2 liters of culture fluid with 50% ammonium sulfate and centrifuged at 8000 rpm for 30 min. The pellet was resuspended in NTE buffer (0.1 *M* NaCl, 0.01 *M* Tris–hydrochloride (pH 7.2), 1 m*M* EDTA) and then placed on a discontinuous sucrose gradient consisting of 60, 50, 30, and 20% (w/w) sucrose in NTE buffer and centrifuged at 26,000 rpm for 13 hr at 4° in a Beckman SW28.1 rotor. The virus band at the interface between 50 and 30% sucrose was collected and diluted threefold with NTE buffer. The diluted virus suspension was centrifuged on a linear sucrose gradient at 26,000 rpm in an SW28.1 rotor for 4 hr at 4°. The virus band was collected and treated with proteinase K (0.2 mg/ml) for 20 min at 37°, followed by 1% SDS for 30 min at 37°. Genomic RNA was extracted with phenol and then with phenol/chloroform, and precipitated with ethanol.

Preparation of intracellular RNA

Monolayers of L132 cells grown in 100 × 20-mm culture dishes were infected with HCV-229E. Cells were incubated in phosphate-free DMEM containing 1% dialyzed fetal calf serum 4 hr prior to RNA extraction. Actinomycin D (1 μg/ml) (Sigma) and [^{32}P]orthophosphate (70 μCi/ml) (ICN Radiochemicals) were added at 3 and 2 hr, respectively, prior to RNA extraction at 15 hr postinfection (p.i.). Cells were collected in cold phosphate-buffered saline and centrifuged at 5000 rpm for 3 min at 4°. The pellet was mixed with cold 0.5% Nonidet-P40 in NTE buffer, incubated for 10 min at 4°, and then centrifuged at 5000 rpm for 3 min. The supernatant was transferred to a fresh tube containing 1/10 vol of 10% SDS at room temperature and vortexed briefly. Intracellular RNA was extracted with phenol and phenol/chloroform and precipitated with ethanol. Poly(A)-containing RNA was selected by oligo(dT)-cellulose chromatography as previously described (Makino *et al.*, 1984).

To examine the kinetics of viral mRNA synthesis, intracellular RNA was extracted from virus-infected L132 monolayers in 60 × 15-mm culture dishes at 7, 21, 29, 46, and 58 hr postinfection.

cDNA cloning

cDNA cloning was performed using a modified method of Gubler and Hoffman (1983). The poly(A)-containing RNA extracted from 229E-infected L132 monolayers was precipitated, dried, and resuspended in 6.72 μl of autoclaved water. The RNA was incubated with 10 m*M* methylmercuric hydroxide in an 8 μl total volume for 10 min at room temperature. First-strand cDNA synthesis was carried out in a 50-μl reaction mixture containing 60 units RNasin (Promega Biotec), 10 m*M* $MgCl_2$, 100 m*M* KCl, 50 m*M* Tris–HCl (pH 8.3 at 42°), 10 m*M* DTT, 1.25 m*M* dNTPs, 40 μCi [α-^{32}P]dATP (3000 Ci/mmol), 28 m*M* β-mercaptoethanol, and 10 ng oligo$(dT)_{12-18}$ primer. After 5 min at room temperature, 40 units of AMV reverse transcriptase (Life Science) was added and the mixture was incubated for 1 hr at 42°. The reaction was stopped by adding 4.4 μl of 250 m*M* EDTA. The products were extracted with phenol/chloroform and precipitated with ethanol containing 0.3 *M* ammonium acetate. For second-strand synthesis, the 100-μl reaction mixture contained 5 m*M* $MgCl_2$, 100 m*M* KCl, 20 m*M* Tris–HCl (pH 7.5), 50 μg/ml bovine serum albumin (BSA), 10 m*M* ammonium sulfate, 0.15 m*M* β-NAD, 100 μ*M* dNTPs, 25 units of *Escherichia coli* DNA polymerase I, 2 units of *E. coli* DNA ligase, and 0.8 units of RNase H. Sequential incubations were for 1 hr at 12° and 1 hr at 22°. The reaction was stopped by the addition of 8.7 μl of 250 m*M* EDTA and the products were extracted with phenol/

chloroform and precipitated with ethanol in the presence of 0.3 *M* ammonium acetate. Homopolymeric tailing of double-stranded cDNA with poly(C) was carried out in a 12-μl reaction mixture containing 10 units of terminal transferase, 200 m*M* potassium cacodylate, 0.5 m*M* $CoCl_2$, 25 m*M* Tris–HCl (pH 6.9), 2 m*M* DTT, 250 μg/ml BSA, and 50 μM dCTP at 37° for 4 min. The dC-tailed double-stranded DNA was annealed to 200 μg of dG-tailed *Pst*I-cut pBR322 plasmid in 20 μl of a buffer containing 10 m*M* Tris–HCl (pH 7.4), 100 m*M* NaCl, and 0.25 m*M* EDTA. The mixture was incubated for 5 min at 68° and then cooled slowly overnight. The annealed molecules were used to transform *E. coli* MCI061 as described (Dagert and Erhlich, 1979).

Colony hybridization

Colonies grown on LB/tetracycline plates were incubated at 37° for 12 hr and transferred to Colony/Plaque Screen disks (New England Nuclear). Bacterial lysis and DNA fixation were carried out according to the methods previously described (Grunstein and Hogness, 1975). The disks were prehybridized in a solution containing 0.2% polyvinylpyrrolidone (MW 40,000), 0.2% Ficoll (MW 400,000), 0.2% BSA, 0.05 *M* Tris–HCl (pH 7.5), 1% SDS, 1 *M* NaCl, 10% dextran sulfate, and 100 μg/ml denatured salmon sperm DNA at 65° for 6-hr. Fragments derived from either the 5′- or 3′-ends of gene 7 were labeled with ^{32}P by nick-translation and added to the solution. Hybridization was carried out for 20 hr at 65°. The disks were then washed twice in 2× SSC (0.3 *M* NaCl, 30 m*M* sodium citrate) at room temperature, twice in 2× SSC containing 1% SDS for 30 min at 65°, and twice in 0.1× SSC at room temperature for 30 min. The disks were air-dried and exposed to X-ray film at −70°.

Northern hybridization

Intracellular RNA from virus-infected cells was denatured by glyoxal treatment and separated by electrophoresis on a 1% agarose gel containing 10 m*M* sodium phosphate (pH 7.0) as described previously (McMaster and Carmichael, 1977). RNA transfer to Biodyne nylon filters (ICN Radiochemicals) and subsequent hybridization were performed according to the method described by Thomas (1980).

Primer extension

A synthetic oligodeoxyribonucleotide was 5′-end-labeled with [γ-^{32}P]ATP by polynucleotide kinase (Pedersen and Haseltine, 1980). The total amount of poly(A)-containing RNA extracted from 229E-infected cell monolayers in three 150 × 20-mm culture dishes was incubated in 8 μl of distilled water containing 10 m*M* methylmercuric hydroxide for 10 min at room temperature. A further incubation was carried out in a 50-μl reaction volume containing 60 units of RNasin (Promega), 10 m*M* $MgCl_2$, 100 m*M* KCl, 50 m*M* Tris–HCl (pH 8.3 at 42°), 10 m*M* DTT, 1.25 m*M* dNTPs, 28 m*M* β-mercaptoethanol, 5′-end-labeled synthetic oligodeoxyribonucleotides, and 20 units of AMV reverse transcriptase (Life Science) for 1 hr at 42°. Reaction products were extracted with phenol/chloroform, precipitated with ethanol, and then analyzed by electrophoresis on a 6% polyacrylamide gel containing 8.3 *M* urea. The primer-extended product was identified by autoradiography and eluted from the gel according to the published procedure (Maxam and Gilbert, 1977).

DNA sequencing

Sequencing was carried out by the dideoxyribonucleotide chain termination method (Sanger *et al.*, 1977) as well as the chemical modification procedure (Maxam and Gilbert, 1977). In the first method, fragments of cDNA inserts generated by various restriction endonucleases were cloned into the M13 vectors mp18 and mp19 (Messing and Vierira, 1982). [α-^{35}S]-dATP was used as a label. Sequence data were also obtained by chemical modification (Maxam and Gilbert, 1977) of various cDNA fragments subcloned into the pT7-3 vector (Tabor and Richardson, 1985). In the second method, cDNA fragments were 3′-end-labeled with Klenow fragment at internal restriction sites or, alternatively, at the polylinker cloning site of pT7-3. End-labeled cDNA restriction fragments were separated by electrophoresis on preparative polyacrylamide gels (Maxam and Gilbert, 1980) and purified as described previously (Hansen *et al.*, 1980; Hansen, 1981). Sequencing of the primer-extended product of mRNA7 was performed by the chemical modification procedure (Maxam and Gilbert, 1977). Sequence analysis was performed by the Intelligenetics and Seqaid programs. Hydropathy profiles were constructed using the PepPlot program of the University of Wisconsin Computer Genetics Group, which employs both the Kyle–Doolittle (KD) and Goldman, Engelman, Steitz (GES) algorithms.

RESULTS

Kinetics of HCV-229E mRNA synthesis

To determine the optimum time for extracting 229E-specific mRNAs, we first studied the kinetics of virus-specific mRNA synthesis. Intracellular RNA was extracted from infected L132 monolayers at specified times p.i. The RNA was separated by agarose gel electrophoresis (Fig. 1). As can be seen, viral mRNA synthesis could be detected as early as 7 hr p.i. and reached maximum at 29 hr p.i. Thereafter, total RNA

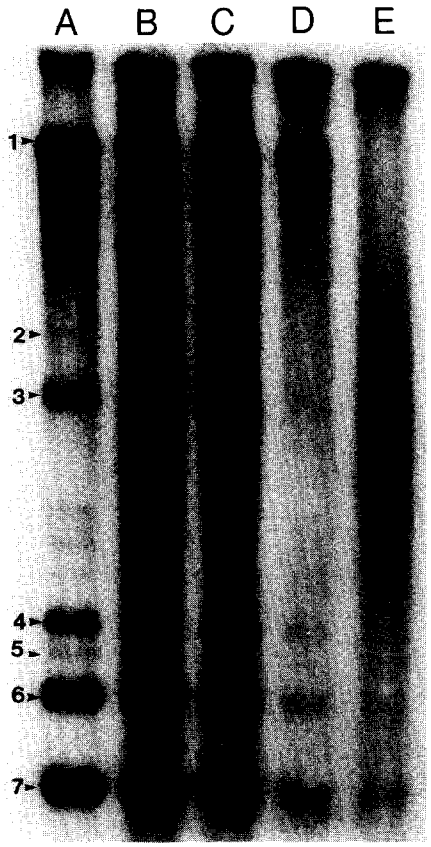

FIG. 1. Kinetics of synthesis of HCV-229E-specific RNAs. Intracellular RNA from HCV-229E-infected L132 cell monolayers was labeled with [^{32}P]orthophosphate, extracted, and separated by agarose gel electrophoresis as described under Materials and Methods. RNA was extracted at the following times: lane A, 7 hr p.i.; lane B, 21 hr p.i.; lane C, 29 hr p.i.; lane D, 46 hr p.i.; lane E, 58 hr p.i. The positions and designations of HCV-229E-specific RNAs are indicated by the numbers on the left side of the figure.

synthesis gradually declined. By 46 hr p.i. only the most abundant mRNA species were evident. The number and size of these mRNA species are comparable to those of MHV mRNAs and are in agreement with previously published results (Weiss and Leibowitz, 1981). Significantly, mRNA 2a, which was previously found only in BCV-infected cells and proposed to encode hemagglutinins (King *et al.*, 1985; Keck *et al.*, 1988), was not present. This is consistent with the finding that HCV-229E does not have hemagglutinating activity (Hierholzer, 1976). The relative amounts of the mRNA species were the same throughout the replication cycle. Therefore, in all of our subsequent experiments, the virus-specific intracellular RNAs were extracted at 15 hr p.i.

Molecular cloning of HCV-229E genomic RNA and intracellular virus-specific mRNAs

cDNA cloning was initially performed using virion genomic RNA as a template. The sizes of inserts in the resultant cDNA clones ranged from 0.2 to 0.5 kb in length. One clone, A34, contained a 0.45-kb insert, which was subsequently characterized by restriction mapping and Northern blot analysis. The 0.45-kb fragment was labeled with ^{32}P by nick-translation and hybridized with intracellular RNA from 229E-infected cells. The result, shown in Fig. 2, revealed that the fragment hybridized to each of the mRNA species. This result suggested that the HCV-229E subgenomic mRNAs possess a nested-set structure similar to other coronaviruses (Lai, 1988) and that A34 represented a cDNA clone of either the 3′-end of the genomic RNA or the leader sequence.

Cloning was subsequently carried out using intracellular RNA from 229E-infected cells as a template. The

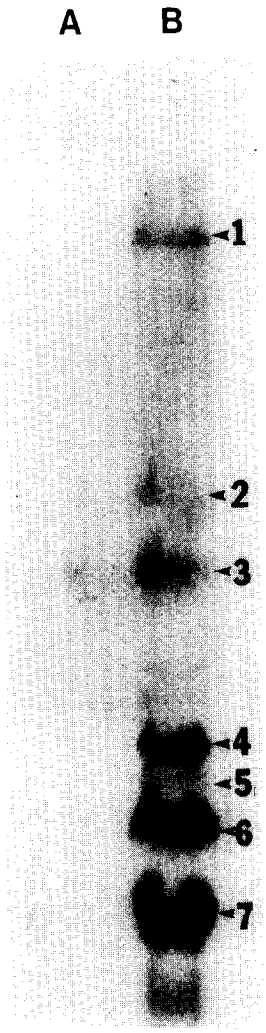

FIG. 2. Northern blot analysis of HCV-229E-specific intracellular RNA hybridized with clone A34. Intracellular RNA from either uninfected (lane A) or HCV-229E-infected (lane B) L132 monolayers was denatured by glyoxal treatment, separated on a 1% agarose gel, and transferred to Biodyne nylon filters as described under Materials and Methods. The positions and designations of the HCV-229E-specific RNAs are indicated by the numbers on the right side of the figure.

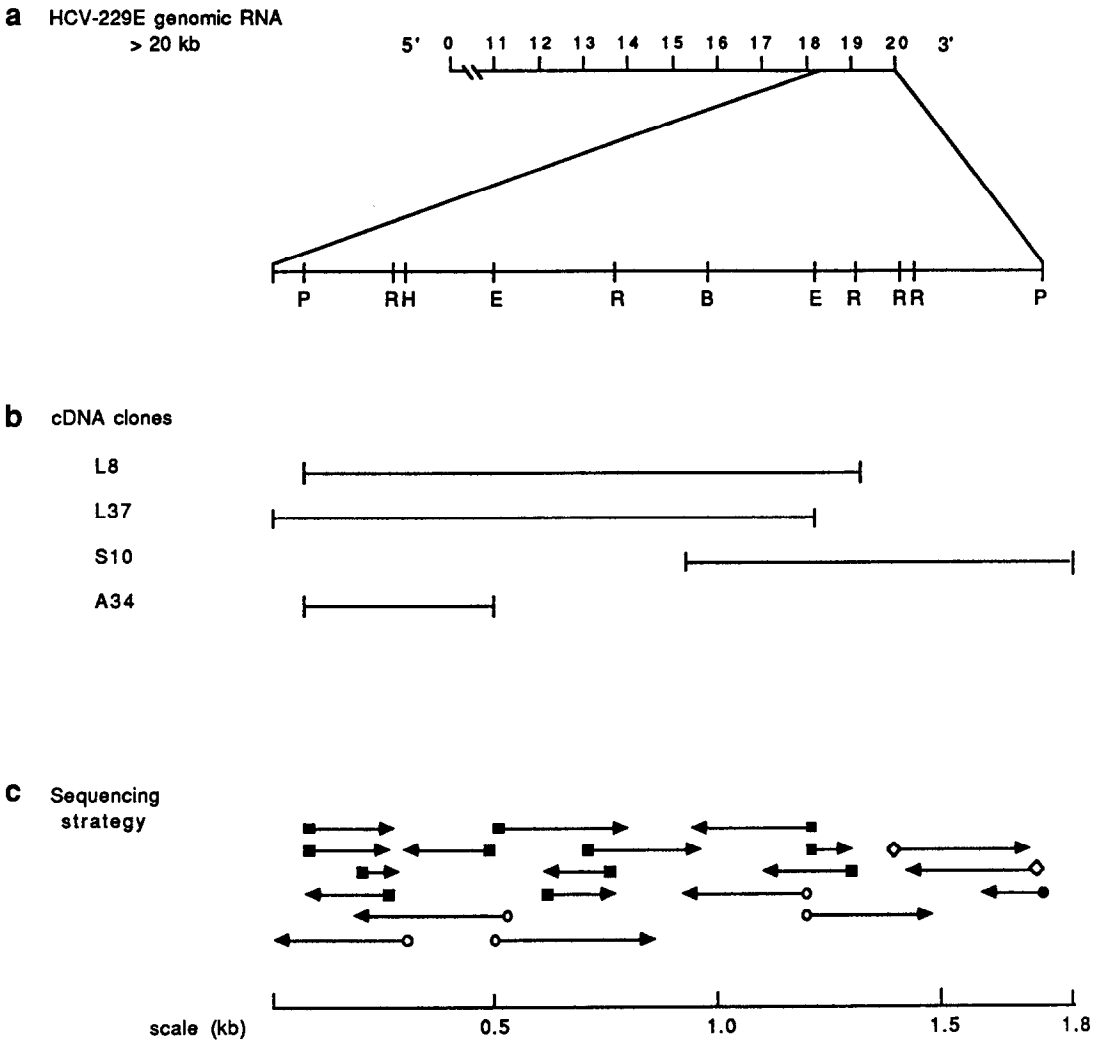

FIG. 3. Diagram of the 3'-end of the HCV-229E genome, including cDNA clones and sequencing strategy. (a) Restriction map of the 3'-end cDNA clones with reference to the entire viral genome. (b) Relative position and size of the cDNA clones which were sequenced. Clones L8 and L37 are shown in part. Clone A34 was used in colony hybridization studies. (c) Direction and extent of sequence data obtained from subcloned fragments. Arrows with solid squares indicate dideoxy sequencing method. Arrows with open circles indicate chemical modification sequencing method. Arrows with open diamonds indicate sequencing, by chemical modification, of fragments subcloned into the *PstI–SmaI* sites of pT7-3 and 3'-end-labeled with Klenow fragment. The arrow with a closed circle indicates a DNA fragment which was labeled at the 3'-end using radiolabeled cordycepin and terminal transferase. Abbreviations: B, *BalI*; E, *Eco*RI; H, *Hind*III; P, *PstI*; R, *RsaI*.

resulting cDNA clones were screened by colony hybridization using the 0.45-kb fragment from clone A34 as a nick-translated probe (Fig. 3). Several positive colonies were identified and characterized further. Clone L8 contained a 3.6-kb insert but lacked a 3'-poly(A) tail. Clone L37, which contained an insert of 1.7 kb, overlapped L8 but was 0.1 kb shorter at the 3'-end. This clone also lacked a poly(A) sequence (see below). Therefore, additional cDNA clones were isolated using a 0.24-kb *Bal* I–*Eco*RI fragment of L8 (Fig. 3a) as a probe. These latter clones were further characterized by Southern blot analysis. Clone S10 contained an insert of 0.8 kb which overlapped the 3'-ends of the two previous clones and extended another 0.4 kb in that direction. Figure 3b shows the orientation and sizes of clones L8, L37, S10, and A34 with reference to the viral genome. Restriction enzyme sites used for sequencing are also shown.

Sequencing of the cDNA clones

To determine the sequence of the 3'-end of HCV-229E genome, various restriction fragments of L8, L37, and S10 were subcloned into M13 vectors. For L8, only the 1.2-kb fragment extending from an internal *PstI* site toward the 3'-end was sequenced. Clone L37 was also sequenced in part. Figure 3c shows the cDNA fragments and strategy used in sequencing. Each region

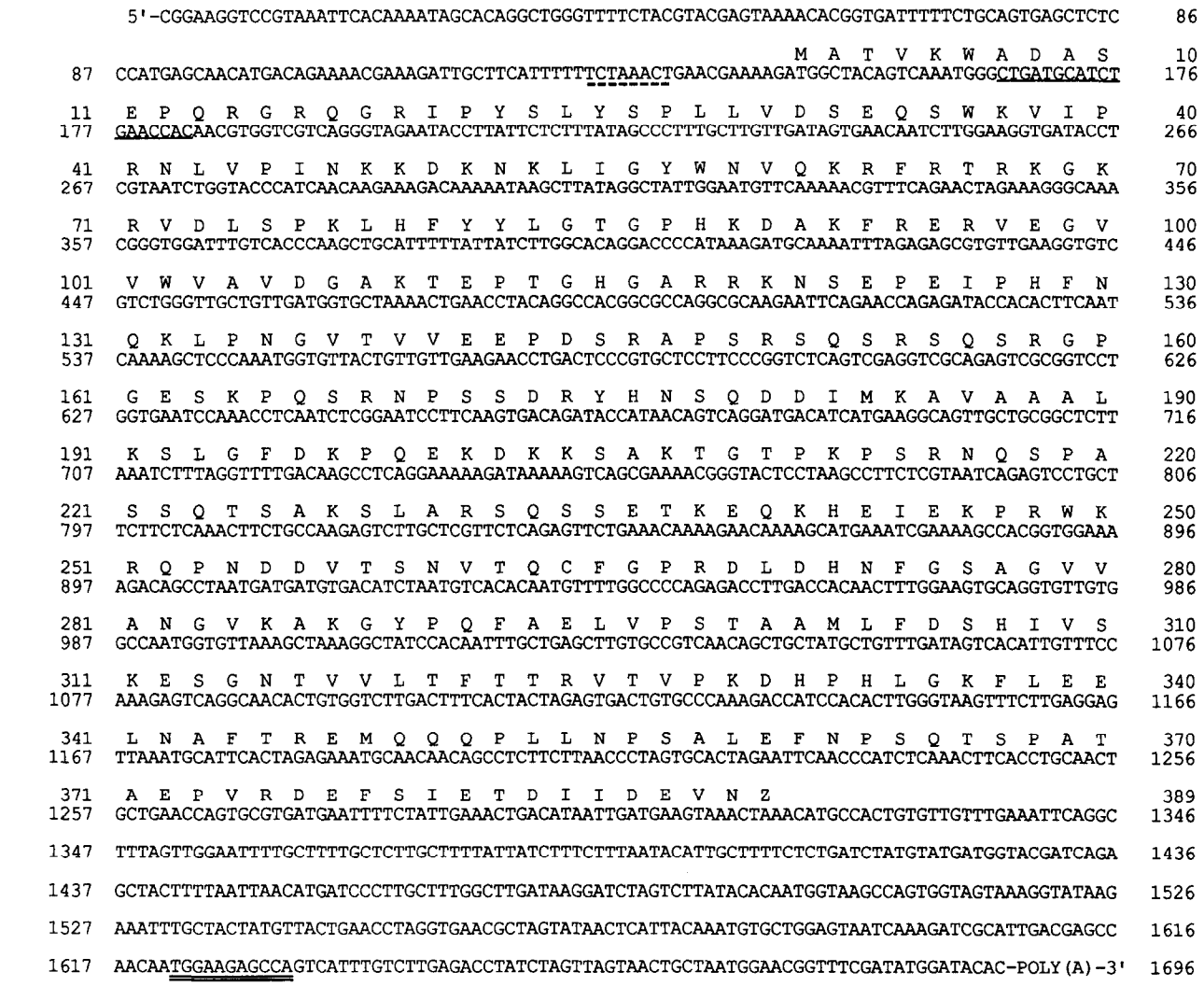

FIG. 4. The primary nucleotide sequence of the 3'-end of HCV-229E RNA and the deduced amino acid sequence of the nucleocapsid protein. A primer extension study was carried out using a synthetic oligodeoxyribonucleotide complementary to an 18-mer sequence underlined near the 5'-end of the gene. The 3'-noncoding region contains a conserved sequence which is shown by the double line. The intergenic conserved sequence, TCTAAACT, is also shown (dotted line).

was verified by dideoxy chain termination sequencing of both strands or by the chemical modification method. Clone S10 was found to have a poly(A) stretch of 34 bases. Figure 4 shows the complete DNA sequence with a translation of the main open reading frame (ORF) in one-letter amino acid code. This ORF extends from base 147 to base 1313 and predicts a 389 amino acid protein with a molecular weight of 43,366 Da. This predicted molecular weight is slightly smaller than the measured molecular weight of the nucleocapsid protein of HCV-229E, which is 50 kDa as determined by SDS–polyacrylamide gel electrophoresis (MacNaughton, 1980). The difference is probably due to phosphorylation or other modification of the protein. The predicted protein shares features with the nucleocapsid proteins of TGEV, MHV, BCV, HCV-OC43, and IBV (Kapke and Brian, 1986; Skinner and Siddell, 1984; Armstrong *et al.*, 1983; Lapps *et al.*, 1987; Kamahora *et al.*, 1988; Boursnell *et al.*, 1985). Namely, the protein is highly basic and rich in serine residues. Sixty percent of the amino acid residues are basic and 12% are acidic. There are 39 serine residues (10% of total), which are presumed to be sites of phosphorylation (Stohlman and Lai, 1979). When compared to TGEV, with which HCV-229E shares antigenic properties, both N proteins have identical amounts of basic and acidic amino acids and serine residues and similar molecular weights (Kapke and Brian, 1986).

Figure 5 shows a schematic diagram of the possible ORFs obtained by translating the nucleotide sequence. The ORF in frame 3 is likely the one which encodes the nucleocapsid protein. In frame 2, the 5'-flanking region probably contains part of the sequence of the matrix protein encoded by gene 6. This possibility is sup-

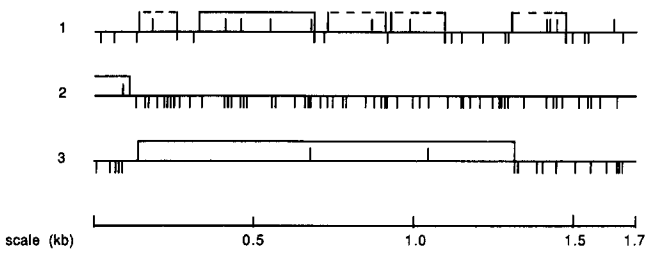

FIG. 5. Schematic diagram of the possible open reading frames obtained when translating the primary nucleotide sequence. Vertical lines above the baseline represent potential initiation codons. Termination codons are indicated by vertical lines below the baseline. Frame 3 depicts a single, long ORF encoding the nucleocapsid protein. ORFs which are greater than 30 amino acids are also shown. Those lacking translation start sites are indicated by dashed lines.

ported by the finding that reading frame 2 remains open at the extreme 5'-end. Furthermore, the sequence TCTAAACT, which is found in the intergenic regions of several other coronaviruses (Kapke and Brian, 1986; Skinner and Siddell, 1984; Armstrong *et al.*, 1983; Lapps *et al.*, 1987; Kamahora *et al.*, 1988; Budzilowicz *et al.*, 1985), is also present between the presumed initiation codon of the main ORF and the 3'-end of gene 6. This sequence is the proposed site of fusion of the leader sequence with the mRNA coding region (Shieh *et al.*, 1987; Makino *et al.*, 1986; Budzilowicz *et al.*, 1985).

The 3'-noncoding region contains the sequence TGGAAGAGCCA, 75 nucleotides from the 3'-end (Fig. 4), which is relatively conserved among coronaviruses and is found at approximately the same location in all of these viral genomes (Kapke and Brian, 1986; Skinner and Siddell, 1984; Armstrong *et al.*, 1983; Lapps *et al.*, 1987; Kamahora *et al.*, 1988; Boursnell *et al.*, 1985) (Table 1). There is only one nucleotide difference in this conserved sequence when it is compared with that of TGEV, BCV, and HCV-OC43. Two and three nucleotide differences are found in IBV and MHV, respectively. This conservation of sequence and location suggests that it may be important for viral RNA replication.

In frame 1, there are several additional ORFs of at least 30 amino acids. Some of these, including one found in the 3'-noncoding region, lack appropriate translation start sites. Another long internal ORF is found from base 322 through 693. This contains an appropriate initiation sequence and encodes a hypothetical protein of 13,974 Da, which is rich in leucine residues (17%). The significance of this ORF remains to be defined.

Leader sequence of HCV-229E

The mRNAs of coronaviruses contain a stretch of leader sequence which is derived from the 5'-end of the viral genome and exhibits homology with the intergenic consensus sequence (Shieh *et al.*, 1987; Budzilowicz *et al.*, 1985). Since our cDNA clones did not appear to contain leader sequences, we used primer extension studies to determine the sequence of the HCV-229E leader RNA. A synthetic oligodeoxyribonucleotide which was complementary to an 18-mer sequence located near the 5'-end of the gene (Fig. 4) was end-labeled and used in a primer extension study with poly(A)-selected intracellular mRNA as a template. The reaction products, separated by agarose gel electrophoresis, revealed six bands (data not shown). Since these bands were most likely to represent the primer-extended products of the individual mRNA species, the smallest and most abundant band, corresponding to the primer-extended product of mRNA7, was eluted and sequenced by the chemical modification method (Maxam and Gilbert, 1977). The sequence of the 3'-end of the primer-extended product was identical to the L8 sequence from nucleotides 129 to 171. At nucleotide 128, immediately 5' to the proposed leader mRNA fusion site, the sequence diverged from the L8 sequence and revealed a putative 60-base leader sequence which is shown in Fig. 6. The figure also shows a degree of homology with the leader sequence of IBV. Considerably less homology exists between the leader sequence of HCV-229E and those of HCV-OC43 and MHV-JHM (data not shown).

DISCUSSION

This report presented the primary sequence of the nucleocapsid gene and leader sequence of HCV-229E. When compared to the known sequences of other coronaviruses (Kapke and Brian, 1986; Skinner and Siddell, 1984; Armstrong *et al.*, 1983; Lapps *et al.*, 1987; Kamahora *et al.*, 1988; Boursnell *et al.*, 1985), common features of coronavirus nucleocapsid proteins emerged; namely, they are highly basic and have a high proportion of serine residues, which have been shown

TABLE 1

CONSERVED SEQUENCE AT THE 3'-NONCODING REGION OF CORONAVIRUS

Virus	3' conserved sequence	
HCV-229E	TGGAAGAGCCA	(75)
TGEV	TGGAAGAGCTA	(76)
BCV	GGGAAGAGCCA	(79)
HCV-OC43	GGGAAGAGCCA	(79)
IBV	GGGAAGAGCTA	(81)
MHV-JHM	GGGAAGAGCTC	(82)
MHV-A59	GGGAAGAGCTC	(82)

Note. Number in parenthesis indicates distance, in nucleotides, of the conserved sequence from the poly(A) region.

```
                          10          20          30          40          50          60
                          |           |           |           |           |           |
HCV-229E    5'-CTTAAG*TACCTTAT*CTATCTA*CAAATAGAAAAG**TTGCTTTTTAGACTTTGTGTC*TA*CTTC
                 :::::: ::  :  :  ::: :: : : :: :  ::  ::::    :::: :::    : :: :
IBV         5'-ACTTAAGATAGATATTAATATATATCTATTACACTAGCCTTGC**GCTAGATTTTTAA*CTTAACAAA.....
```

FIG. 6. HCV-229E mRNA leader sequence compared to the leader sequence of IBV. The IBV leader extends for at least 16 nucleotides in the 3′ direction.

to be sites of phosphorylation (Stohlman and Lai, 1979). The relationship between the nucleocapsid genes of HCV-229E and TGEV is particularly interesting since the viruses are antigenically related (MacNaughton, 1981). The predicted molecular weights of the N protein and the number of potential phosphorylation sites of both viruses are almost identical. Although these two viruses have little nucleotide sequence homology between their nucleocapsid genes, the amino acid sequences are homologous within a limited region. Amino acid sequence analysis revealed several structural features common to both viruses, which may have functional significance. For instance, there is a region of 46% homology within the amino-terminal one-third of the protein which extends from residues 29 to 134 in HCV-229E, and 41 to 146 in TGEV. Furthermore, approximately 10 amino acids downstream from the homologous region in both proteins lies an area which is abundant in serine residues, suggesting that this may be an important functional domain of the molecule. To further examine such functional homology between the two proteins, hydropathy profiles were constructed (Fig. 7). The contour of these plots suggests that a certain degree of functional homology exists within the first and last one-third of each molecule, with an additional region around position 200.

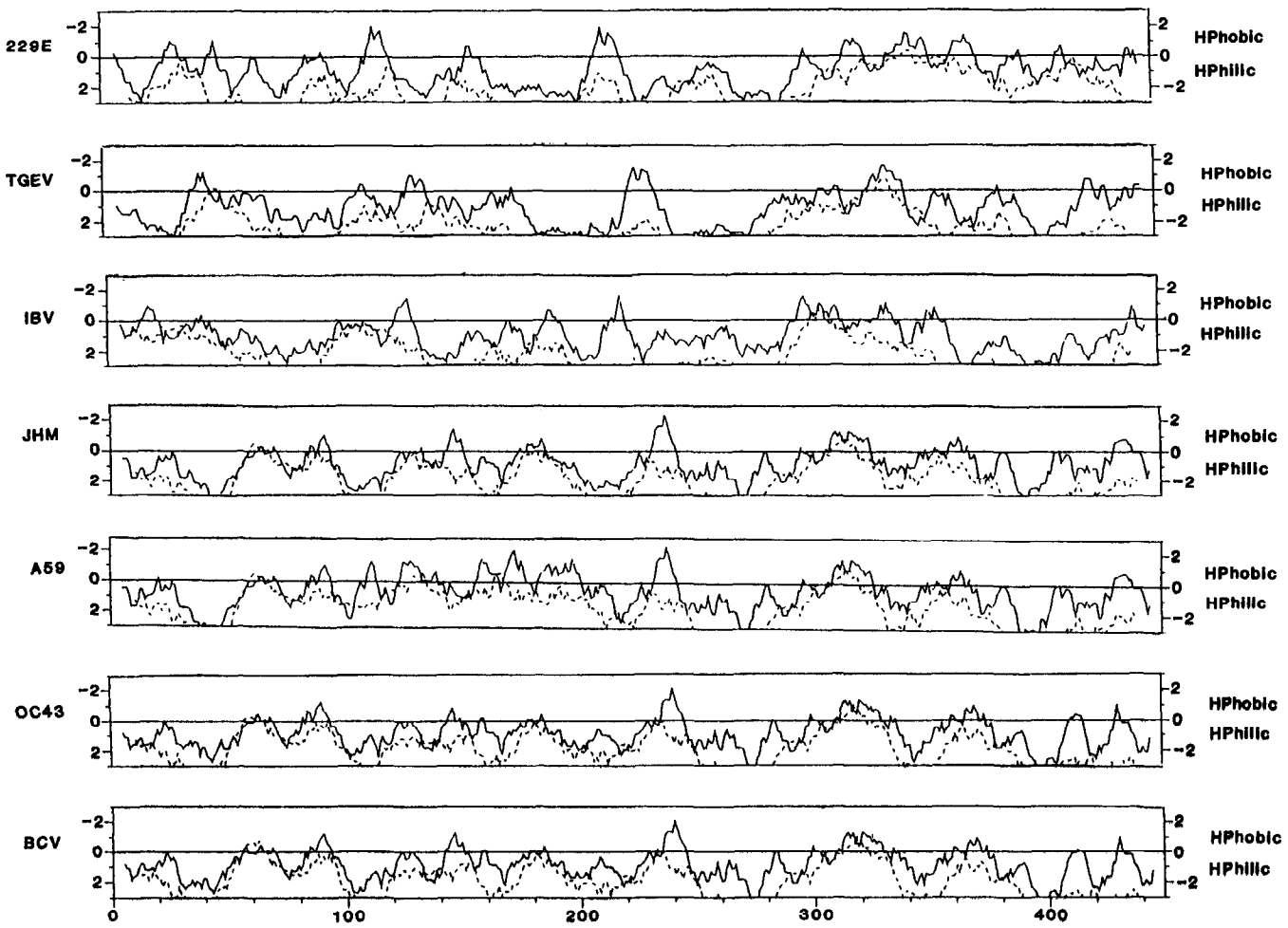

FIG. 7. Hydropathy profiles of coronavirus N proteins. Both the K-D (solid line) and GES (dashed line) curves are depicted with scales on the right and left, respectively.

The peak around position 200 occurs just after the serine-rich region of the molecule. The relative conservation of these regions suggests a possible role in the interaction of the N protein with the viral genome. Similar structural features exist among the N proteins of HCV-229E, IBV, MHV, HCV-OC43, and BCV (Skinner and Siddell, 1984; Lapps *et al.*, 1987; Kamahora *et al.*, 1988; Boursnell *et al.*, 1985). This is demonstrated by the hydropathy profiles of these proteins, which are also shown in Fig. 7. Further studies are required to reveal the functional significance of the conserved domains.

Another interesting finding is the open reading frame internal to the main coding region of the HCV-229E N gene. Thus far, two other coronaviruses, BCV and MHV-JHM, have been found to contain internal ORFs in gene 7 (Skinner and Siddell, 1984; Lapps *et al.*, 1987) which are preceded by optimum translation initiation signals according to Kozak's consensus sequence (Kozak, 1983). The predicted amino acid sequences could encode hypothetical proteins of molecular weights 13,973; 14,842; and 23,057 for HCV-229E, MHV-JHM, and BCV, respectively. Interestingly, all three sequences are abundant in leucine residues (17 to 19%). HCV-OC43 also has two smaller internal ORFs encoding potential leucine-rich proteins of 8830 and 16,297 molecular weights (Kamahora *et al.*, 1988). Further studies to determine whether this hypothetical protein can be detected in 229E-infected cells or by *in vitro* translation of a full-length cDNA clone (i.e., L8) are in progress.

Finally, the 3′-noncoding conserved sequence of gene 7 lends additional support to a common ancestry for coronaviruses, regardless of antigenic subgroup. This sequence has been proposed as a recognition site for the virus-encoded RNA-dependent RNA polymerase prior to negative-strand synthesis (Kapke and Brian, 1986). Certainly future studies must focus on examining the role of this conserved region in the viral replication cycle.

ACKNOWLEDGMENTS

We thank Carol Flores for assistance in preparation of the manuscript. This work was supported by Public Health Service Research Grants NS18146 and AI19244 from the National Institutes of Health and Grant 1449 from the National Multiple Sclerosis Society. S.S.S. is supported by a postdoctoral training fellowship from the National Institutes of Health Grant NS07149.

REFERENCES

ARMSTRONG, J., SMEEKENS, S., and ROTTIER, P. (1983). Sequence of the nucleocapsid gene from murine coronavirus MHV-A59. *Nucleic Acids Res.* **11,** 883–891.

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

BOURSNELL, M. E. G., BINNS, M. M., FOULDS, I. J., and BROWN, T. D. K. (1985). Sequences of the nucleocapsid genes from two strains of avian infectious bronchitis virus. *J. Gen. Virol.* **66,** 573–580.

BOURSNELL, M. E. G., 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,** 57–77.

BRAYTON, P. R., LAI, M. M. C., PATTON, C. D., and STOHLMAN, S. A. (1982). Characterization of two RNA polymerase activities induced by mouse hepatitis virus. *J. Virol.* **42,** 847–853.

BUDZILOWICZ, C. J., WILCZYNSKI, S. P., and WEISS, S. R. (1985). Three intergenic regions of coronavirus mouse hepatitis virus strain A59 genome RNA contain a common nucleotide sequence that is homologous to the 3′ end of the viral mRNA leader sequence. *J. Virol.* **53,** 834–840.

BURKS, J., DEVALD, B. L., JANKOVSKY, L. D., and GERDES, J. C. (1980). Two coronaviruses isolated from central nervous system tissue of two multiple sclerosis patients. *Science* **209,** 933–934.

DAGERT, M., and EHRLICH, S. D. (1979). Prolonged incubation in calcium chloride improves the competence of Escherichia coli cells. *Gene* **6,** 23–29.

GRUNSTEIN, M., and HOGNESS, D. S. (1975). Colony hybridization: A method for the isolation of cloned DNAs that contain a specific gene. *Proc. Natl. Acad. Sci. USA* **72,** 3961–3965.

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

HANSEN, J. N. (1981). Use of solubilizable acrylamide disulfide gels for isolation of DNA fragments suitable for sequence analysis. *Anal. Biochem.* **116,** 146–151.

HANSEN, J. N., PHEIFFER, B. H., and BOEHNERT, J. A. (1980). Chemical and electrophoretic properties of solubilizable disulfide gels. *Anal. Biochem.* **105,** 192–201.

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

HIERHOLZER, J. C., KEMP, M. C., and TANNOCK, G. A. (1981). The RNA and proteins of human coronaviruses. *In* "Biochemistry and Biology of Coronaviruses" (V. ter Meulen, S. Siddell, and H. Wege, Eds.), Vol. 142, pp. 43–69. Plenum, New York/London.

KAMAHORA, T., SOE, L. H., and LAI, M. M. C. (1988). Sequence analysis of nucleocapsid gene and leader RNA of human coronavirus OC43. *Virus Res.*, in press.

KAPKE, P. A., and BRIAN, D. A. (1986). Sequence analysis of the porcine transmissable gastroenteritis coronavirus nucleocapsid protein gene. *Virology* **151,** 41–49.

KECK, J. G., HOGUE, B. G., BRIAN, D. A., and LAI, M. M. C. (1988) Temporal regulation of bovine coronavirus RNA synthesis. *Virus Res.* **9,** 343–356.

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. *In* "Molecular Biology and Pathogenesis of Coronaviruses" (P. J. M. Rottier, B. A. M. van der Zeijst, W. J. M. Spaan, and M. C. Horzinek, Eds.), Vol. 173, pp. 65–79. Plenum, New York/London.

KENNEDY, D. A., and JOHNSON-LUSSENBERG, C. M. (1975/76). Isolation and morphology of the internal component of human coronavirus, strain 229E. *Intervirology* **6,** 197–206.

KING, B., POTTS, B. J., and BRIAN, D. A. (1985). Bovine coronavirus hemagglutinin protein. *Virus Res.* **2,** 53–59.

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

LAI, M. M. C. (1988). Replication of coronavirus RNA. *In* "RNA Genetics" (J. Holland, P. Ahlquist, and E. Domingo, Eds.), Vol. 1, pp. 115–136. CRC Press, Boca Raton, FL.

LAI, M. M. C., BARIC, R. S., BRAYTON, P. R., and STOHLMAN, S. A. (1984). Characterization of leader RNA sequences on the virion and mRNAs of mouse hepatitis virus, a cytoplasmic virus. *Proc. Natl. Acad. Sci. USA* **81,** 3626–3630.

LAI, M. M. C., BRAYTON, P. R., ARMEN, R. C., PATTON, C. D., PUGH, C., and STOHLMAN, S. A. (1981). Mouse hepatitis virus A59: Messenger RNA structure and genetic localization of the sequence divergence from the hepatotropic strain MHV-3. *J. Virol.* **39,** 823–834.

LAI, M. M. C., MAKINO, S., SOE, L. H., SHIEH, C.-K., KECK, J. G., and FLEMING, J. O. (1987). Coronavirus: A jumping RNA transcription. *Cold Spring Harbor Symp. Quant. Biol.* **52,** 359–365.

LAI, M. M. C., PATTON, C. D., BARIC, R. S., and STOHLMAN, S. A. (1983). Presence of leader sequences in the mRNA of mouse hepatitis virus. *J. Virol.* **46,** 1027–1033.

LAI, M. M. C., and STOHLMAN, S. A. (1978). The RNA of mouse hepatitis virus. *J. Virol.* **26,** 236–242.

LAPPS, W., HOGUE, B. G., and BRIAN, D. A. (1987). Sequence analysis of the bovine coronavirus nucleocapsid and matrix protein genes. *Virology* **157,** 47–57.

MACNAUGHTON, M. R. (1980). The polypeptides of human and mouse coronaviruses. *Arch. Virol.* **63,** 75–80.

MACNAUGHTON, M. R. (1981). Structural and antigenic relationships between human, murine and avian coronaviruses. *In* "Biochemistry and Biology of Coronaviruses (V. ter Meulen, S. Siddell, and H. Wege, Eds.), Vol. 142, pp. 19–29. Plenum, New York/London.

MAKINO, S., STOHLMAN, S. A., and 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. *Proc. Natl. Acad. Sci. USA* **83,** 4204–4208.

MAKINO, S., TAGUCHI, F., HIRANO, N., and FUJIWARA, K. (1984). Analysis of genomic and intracellular viral RNAs of small plaque mutants of mouse hepatitis virus, JHM strain. *Virology* **139,** 138–151.

MAXAM, A. M., and GILBERT, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74,** 560–564.

MAXAM, A. M., and GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *In* "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 499–560. Academic Press, Orlando, FL.

MCINTOSH, K., CHAO, R. K., KRAUSE, H. E., WASIL, R., MOCEGA, H. E., and MUFSON, M. A. (1974). Coronavirus infection in acute lower respiratory tract disease of infants. *J. Infect. Dis.* **139,** 502–510.

MCMASTER, G. K., and CARMICHAEL, G. G. (1977). Analysis of single- and double-stranded nucleic acids on polyacrylamide and agarose gels by using glyoxal and acridine orange. *Proc. Natl. Acad. Sci. USA* **74,** 4835–4838.

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

PEDERSEN, F. S., and HASELTINE, W. A. (1980). A micromethod for detailed characterization of high molecular weight RNA. *In* "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 680–687. Academic Press, Orlando, FL.

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

RESTA, S., LUBY, J. P., ROSENFELD, C. R., and SIEGEL, J. D. (1985). Isolation and propagation of a human enteric coronavirus. *Science* **229,** 978–981.

ROTTIER, P. J. M., SPAAN, W. J. M., HORZINEK, N. C., and VAN DER ZEIJST, B. A. M. (1981). Translation of three mouse hepatitis virus strain A59 subgenomic RNAs in *Xenopus laevis* oocytes. *J. Virol.* **38,** 20–26.

SANGER, F., NICKLEN, S., and COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.

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.

SKINNER, M., and SIDDELL, S. (1984). Nucleotide sequencing of mouse hepatitis virus strain JHM messenger RNA 7. *In* "Molecular Biology and Pathogenesis of Coronaviruses" (P. J. M. Rottier, B. A. M. van de Zeijst, W. J. M. Spaan, and M. C. Horzinek, Eds.), Vol. 173, pp. 163–173. Plenum, New York/London.

STOHLMAN, S. A., and LAI, M. M. C. (1979). Phosphoproteins of murine hepatitis virus. *J. Virol.* **32,** 672–675.

TABOR, S., and RICHARDSON, C. C. (1985). A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. *Proc. Natl. Acad. Sci. USA* **82,** 1074–1078.

THOMAS, P. S. (1980). Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci. USA* **77,** 5201–5205.

WEGE, H., SIDDELL, S., and TER MEULEN, V. (1982). The biology and pathogenesis of coronaviruses. *Curr. Top. Microbiol. Immunol.* **99,** 165–200.

WEINER, L. P., and STOHLMAN, S. A. (1978). Viral models of demyelination. *Neurology* **28,** 111–114.

WEISS, S. R., and LEIBOWITZ, J. L. (1981). Comparison of the RNAs of murine and human coronaviruses. *In* "Biochemistry and Biology of Coronaviruses" (V. ter Meulen, S. Siddell, and H. Wege, Eds.), Vol. 142, pp. 43–69. Plenum, New York/London.