Virus Genes 9:1, 23–32, 1994

A New Chinese Isolate of Hepatitis E Virus: Comparison with Strains Recovered from Different Geographical Regions*

SHURONG YIN[1,2] ROBERT H. PURCELL[1], & SUZANNE U. EMERSON[1]

[1]*Hepatitis Viruses Section, Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland and* [2]*Institute of Microbiology and Epidemiology, Beijing, Peoples' Republic of China*

Received October 15, 1993; accepted January 7, 1994

Requests for reprints should be addressed to Shurong Yin, LTVB/DCE/NCI/NIH, Building 41, Room D503, 9000 Rockville Pike, Bethesda, MD20892.

Abstract. The full-length cDNA of a new Chinese strain (KS2-87) of Hepatitis E virus (HEV) has been constructed and sequenced. The 5′ noncoding region of KS2-87 is 26 nucleotides in length, which is one nucleotide shorter than that of HEV (B1) (Burma) and 23 nucleotides longer than that of HEV (Mexico). Comparison of the nucleotide and amino acid sequences of KS2-87 with all other published HEV sequences showed that KS2-87 was closer to two other Chinese strains (CHT-88, CHT-87) and SAR-55 (Pakistan) than to HEV (B1) and HEV (B2) (Burma) or HEV (Mexico). Comparisons of partial sequences of genes encoding a nonstructural and a structural protein revealed the existence of genetically related groups of HEV within geographical regions, whereas larger nucleotide differences were seen among isolates that were more geographically and epidemiologically distant.

Key words: HEV genome, PCR, genetic heterogeneity

Introduction

Viral hepatitis E, formerly called epidemic or enterically transmitted non-A, non-B hepatitis, is prevalent in many developing countries in Asia and Africa, and has been found in Mexico (1,2). In Asia and Africa, hepatitis E epidemics occur frequently and sporadic hepatitis E has been found in the period between outbreaks. Hepatitis E is sporadically seen also in industrialized countries as an "imported hepatitis" via immigrants and tourists (3). Hepatitis E virus (HEV), the major etiologic agent of enterically transmitted non-A, non-B hepatitis is a 27–34 nm diameter, nonenveloped virus with physicochemical properties similar to caliciviruses. Cynomolgus and rhesus macaques have been infected with HEV of human origin and have been extensively used for the generation of virus-positive source materials (4,5).

The full-length genome of a Burmese strain [HEV(B1)] (Rangoon, Burma, 1982) was the first to be identified and sequenced (6). There was little similarity between the predicted amino acid sequences of HEV and calciviruses, Norwalk virus, or other human viruses. The HEV genome consists of a positive-sense, single-stranded, polyadenylated RNA molecule of approximately 7.5 kb and encodes three open reading frames (ORF1, ORF2, and ORF3) (7). The 5′ ORF (ORF1) of the HEV genome appears to encode the nonstructural proteins and the 3′ ORF (ORF2) encodes the viral structural proteins (8).

More recently, strains of HEV from Pakistan (SAR-55; Sargohda, Pakistan, 1987) (9,10) and Mexico [HEV (M); Telixtac, Mexico, 1986] (11,12) and an additional strain from Burma

*Sequence accession No. L25595.

[HEV (B2); Rangoon, Burma, 1986] (13,14) have been molecularly cloned and sequenced, and two strains of HEV from China (CHT-88; Hetian, China, 1988) (15) and (CHT-87; Hetian, China, 1987) (16) have been directly sequenced from PCR products. Limited sequence comparison of eight new Chinese HEV isolates previously suggested significant homology of all eight to SAR-55 (17). However, sequence comparisons among three strains of HEV had identified a potential hypervariable region in ORF1 that exhibited only 58% and 54% similarity of nucleic acid sequence between HEV (M) and HEV (B1) and SAR-55, respectively. Furthermore, there was only 13% similarity of amino acid sequence between HEV (M) and the other two strains.

Strain variation in HEV could impose some problems in the diagnosis of hepatitis E and in the development of vaccines. It is therefore of biological and epidemiological importance to analyze the extent of sequence diversity of HEV from multiple sources. In the work described here, the complete nucleotide sequence of a new Chinese strain of HEV (KS2-87; Kashi, China, 1987) was determined. This sequence was then compared with the published complete sequences of the other hepatitis E viruses in order to identify areas of conservation and diversity that exist in the hepatitis E viruses in order to identify areas of conversation and diversity that exist in the hepatitis E virus pool. We also analyzed the genetic relatedness of selected regions of ORF1 and ORF2 of 10 Chinese isolates, HEV (B1), HEV (B2), SAR-55, and HEV (M).

Materials and Methods

Virus

The virus used in this experiment was from a bile specimen collected by autopsy from an HEV-infected rhesus monkey representing the second rhesus passage of virus. The virus was originally obtained from the feces of a patient who was infected during a hepatitis E epidemic in Kashi prefecture in the Xinjiang Uighur Autonomous Region (XUAR) in China in 1987. When we began this work sequence data for other Chinese strains were not available.

Synthesis of cDNA Clones

Eleven overlapping cDNA clones spanning the entire HEV genome were generated by the insertion of reverse transcription-polymerase chain reaction (RT-PCR)-derived products into pGEM-3Z vectors (Promega). The information required for the synthesis of the oligonucleotide primers used in the generation of individual cDNA clones was obtained from the nucleotide sequence of the SAR-55 and CHT-88 strains of HEV. Conditions for RNA extraction and RT-PCR have been described (17). Amplification was performed for 35 cycles in an automatic thermocycler (Perkin-Elmer Cetus) using unique conditions for each region to be amplified. Two rounds of PCR were used with the same conditions except that a 5 μl aliquot of the first PCR reaction was amplified in the second PCR reaction using internal sense and antisense primers.

The PCR products were phenol:chloroform extracted and ethanol precipitated, digested with the appropriate restriction endonucleases (New England Biolabs, Beverly, MA), separated by electrophoresis of an 0.8% agarose gel, and purified using Geneclean (Bio101, La Jolla, CA). Purified cDNA fragments were ligated into the appropriate sites of the pGEM-3Z vector. Double-stranded cDNA clones were sequenced with the Sequenase version 2.0 kit (United States Biochemicals, Cleveland, OH) following alkali denaturation.

Cloning of 3′ and 5′ Termini

The 3′ end of the HEV genome was isolated by using oligo(dT) with a 5′ artificial restriction site (XbaI) as the reverse primer and the forward primers 5′-GGTCACTCGACTGGATCCAGGTCACAC-3′. At positions 6680–6706 and 5′-CAGGAATTCCTAAAGCCGGGTACC-3′ at positions 6794–6817 for RT-PCR. The 5′ end of the HEV genome was isolated by using a 5′ RACE system kit (Gibco BRL, Gaithersburg, MD). A 540 nucleotide fragment of the 5′ end of the virus genome was reverse transcribed from RNA extracted from 50 μl of bile. After removal of RNA templates with RNase H and purification of the first strand products with the Glassmax cartridge, poly(dC), poly(dG), or poly(dA) tails

were added using terminal deoxy nucleotidyl transferase according to the instruction manual of the 5′ Race system kit. The tailed product was amplified by PCR with anchor primer G (supplied in the kit), anchor primer C, or anchor primer T (synthesized in our laboratory). The second round of PCR amplification was conducted as already described. PCR products were cloned into the pGEM-3Z vector, and at least 10 cDNA clones from each homopolymeric tailing were then sequenced.

Computer Analysis

Computer analysis was through the BIONET National Computer Resource for Molecular Biology. The data for nucleotide sequences of HEV (B1), HEV (B2), HEV (M), CHT-88, and CHT-87 were from GenBank nucleotide databases. The nucleotide sequence of SAR-55 was provided by Dr. S. Tsarev. PCGENE was used to study genetic relationships. cDNA sequences were aligned in all pairwise combinations, and the respective number of identical nucleotides was calculated. A dendrogram was then constructed using the CLUSTAL program.

Results

Nucleotide Sequence of the KS2-87 Strain of HEV

Eleven overlapping cDNA clones that spanned the entire HEV genome were obtained by RT-PCR amplification of bile from RHE-2 infected with the KS2-87 strain of HEV (17) (Fig. 1). The clones, which ranged in size from 263 to 1658 base pairs, were sequenced. The final nucleotide sequence was based on information derived from at least two independent cDNA clones. Differences between two clones were resolved by sequencing a third clone. The region between nucleotides 5081 and 5829 was amplified by two independent RT-PCR with different primer sets and cloned and sequenced. The consensus sequence of some fragments was confirmed by the direct sequencing of the PCR products. A total of 7193 nucleotides, except adenosine residues located at the very 3′ end, comprised the KS2-87 genome, compared to 7194 nucleotides for the genome of HEV (B1) and 7170 nucleotides for that of HEV (M). The nucleotide composition of the KS2-87 genomic RNA was 17% A, 32% C, 26% G and 25% U, conferring an overall G + C content of 58%. Comparison of nucleotide sequences of the entire genome showed that KS2-87 was much more closely related to CHT-88, CHT-87, and SAR-55 (98.7%, 98.7%, and 98.9% homology, respectively) than to HEV (B1), HEV (B2), and HEV (M) (93.6%, 93.0%, and 71.3%).

Analysis of the Noncoding Regions of the Virus Genomes

The 5′ noncoding region of KS2-87 strain was obtained using three different nucleotides (A, C, or G) for homopolymeric tailing in order to confirm the identity of the first nucleotide. With all three homopolymers, a 5′ noncoding region 26 nucleotides in length was identified, which is 1 nucleotide shorter than that of HEV (B1) and 23 nucleotides longer than that reported for HEV (M) (Fig. 2).

There were 65 nucleotides in addition to the poly(A) tail in the 3′ noncoding region of KS2-87. There were three nucleotide differences between KS2-87 and HEV (B1) at positions 7151 (U > C), 7164 (U > A), and 7183 (U > C). There were no deletions or insertions in KS2-87 when compared with HEV (B1), but there were 15 nucleotide substitutions and 9 nucleotide insertions in the 3′ noncoding region of HEV (M) compared to KS2-87 and HEV (B1) (Fig. 2). The complete sequence of these regions was not available for HEV (B2), SAR-55, or the other two Chinese strains.

Analysis of the Three Open Reading Frames

Computer analysis identified the same three ORFs in KS2-87 as were found in all six of the other HEV strains that have been almost totally sequenced. All of the translation start and stop signals were conserved. The percent nucleotide and amino acid identities in each of three open reading frames among KS2-87, CHT-88, CHT-87, HEV (B1), HEV (B2), SAR-55, and HEV (M) are given in Table 1. The ORF1 of

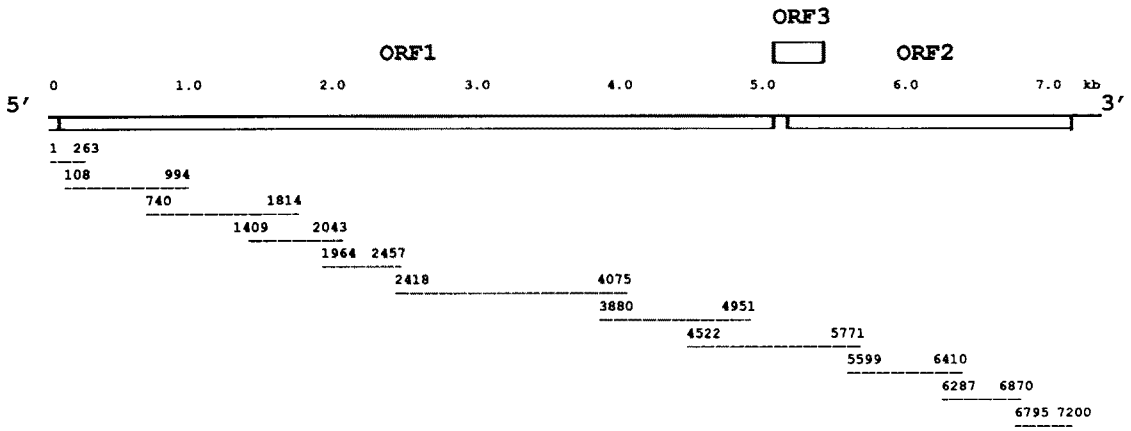

Fig. 1. Map of HEV genome and clones of PCR fragments that include the length of oligonucleotide primers. The putative RNA-dependent RNA polymerase and helicase domains are located between nt 3645 and 5105 and nt 2903 and 3602, respectively. The putative hypervariable region described by Tsarev et al. (9) is located between nt 2010 and 2330.

HEV appears to encode an RNA-dependent RNA polymerase, since it contains the canonical gly-asp-asp (GDD) tripeptide found in all positive-stranded animal and plant RNA viruses (18). In addition, ORF1 also contains conserved motifs (GVPGSGKS and DEAP) associated with purine NTPase activity found in a variety of cellular and viral helicases (19).

Comparison of the sequences of ORF1 of the seven HEV strains listed in Table 1 showed that the cDNA nucleotide (nt) sequences differed by 1.0–25.7%. However, the degree of sequence variation was not uniform throughout the whole open reading frame, and the highest degree of relatedness was at the extreme 5′ end. The highest degree of sequence diversity was between nt 2010 and 2330 of KS2-87; between nt 2011 and 2331 of HEV (B1), HEV (B2), SAR-55, CHT-88, and CHT-87; and between nt 2121 and 2329 of HEV (M) (11). The alignment for ORF1 showed that the KS2-87 sequence was very similar to that of CHT-88 and, CHT-87, and closer to that of SAR-55 than to that of HEV (B1), HEV (B2), or HEV (M). There were 10 (0.6%), 15 (0.9%), 7 (0.4%), 26 (1.5%), 38 (2.2%), or 280 (16.5%) predicted amino acid differences in ORF1 between KS2-87 and CHT-88, CHT-87, SAR-55, HEV (B1), HEV (B2), or HEV (M), respectively. Many of these differences [2 for CHT-87, 4 for SAR-55, 13 for HEV (B1), 16 for HEV (B2) and 61 for HEV (M)] clustered in the region previously termed *hypervariable* [between nt 2011 and 2331 of Asian strains and between nt 1987 and 2307 of HEV (M)] to give predicted amino

```
5′noncoding region

HEV(B):     AGGCAGACCACATATGTGGTCGATGCC
KS2-87:      GGCAGACCACATATGTGGTCGATGCC
HEV(M):                              GCC

3′noncoding region

HEV(B1) TTTATTTGCTTGTGCCC   CCCTTCTTTCTG       TTGCTTATTTCTCATTTCTGCGTTCC
KS2-87  .................   .....T......       ......T..................T
HEV(M)  ........GC.......ACCTA...A.A....CTGATT.CCT.......CTT.....CG..C..

HEV(B1) GCGCTCCCTGAAAAA
KS-87   ...............
HEV(M)  ...............
```

Fig. 2. Comparison of nucleotide sequence in 5′ and 3′ noncoding regions of HEV (B1), KS2-87, and HEV (M).

Table 1. Percent amino acid and nucleotide identity

	Length								
	nt	aa	Sar-55	HEV (B1)	HEV (B2)	CHT-88	CHT-87	KS2-87	HEV (M)
			Amino Acids						
Sar-55									
ORF1	5079	1693		98.4	97.7	99.2	99.1	99.6	83.8
ORF2	1980	660		99.4	98.6	99.7	99.1	99.4	93.3
ORF3	369	123		100.0	98.4	100.0	98.4	98.4	87.0
HEV (B1)									
ORF1	5079	1693	93.1		98.6	98.2	98.0	98.5	83.8
ORF2	1980	660	93.8		98.9	99.7	99.1	99.1	93.2
ORF3	369	123	98.9		98.4	100.0	98.4	98.4	87.0
HEV (B2)									
ORF1	5079	1693	92.6	98.5		97.5	97.2	97.8	83.2
ORF2	1980	660	93.1	98.5		98.9	98.3	98.3	92.4
ORF3	369	123	98.1	99.2		98.4	96.7	96.7	87.0
CHT-88									
ORF1	5079	1693	98.2	93.5	93.1		98.8	99.4	83.4
ORF2	1980	660	98.2	94.4	93.7		99.1	99.4	93.3
ORF3	369	123	99.7	99.2	98.4		98.4	98.4	87.0
CHT-87									
ORF1	5079	1693	98.4	93.3	92.8	98.1		99.1	83.6
ORF2	1980	660	98.1	94.4	93.5	98.6		98.8	92.9
ORF3	369	123	99.2	98.1	97.3	98.9		96.7	85.4
KS2-87									
ORF1	5079	1693	99.0	93.1	92.7	98.6	98.7		83.6
ORF2	1980	660	98.3	94.3	93.6	98.8	98.7		93.0
ORF3	369	123	99.2	98.1	97.3	98.9	98.4		85.4
HEV (M)									
ORF1	5073	1691	74.5	74.4	74.3	74.4	74.6	74.5	
ORF2	1977	659	81.5	81.2	81.1	81.3	81.1	81.4	
ORF3	369	123	90.8	89.7	90.1	90.5	90.1	89.9	
			Nucleotides						

acid differences from KS2-87 of 1.9% for CHT-87, 3.7% for SAR-55, 12.1% for HEV (B1), 15.0% for HEV (B2), and 57.0% for HEV (M). In contrast, there were no differences predicted in this region between KS2-87 and CHT-88. Of interest is that the nucleotide sequence of the region between nt 2681 and 2981 of the CHT-88 strain was almost the same as that of HEV (B1) and HEV (B2) and very different from those of other Chinese strains and the SAR-55 strain, although the rest of the genome of the CHT-88 strain was very similar to that of CHT-87, KS2-87, and SAR-55 (Fig. 3; Table 1).

The computed translation of ORF2 of HEV previously predicted a novel polypeptide not identified in the PIR protein database (7). Since this polypeptide contains a high basic amino acid content, similar to that seen with other viral capsid proteins, it is thought to be involved in the encapsidation of the genomic RNA (20). The complete open reading frame (ORF2) for KS2-87, CHT-88, CHT-87, SAR-55, HEV (B1), and HEV (B2) consists of 1980 bases encoding 660 amino acids, while that for the HEV (M) consists of 1977 bases encoding 659 amino acids. The alignment of the nucleotide sequences of ORF2 shows that although there were similar nucleotide identities in ORF2 and ORF1 between KS2-87 and the other strains, there were fewer changes of predicted amino acids in ORF2 compared to ORF1 (Table 1).

ORF3 partially overlaps ORF1 and ORF2, and

```
SAR-55    CTCCTCGGGACCGGCATATACCAGGTGCCGATCGGTCCCAGTTTTGACGCCTGGGAGCGG
HEV(B1)   ..........................................C.....................
HEV(B2)   ..........................................C.....................
CHT-88    ..........................................C.....................
CHT-87    ..............................................................
KS2-87    ..............................................................
HEV(M)    ...T.A..C..T.....T...........TG.TA..TTG........T...........

SAR-55    AATCACCGCCCCGGGGACGAGTTGTACCTTCCTGAGCTTGCTGCCAGATGGTTCGAGGCC
HEV(B1)   ..C.............T.....................................T......
HEV(B2)   ..C.............T.....................................T......
CHT-88    ..C.....T........T.........A..........................T......
CHT-87    ..............T.....................................T.....
KS2-87    ..............................................................
HEV(M)    ..C........GTTT......C.T.....AA.A.....G..G..TC.G.....T..AT..

SAR-55    AATAGGCCGACCTGCCCAACTCTCACTATAACTGAGGATGTTGCGCGGACAGCAAATCTG
HEV(B1)   ............C....G.........................A........G......
HEV(B2)   ............C....G.........................A........G......
CHT-88    ............C....G.........................A........G......
CHT-87    ..........................................A.....................
KS2-87    ..........................................A.....................
HEV(M)    ..CC.C..CGGTCAG..C..GT.G.AC............ACC..C..TG.G..C..C...

SAR-55    GCTATCGAACTTGACTCAGCCACAGACGTCGGCCGGGCCTGTGCCGGCTGTCGAGTCACC
HBV(B1)   ..C.....G..................T.........................G......
HBV(B2)   ..C.....G........T........T..............C...........G......
CHT-88    ..C.....G..................T.........................G......
CHT-87    ..............................................................
KS2-87    ..............................................................
HEV(M)    .CCC.G..G........C.GG.GT..A..A.....C..A........G...AA....GAG

SAR-55    CCCGGCGTTGTGCAGTACCAGTTTACCGCAGGTGTGCCTGGATCCGGCAAGTCCCGCTCTAT
HEV(B1)   ...........T.............T..................................
HEV(B2)   ...........T.............T..................................
CHT-88    ...........T.............T..................................
CHT-87    ..............................................................
KS2-87    ..............................................................
HEV(M)    ..T.........G...T........A.....................................
```

Fig. 3. Comparison of the nucleotide sequence of nt 2680–2981 [Sar-55, HEV (B1), HEV (B2), CHT-88, CHT-87, and KS2-87] and nt 2650–2951 of HEV (M).

is the smallest (369 bp) and most highly conserved open reading frame. Several experiments confirmed the substitution of T for C at position 5384 of KS2-87. This change, which was not present in any of the other strains, including CHT-88 and CHT-87, should induce amino acid changes in both ORF2 (A > V) and ORF3 (P > S). A large 246 bp deletion in ORF3 has been reported for some Indian strains of HEV (21). However, Sergei Tsarev had found that PCR artifacts were common in this region but that DNA products of the predicted length could be obtained for the SAR-55 strain if 10% glycerol or DMSO was incorporated into the PCR reaction (Tsarev, personal communication). When standard PCR conditions were used, we detected short PCR products from this region consistent with a deletion, but we found that inclusion of 5% (final concentration) formamide (Fluka, Switzerland) in the PCR reaction resulted in longer PCR products that did not contain a deletion or insertion of nucleotides in ORF3 for KS2-87 compared to HEV (B1), HEV (B2), SAR-55, CHT-88, CHT-87, and HEV (M) (Fig. 4).

Genetic Relatedness of HEV Strains

RNA sequences in a nonstructural region (nt 2001–2420) and in a structural region (nt 5651–6190) of the viruses isolated during the epidemics in the XUAR in China in 1987 and 1990 (12) were compared with those of independent isolates from the XUAR of China (CHT-88 and CHT-87) and with those of strains isolated in other countries [HEV (B1), HEV (B2), SAR-55, and HEV (M)] (Fig. 5). In general, isolates from a distinct geographic region clustered together with <2.0% genetic divergence between isolates irrespective of the year of isolation, whereas much larger nucleotide differences were seen between epidemiologically unrelated strains. The intraepidemic variation of nucleotides among KS1-87, KS2-87, CHT-88, and CHT-87 in these two regions of the genome was from 0.7% to 1.4%, and that be-

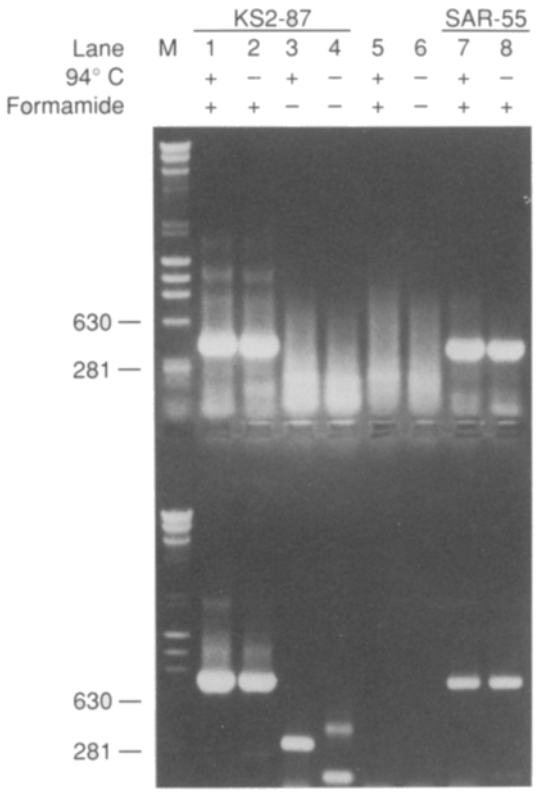

Fig.4. Agarose gel showing that formamide was required for accurate amplification of ORF-3 PCR products. RT was performed with (+) or without (−) prior denaturation of RNA at 94°C, and PCR was performed as usual with (+) or without (−) the addition of formamide to a 5% final concentration. One set of primers was designed to amplify the 413 base long region of ORF-3 between nt 5081 and 5493 (upper panel), and the other set was designed to amplify the 749 base long region of ORF-3 between nt 5081 and 5829 (lower panel). RNA was from KS2-87 virus (lanes 1–4) or Sar-55 (lanes 7, 8). Viral RNA was not added to the control (lanes 5, 6).

tween TU1-90, TU2-90, TU3-90, TU4-90, TU5-90, and TU6-90 was from 0% to 1.7%. The sequence of the non-Chinese strains differed from the consensus sequence of the nine Chinese isolates in the nonstructural region and in the structural region, respectively, by 2.1% and 2.6% in SAR-55, by 7.9% and 7.4% in HEV (B1), by 8.3% and 8.1% in HEV (B2), and by 36.5% and 20.1% in HEV (M).

Two dendrograms were constructed using the CLUSTAL program based on the absolute number of nucleotide substitutions in these two regions (Fig. 6). We identified three distinct branches within the dendrogram of the structural region (Fig. 6b). One branch consisted of SAR-55 and the Chinese strains of HEV, among which there was ~3% genetic difference. Within this branch there were three clusters: one contained the six TU strains isolated from the sera collected from six patients with acute hepatitis E in Turfan prefecture in XUAR of China in 1990, the second contained the CHT-88, KS1-87, and KS2-87 strains; and the third contained the SAR-55 and CHT-87 strains. The general structure of the dendrogram based on the sequence data from the putative hypervariable region (Fig. 6a) of the HEV genome was somewhat different from that constructed for the structural region (Fig. 6b) and did not show distinct sub-branch formation. HEV (B1) and HEV (B2), which differed by 1.7%, comprised a branch differing from the Chinese strains and SAR-55 by ~8.0%. The third branch, HEV (M), differed from the other groups of HEV isolates by ~35% in the nonstructural region and by ~20% in the structural region. Previous sequence analysis showed that the Mexican strain belonged to another genetic group of HEV characterized by a short 5′ noncoding region and many deletions and insertions when compared to the Asian strains of HEV (22).

Discussion

We have cloned and sequenced the full-length genome of a Chinese strain of HEV (KS2-87) recovered from a patient during a hepatitis epidemic in the Kashi prefecture of the XUAR of China. Although the genomic coding regions of five other HEV strains have been sequenced, the sequence of the 5′ noncoding region has been reported only for the HEV (B1) and HEV (M) strains. The ninefold difference reported in the length of the 5′ noncoding region of these two strains [27 vs. 3 nucleotides for HEV (B1) and HEV (M), respectively] is highly unusual, since the 5′ terminus of related single-strand RNA virus strains is generally conserved. For this reason, particular effort was expended to confirm the sequence of the 5′ noncoding region of KS2-87. Several experiments confirmed that the 5′ noncoding region of KS2-87 was identical to that of HEV (B1), except for containing 26 rather

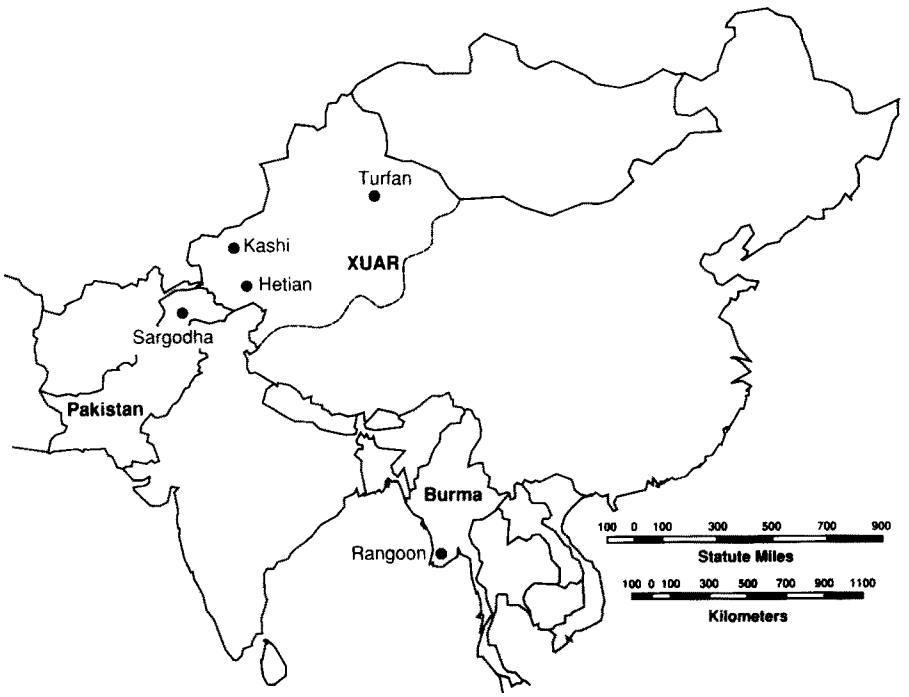

Fig. 5. Map of China where the samples studied were collected. The Sar-55 strain was isolated in Sargodha in Pakistan. HEV (B1) and HEV (B2) were isolated in Rangoon, Burma. The Mexican strain, HEV (M), is not shown.

than 27 nucleotides (Fig. 2). Since the nucleotide sequence of HEV (B1) and KS2-87 differ by 6.9% in ORF1 (Table 1), the viruses have accumulated sufficient numbers of mutations in the coding region to exhibit divergence, and the strict sequence conservation of the 5′ noncoding region most likely reflects strong selective pressures for translation or replication functions. This 5′ noncoding sequence conservation also extends to the three nucleotides identified for HEV (M). These comparisons provide a basis for suspecting that a large portion of the 5′ terminus of HEV (M) was missed during the cDNA cloning procedure and that the 5′ noncoding region of all HEV strains is similar in length and sequence to that of HEV (B1).

The sequence analysis showed that the KS2-87 strain was more closely related to the CHT-88, CHT-87, and SAR-55 strains of HEV than to HEV (B1), HEV (B2), and HEV (M). The comparison of partial nucleic acid sequences of 12 HEV isolates demonstrated the existence of genetically related groups of HEV strains within certain geographical regions. Although isolates KS1-87, KS2-87, CHT-88, CHT-87,

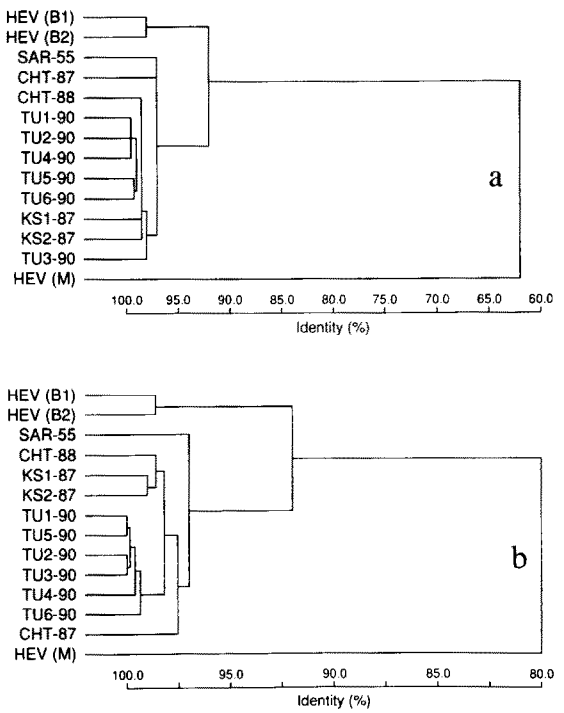

Fig. 6. Percent identity between the HEV nucleotide sequences. a: Hypervariable region (nt 2001–2420). b: Structural region (nt 5651–6190).

TU1-90, TU2-90, TU3-90, TU4-90, TU5-90, and TU6-90 were all recovered in the XUAR in China, the TU isolates were all from an area of the Turfan Prefecture that is far from the regions where the other four Chinese strains were isolated (Fig. 5). KS1-87 and KS2-87 viruses were isolated in the same area in the Kashi prefecture in 1987, so it is not surprising that they are similar. Although CHT-88 and CHT-87 were isolated in the Hetian prefecture in 1988 and 1987, they have been epidemiologically linked to strains KS1-87 and KS2-87 (23), so their similarity is also not surprising.

The SAR-55 strain that was isolated in Pakistan appears by sequence analysis to be more closely related to the Chinese strains than to either of the Burmese strains or the Mexican strain. The geographical region where the SAR-55 strain was isolated is much closer to the Chinese sites (Hetian prefecture in the XUAR) of large HEV epidemics (24) than to those where HEV (B1), HEV (B2), or HEV (M) were isolated. There are some religious and business relationships between the Pakistanis and the Uighur Tribe in the XUAR, which may explain why the SAR-55 and Chinese strains are so closely related. A similar genetic analysis of poliovirus isolates has shown that poliovirus could be segregated into epidemiologically distinct groupings that paralleled the geographic distribution of the isolates (25). Additional sequence data for HEV strains from other continents such as Africa and Europe may reveal more subtle evolutionary changes. Data such as these will be useful in determining the epidemiology of this newly defined virus.

Previously we had suggested that the HEV viral genome is relatively stable, since there were no nucleotide changes introduced into three large portions of the KS2-87 genome during three serial passages in rhesus monkeys (17). In addition, there were only 1.5% nucleotide and 1.3% amino acid differences found between the entire coding regions of the two Burmese strains that were isolated 4 years apart (7,13), and the three Chinese strains also displayed low levels of differences among themselves. Exceptions to this sequence conservation in the coding region were noted in two cases. The region designated a hypervariable region by Tsarev et al. (9) did show moderate amino acid divergence between strains of the Burmese branch compared to those of the Sar-55 branch, although the sequence of viruses within a branch was relatively well conserved. However, the amino acid sequence of the HEV (M) strain in this region was so different that there was only 43.0% homology in a 107 amino acid long sequence [nt 2010–2330 for KS2-87, nt 1987–2307 for HEV (M)] between KS2-87 and HEV (M).

Overall the data suggest that the HEV (M) strain diverged from the others a long time ago and indicate that this region may be exceptionally plastic. It will be interesting to determine whether this region of the genome encodes a functional protein. A second limited area of sequence heterogeneity was noted when comparing the coding sequences of CHT-88 with those of CHT-87 or KS2-87 (Fig. 3). For the majority of the genome, the sequence of all three viruses was closer to that of Sar-55 than to that of either Burmese strain. However, while the CHT-87 and KS2-87 strains were almost identical to Sar-55 in the region between nucleotides 2680 and 2981, the CHT-88 strain contained 16 nucleotide differences from Sar-55, and 14 of the 16 were also in both Burmese strains. The same mutations might have been selected in the Burmese and CHT-88 strains if those mutated sites could tolerate only the Sar-55 or the Burmese sequence. However, the sequence comparison with the HEV (M) strain (Fig. 3) shows that for four of the sites at least, a third nucleotide could be incorporated. Therefore, the data are most suggestive of genetic recombination between a Chinese and Burmese strain. Recombination is quite frequent for RNA viruses such as poliovirus (26), but additional studies will be required to determine whether HEV genomes undergo recombination.

The sequence comparisons to date have yielded valuable information about the distribution and diversity of HEV. The data are important for vaccine development but also may be useful for analyzing the molecular biology of HEV. Since attempts to grow HEV in cell culture have met with limited success (27), it may be that an infectious cDNA clone such as was used for hepatitis A virus (28) could be used to bypass a cell culture step and to directly transfect an animal. The sequence comparisons

of multiple strains provide a rational basis for designing such a full-length clone.

References

1. Gust I.D. and Purcell R.H., J Infect Dis *156*, 630–635, 1987.
2. Purcell R.H., and Ticehurst J.R. in Zuckerman A.J. (ed) *Viral Hepatitis and Liver Disease*. Alan R. Liss, New York, 1988, pp. 131–137.
3. De cock K.M., Bradley D.W., Sandford N.L., Govindarajan S., Maynard J.E., and Redeker A.G., Ann Intern Med *106*, 227–230, 1987.
4. Bradley D.W., Krawczynski K., Cook, Jr. E.H., McCaustland K.A., Humphrey C.D., Spelbring J.E., Myint H., and Maynard J.E., Proc Natl Acad Sci USA *84*, 6277–6281, 1987.
5. Panda S.K., Datta R., Kaur J., Zuckerman A.J., and Nayak N.C., Hepatology *10*, 466–472, 1989.
6. Reyes G.R., Purdy M.A., Kim J., Luk K.-C., Young L.M., Fry K.E., and Bradley D.W., Science *247*, 1335–1339, 1990.
7. Tam A.W., Smith M.M., Guerra M.E., Huang C.C., Bradley D.W., Fry K.E., and Reyes G.R., Virology *185*, 120–131, 1991.
8. Yarbough P.O., Tam A.W., Fry K.E., Krawczynski K., McCaustland K.A., Bradley D.W., and Reynes G.R., J Virol *65*, 5790–5797, 1991.
9. Tsarev S.A., Emerson S.U., Reyes G.R., Tsareva T.S., Legters L.J., Malik I.A., Iqbal M., and Purcell R.H., Proc Natl Acad Sci USA *89*, 559–563, 1992.
10. Ticehurst J., Popkin T.J., Bryan J.P., Innis B.L., Duncan J.F., Ahmed A., Iqbal M. Malik I., Kapikian A.Z., Legters L.J., and Purcell R.H., J Med Virol *36*, 84–92, 1992.
11. Huang C.-C., Nguyen D., Fernandez J., Yun K.Y., Fry K.E., Bradley D.W., Tam A.W., and Reyes G.R., Virology *191*, 550–558, 1992.
12. Bradley D.W., Andjaparidze A.G., Cook Jr. E.H., McCaustland K.A., Balayan M., Stetler H., Velazquez O., Robertson B., Humphrey C.D., Kane M., and Weisfuse I., J Gen Virol *69*, 731–738, 1988.
13. Aye T.T., Uchida T., Ma X., Iida F., Shikata T., Ichikawa M., Rikihisa T., and Win K.M., Virus Genes *7*, 95–110, 1993.
14. Uchida T., Win K.M., Suzuki K., Komatsu K., Iida F., Shikata T., Rikihisa T., Mizuno K., Soe S., Myint H., Tin K.M., and Nakane K., Jpn J Exp Med *60*, 13–21, 1990.
15. Aye T.T., Uchida T., Ma X.Z., Iida F., Shikata T., Zhuang H., and Win K.M., Nucleic Acids Res *20*, 3512–3512, 1992.
16. Bi S.L., Purdy M.A., McCaustland K.A., Margolis H.S., and Bradley D.W., Virus Res *28*, 233–247, 1993.
17. Yin S.R., Tsarev S.A., Purcell R.H., and Emerson S.U., J Med Virol *41*, 230–241, 1993.
18. Kramer G. and Argos P., Nucleic Acids Res *12*, 7269–7282, 1984.
19. Gorbalenya A.E., Koonin E.V., Donchenko A.P., and Blinov V.M., Nucleic Acids Res *17*, 4713–4730, 1989.
20. Rice C.M., Lenches E.M., Eddy S.R., Shin S.J., Sheets R.L., and Strauss J.H., Science *229*, 726–733, 1985.
21. Ray R., Jameel S., Manivel V., and Ray R., Virology *189*, 359–362, 1992.
22. Reyes G.R., Huang C.-C. Yarbough P.O., and Tam A.W., Hepatology *13*, S155–S161, 1991.
23. Cao X.Y., Ma X.Z., Liu Y.Z., Jin X.M., Gao Q., Dong H.J., Zhuang H, Liu C.B., and Wang G.M. in Shikata T., Purcell R.H., and Uchida T. (eds). *Viral Hepatitis C, D and E*. Elsevier Science Publishers, Amsterdam, The Netherlands, 1991, pp. 297–312.
24. Zhuang H., Cao X.Y., Liu C.B., and Wang G.M. in Shikata T., Purcell R.H., and Uchida T. (eds). *Viral Hepatitis C, D and E*. Elsevier Science Publishers, Amsterdam, The Netherlands, 1991, pp. 275–285.
25. Rico-Hesse R., Pallansch M.A., Nottay B.K., and Kew O.W., Virology *160*, 311–322, 1987.
26. Jarvis T.C. and Kirkegaard K., EMBO J *11*, 3135–3145, 1992.
27. Kazachkov Y.A., Balayan M.S., Ivannikova T.A., Panina L.I., Orlova T.M., Zamyatina N.A., and Kusov Y.Y. Arch Virol *127*, 399–402, 1992.
28. Emerson S.U., Lewis M., Govindarajan S., Shapiro M., Moskal T., and Purcell R.H. J Virol *66*, 6649–6654, 1992.