Hepatitis E Virus (HEV): Molecular Cloning and Sequencing of the Full-Length Viral Genome

ALBERT W. TAM, MATTHEW M. SMITH, MARTHA E. GUERRA, CHIAO-CHAIN HUANG,
DANIEL W. BRADLEY*, KIRK E. FRY, AND GREGORY R. REYES[1]

*Molecular Virology Department, Genelabs Incorporated, Redwood City, California 94063; and *Hepatitis Branch,
Division of Viral Diseases, Centers for Disease Control, Atlanta, Georgia 30333*

We have recently described the cloning of a portion of the hepatitis E virus (HEV) and confirmed its etiologic association with enterically transmitted (waterborne, epidemic) non-A, non-B hepatitis. The virus consists of a single-stranded, positive-sense RNA genome of approximately 7.5 kb, with a polyadenylated 3' end. We now report on the cloning and nucleotide sequencing of an overlapping, contiguous set of cDNA clones representing the entire genome of the HEV Burma strain [HEV(B)]. The largest open reading frame extends approximately 5 kb from the 5' end and contains the RNA-directed RNA polymerase and nucleoside triphosphate binding motifs. The second major open reading frame (ORF2) begins 37 bp downstream of the first and extends approximately 2 kb to the termination codon present 65 bp from the 3' terminal stretch of poly(A) residues. ORF2 contains a consensus signal peptide sequence at its amino terminus and a capsid-like region with a high content of basic amino acids similar to that seen with other virus capsid proteins. A third open reading frame partially overlaps the first and second and encompasses only 369 bp. In addition to the 7.5-kb full-length genomic transcript, two subgenomic polyadenylated messages of approximately 3.7 and 2.0 kb were detected in infected liver using a probe from the 3' third of the genome. The genomic organization of the virus is consistent with the 5' end encoding nonstructural and the 3' end encoding the viral structural gene(s). The expression strategy of the virus involves the use of three different open reading frames and at least three different transcripts. HEV was previously determined to be a nonenveloped particle with a diameter of 27–34 nm. These findings on the genetic organization and expression strategy of HEV suggest that it is the prototype human pathogen for a new class of RNA virus or perhaps a separate genus within the Caliciviridae family. © 1991 Academic Press, Inc.

INTRODUCTION

Viral hepatitis results from infection with one of at least four very different viral agents. Available serological tests allow the diagnosis of acute hepatitis due to infection with hepatitis A virus (HAV) and hepatitis B virus (HBV). HBV is required for propagation of the delta agent, or hepatitis D virus (HDV); this co-infection results in a high proportion of cases progressing to chronic active hepatitis. The clinical and diagnostic exclusion of HAV and HBV led to the recognition of other viral hepatitides that were formerly grouped together as non-A, non-B hepatitis (NANBH) (Prince *et al.,* 1974; Feinstone *et al.,* 1975; Tabor, 1985).

NANBH is caused by more than one viral agent and can be transmitted by either parenteral or fecal/oral routes (Bradley, 1990a; Reyes and Baroudy, 1991). The cloning of a blood-borne agent, termed hepatitis C virus (HCV) by us and others led to the development of a specific assay for circulating antibody to HCV (Choo *et al.,* 1989; Kuo *et al.,* 1989; Kubo *et al.,* 1989; Maeno *et al.,* 1990; Reyes *et al.,* 1991d). This assay predomi-

nantly detects infections at the chronic stage, but has facilitated the identification of HCV as the cause of up to 90% of parenterally transmitted NANBH. A second epidemiologically distinct form of NANBH was shown to occur in both epidemic and sporadic patterns in developing countries and is referred to as enterically transmitted non-A, non-B hepatitis (ET-NANBH) due to its water-borne mode of virus transmission and presumed enteric route of infection (Khuroo, 1980; Wong *et al.,* 1980). ET-NANBH has been documented in India, Pakistan, Burma, USSR, Costa Rica, Mexico, and countries in Africa, where epidemic outbreaks can generally be traced to fecal contamination of drinking water (Bradley and Maynard, 1986; Bradley, 1990b). The causative viral agent was previously shown to passage successfully in cynomolgus macaques (cyno) and tamarins with typical liver enzyme elevations and recovery of morphologically similar 27- to 34-nm virus-like particles from the feces of clinical specimens and experimental animals (Balayan *et al.,* 1983; Anjaparidze *et al.,* 1986; Bradley *et al.,* 1987; Arankalle *et al.,* 1988).

We recently reported the isolation of a partial cDNA clone from the virus responsible for ET-NANBH, and have termed the newly identified agent the hepatitis E virus (HEV) (Reyes *et al.,* 1990). The clone was from a

---

[1] To whom correspondence and reprint requests should be addressed at Molecular Virology Department, Genelabs Incorporated, 505 Penobscot Dr., Redwood City, CA 94063.

Burma isolate of HEV and hybridized with cDNA made from five other distinct geographic isolates. These molecular epidemiological findings are consistent with the available serologic data based on the use of immune electron microscopy and immunofluorescence blocking studies that indicate a single major agent is responsible for the majority of ET-NANBH seen worldwide (Purcell and Ticehurst, 1988; Bradley et al., 1988a; Krawczynski and Bradley, 1989). We now report on the molecular cloning and sequencing of the complete HEV (Burma; B) viral genome together with the deduced amino acid sequences of viral-encoded proteins. General perspectives on the genetic organization of the virus, as deduced from sequence and open reading frame analyses, indicate that HEV bears some similarity to the caliciviridae but may represent a new class of nonenveloped RNA virus.

MATERIALS AND METHODS

RNA purification. Total cellular RNA was isolated from normal and HEV(B)-infected cyno livers by the guanidinium-LiCl precipitation method (Cathala et al., 1983), and poly(A)$^+$ RNA was selected by one round of oligo(dT) cellulose chromatography (Aviv and Leder, 1972).

cDNA library construction and screening. Synthesis and screening of the infectious bile cDNA library had previously been described (Reyes et al., 1990). Oligo(dT)-, random hexamer-, and HEV sequence-specific oligomer-primed (primer A, see Fig. 1) cDNA were synthesized using a commercially available cDNA synthesis kit (Boehringer-Mannheim Biochemicals, Indianapolis, IN), ligated to EcoRI linker-adapters and cloned into λ gt10 (Stratagene, San Diego, CA). G-tailed cDNA was made essentially as described before (Tam et al., 1989). Briefly, first strand cDNA primed with HEV sequence-specific primer C (see Fig. 1) was tailed with dGTP using terminal deoxynucleotidyl transferase. The modified cDNA was then amplified in a polymerase chain reaction (PCR) (Saiki et al., 1985; Mullis and Faloona, 1987) employing the same synthetic HEV primer and an oligo(dC) primer, both of which contained an EcoRI cloning site at the 5' end. All four cDNA libraries were screened with appropriate synthetic oligomer probes (Applied Biosystems, Foster City, CA) described under Results. Hybridizations were generally performed in duplicate (using ^{32}P kinased probes) at 42° in 30% formamide, 5× SSC, 5× Denhardt's (0.1% ficoll, 0.1% polyvinylpyrrolidone, 0.1% BSA), 50 mM sodium phosphate, pH 7.0, and 50 μg/ml salmon sperm DNA. After an overnight hybridization, filters were washed three times with 0.2× SSC and 0.1%

SDS at 37–42° depending on the length of the oligomer probe.

Primer extension analysis. Primer extension studies were carried out using oligonucleotide primers kinased to a specific activity greater than 3×10^9 cpm/μg with [γ-^{32}P]ATP (ICN Radiochemicals, Irvine, CA) essentially as described (McKnight et al., 1981). Extension products were separated on a 6% polyacrylamide-8 M urea sequencing gel that was subsequently dried and autoradiographed. The sequences for the HEV primers used in these studies are:

Primer A: 5'-CCCGATAAGCAGCCTCAAGCCTC-3'
Primer B: 5'-CCGCGTACACACTAACCCCCCGGC-CAATAATTCACGCTGG-3'
Primer C: 5'-CAAGCTGGCGAGGTTGCATTAGG-3'
Primer D: 5'-ACAGCATTCGCCAGGGCAGAGTT-3'

Northern blot analysis. Four micrograms of HEV(B)-infected cyno liver poly(A)$^+$ RNA was electrophoresed on a 1.2% agarose gel containing 2.2 M formaldehyde and transferred onto a nitrocellulose filter. The filter was hybridized under high stringency conditions with a radiolabeled BET6-1 EcoRI fragment insert (5×10^8 cpm/μg).

DNA nucleotide sequencing. DNA sequencing was performed by the dideoxynucleotide method (Sanger et al., 1977) using 7-deaza-dGTP (Pharmacia, Piscataway, NJ). All sequencing reactions were carried out on both strands using Bluescript plasmid (Stratagene, San Diego, CA) subclones obtained from HEV λgt10 phage clones. Appropriate overlapping subfragments were exploited wherever possible, or adjoining dissimilar-end subclones were employed for unambiguous orientation. Sequencing primers were commercially available or synthesized based on derived HEV sequences. 7-deaza-dGTP eliminated areas of compression due to the high G + C content of the viral genome (see Results).

Computer analyses of nucleotide and amino acid sequences. Computer programs for manipulation of nucleic acid and protein sequences were obtained from Intelligenetics (Mountain View, CA).

RESULTS

cDNA clone derivation

A partial HEV cDNA clone, ET1.1, was isolated by differential screening of a cDNA library constructed from infectious bile collected from a third-passage cyno inoculated with subpassaged fecal suspensions originally derived from Burma patients with well-defined ET-NANBH (Reyes et al., 1990). Bile was chosen as the RNA source for cDNA synthesis because it contained relatively large numbers of virus particles when
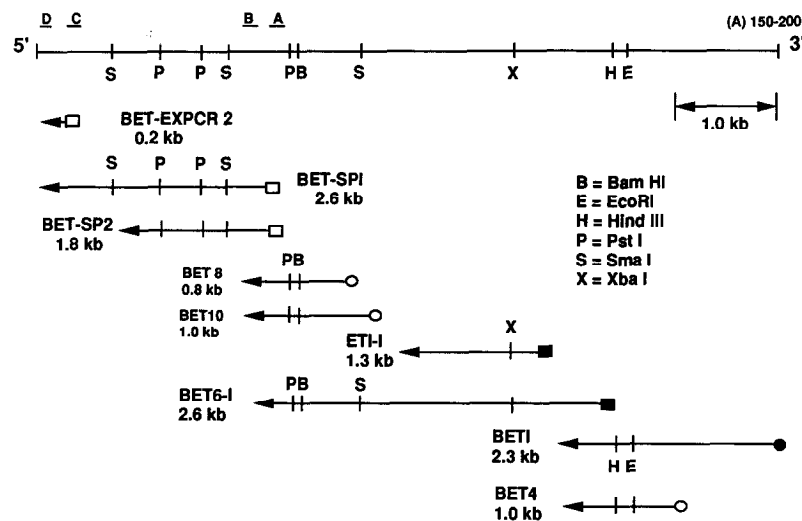
Fig. 1. Schematic representation of hepatitis E virus (HEV) cDNA clones. HEV cDNA clones were identified from libraries made from randomly primed cyno bile (solid square), or from cyno liver after priming by oligo-dT (solid circle), random sequence hexamers (open circle) and HEV-sequence specific oligonucleotides (open square). The designations given to the various clones are indicated together with their sizes and relative position and overlap along the ~7.5 kb genome. A and B represent synthetic oligonucleotides used for the generation and screening, respectively, of specifically primed cDNA libraries. The anchor PCR strategy using G-tailing and PCR (primer C) was used in the synthesis of primer extension libraries for the extreme 5' end. The procedure yielded numerous clones by hybridization with primer D, of which BET-EXPCR2 is a representative example. Primer extension studies confirmed the 5' extent of the viral genome (see Fig. 2). The BET1 clone contained a long stretch of poly(A) residues at its 3' end indicating its position at the 3' terminus of the viral genome.

compared with fecal preparations. It was also expected that the lower sequence complexity would enhance the sensitivity of the differential (plus/minus) screening protocol used for clone identification. ET1.1 contained a 1.3-kb EcoRI fragment that was exogenous to both human and cyno genomic DNA and specifically hybridized to cDNA derived only from infected sources (Reyes et al., 1990). Oligonucleotides based on the end sequences of ET1.1 were used as hybridization probes to rescreen the original bile-derived cDNA library. The largest identified clone, BET6-1, contained a 2.6-kb EcoRI insert. Restriction mapping revealed that the original ET1.1 clone was contained within the larger BET6-1 (Fig. 1; Fry et al., 1991).

The same end-probe strategy was used with oligonucleotides derived from BET6-1 to screen oligo(dT)-primed and random hexamer-primed HEV(B)-infected cyno liver cDNA libraries. A collection of overlapping clones was identified from both libraries (Fig. 1). One of the oligo(dT)-primed clones, BET1 contained two EcoRI fragments that comprised 2.4 kb in total length. The authenticity of the EcoRI site was strengthened by its presence in another clone, BET4, isolated from the random-primed cDNA library. A long poly(A) stretch of ~150–200 adenosine residues was located at the 3' end of BET1 confirming the original observation that genomic RNA could be selected on oligo-dT cellulose (Reyes, 1991a). This result indicated that the 3' end of the viral genome was present in the BET1 clone.

The 5' end of the viral genome was isolated from a cDNA library made by primer extension using a synthetic 23-bp oligonucleotide complementary to the 5' end of clone BET8 (primer A, see Fig. 1). One of two positive clones identified by an oligonucleotide probe (primer B), located 5' to the specific primer, was clone BET-SP1. This clone contained a single large insert of 2.6 kb. With the acquisition of BET-SP1, the composite cDNA map (omitting overlaps) spanned approximately 7.4 kb from the 5' end of BET-SP1 to the polyadenylated 3' end of clone BET1; in good agreement with the maximum length of HEV RNA as detected on Northern blots. The 5' end of BET-SP1 was therefore believed to be in close proximity to the putative 5' end of the viral genome.

Primer extension analysis and 5' end cloning

Primer extension studies using poly(A)-selected RNA from infected cyno liver were performed in order to firmly establish the distance from the existing 5' end of BET-SP1 to the end of the genome (Fig. 2). Two specific oligonucleotide primers (primers C and D, see Fig. 1) were synthesized 143 and 72 bp from the 5' end of BET-SP1 and used to prime cDNA synthesis after ^{32}P labeling their 5' ends with polynucleotide kinase. The resulting extension products for each synthesis reaction were, respectively, 50 and 51 bp longer than the expected product, thereby suggesting that the 5' end
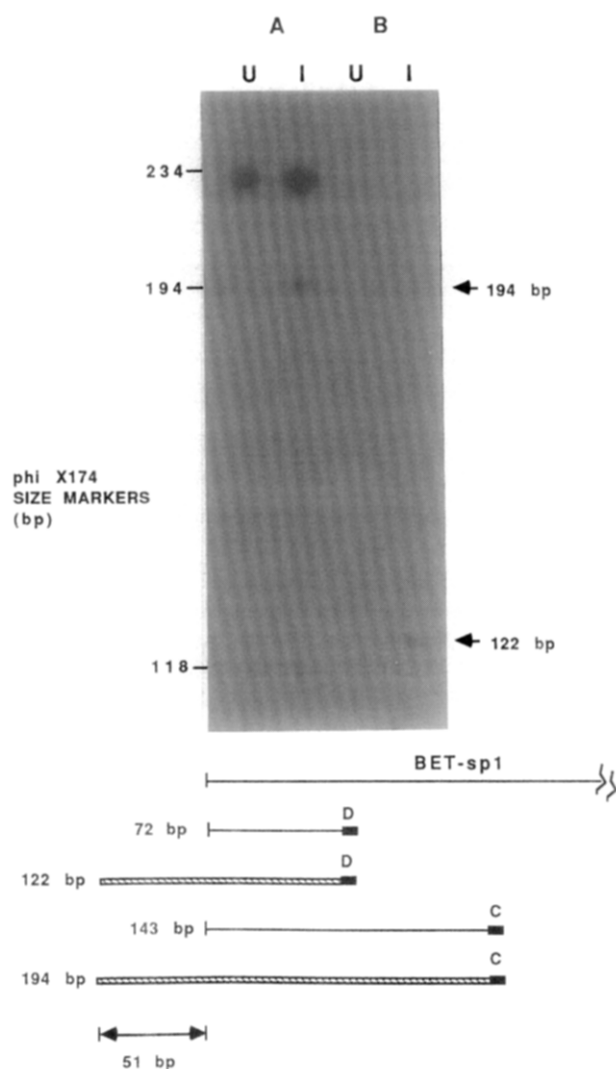
FIG. 2. Primer extension studies establish the 5' end of the HEV genome. Primer extension was performed using poly(A) RNA from uninfected (U) and infected (I) cyno liver. Two different HEV-specific oligonucleotide primers were used after being labeled with ^{32}P: primer D located 72 bp downstream from the 5' end of the clone BET-SP1 and primer C located 143 bp from the 5' end of BET-SP1 (bottom of figure). The extension reaction product synthesized using primer C is indicated under A where a unique fragment of 194 nucleotides was observed in only the infected RNA source. A similarly unique band of 122 nucleotides was found with primer D (under B). The lengths of the extension products were consistent with the 5' end of clone BET-SP1 mapping ~50 bp downstream of the 5' end of the viral genome. No specific bands were observed for the uninfected controls. The upper band present in both the infected and uninfected samples (see A) resulted from oligonucleotide C priming of an RNA unrelated to HEV.

of BET-SP1 was about 50 nucleotides from the 5' end of the virus (Fig. 2).

After several failed attempts at cloning the remaining 5' end sequences by oligonucleotide hybridization of specifically primed cDNA libraries, an alternative expansion/enrichment procedure of PCR amplification of specifically primed G-tailed cDNA was applied (Tam et al., 1989). An aliquot of the amplified material was EcoRI digested, electrophoresed, blotted, and probed with a 5' internal HEV oligomer (primer D). This hybridization study confirmed the amplification of the desired HEV extension products (data not shown). This same DNA, after preparative gel electrophoresis, was recovered and ligated into λgt10. The specific priming procedure (followed by PCR amplification) resulted in a high percentage (over 10%) of HEV-positive recombinants in the enriched library. BET-EXPCR2 is a representative clone from over 50 analyzed; all of these were 50 bp in length and therefore in agreement with the primer extension experiment. The isolation of BET-EXPCR2 completed the HEV genomic cDNA cloning.

HEV genome sequencing and analysis

The entire nucleotide and deduced amino acid sequence of HEV are presented in Fig. 3. The nucleotide composition of the HEV genomic RNA is 17% A, 32% C, 26% G, and 25% U, conferring an overall G + C content of 58%. Sequence homology to any nucleotide sequences contained in the GenBank database could not be detected when the HEV sequence was searched in either the forward or reverse orientation. Only two regions were identified that had homology with previously described nonstructural gene elements present in other positive strand RNA viruses (see below; Fry et al., 1991).

Using the COD RNY sequence analysis program, the ~7.2-kb of HEV sequence, exclusive of the 3' poly(A) tract, was analyzed for the presence of open reading frames (ORF) in the six possible translation frames (Fig. 4). The identification of the RNA-dependent RNA polymerase in the original ET1.1 clone (Reyes et al., 1990) and strand-specific probe hybridization (Reyes et al., 1991b) established the positive-sense orientation of the HEV genome. A representation of the potential ORFs and stop codons in the three positive-polarity frames is presented in Fig. 4a. Two large potential ORFs were found in the first and second reading frames. ORF1 begins at the 5' end of the viral genome after 27 bp of apparent noncoding sequence at the 5' end, and then extends 5079 bp before termination at nucleotide position 5107. The second major ORF (ORF2) begins at nucleotide position 5147 and extends 1980 bp before terminating 65 bp upstream of the poly(A) tail. The termination of ORF1 and the transition into ORF2 was confirmed by sequencing the region in question five times using two different HEV sequence-specific primers. The sequence of a second clone in this region yielded the same results. Furthermore, cDNA clones isolated directly from infected human
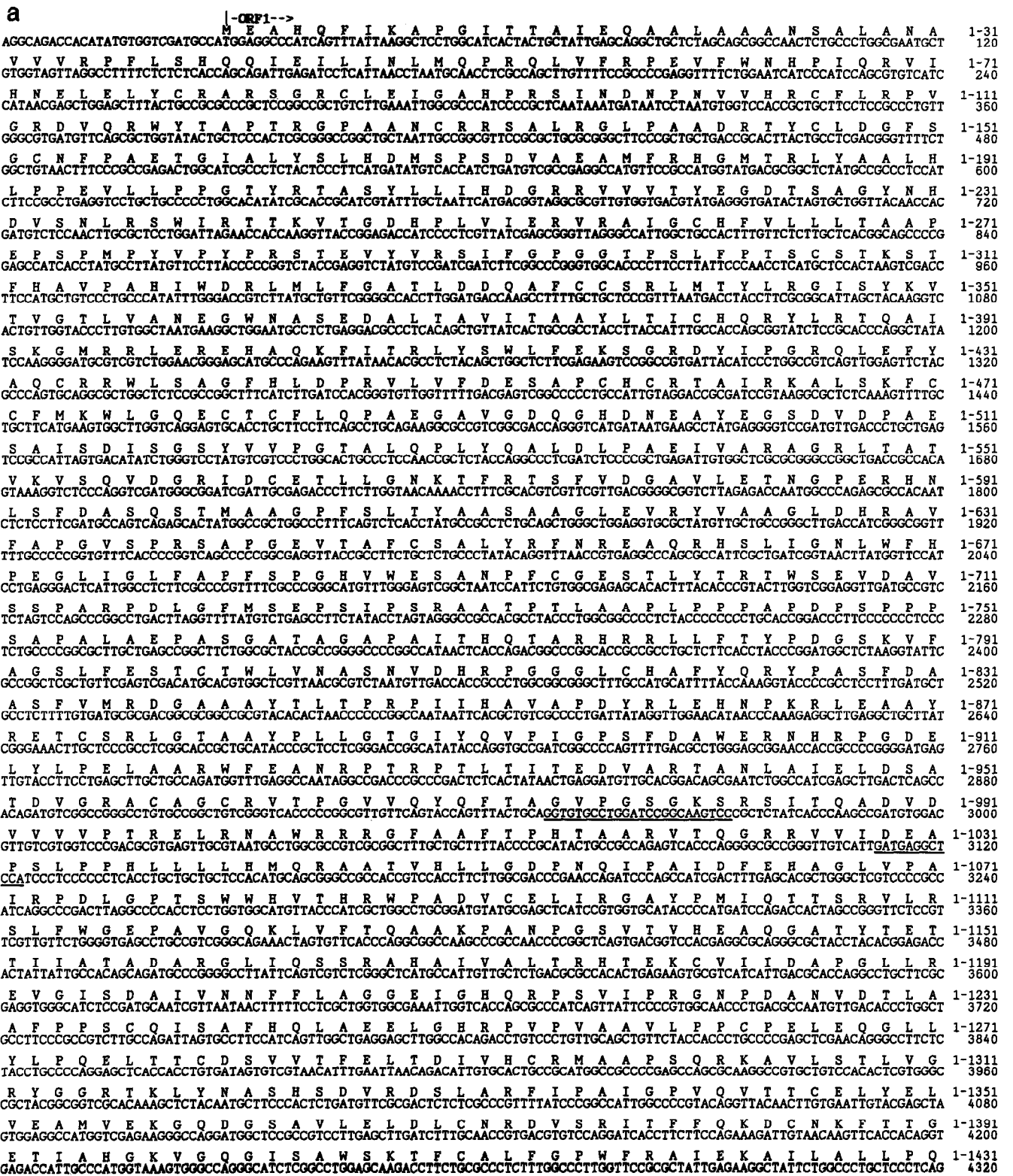
a

```
                   |-ORF1-->
       M  E  A  H  Q  F  I  K  A  P  G  I  T  T  A  I  E  Q  A  A  L  A  A  A  N  S  A  L  A  N  A      1-31
AGGCAGACCACATATGTGGTCGATGCCATGGAGGCCCATCAGTTTATTAAGGCTCCTGGCATCACTACTGCTATTGAGCAGGCTGCTCTAGCAGCGGCCAACTCTGCCCTGGCGAATGCT   120

 V  V  V  R  P  F  L  S  H  Q  Q  I  E  I  L  I  N  L  M  Q  P  R  Q  L  V  F  R  P  E  V  F  W  N  H  P  I  Q  R  V  I      1-71
GTGGTAGTTAGGCCTTTTCTCTCTCACCAGCAGATTGAGATCCTCATTAACCTAATGCAACCTCGCCAGCTTGTTTTCCGCCCCGAGGTTTTCTGGAATCATCCCATCCAGCGTGTCATC   240

 H  N  E  L  E  L  Y  C  R  A  R  S  G  R  C  L  E  I  G  A  H  P  R  S  I  N  D  N  P  N  V  V  H  R  C  F  L  R  P  V      1-111
CATAACGAGCTGGAGCTTTACTGCCGCGCCCGCTCCGGCCGCTGTCTTGAAATTGGCGCCCATCCCCGCTCAATAAATGATAATCCTAATGTGGTCCACCGCTGCTTCCTCCGCCCTGTT   360

 G  R  D  V  Q  R  W  Y  T  A  P  T  R  G  P  A  A  N  C  R  R  S  A  L  R  G  L  P  A  A  D  R  T  Y  C  L  D  G  F  S      1-151
GGGCGTGATGTTCAGCGCTGGTATACTGCTCCCACTCGCGGGCCGGCTGCTAATTGCCGGCGTTCGCGCTGCGCGGGCTTCCCGCTGCTGACCGCACTTACTGCCTCGACGGGTTTTCT   480

 G  C  N  F  P  A  E  T  G  I  A  L  Y  S  L  H  D  M  S  P  S  D  V  A  E  A  M  F  R  H  G  M  T  R  L  Y  A  A  L  H      1-191
GGCTGTAACTTTCCCGCCGAGACTGGCATCGCCCTCTACTCCCTTCATGATATGTCACCATCTGATGTCGCCGAGGCCATGTTCCGCCATGGTATGACGCGGCTCTATGCCGCCTCCAT   600

 L  P  P  E  V  L  L  P  P  G  I  Y  R  T  A  S  Y  L  L  I  H  D  G  R  R  V  V  V  T  Y  E  G  D  T  S  A  G  Y  N  H      1-231
CTTCCGCCTGAGGTCCTGCTGCCCCCTGGCACATATCGCACCGCATCGTATTTGCTAATTCATGACGGTAGGCGCGTTGTGGTGACGTATGAGGGTGATACTAGTGCTGGTTACAACCAC   720

 D  V  S  N  L  R  S  W  I  R  T  T  K  V  T  G  D  H  P  L  V  I  E  R  V  R  A  I  G  C  H  F  V  L  L  L  T  A  A  P      1-271
GATGTCTCCAACTTGCGCTCCTGGATTAGAACCACCAAGGTTACCGGAGACCATCCCCTCGTTATCGAGCGGGTTAGGGCCATTGGCTGCCACTTTGTTCTCTTGCTCACGGCAGCCCCG   840

 E  P  S  P  M  P  Y  V  P  Y  P  R  S  T  E  V  V  V  R  S  I  F  G  P  G  G  T  P  S  L  F  P  T  S  C  S  T  K  S  T      1-311
GAGCCATCACCTATGCCTTATGTTCCTTACCCCCGGTCTACCGAGGTCTATGTCCGATCGATCTTCGGCCCGGGTGGCACCCCTTCCTTATTCCCAACCTCATGCTCCACTAAGTCGACC   960

 F  H  A  V  P  A  H  I  W  D  R  L  M  L  F  G  A  T  L  D  D  Q  A  F  C  C  S  R  L  M  T  Y  L  R  G  I  S  Y  K  V      1-351
TTCCATGCTGTCCCTGCCCATATTTGGGACCGTCTTATGCTGTTCGGGGCCCACCTTGGATGACCAAGCCTTTTGCTGCTCCCGTTTAATGACCTACCTTCGCGGCATTAGCTACAAGGTC   1080

 T  V  G  T  L  V  A  N  E  G  W  N  A  S  E  D  A  L  T  A  V  I  T  A  A  Y  L  T  I  C  H  Q  R  Y  L  R  T  Q  A  I      1-391
ACTGTTGGTACCCTTGTGGCTAATGAAGGCTGGAATGCCTCTGAGGACGCCCTCACAGCTGTTATCACTGCCGCCTACCTTACCATTTGCCACCAGCGGTATCTCCGCACCCAGGCTATA   1200

 S  K  G  M  R  R  L  E  R  E  H  A  Q  K  F  I  T  R  L  Y  S  W  L  F  E  K  S  G  R  D  Y  I  P  G  R  Q  L  E  F  Y      1-431
TCCAAGGGGATGCGTCGTCTGGAACGGGAGCATGCCCAGAAGTTTATAACGACGCCTCTACAGCTGGCTCTTCGAGAAGTCCGGCCGTGATTACATCCCTGGCCGTCAGTTGGAGTTCTAC   1320

 A  Q  C  R  R  W  L  S  A  G  F  H  L  D  P  R  V  L  V  P  D  E  S  A  P  C  H  C  R  T  A  I  R  K  A  L  S  K  F  C      1-471
GCCCAGTGCAGGCGCTGGCTCTCCGCCGGCTTTCATCTTGATCCACGGGTGTTGGTTTTTGACGAGTCGGCCCCCTGCCATTGTAGGACCGCGATCCGTAAGGCGCTCTCAAAGTTTTGC   1440

 C  F  M  K  W  L  G  Q  E  C  T  C  F  L  Q  P  A  E  G  A  V  G  D  Q  G  H  D  N  E  A  Y  E  G  S  D  V  D  P  A  E      1-511
TGCTTCATGAAGTGGCTTGGTCAGGAGTGCACCTGCTTCCTTCAGCCTGCAGAAGGCGCCGTCGGCGACCAGGGTCATGATAATGAAGCCTATGAGGGGTCCGATGTTGACCCTGCTGAG   1560

 S  A  I  S  D  I  S  G  S  Y  V  V  P  G  T  A  L  Q  P  L  Y  Q  A  L  D  L  P  A  E  I  V  A  R  A  G  R  L  T  A  T      1-551
TCCGCCATTAGTGACATATCTGGGTCCTATGTCGTCCCTGGCACTGCCCTCCAACCGCTCTACCAGGCCCTCGATCTCCCCGCTGAGATTGTGGCTCGCGCGGGCCGGCTGACCGCCACA   1680

 V  K  V  S  Q  V  D  G  R  I  D  C  E  T  L  L  G  N  K  T  F  R  T  S  F  V  D  G  A  V  L  E  T  N  G  P  E  R  H  N      1-591
GTAAAGGTCTCCCAGGTCGATGGGCGGATCGATTGCGAGACCCTTCTTGGTAACAAAACCTTTCGCACGTCGTTCGTTGACGGGGCGGTCTTAGAGACCAATGGCCCAGAGCGCCACAAT   1800

 L  S  F  D  A  S  Q  S  T  M  A  A  G  P  F  S  L  T  Y  A  A  S  A  A  G  L  E  V  R  Y  V  A  A  G  L  D  H  R  A  V      1-631
CTCTCCTTCGATGCCAGTCAGAGCACTATGGCCGCTGGCCCTTTCAGTCTCACCTATGCCGCCTCTGCAGCTGGGCTGGAGGTGCGCTATGTTGCTGCCGGGCTTGACCATCGGGCGGTT   1920

 F  A  P  G  V  S  P  R  S  A  P  G  E  V  T  A  F  C  S  A  L  Y  R  F  N  R  E  A  Q  R  H  S  L  I  G  N  L  W  F  H      1-671
TTTGCCCCCGGTGTTTCACCCCGGTCAGCCCCCGGCGAGGTTACCGCCTTCTGCTCTGCCCTATACAGGTTTAACCGTGAGGCCCAGCGCCATTCGCTGATCGGTAACTTATGGTTCCAT   2040

 P  E  G  L  I  G  L  F  A  P  F  S  P  G  H  V  W  E  S  A  N  P  F  C  G  E  S  T  L  Y  T  R  T  W  S  E  V  D  A  V      1-711
CCTGAGGGACTCATTGGCCTCTTCGCCCCGTTTTCGCCCGGCATGTTTGGGAGTCGGCTAATCCATTCTGTGGCGAGAGCACACTTTACACCCGTACTTGGTCGGAGGTTGATGCCGTC   2160

 S  S  P  A  R  P  D  L  G  F  M  S  E  P  S  I  P  S  R  A  A  T  P  T  L  A  A  P  L  P  P  P  A  P  D  P  S  P  P  P      1-751
TCTAGTCCAGCCCGGCCTGACTTAGGTTTTATGTCTGAGCCTTCTATACCTAGTAGGGCCGCCACGCCTACCCTGGCGGCCCCTCTACCCCCCCCTGCACCGGACCCTTCCCCCCCTCCC   2280

 S  A  P  A  L  A  E  P  A  S  G  A  T  A  G  A  P  A  I  T  H  Q  T  A  R  H  R  R  L  L  F  T  Y  P  D  G  S  K  V  F      1-791
TCTGCCCCGGCGCTTGCTGAGCCGGCTTCTGGCGCTACCGCCGGGGGCCCCGGCCATAACTCACCAGACGGCCCGGCACCGCCGCCTGCTCTTCACCTACCCGGATGGCTCTAAGGTATTC   2400

 A  G  S  L  F  E  S  T  C  T  W  L  V  N  A  S  N  V  D  H  R  P  G  G  G  L  C  H  A  F  Y  Q  R  Y  P  A  S  F  D  A      1-831
GCCGGCTCGCTGTTCGAGTCGACATGCACGTGGCTCGTTAACGCGTCTAATGTTGACCACCGCCCTGGCGGCGGGCTTTGCCATGCATTTTACCAAAGGTACCCCGCCTCCTTTGATGCT   2520

 A  S  F  V  M  R  D  G  A  A  A  Y  T  L  T  P  R  P  I  I  H  A  V  A  P  D  Y  R  L  E  H  N  P  K  R  L  E  A  A  Y      1-871
GCCTCTTTTGTGATGCGCGACGGCGCGGCCGCGTACACACTAACCCCCCCGGCCAATAATTCACGCTGTCGCCCCTGATTATAGGTTGGAACATAACCCAAAGAGGCTTGAGGCTGCTTAT   2640

 R  E  T  C  S  R  L  G  T  A  A  Y  P  L  L  G  T  G  I  Y  Q  V  P  I  G  P  S  F  D  A  W  E  R  N  H  R  P  G  D  E      1-911
CGGGAAACTTGCTCCCGCCTCGGCACCGCTGCATACCCGCTCCTCGGGACCGGCATATACCAGGTGCCGATCGGCCCCAGTTTTGACGCCTGGGAGCGGAACCACCGCCCCGGGGATGAG   2760

 L  Y  L  P  E  L  A  A  R  W  F  E  A  N  R  P  T  R  P  T  L  T  I  T  E  D  V  A  R  T  A  N  L  A  I  E  L  D  S  A      1-951
TTGTACCTTCCTGAGCTTGCTGCCCGATGGTTTGAGGCCAATAGGCCGACCCGCCCGACTCTCACTATAACTGAGGATGTTGCACGGACAGCGAATCTGGCCATCGAGCTTGACTCAGCC   2880

 T  D  V  G  R  A  C  A  G  C  R  V  T  P  G  V  Q  Y  Q  F  T  A  G  V  P  G  S  G  K  S  R  S  I  T  Q  A  D  V  D      1-991
ACAGATGTCGGCCGGGCCTGTGCCGGCTGTCGGGTCACCCCCGGCGTTGTTCAGTACCAGTTTACTGCAGGTGTGCCTGGATCCGGCAAGTCCCGCTCTATCACCCAAGCCGATGTGGAC   3000

 V  V  V  V  P  T  R  E  L  R  N  A  W  R  R  R  G  F  A  A  F  T  P  H  T  A  A  R  V  T  Q  G  R  R  V  V  I  D  E  A      1-1031
GTTGTCGGTGGTCCCGACGCGTGAGTTGCGTAATGCCTGGCGCCGTCGCGGCGTTTGCTGCTTTTACCCCGCATACTGCCGCCAGAGTCACCCAGGGGCGCCGGGTTGTCATTGATGAGGCT   3120

 P  S  L  P  H  L  L  L  L  H  M  Q  R  A  A  T  V  H  L  L  G  D  P  N  Q  I  P  A  I  D  F  E  H  A  G  L  V  P  A      1-1071
CCATCCCTCCCCCCTCACCTGCTGCTGCTCCACATGCAGCGGGCCGCCACCGTCCACCTTCTTGGCGACCCGAACCAGATCCCAGCCATCGACTTTGAGCACGTGGGCTCGTCCCCGCC   3240

 I  R  P  D  L  G  P  T  S  W  W  H  V  T  H  R  W  P  A  D  V  C  E  L  I  R  G  A  Y  P  M  I  Q  T  T  S  R  V  L  R      1-1111
ATCAGGCCCGACTTAGGCCCCACCTCCTGGTGGCATGTTACCCATCGCTGGCCTGCGGATGTATGCGAGCTCATCCGTGGTGCATACCCCATGATCCAGACCACTAGCCGGGTTCTCCGT   3360

 S  L  F  W  G  E  P  A  V  G  Q  K  L  V  F  T  Q  A  A  K  P  A  N  P  G  S  V  T  V  H  E  A  Q  G  A  T  Y  T  E  T      1-1151
TCGTTGTTCTGGGGTGAGCCTGCCGTCGGGCAGAAACTAGTGTTCACCCAGGCGGCCAAGCCCGCCAACCCCGGCTCAGTGACGGTCCACGAGGCGCAGGGCGCTACCTACACGGAGACC   3480

 T  I  I  A  T  A  D  A  R  G  L  I  Q  S  S  R  A  H  A  I  V  A  L  T  R  H  T  E  K  C  V  I  I  D  A  P  G  L  L  R      1-1191
ACTATTATTGCCACAGCAGATGCCCGGGGGCCTTATTCAGTCGTCTCGGGGTCATGCCATTGTTGCTCTGACGCGCCACACTGAGAAGTGCGTCATCATTGACGCACCAGGCCTGCTTCGC   3600

 E  V  G  I  S  D  A  I  V  N  N  F  F  L  A  G  G  E  I  G  H  Q  R  P  S  V  I  P  R  G  N  P  D  A  N  V  D  T  L  A      1-1231
GAGGTGGGCATCTCCGATGCAATCGTTAATAACTTTTTCCTCGCTGGTGGCGAAATTGGTCACCAGCGCCCATCAGTTATTCCCCGTGGCAACCCTGACGCCAATGTTGACACCCTGGCT   3720

 A  F  P  P  S  C  Q  I  S  A  F  H  Q  L  A  E  E  L  G  H  R  P  V  P  V  A  A  V  L  P  P  C  P  E  L  E  Q  G  L  L      1-1271
GCCTTCCCGCCGTCTTGCCAGATTAGTGCCTTCCATCAGTTGGCTGAGGAGCTTGGCCACAGACCTGTCCCTGTTGCAGCTGTTCTACCACCCTGCCCCGAGCTCGAACAGGGCCTTCTC   3840

 Y  L  P  Q  E  L  T  T  C  D  S  V  V  T  F  E  L  T  D  I  V  H  C  R  M  A  A  P  S  Q  R  K  A  V  L  S  T  L  V  G      1-1311
TACCTGCCCCAGGAGCTCACCACCTGTGATAGTGTCGTAACATTTGAATTAACAGACATTGTGCACTGCCGCATGGCCGCCCCGAGCCAGCGCAAGGCCGTGCTGTCCACACTCGTGGGC   3960

 R  Y  G  G  R  T  K  L  Y  N  A  S  H  S  D  V  R  D  S  L  A  R  F  I  P  A  I  G  P  V  Q  V  T  T  C  E  L  Y  E  L      1-1351
CGCTACGGCGGTCGCACAAAGCTCTACAATGCTTCCCACTCTGATGTTCGCGACTCTCTCGCCCGTTTTATCCCGGCCATTGGCCCCGTACAGGTTACAACTTGTGAATTGTACGAGCTA   4080

 V  E  A  M  V  E  K  G  Q  D  G  S  A  V  L  E  L  D  L  C  N  R  D  V  S  R  I  T  F  F  Q  K  D  C  N  K  F  T  T  G      1-1391
GTGGAGGCCATGGTCGAGAAGGGCCAGGATGGCTCCGCCGTCCTTGAGCTTGATCTTTGCAACCGTGACGTGTCCAGGATCACCTTCTTCCAGAAAGATTGTAACAAGTTCACCACAGGT   4200

 E  T  I  A  H  G  K  V  G  Q  G  I  S  A  W  S  K  T  F  C  A  L  F  G  P  W  F  R  A  I  E  K  A  I  L  A  L  L  P  Q      1-1431
GAGACCATTGCCCATGGTAAAGTGGGCCAGGGCATCTCGGCCTGGACAAGACCTTCTGCGCCCTCTTTGGCCCTTGGTTCCGCGCTATTGAGAAGGCTATTCTGGCCCTGCTCCCTCAG   4320
```

FIG. 3. Complete nucleotide and deduced amino acid sequence of HEV. The full-length nucleotide sequence of the HEV (Burma) genome is presented. The first methionine residue found within each of the three forward open reading frames (ORFs) was taken as the start of the ORF and used as the basis for the ORF amino acid numbering scheme whereby the amino acids in each translated ORF are prefixed by the number of that ORF (1, 2, or 3). Within ORF1, the conserved motifs for the NTP binding site and the Mg+2 binding site of the helicase RNA or DNA-dependent NTPase activity are underlined. The canonical GDD (gly-asp-asp) tripeptide sequence found in all viral RNA-dependent RNA polymerases is marked with an (*) at amino acid position 1550 of ORF1. The hydrophobic stretch of amino acids, marking a putative signal peptide, is highlighted at the beginning of ORF2. Note the overlap of ORF3 with the major 5' ORF1 and the 3' ORF2. The GenBank accession number for this sequence is M73218.
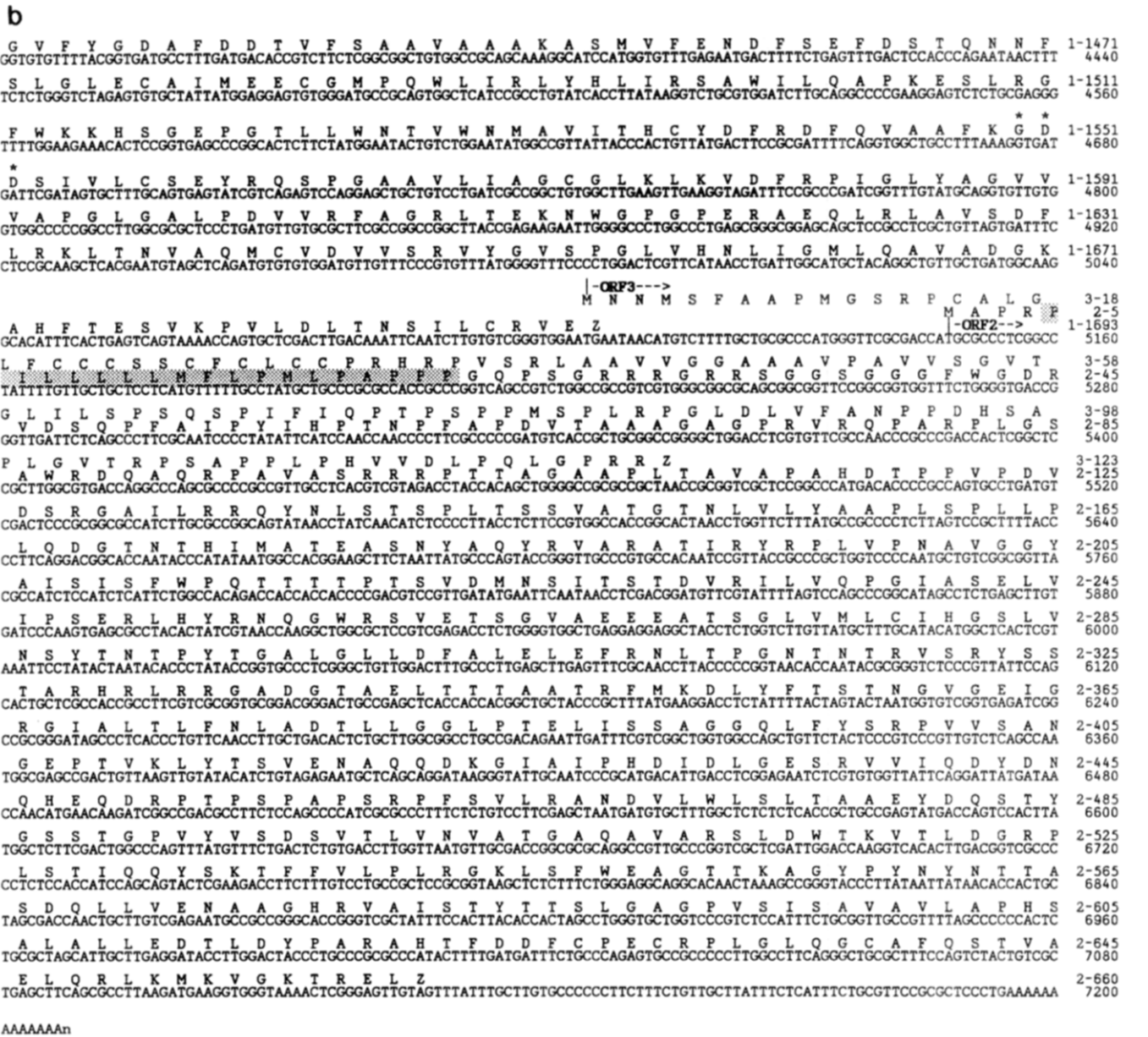
b

```
        G V F Y G D A F D D T V F S A A V A A A K A S M V F E N D F S E F D S T Q N N F    1-1471
GGTGTGTTTTACGGTGATGCCTTTGATGACACCGTCTTCTCGGCGGCTGTGGCCGCAGCAAAGGCATCCATGGTGTTTGAGAATGACTTTTCTGAGTTTGACTCCACCCAGAATAACTTT    4440

        S L G L E C A I M E E C G M P Q W L I R L Y H L I R S A W I L Q A P K E S L R G    1-1511
TCTCTGGGTCTAGAGTGTGCTATTATGGAGGAGTGTGGGATGCCGCAGTGGCTCATCCGCCTGTATCACCTTATAAGGTCGCGTGGATCTTGCAGGCCCCGAAGGAGTCTCTGCGAGGG    4560

                                                                                      * *
        F W K K H S G E P G T L L W N T V W N M A V I T H C Y D F R D F Q V A A F K G D    1-1551
TTTTGGAAGAAACACTCCGGTGAGCCCGCACTCTTCTATGGAATACTGTCTGGAATATGGCCGTTATTACCCACTGTTATGACTTCCGCGATTTTCAGGTGGCTGCCTTTAAAGGTGAT    4680

        *
        D S I V L C S E Y R Q S P G A A V L I A G C G L K L K V D F R P I G L Y A G V V    1-1591
GATTCGATAGTGCTTTGCAGTGAGTATCGTCAGAGTCCAGGAGCTGCTGTCCTGATCGCCGGCTGTGGCTTGAAGTTGAAGGTAGATTTCCGCCCGATCGGTTTGTATGCAGGTGTTGTG    4800

        V A P G L G A L P D V V R F A G R L T E K N W G P G P E R A E Q L R L A V S D F    1-1631
GTGGCCCCGGCCTTGGCGCGCTCCCTGATGTTGTGCGGCTTCGCCGGCCGGCTTACCGGAGAAGAATTGGGGCCCTGGCCCTGAGCGGGCGGAGCAGCTCCGCCTCGCTGTTAGTGATTTC    4920

        L R K L T N V A Q M C V D V V S R V Y G V S P G L V H N L I G M L Q A V A D G K    1-1671
CTCCGCAAGCTCACGAATGTAGCTCAGATGTGTGTGGATGTTGTTTCCCGTGTTTATGGGGTTTCCCCTGGACTCGTTCATAACCTGATTGGCATGCTACAGGCTGTTGCTGATGGCAAG    5040

                                              |-ORF3--->
                                              M N N M S F A A P M G S R P C A L G      3-18
                                                                          M A P R P    2-5
        A H F T E S V K P V L D L T N S I L C R V E Z            |-ORF2-->              1-1693
GCACATTTCACTGAGTCAGTAAAACCAGTGCTCGACTTGACAAATTCAATCTTGTGTCGGGTGGAATGAATAACATGTCTTTTGCTGCGCCCATGGGTTCGCGACCATGCGCCCTCGGCC    5160

        L F C C C S S C F C L C C P R H R P V S R L A A V V G G A A A V P A V V S G V T    3-58
        I L L L L L L M F L P M L P A P F P G Q P S G R R R G R R S G G S G G G F W G D R  2-45
TATTTTGTTGCTGCTCCTCATGTTTTTGCCTATGCTGCCCGCGCCACCGCCCGGTCAGCCGTCTGGCCGCCGTGTGGGCGGCGGCAGCGGCGGTTCCGGCGGTGGTTTCTGGGGTGACCG   5280

                V D S Q P F A I P Y I H P T N P F A P D V T A A A G A G P R V R Q P A R P L G S  2-85
        G L I L S P S Q S P I F I Q P T P S P P M S P L R P G L D L V F A N P P D H S A    3-98
GGTTGATTCTCAGCCCTTCGCAATCCCCTATATTCATCCAACCAACCCCTTCGCCCCCGATGTCACCGCTGCGGCCGGGGCTGGACCTCGTGTTCGCCAACCCGCCGACCACTCGGCTC    5400

                A W R D Q A Q R P A V A S R R R P T T A V A P L T A V A P A H D T P P V P D V  2-125
        P L G V T R P S A P P L P H V V D L P Q L G P R R Z                              3-123
CGCTTGGCGTGACCAGGCCCAGCGCCCCGCCGTTGCCTCACGTCGTAGACCTACCAGCTGGGGCCGCGCTAACCGCGGTCGCTCCGGCCCATGACACCCCGCCAGTGCCTGATGT        5520

        D S R G A I L R R Q Y N L S T S P L T S S V A T G T N L V L Y A A P L S P L L P    2-165
CGACTCCCGCGGCGCCATCTTGCGCCGGCAGTATAACCTATCAACATCTCCCCTTACCTCTTCCGTGGCCACCGGCACTAACCTGGTTCTTTATGCCGCCCCTCTTAGTCCGCTTTTACC   5640

        L Q D G T N T H I M A T E A S N Y A Q Y R V A R A T I R Y R P L V P N A V G G Y    2-205
CCTTCAGGACGGCACCAATACCCATATAATGGCCACGGAAGCTTCTAATTATGCCCAGTACCGGGTTGCCCGTGCCACAATCCGTTACCGCCGCTGGTCCCCAATGCTGTCGGCGGTTA   5760

        A I S I S F W P Q T T T T T P T S V D M N S I T S T D V R I L V Q P G I A S E L V    2-245
CGCCATCTCCATCTCATTCTGGCCACAGACCACCACCACCCCGACGTCCGTTGATATGAATTCAATAACCTCGACGGATGTTCGTATTTTAGTCCAGCCCGGCATAGCCTCTGAGCTTGT   5880

        I P S E R L H Y R N Q G W R S V E T S G V A E E E A T S G L V M L C I H G S L V    2-285
GATCCCAAGTGAGCGCCTACACTATCGTAACCAAGGCTGGCGCTCCGTCGAGACCTCTCGGGGTGGCTGAGGAGGAGGCTACCTCTGGTCTTGTTATGCTTTGCATACATGGCTCACTCGT   6000

        N S Y T N T P Y T G A L G L L D F A L E L E F R N L T P G N T N T R V S R Y S S    2-325
AAATTCCTATACTAATACACCCTATACCGGTGCCCTCGGGCTGTTGGACTTTGCCCTTGAGCTTGAGTTTCGCAACCTTACCCCCGGTAACACCAATACGCGGGTCTCCCGTTATTCCAG   6120

        T A R H R L R R G A D G T A E L T T T A A T R F M K D L Y F T S T N G V G E I G    2-365
CACTGCTCGCCACCGCCTTCGTCGCGGTGCGGACGGGACTGCCGAGCTCACCACCACGGCTGCTACCCGCTTTATGAAGGACCTCTATTTTACTAGTACTAATGGTGTCGGTGAGATCGG   6240

        R G I A L T L F N L A D T L L G G L P T E L I S S A G G Q L F Y S R P V V S A N    2-405
CCGCGGGATAGCCCTCACCCTGTTCAACCTTGCTGACACTCTGCTTGGCGGCCTGCCGACAGAATTGATTTCGTCGGCTGGTGGCCAGCTGTTCTACTCCCGTCCCGTTGTCTCAGCCAA   6360

        G E P T V K L Y T S V E N A Q Q D K G I A I P H D I D L G E S R V V I Q D Y D N    2-445
TGGCGAGCCGACTGTTAAGTTGTATACATCTGTAGAGAATGCTCAGCAGGATAAGGGTATTGCAATCCCGCATGACATTGACCTCGGAGAATCTCGTGTGGTTATTCAGGATTATGATAA   6480

        Q H E Q D R P T P S P A P S R P F S V L R A N D V L W L S L T A A E Y D Q S T Y    2-485
CCAACATGAACAAGATCGGCCGACGCCTTCTCCAGCCCCATCGCGCCCTTTCTCTGTCCTTCGAGCTAATGATGTGCTTTGGCTCTCTCTCACCGGCTGCCGAGTATGACCAGTCCACTTA   6600

        G S S T G P V Y Y V S D S V T L V N V A T G A Q A V A R S L D W T K V T L D G R P    2-525
TGGCTCTTCGACTGGCCCAGTTTATGTTTCTGACTCTGTGACCTTGGTTAATGTTGCGACCGGCGCGCAGGCCGTTGCCCGGTCGCTCGATTGGACCAAGGTCACCTTGACGGTCGCCC   6720

        L S T I Q Q Y S K T F F V L P L R G K L S F W E A G T T K A G Y P Y N Y N T T A    2-565
CCTCTCCACCATCCAGCAGTACTCGAAGACCTTCTTTGTCCTGCCGCTCCGCGGTAAGCTCTCTTTCTGGGAGGCAGGCACAACTAAAGCCGGGTACCCTTATAATTATAACACCACTGC   6840

        S D Q L L V E N A A G H R V A I S T Y T T S L G A G P V S I S A V A V L A P H S    2-605
TAGCGACCAACTGCTTGTCGAGAATGCCGCCGGGCACCGGGTCGCTATTTCCACTTACACCACTAGCCTGGGTGCTGGTCCCGTCTCCATTTCTGCGGTTGCCGTTTTAGCCCCCACTC   6960

        A L A L L E D T L D Y P A R A H T F D D F C P E C R P L G L Q G C A F Q S T V A    2-645
TGCGCTAGCATTGCTTGAGGATACCTTGGACTACCCTGCCCGCGCCCATACTTTTGATGATTTCTGCCCAGAGTGCCGCCCCCTTGGCCTTCAGGGCTGCGCTTTCCAGTCTACTGTCGC   7080

        E L Q R L K M K V G K T R E L Z                                                  2-660
TGAGCTTCAGCGCCTTAAGATGAAGGTGGGTAAAAACTCGGGAGTTGTAGTTTATTTGCTTGTGCCCCCCTTCTTTCTGTTGCTTATTTCTCATTTCTGCGTTCCGCGCTCCCTGAAAAAA   7200

AAAAAAAn
```

<center>Fig. 3—Continued</center>

feces of an altogether different strain (Mexico) demonstrated the same shift in open reading frame (Huang *et al.*, 1991, data not shown). A third positive-polarity reading frame of 369 bp (ORF3) overlaps both ORF1 and ORF2 and was found by independent experiments to encode an immunoreactive epitope recognized by sera from HEV-infected humans and animals (Yarbough *et al.*, 1991; Reyes *et al.*, 1991b). No ORFs greater than 590 bp were identified by computer search of the negative-polarity RNA strand (Fig. 4b).

The nucleotide frequencies at each codon position were also analyzed and a comparison was made with two other hepatotropic positive-strand RNA viruses (Ta-

ble 1). The overall frequencies for HEV and HCV are similar (~58% G + C), but differ markedly from that of HAV (37% G + C). The relatively high G + C content results in a higher overall frequency of codons containing G + C throughout the HEV coding sequence. This contrasts with the CG dinucleotide discrimination in the second and third position that has been noted in human coding sequences (Nussinov, 1981). There appears to be a slight selection for codons ending in C, which is also seen with HCV; however, the discrimination against codons ending in A is far more apparent (~9%) and is a unique feature of HEV when compared to HAV and HCV. The third position discrimination
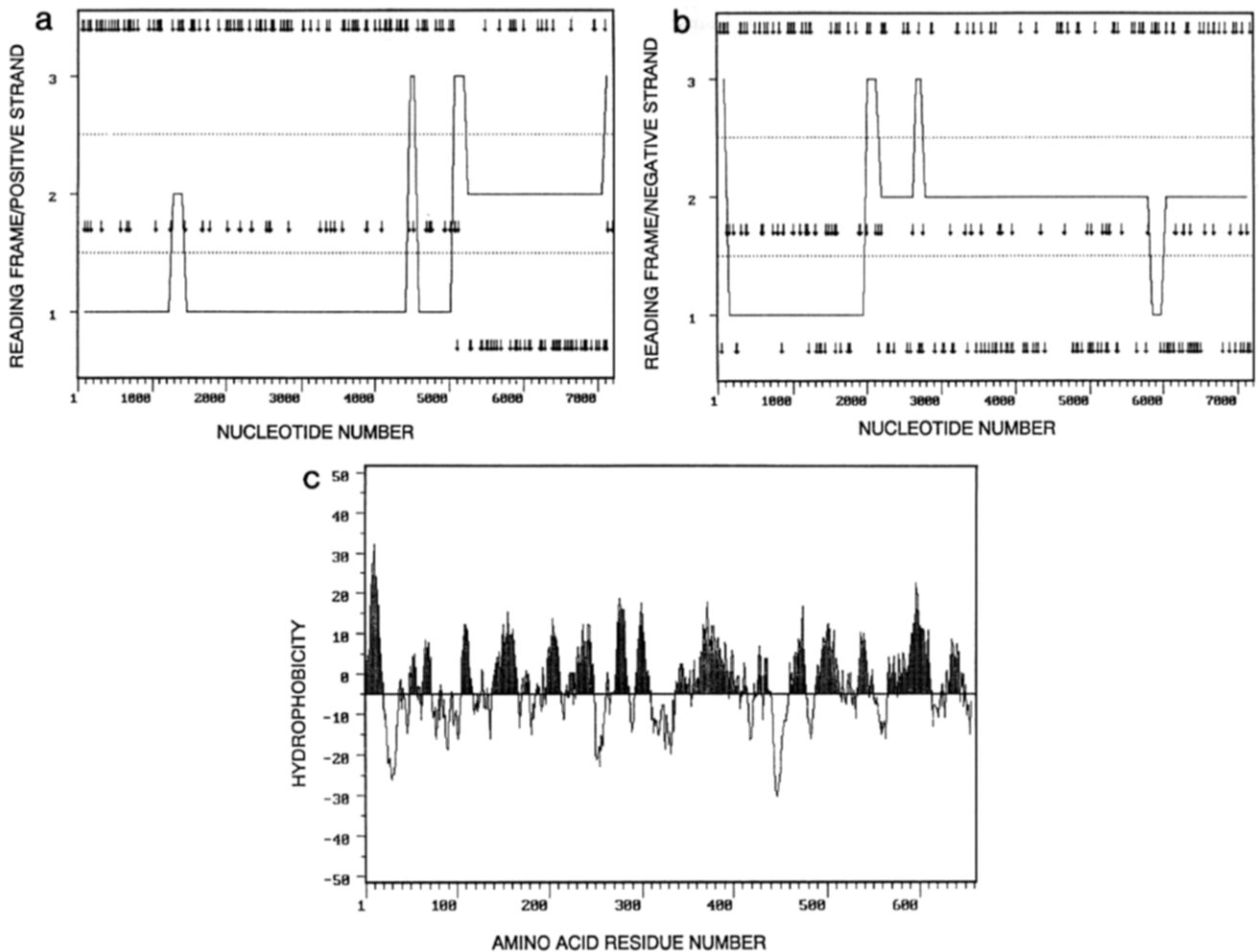
Fig. 4. Open reading frame analysis and hydrophilicity plot. Computer generated open reading frame analysis of the entire HEV nucleotide sequence is presented in both the forward (a) and reverse directions (b). The positions of all termination codons are depicted by arrows. The three forward ORFs are numbered 1, 2, and 3 and those on the opposite strand similarly labeled. The forward (positive-sense) orientation was defined by strand-specific hybridization of genomic RNA (Reyes *et al.*, 1991b) and the identification of consensus sequence motifs related to nonstructural gene products in ORF1 (Kamer and Argos, 1984). The horizontal line running through the various ORFs indicates that ORF with the highest probability of encoding protein as predicted by the algorithm devised by Shepherd based on the RNY codon analysis in all three ORFs (Shepherd, 1981). The hydrophilicity plot of ORF2 is presented in (c). The dotted line plotted at the −5 value on the *y*-axis represents the midpoint with hydrophobic domains above and hydrophilic domains below the dotted line. Note the large hydrophobic region at the beginning of the sequence that marks the putative signal sequence highlighted in Fig. 3.

against A is shared by the structural ORF region of another positive-strand RNA virus, rubella virus, where the frequency of A is only 7% (Frey and Marr, 1988). HEV and HCV are also similar in their apparent preference for G in the first coding position (35 and 33%, respectively).

Genomic organization and expression strategy

The computer translation of the partial nucleotide sequence from clone ET1.1 led to the detection of a conserved amino acid motif recognized in all positive-strand RNA viruses (Reyes *et al.*, 1990; Fry *et al.*,

1991). The canonical Gly-Asp-Asp (GDD) tripeptide (amino acids 1550-1552, identified by asterisks in Fig. 3) is believed to encode a portion of the RNA-dependent RNA polymerase (RDRP) gene critical to viral replication (Kamer and Argos, 1984). Translation of the complete ORF1 revealed a second region 5′ to the RDRP gene bearing similarity to another nonstructural gene product (Fry *et al.*, 1991). Two well-conserved sequence motifs have been found in association with purine nucleoside triphosphate (NTP)-binding activity (Gentry, 1985; Strauss and Strauss, 1988). The first, site A (G/AXXXXGKS/T), is represented in the HEV sequence by GVPGSGKS at amino acid position 975-982 (underlined in Fig. 3). A version of the second NTP-

TABLE 1

Nucleotide Frequencies at Codon Positions in HEV ORFs

| | Nucleotide frequencies | | | | | | | | | | | |
| | A | | | C | | | G | | | U | | |
	HA	HC	HE	HA	HC	HE	HA	HC	HE	HA	HC	HE
Overall	0.30	0.20	0.17	0.15	0.31	0.32	0.22	0.28	0.26	0.33	0.21	0.25
1st Position	0.31	0.24	0.19	0.14	0.23	0.28	0.30	0.33	0.35	0.25	0.19	0.18
2nd Position	0.32	0.22	0.23	0.22	0.29	0.32	0.15	0.22	0.19	0.31	0.27	0.26
3rd Position	0.28	0.14	0.09	0.10	0.39	0.37	0.20	0.30	0.25	0.42	0.17	0.29

Note. The overall nucleotide composition of the genome (Overall) as well as frequencies in each of the three codon positions (1st through 3rd) are presented for three different hepatotropic RNA viruses: HA, hepatitis A virus; HC, hepatitis C virus; HE, hepatitis E virus.

binding motif, site B (DEAD), occurs approximately 46 amino acids downstream (3') from site A and is represented in HEV by the partially conserved amino acid sequence DEAP at position 1029–1032 (underlined in Fig. 3). The latter site is believed to interact with the Mg+2 cation of the Mg-NTP complex for RNA- or DNA-dependent NTPase activity. A superfamily of helicases involved in replication, recombination, and DNA repair has been described with consensus features similar to those described here for NTP-binding (Gorbalenya *et al.*, 1989). These nonstructural gene similarities are seen in other geographically distinct isolates of HEV (Fry *et al.*, 1991) and may indicate a putative helicase function for this region. The localization of an NTP-binding domain and the RDRP gene to ORF1 is consistent with a genomic organization where the nonstructural genes are expressed from the 5' end of the viral genome.

Translation of the second major open reading frame, ORF2, indicated a novel polypeptide not present in the PIR protein database. The hydropathicity plot of the sequence indicated a large hydrophobic domain at the amino terminus of ORF2 followed by a hydrophilic electropositive peak (Fig. 4c). The hydrophobic region marks a typical signal sequence (amino acids 5 to 22) and contains a potential cleavage site (PA/PPP) as predicted by the Intelligenetics eukaryotic secretory signal sequence program. In ORF2, between residues 22 and 322, nearly 10% of the amino acids are arginine conferring a high isoelectric point (p*I* = 10.35) to the first half of the ORF2 polypeptide. The basic charge of capsid proteins is believed to indicate their involvement in the encapsidation of the genomic transcript by effectively neutralizing the electronegatively charged RNA (Dalgarno *et al.*, 1983; Rice *et al.*, 1985). The mechanism of capsid assembly in HEV, and the exact nature of the membrane targeting (if any) of the ORF2 polypeptide, will require further study. Such studies will be facil-

itated by the availability of an appropriate *in vitro* propagation system for HEV and immunospecific anti-HEV reagents.

The utilization of ORF2 was substantiated by the independent isolation of a cDNA clone by immunoscreening of a λgt11 cDNA expression library made from the HEV (Mexico) isolate (Yarbough *et al.*, 1991). That λgt11 clone mapped to the 3' end of ORF2. These same experiments identified a second cDNA epitope clone that was localized to ORF3: the third positive-polarity open reading frame that overlaps both ORF1 and ORF2. The fact that sera from acutely infected humans and animals detected HEV antigens encoded by ORF2 and ORF3 confirmed their expression and established that the virus utilized all three positive-polarity reading frames. The presence of a consensus signal sequence motif in ORF2, together with the immunodominant seroreactivity of an identified epitope (Yarbough *et al.*, 1991), suggested that the viral structural protein(s) were encoded by this region of the genome.

The mechanism by which ORF2 and ORF3 are expressed was suggested by a Northern blot hybridization using the BET6-1 clone as probe (Fig. 5). In addition to the previously identified poly(A) transcript of ~7.5 kb, the probe also hybridized to subgenomic messages of 2.0 and 3.7 kb present in the infected cyno liver. It is of note that ET1.1 did not originally identify these subgenomic messages (Reyes *et al.*, 1990) and other Northern blot studies using probes located 5' to ET1.1 also did not hybridize to these viral-specific transcripts (data not shown). These same subgenomic messages were identified in poly(A)-selected RNA from HEV(M)-infected cyno liver when the epitope-encoding clones were used as probes (Yarbough *et al.*, 1991). The ORF2 epitope is located at the extreme 3' end of that reading frame (Yarbough *et al.*, 1991), therefore indicating that these messages may be co-terminal with the 3' end of the viral genomic transcript. It is pos-
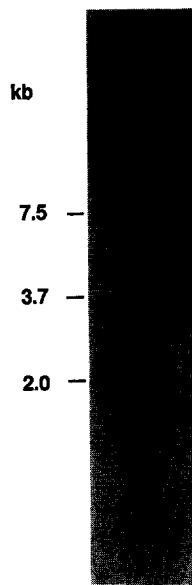
FIG. 5. Northern blot analysis of HEV (Burma)-infected cyno liver RNA. Three HEV transcripts were detected using the 2.6-kb EcoRI insert from BET6-1 as probe. Numbers to the left represent the sizes of the three hybridizing RNA species as determined relative to RNA size markers. HEV cDNA probes were negative against similarly prepared RNA from uninfected liver (data not shown).

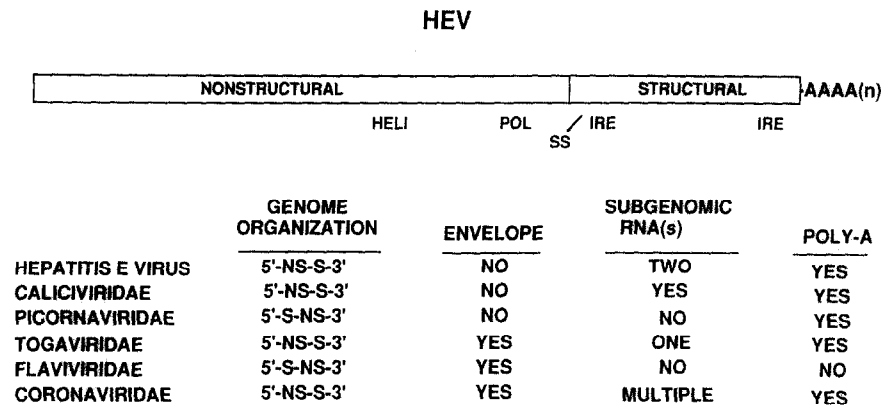sible that these polyadenylated subgenomic messages are used in the expression of ORF2 and ORF3.

DISCUSSION

ET-NANBH has been well-documented in both sporadic and epidemic outbreaks throughout the developing world. Hepatitis E virus has been established as the major causative agent of ET-NANBH by the association of HEV-specific sequences with human specimens derived from six geographically diverse epidemics and also through the detection of these same sequences in various specimens derived from experimentally infected animals (Reyes et al., 1990). HEV viral particles recovered from infected patients are similar to those recovered from infected primates. The virus contains a single-strand, positive-sense RNA genome of approximately 7.5 kb. The nucleotide sequence described here comprises 7194 bases excluding the poly(A) tail. If the 3' stretch of adenosine residues (at least 150–200 nucleotides) is included, the determined sequence agrees well with the genome size originally estimated by Northern hybridization studies (Reyes et al., 1990).

Open reading frame analysis of the nucleotide sequence revealed two major positive-polarity ORFs. A portion of ORF1 appears to encode the RDRP gene of the virus. The highly conserved amino acid residues, including the invariant GDD tripeptide found in all posi-

tive-strand animal and plant RNA viruses, can be located in the deduced amino acid sequence (Reyes et al., 1990; Fry et al., 1991). Additional evidence for the encoded polyprotein having a function in viral replication is provided by the presence of conserved motifs involved in purine NTPase activity found in a variety of cellular and viral helicases (Geider and Hoffman-Berling, 1981). These helicases promote the unwinding of DNA, RNA, or DNA-RNA duplexes required for genome replication, recombination, repair, and transcription. The deduced amino acid sequence of ORF2 suggests that it encodes a capsid-like peptide following the canonical signal sequence at its 5' end. ORF2 would appear to be the major ORF encoding the viral structural protein(s).

An identified immunoreactive epitope in ORF3 indicates that the virus utilizes all three positive-polarity frames for encoding viral proteins (Yarbough et al., 1991; Reyes et al., 1991c). This pattern of gene expression employed by HEV has not been described in the various families of single-stranded positive-sense, nonenveloped RNA viruses affecting humans or animals. Among the enveloped RNA viruses, the structural proteins of rubella virus and certain alphaviruses are found in a different reading frame from those encoding the nonstructural proteins and are also expressed from a subgenomic 3' end transcript (Ou et al., 1982). The presence of HEV-specific subgenomic RNAs localized to the 3' one-third of the genome suggests that these may be the transcripts from which these 3' end ORFs are expressed and is indicative of a unique expression strategy among nonenveloped positive-sense RNA viruses infecting humans. The mechanism by which these subgenomic transcripts are generated is unknown. The differential abundance of the various messages (i.e., 7.6 kb > 2 kb > 3.7 kb; see Fig. 5) does, however, suggest active transcriptional regulation rather than genomic RNA fragmentation as the means by which these subgenomic messages are generated. This would in turn imply the existence of an internal RNA initiation sequence and expression from the anti-genomic strand. Experiments are currently in progress to map the 5' ends of these subgenomic transcripts. We at this time, however, cannot exclude other mechanisms of expression for ORF2 and ORF3 including frameshifting or internal translation initiation, although there is little evidence for the latter among other positive-sense RNA viruses (Morch and Haenni, 1987). It is also possible that complex RNA splicing could account for these subgenomic messages although there is evidence that Northern hybridization probes from the extreme 5' end failed to detect hybridization to the 3.7- and 2.0-kb messages (A. W. Tam, unreported).

HEV

| NONSTRUCTURAL | | STRUCTURAL | AAAA(n) |

HELI   POL / IRE   IRE
         SS

	GENOME ORGANIZATION	ENVELOPE	SUBGENOMIC RNA(s)	POLY-A
HEPATITIS E VIRUS	5'-NS-S-3'	NO	TWO	YES
CALICIVIRIDAE	5'-NS-S-3'	NO	YES	YES
PICORNAVIRIDAE	5'-S-NS-3'	NO	NO	YES
TOGAVIRIDAE	5'-NS-S-3'	YES	ONE	YES
FLAVIVIRIDAE	5'-S-NS-3'	YES	NO	NO
CORONAVIRIDAE	5'-NS-S-3'	YES	MULTIPLE	YES

FIG. 6. HEV genomic organization: The proposed organization of the HEV genome is presented with the nonstructural genes encoded by the 5' ORF1 and the structural genes located at the 3' end of the genome (ORF2 and possibly ORF3). The genomic organization, nature of the virus particle (enveloped or nonenveloped), presence of subgenomic messages, and the presence of a 3' terminal poly(A) addition is compared for the various positive-sense, single-stranded RNA virus families. The relative locations of the various virus sequence motifs is also indicated, including the: HEL, putative helicase motif or NTP binding domain; POL, RNA-directed RNA polymerase; SS, signal sequence; IRE, immunoreactive epitope. S, structural gene coding region; NS, nonstructural gene coding region.

It is postulated from the proposed genomic organization of HEV, as presented in Fig. 6, that the nonstructural viral proteins are translated from the full-length genomic RNA. The 5' nonstructural/3' structural genomic organization of HEV is similar to that found in the alphavirus, rubivirus, and coronavirus families (see Fig. 6). There is an absence of any significant homology with these enveloped viruses at both the nucleotide and amino acid levels (excluding the canonical amino acid residues noted above for the nonstructural gene products). Immune electron microscopy has clearly established that the virions of HEV are 27- to 34-nm nonenveloped viral particles and are therefore clearly distinguished from these enveloped viruses. Picornaviruses are small nonenveloped, single-stranded, positive-sense, polyadenylated RNA viruses. The various members of the picornaviridae, however, exhibit vastly different genomic organization (Siddell, 1987). HEV has been shown to be unrelated antigenically and biophysically to picornaviruses (Arankalle, 1988).

It was previously hypothesized that HEV is calicivirus-like based on the biophysical characterization of viral particles (Bradley and Balayan, 1988; Bradley et al., 1988a,b). Recently the nucleotide sequence of a large portion of the nonstructural gene region of feline calicivirus (FCV) has become available for comparison to HEV (Neill, 1990). Although having a similar overall genomic organization to that of HEV with 5'-nonstructural and 3'-structural genes, it is clear that FCV shares a higher degree of similarity with picornaviruses in the recognized nonstructural gene motifs. The proposed gene order for the nonstructural polypeptides in FCV is 2C (NTP-binding), 3C (cysteine protease), followed by the 3D gene (RDRP). The distance between the NTP-

binding site motif A and the GDD triplet of the RDRP is 1100 amino acids in FCV compared to the 568 amino acids in HEV. In addition there is no evidence in the HEV sequence for an intervening cysteine protease-like region as recognized in FCV. These findings would further suggests that HEV represents either the prototype member of an as yet unclassified novel virus family or perhaps a separate genus within the caliciviridae. It is too early, however, to propose a definitive classification of HEV beyond the hypothesis presented here based on the proposed genetic organization and expression strategy of the virus.

ACKNOWLEDGMENT

REFERENCES

ANDJAPARIDZE, A. G., BALAYAN, M. S., SAVINOV, A. P., BRAGINSKY, D. M., POLESCHUK, V. F., and ZAMYATINA, N. A. (1986). Fecal-oral transmitted non-A, non-B hepatitis induced in monkeys. Vopr. Virusol. 1, 73–80.

ARANKALLE, V. A., TICEHURST, J., SREENIVASAN, M. A., et al. (1988). Aetiological association of a virus-like particle with enterically transmitted non-A, non-B hepatitis. Lancet 1, 550–554.

AVIV, H., and LEDER, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligo-thymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA 69, 1408–1412.

BALAYAN, M. S., ANDJAPARIDZE, A. G., SAVINSKAYA, S. S., et al. (1983). Evidence for a virus in non-A, non-B hepatitis transmitted via the fecal-oral route. Intervirology 20, 23–31.

BRADLEY, D. W. (1990a). Hepatitis non-A, non-B viruses become identified as hepatitis C and E viruses. Prog. Med. Virol. 37, 101–135.

BRADLEY, D. W. (1990b). Enterically-transmitted non-A, non-B hepatitis. Br. Med. Bull. 46, 442–461.

BRADLEY, D. W., and BALAYAN, M. S. (1988). Virus of enterically transmitted non-A, non-B hepatitis. Lancet 1, 819.

BRADLEY, D. W., and MAYNARD, J. E. (1986). Etiology and natural history of post-transfusion and enterically-transmitted non-A, non-B hepatitis. Sem. Liver Dis. 6, 56–66.

BRADLEY, D. W., KRAWCZYNSKI, K., COOK, E. H., JR., et al. (1987). Enterically transmitted non-A, non-B hepatitis: serial passage of disease in cynomolgus macaques and tamarins and recovery of disease-associated 27- to 34-nm viruslike particles. Proc. Natl. Acad. Sci. USA 84, 6277–6281.

BRADLEY, D. W., KRAWCZYNSKI, K., COOK, E. H., JR., et al. (1988a). Enterically transmitted non-A, non-B hepatitis: Etiology of disease and laboratory studies in nonhuman primates. In "Viral Hepatitis and Liver Disease" (A. J. Zuckerman, Ed.), pp. 138–147, A. R. Liss, New York.

BRADLEY, D. W., ANDJAPARIDZE, A., COOK, E. H., JR., et al. (1988b). Aetiological agent of enterically transmitted non-A, non-B hepatitis. J. Gen. Virol. 69, 731–738.

CATHALA, G., SAVONRET, J.-F., MENDEZ, R., et al. (1983). A method for isolation of intact translationally active ribonucleic acid. DNA 2, 329–335.

CHOO, Q.-L., KUO, G., WEINER, A. J., OVERBY, L. R., BRADLEY, D. W., and HOUGHTON, M. (1989). Isolation of cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science 244, 359–362.

DALGARNO, L., RICE, C. M., and STRAUSS, J. H. (1983). Ross river virus 26S RNA: Complete nucleotide sequence and deduced sequence of the encoded structural proteins. Virology 129, 170–187.

FEINSTONE, S. M., KAPIKIAN, A. Z., PURCELL, R. H., ALTER, H. J., and HOLLAND, P. V. (1975). Transfusion-associated hepatitis not due to viral hepatitis type A or B. N. Engl. J. Med. 292, 767–770.

FREY, T. K., and MARR, L. D. (1988). Sequence of the region coding for virion proteins C and E2 and the carboxy terminus of the nonstructural proteins of rubella virus: Comparison with alphaviruses. Gene 62, 85–99.

FRY, K. E., TAM, A. W., SMITH, M. M., et al. (1991). Hepatitis E virus (HEV): Strain variation in the nonstructural gene region encoding consensus motifs for an RNA-dependent RNA polymerase and an ATP/GTP binding site, Virus Genes, in press.

GEIDER, K., and HOFFMAN-BERLING, H. (1981). Proteins controlling the helical structure of DNA. Annu. Rev. Biochem. 50, 233–260.

GENTRY, G. A. (1985). Locating a nucleotide binding site in the thymidine kinase of vaccinia virus and of herpes simplex virus by scoring triply aligned protein sequences. Proc. Natl. Acad. Sci. USA 82, 6815–6819.

GORBALENYA, A. E., KOONIN, E. V., DONCHENKO, A. P., and BLINOV, V. M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. Nucleic Acids Res. 17, 4713–4730.

HUANG, C.-C., NGUYEN, D., FERNANDEZ, J., et al. (1991). Hepatitis E virus: Cloning of a Mexican strain and comparison to the Burmese strain, in preparation.

KAMER, G., and ARGOS, P. (1984). Primary structural comparison of RNA-dependent polymerases from plant, animal, and bacterial viruses. Nucleic Acids Res. 12, 7269–7282.

KHUROO, M. S. (1980). Study of an epidemic of non-A, non-B hepatitis: Possibility of another human hepatitis virus distinct from post-transfusion non-A, non-B type. Am. J. Med. 68, 818–824.

KLEIN, P., KANEHISA, M., and DELISI, C. (1985). The detection and classification of membrane-spanning proteins. Biochem. Biophys. Acta 815, 468–476.

KRAWCZYNSKI, K., and BRADLEY, D. W. (1989). Enterically transmitted non-A, non-B hepatitis: Identification of virus-associated antigen in experimentally infected cynomolgus macaques. J. Infect. Dis. 159, 1042–1049.

KUBO, Y., TAKEUCHI, K., BOONMAR, S., et al. (1989). A cDNA fragment of hepatitis C virus isolated from an implicated donor of post-transfusion non-A, non-B hepatitis in Japan. Nucleic Acids Res. 17, 10,367–10,372.

KUO, G., CHOO, Q.-L., ALTER, H. J., et al. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. Science 244, 362–364.

MAENO, M., KAMINAKA, K., SUGIMOTO, H., et al. (1990). A cDNA clone closely associated with non-A, non-B hepatitis. Nucleic Acids Res. 18, 2685–2689.

MCKNIGHT, S. L., GAVIS, E. R., and KINGSBURY, R. (1981). Analysis of transcriptional regulatory signals of the HSV thymidine kinase gene: Identification of an upstream control region. Cell 25, 385–398.

MORCH, M.-D., and HAENNI, A. L. (1987). Organization of plant virus genomes that comprise a single RNA molecule. In "The Molecular Biology of Positive Strand RNA Viruses" (D. J. Rowlands, M. A. Mayo, and B. W. J. Mahy, Eds.), pp. 153–175. Academic Press, London.

MULLIS, K. B., and FALOONA, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase catalysed chain reaction. In "Methods in Enzymology" (R. Wu, Ed.), Vol 155, pp. 335–350. Academic Press, New York.

NEILL, J. D. (1990). Nucleotide sequence of a region of the feline calicivirus genome which encodes picornavirus-like RNA-dependent RNA polymerase, cysteine protease and 2C polypeptides. Virus Res. 17, 145–160.

NUSSINOV, E. (1981). Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J. Mol. Biol. 149, 125–131.

OU, J.-H., RICE, C. M., DALGARNO, L., et al. (1982). Sequence studies of several alphavirus genomic RNAs in the region containing the start of the subgenomic RNA. Proc. Natl. Acad. Sci. USA 79, 5235–5239.

PRINCE, A. M., BROTMAN, B., and GRADY, G. F. (1974). Long-incubation post-transfusion hepatitis without serological evidence of exposure to hepatitis B virus. Lancet 2, 241–246.

PURCELL, R. H., and TICEHURST, J. R. (1988). Enterically transmitted non-A, non-B hepatitis: epidemiology and clinical characteristics. In "Viral Hepatitis and Liver Disease" (A. J. Zuckerman, Ed.), pp. 131–137. A. R. Liss, New York.

REYES, G. R., and BAROUDY, B. M. (1991). Molecular biology of non-A, non-B hepatitis agents: the hepatitis C and hepatitis E viruses. In "Advances in Virus Research" (K. Maramorosch, F. A. Murphy, and A. J. Shatkin, Eds.), Vol. 40, pp. 55–101. Academic Press, Orlando.

REYES, G. R. (1991a). Molecular Cloning of the hepatitis E Virus. In "Viral Hepatitis and Liver Disease" (F. B. Hollinger, S. M. Lemon, and H. S. Margolis, Eds.), pp. 514–517. Williams & Wilkens, Baltimore.

REYES, G. R., HUANG, C.-C., YARBOUGH, P. O., et al. (1991b). Hepatitis E virus (HEV): Epitope mapping and detection of strain variation. In "Viral Hepatitis C, D and E" (T. Shikata, R. H. Purcell and T. Uchida, Eds.), pp. 237–245. Elsevier, Amsterdam.

REYES, G. R., YARBOUGH, P. O., TAM, A. W., PURDY, M. A., HUANG, C.-C., KIM, J. K., BRADLEY, D. W., and FRY, K. E. (1991c). Hepatitis E Virus (HEV): The novel agent responsible for enterically transmitted non-A, non-B hepatitis. Gastroenterol. Jpn. 26(suppl. 3), 142–147.

REYES, G. R., KIM, J. P., et al. (1991d), in preparation.

REYES, G. R., PURDY, M. A., KIM, J. P., et al. (1990). Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. Science 247, 1335–1339.

RICE, C. M., LENCHES, E. M., EDDY, S. R., et al. (1985). Nucleotide sequence of yellow fever virus: Implications for flavivirus gene expression and evolution. Science 229, 726–733.

SAIKI, R. K., SCHARF, S., FALOONA, F., *et al.* (1985). Enzymatic amplification of b-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354.

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

SHEPHERD, J. C. W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA* **78**, 1596–1600.

SIDDELL, S. (1987). The organization and expression of coronavirus genomes. *In* "The Molecular Biology of the Positive Strand RNA Viruses" (D. J. Rowlands, M. A. Mayo, and B. W. J. Mahy, Eds.), pp. 117–127. Academic Press, London.

STRAUSS, J. H., and STRAUSS, E. G. (1988). Evolution of RNA viruses. *Annu. Rev. Microbiol.* **42**, 657–683.

TABOR, E. (1985). The three viruses of non-A, non-B hepatitis. *Lancet* **1**, 743–745.

TAM, A. W., SMITH, M. M., FRY, K. E., and LARRICK, J. W. (1989). Construction of cDNA libraries from small numbers of cells using sequence independent primers. *Nucleic Acids Res.* **17**, 1269.

WONG, D. C., PURCELL, R. H., SREENIVASAN, M. A., PRASAD, S. R., and PAVRI, K. M. (1980). Epidemic and endemic hepatitis in India: evidence for a non-A, non-B hepatitis virus aetiology. *Lancet* **2**, 876–879.

YARBOUGH, P. O., TAM, A. W., FRY, K. E., *et al.* (1991). Hepatitis E virus: Identification of type common epitopes, *J. Virol.* in press.