

# Full-length genome of wild-type hepatitis A virus (DL3) isolated in China

Guo-Dong Liu, Ning-Zhu Hu, Yun-Zhang Hu

**Guo-Dong Liu, Ning-Zhu Hu, Yun-Zhang Hu**, Department of Vaccine Research, Institute of Medical Biology, Chinese Academy of Medical Sciences, Peking Union of Medical College, Kunming 650118, Yunnan Province, China

**Correspondence to:** Yun-Zhang Hu, Department of Vaccine Research, Institute of Medical Biology, Chinese Academy of Medical Sciences, 379 Jiaoling Road, Kunming 650118, Yunnan Province, China. huyunz@21cn.com

**Telephone:** +86-871-8335334 **Fax:** +86-871-8334483

**Received:** 2002-10-09 **Accepted:** 2002-11-08

## Abstract

**AIM:** To characterize the genome of an wild-type HAV isolate (DL3) in China.

**METHODS:** A stool specimen was collected from hepatitis A patient from Dalian, China. HAV (DL3) was isolated and viral RNA was extracted. The genome of DL3 was amplified by reverse transcription and polymerase chain reaction (RT-PCR), followed by cloning into pGEM-T vector. The positive colonies were selected and sequenced. The full-length genome of DL3 was analyzed and compared with other wild-type HAV isolates.

**RESULTS:** The genome of DL3 was 7 476 nucleotides (nt) in size, containing 732-nt 5' untranslated region (UTR), 6 681-nt open reading frame (ORF) which encoded a polyprotein of 2 227 amino acids (aa), and 63-nt 3' UTR. The base composition was 28.96 % A (2 165), 16.08 % C (1 202), 22.11 % G (1 653) and 32.85% U (2 456). Genomic comparisons with wild-type HAV isolates revealed that DL3 had the highest identity of 97.5 % for nt (185 differences) with AH1, the lowest identity of 85.7 % (1 066 differences) with SLF88. The highest identity of 99.2 % for amino acid (18 differences) appeared among DL3, AH2 and FH3, and the lowest identity of 96.8 % (72 differences) between DL3 and SLF88. Based upon comparisons of the VP1/2A junction and the VP1 amino terminus, DL3 was classified as subgenotype IA. Phylogenetic analysis showed that DL3 was closest to the isolates in Japan.

**CONCLUSION:** The sequence comparison and phylogenetic analysis revealed that DL3 is most similar to the isolates in Japan, suggesting the epidemiological link of hepatitis A happened in China and Japan.

Liu GD, Hu NZ, Hu YZ. Full-length genome of wild-type hepatitis A virus (DL3) isolated in China. *World J Gastroenterol* 2003; 9 (3): 499-504

<http://www.wjgnet.com/1007-9327/9/499.htm>

## INTRODUCTION

Hepatitis A virus (HAV) is an important human pathogen causing hepatitis, with a higher incidence in developing countries than that in developed countries. Direct person-to-person spread by the fecal/oral route is the most important means of transmission of hepatitis A, and infection with HAV

can cause sporadic and epidemic acute hepatitis in humans<sup>[1,2]</sup>. Although improvements in sanitation have led to a significant reduction in the endemicity of hepatitis A virus infection, hepatitis A is still the most common viral hepatitis infection and a cause of substantial morbidity in China.

HAV is classified as one of two members of the genus *Hepatovirus* within the family *Picornaviridae*<sup>[3,4]</sup>. HAV virion is a naked, spherical particle with a diameter of 27-32 nm. The virion consists of a genome of a linear, single-stranded, 7.5-kb positive-sense RNA and of a protein shell made up of three major proteins, VP1- VP3<sup>[1,5]</sup>. The genome can be divided into a long 5' terminal untranslated region (5' UTR) of about 735 nucleotides, a large open reading frame encoding a polyprotein of 2 227 amino acids, a short 3' UTR with a polyA tail. The HAV polyprotein is co- and posttranslationally cleaved into smaller structural proteins (VP1, VP2, VP3 and a putative VP4) and nonstructural proteins (2A, 2B, 2C, 3A, 3B, 3C and 3D) by the virus-encoding proteinase<sup>[6]</sup>.

Human isolates of HAV possess a single serotype, and monoclonal antibodies raised to various isolates of human HAV have failed to distinguish between individual isolates. However, the nucleotide sequencing of selected genome regions that encode the putative VP1/2A junction region of wild-type HAV isolates present in human specimens has demonstrated significant sequence heterogeneity<sup>[7]</sup>. Using this approach, HAV isolates could be differentiated into seven unique genotypes based upon the sequence of the VP1/2A junction region. A genotype is defined as a group of viruses that differed from each other by no more than 15 %, and a subgenotype as a group of viruses with >92.5 % nucleotide sequence identity<sup>[7]</sup>.

Until now, complete nucleotide sequences of eleven different human wild-type HAV isolates and partial nucleotide sequences of wild-type isolates or cell-adaptative variants of HAV have been reported<sup>[8-19]</sup>. These isolates were isolated from hepatitis A epidemic of diverse geographic origin, including partial sequence of the isolates from fulminant hepatitis A in Shanghai in 1988<sup>[7]</sup>. But the complete nucleotide sequence of genome of the isolate from China remains unknown. In order to elucidate the genetic characteristics and molecular epidemiology and evolution of wild-type HAV in China, we determined the complete nucleotide sequence of genome of an acute HAV isolate (DL3) from Dalian, and compared the complete genome sequences and deduced amino acid sequence of isolate DL3 with those of other eleven wild-type HAV isolates. There are significant differences and identities among the twelve independent isolates.

## MATERIALS AND METHODS

### *Virus sample*

Wild-type human HAV isolate DL3 was recovered from stool specimens stored at -70 °C from an patient with acute hepatitis A, infected during an outbreak at Dalian, Liaoning Province, and was named as DL3. HAV contained in stool were identified by ELISA, immune precipitation and immune electron microscopy. The concentrated and purified HAV for RT-PCR were prepared by chloroform extraction for three times from 10 % of stool supernatant, followed by discontinuous sucrose/glycerol density gradient ultracentrifugation<sup>[20]</sup>.

### cDNA synthesis and cloning

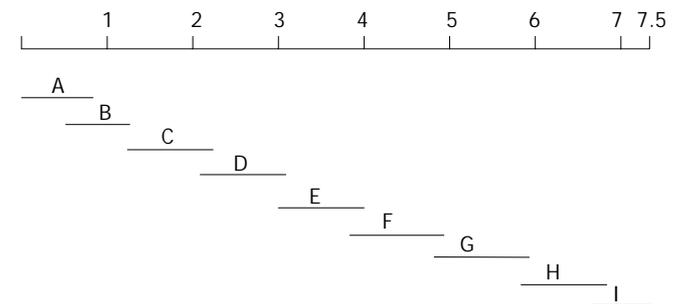
Antigen-capture RT-PCR was used to prepare cDNA of DL3 genome<sup>[21]</sup>, with some modifications. Sterile 0.5-ml conical tube (Eppendorf) was coated with 100 µl of human anti-HAV IgG diluted 1:1 000 in 50 mM sodium carbonate buffer (pH 9.6). After 4 h of incubation at 37 °C, the unbound IgG was removed, and 150 µl of 1 % bovine serum albumin (Sigma) diluted in the buffer was added. After 1 h at 37 °C, the tube was washed three times with 300 µl of PBS (pH 7.4) containing 0.05 % Tween 80. Purified HAV (100 µl) was added, and the preparation was incubated overnight at 4 °C. The tube was washed six times with 500 µl of a 40 mM Tris (pH 8.4) -40 mM KCL-7 mM MgCl<sub>2</sub> solution. Then 100 µl of water was added and tube was heated to 95 °C for 5 min to disrupt captured viruses and melt any secondary structures within the viral RNA. The first strand cDNA was synthesized using Superscript™ First-Strand Synthesis System for RT-PCR kit (Gibco, Life Technologies), following the instruction by manufacturer. Oligo (dT)<sub>18</sub> was used as primer. The partial sequencing showed that subgenotype of isolate from Shanghai in 1988 was IA<sup>[7]</sup>, the oligonucleotide primers (Table 1) corresponding to the nucleotide sequence of wild-type HAV isolate GBM<sup>[16]</sup> were designed to produce subgenomic overlapping HAV fragments with an average length of about 1 000bp which cover the entire HAV genome. The clones of different fragments were performed by PCR in a mixture (50 µl) including 5 µl 10×LA PCR buffer, 8 µl 2.5mM dNTPs, 2 µl template of RT-PCR products, 300 nM positive-sense primer, 300 nM negative-sense primer and 2.5U *Taq* DNA polymerase (TaKaRa). The reaction mixture was subjected to 95 °C for 5 min, then 30 automated cycles of denaturation at 95 °C for 30 sec, annealing at 50 °C for 30 sec, and extension at 72 °C for 1 min or 1 min 30 sec. After the PCR products were recovered and purified, the fragments were ligated into pGEM®-T Vector (Promega). The resulting products were transformed into competent *E. coli* DH5α cells. Three ampicillin-resistant clones were picked out for each fragment. The size of inserts in positive clones was estimated with restriction enzyme site at either side of the inserted fragment. Rapid plasmid preparations were made with the Wizard plasmid purification kit (Promega).

**Table 1** The primers used for amplification of DL3 genomic RNA

Clones	Primers	Sequences
A (0.8kb)	5' RACE RT-primer	5-(P)GCAGAATGAATC-3
	5' RACE A1	5-AGTCCGTTGATAGGACTGAG-3
	5' RACE S1	5-TGTTCTTCCTCAATATCTGCC-3
	5' RACE A2	5-TTCTAAGAAGACTCAGGGGG-3
	5' RACE S2	5-CTGAAAATTCCTTGTGGCC-3
B (0.5kb)	B1	5-GCTGAGGTACTCAGGGGC-3
	B2	5-AGGATAAACAGTCAAGGATGC-3
C (1.1kb)	C1	5-ACATATGCAAGATTGGCATTG-3
	C2	5-ATCCATAGCATGATAAAGAGG-3
D (1.0kb)	D1	5-CCTGGATTCTGACTCC-3
	D2	5-CAGTGGATAACATGGCATTG-3
E (1.1kb)	E1	5-TCTGTACAGAAACATCAGAG-3
	E2	5-AATCCCTGAACAAATGTCTCC-3
F (1.2kb)	F1	5-TCCAGAATGATGGAGCTGAG-3
	F2	5-CTTCGACAAGCACTCCAAG-3
G (1.2kb)	G1	5-AGTTCCTTAGTAATGACAGTTG-3
	G2	5-GCCATTGGATCAATCTCAGC-3
H (1.1kb)	H1	5-AAGTGGAAATTTCTCAGTGTTC-3
	H2	5-GTCCAATCAAGTCAAGATTATC-3
I (0.5kb)	I1	5-GATTCTCTGTTATGGAGATG-3
	I2	5-TTTTTTTTTTTTTTTTTTATT-3

### DNA sequencing and analysis

Sequencing strategy of genome of DL3 was showed in Figure 1. Oligonucleotide primers specific for HAV and primers corresponding to the T7/SP6 promoter region of pGEM®-T Vector were used to sequence the inserted and identified HAV fragment. A *Taq* DyeDeoxy Terminator Cycle sequencing kit and a 377 DNA sequencer (Perkin Elmer) were used to determine nucleotide sequences. To eliminate the possibility of errors in the sequence due to *Taq* polymerase for PCR, at least three clones of each amplified fragment, derived from two individual PCR products, were sequenced. Also to correctly determine the sequence of extreme 5' terminus of HAV genome, a 5' RACE reaction<sup>[22]</sup> was used to obtain the a cDNA fragment from 5' UTR of genome with 5' -Full RACE Core Set (TaKaRa). Analysis, alignment and translation in the amino acids of the obtained nucleotide sequences were done using the sequence analysis program OMEGA2.0 (Oxford Molecular).



**Figure 1** Sequencing strategy of genome of DL3.

### Phylogenetic analysis

Multiple alignments of genome sequences of twelve HAV isolates were done using the Clustal W Program<sup>[23]</sup>. Phylogenetic tree was calculated from genomes to determine the association of DL3 with other eleven wild-type HAV isolates with Vector NTI Suite6.0 software using the Neighbor Joining method<sup>[24]</sup>.

### Reference isolates

The accession numbers of the sequences included in the analysis were as follows: LA, K02990; HM175, M14707; MBB, M20273; GBM, X75215; AH1: AB020564; AH2: AB020564 AB020565; AH3: AB020564 AB020566; FH1: AB020567; FH2: AB020568; FH3: AB020569; SLF88, AY032861.

## RESULTS

### Complete nucleotide sequence of DL3 genome and deduced amino acids

The Complete nucleotide sequence of DL3 has been deposited into GenBank under the accession no. AF512536.

The genome was 7 476-nucleotide long and encoded a polyprotein of 2 227 amino acids. The genome contained a 5' UTR of 732 nucleotides that was two nucleotides shorter than that of HAV HM175/WT. The long open reading frame coding for polyprotein was started at base 733 by codon AUG and terminated at base 7 416 by codon UGA followed by a second stop codon (UAA) 7 nucleotides downstream. The 3' UTR consisted of 63 nucleotides. The base composition was 28.96 % A (2165), 16.08 % C (1 202), 22.11 % G (1 653) and 32.85 % U (2 456). The G+C content is 38.19 %.

The 5' UTR contained three main pyrimidine-rich tracts. The first region, near the 5' terminus (nucleotide99- 138), had

a 92.3 % pyrimidine content. The second region (nucleotide 204- 250) had a 83 % pyrimidine content. The third tract (nucleotide 711- 725) with a 93.9 % pyrimidine content lay immediately before the initiation codon. The 5' UTR had a G+C content of 46.29 % which is higher than that of the genome.

### Comparisons with other isolates

The complete nucleotide sequence and deduced amino acid

sequences of HAV DL3 have been compared with those of reported wild-type isolates (Table 2, Table3). DL3 had the highest identity of 97.5 % for nt (185 differences) with AH1, but the highest identity of 99.2 % for amino acid (18 differences) with AH2 and FH3. DL3 shared the lowest identities of 85.7 % for nucleotide (1 066 differences) and 96.8 % for amino acid (72 differences) with SLF88. Most of the changes were nucleotide transitions.

**Table 2** Nucleotide identities of full-length genomes between DL3 and reference isolates

	Numbers of nucleotide differences/identity										
	AH1	AH2	AH3	FH1	FH2	FH3	GBM	LA	HM175	MBB	SLF88
Full	185/97.5	263/96.5	307/95.9	282/96.2	200/97.3	199/97.3	323/95.7	344/95.4	642/91.4	659/91.2	1066/85.7
5' NTR	15/98.0	29/96.0	25/96.6	16/97.8	13/98.2	13/98.2	18/97.4	14/98.1	38/94.8	50/93.1	84/88.5
VP4	1/98.6	2/97.1	4/94.2	2/97.1	2/97.1	2/97.1	3/95.7	2/97.1	7/89.9	7/89.9	8/88.4
VP2	15/97.7	25/96.2	28/95.8	24/96.4	18/97.3	24/96.4	39/94.1	28/95.8	58/91.3	58/91.3	98/85.3
VP3	15/98.0	25/96.6	28/96.2	27/96.3	13/98.2	23/96.9	35/95.3	40/94.6	67/90.9	63/91.5	114/84.6
VP1	23/97.2	27/96.7	25/97.0	23/97.2	22/97.3	13/98.4	30/96.4	44/94.6	76/90.8	78/90.5	124/84.9
2A	5/97.7	10/95.3	11/94.8	9/95.8	3/98.6	9/95.8	9/95.8	12/94.4	17/92.0	14/93.4	27/87.3
2B	13/98.3	26/96.5	29/96.1	46/93.9	18/97.6	28/96.3	33/95.6	37/95.1	75/90.0	73/90.3	99/86.9
2C	30/97.0	42/95.8	42/95.8	44/95.6	35/96.5	35/96.5	52/94.8	55/94.5	114/88.7	120/88.1	152/84.9
3A	1/99.5	4/98.2	4/98.2	9/95.9	4/98.2	3/98.6	13/94.1	8/96.4	14/93.7	18/91.9	23/89.6
3B	1/98.6	1/98.6	2/97.1	3/95.7	3/95.7	0	2/97.1	3/95.7	6/91.3	7/89.9	12/82.6
3C	14/97.9	20/97.0	24/96.3	25/96.2	16/97.6	21/96.8	30/95.4	27/95.9	56/91.5	50/92.4	98/85.1
3D	51/96.5	51/96.5	84/94.3	53/96.4	52/96.5	27/98.2	57/96.1	64/95.6	113/92.3	114/92.2	209/85.8
3' NTR	1/98.4	1/98.4	1/98.4	1/98.4	1/98.4	1/98.4	2/96.8	10/84.4	1/98.4	7/88.9	18/71.4

**Table 3** Amino acid identities of full-length genomes between DL3 and reference isolates

	Numbers of amino acid differences/identity										
	AH1	AH2	AH3	FH1	FH2	FH3	GBM	LA	HM175	MBB	SLF88
Full	21/99.1	18/99.2	36/98.4	28/98.7	21/99.1	18/99.2	36/98.4	34/98.5	35/98.4	49/97.8	72/96.8
VP4	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	1/95.7	2/91.3
VP2	0	0	0	1/99.5	0	0	1/99.5	0	2/99.1	2/99.1	1/99.5
VP3	1/99.6	1/99.6	1/99.6	2/99.2	1/99.6	1/99.6	1/99.6	1/99.6	2/99.2	2/99.2	4/98.4
VP1	0	0	0	1/99.6	0	1/99.6	2/99.3	13/95.3	1/99.6	2/99.3	2/99.3
2A	1/98.6	1/98.6	1/98.6	1/98.6	1/98.6	0	1/98.6	0	1/98.6	2/97.2	1/98.6
2B	2/99.2	2/99.2	3/98.8	5/98.0	5/98.0	5/98.0	3/98.8	4/98.4	4/98.4	6/97.6	6/97.6
2C	6/98.2	5/98.5	4/98.8	6/98.2	5/98.5	4/98.8	7/97.9	5/98.5	11/96.7	17/94.9	14/95.8
3A	0	1/98.6	1/98.6	1/98.6	0	0	5/93.2	1/98.6	1/98.6	3/95.9	1/98.6
3B	0	0	0	0	0	0	1/95.7	0	0	0	0
3C	2/99.1	2/99.1	3/98.6	2/99.1	2/99.1	2/99.1	3/98.6	2/99.1	2/99.1	2/99.1	8/96.3
3D	8/98.4	5/99.0	22/95.5	8/98.4	6/98.8	4/99.2	11/97.8	7/98.6	10/98.0	12/97.5	33/93.3

### 5' Untranslated region

The 5' UTR of HAV forms conserved and highly ordered secondary structure and plays an important role in controlling viral translation<sup>[25, 26]</sup>. Comparisons with eleven other isolates in 5' UTR showed that DL3 shared the lowest identity of 88.5 % (84 nucleotide changes) with SLF88, the highest identity of 98.2 % (13 nucleotide changes) with FH2 and FH3. There were a G deletion at position 28 and a T deletion at position 102 compared with HM175 and MBB. In contrast to the comparison with complete nucleotide sequence, identities of 5' UTR for DL3 with other isolates were higher than that of complete genomes except the identity with AH2, which suggested that the 5' UTR of HAV was conserved. In contrast to ribosomes scanning from the capped 5' ends of the majority of cellular mRNA, the long and highly structured 5' UTR of HAV mediated the binding of ribosomal subunits internally at internal ribosome entry site (IRES) of nucleotide 324 to nucleotide 692 (base position was corresponding to HM175/WT) that directed cap-independent initiation of viral translation at correct AUG codon<sup>[25, 26]</sup>. In this region DL3 shared higher identities of 97.9 %, 98.4 %, 99.2 %, 98.6 %, 98.6 %, 95.9 %, 95.4 % and 94.9 % with AH2, FH1, FH2, FH3, LA, HM175, MBB and SLF88 than that in whole 5' UTR. But DL3 was identical to SLF88 between nucleotide 223 to nucleotide 371 (base position is corresponding to HM175/WT). The result suggested higher conservative and importance of common secondary structure of IRES in initiation of viral translation.

### Coding region

The open reading frame of HAV RNA encoding a large polyprotein with 2 227 amino acids was divided into P1, P2 and P3 region. The P1 region encodes structural or capsid proteins, VP4, VP2, VP3 and VP1. The P2 and P3 region encoded nonstructural proteins, 2A, 2B, and 2C, and 3A, 3B, 3C and 3D, respectively. In coding region, the amino acid differences for the structural and nonstructural proteins of twelve isolates are showed in Table 3. In contrast to the high heterogeneous of nucleotide sequences, the amino acid sequences were highly conserved among twelve wild-type isolates. Comparisons within the coding region (6 681 nucleotides) of DL3 and SLF88 yielded 964 nucleotide changes (14.4 % difference), resulting in only 72 amino acid substitutions (3.2 % difference). In comparison with eleven isolates, DL3 had the highest identity (99.2 %) of amino acids with AH2 and FH3, lowest identity (96.8 %) with SLF88.

P1 region, the structural protein region of HAV, consisted of 2 295 nucleotides and encodes 765-amino acid polypeptide which was processed into four structural proteins, VP4, VP2, VP3 and VP1. The capsid protein of HAV consisted of VP1, VP2 and VP3. The presence of a fourth protein VP4 has been described repeatedly, but the reported apparent molecular weights (7- 14kD) contrasted sharply with those predicted from nucleic acid sequence data (1.5 or 2.3kD). Conclusive physical identification of VP4 was still unavailable. The alignment of amino acid sequence indicated that no amino acid mutation of polyprotein of DL3 occurred at cleavage site of VP4/VP2, VP2/VP3, VP3/VP1, implying the importance of conserved cleavage sites in entire viral structure. In VP4 DL3 shared lowest amino acid sequence identities with other isolates in structural region, 91.3 % identity with SLF88, 95.7 % identity with other ten wild-type isolates. DL3 had a unique amino acid mutation (S to A) at position 4 different from all the other isolates. In VP2 region amino acid sequences were highly conserved. There was 100 % identity between DL3 and AH1, AH2, AH3, FH2, FH3 and LA. And even for DL3 and SLF88, the identity was also up to 99.5 %. In VP3 region DL3 had another unique amino acid mutation (K to R) at position 47 different from all the other isolates. It had been reported that

VP1 had the most amino acid diversity of the capsid proteins<sup>[16]</sup>. But it was not the case for DL3. In VP1 region there was 100 % amino acid sequence identity between DL3 and AH1, AH2, AH3 and FH2. The highest difference was between DL3 and LA (13 amino acid differences, 95.3 % identity), just because one amino acid was inserted into the LA between amino acids 540-548, at which a cluster of 8 amino acids was altered due to 3 changes in the open reading frame resulting from the insertion of 3 nucleotides.

The HAV P2 region encoded nonstructural proteins 2A, 2B and 2C. The proposed 2A region encoded for 71 amino acids, but the exact function of 2A was not yet determined. Recent study revealed that 2A protein might participate in virion morphogenesis<sup>[27]</sup>. 2B and 2C proteins were found to have 251 and 335 amino acids residues, respectively. Both proteins playing important roles in the replication of viral RNA were considered significant in host-dependent adaptation since many mutations had been detected in both regions of adapted variants<sup>[28]</sup>. HAV 2C protein was considered to have helicase and NTPase activities. With some variants, multiple mutations in P2 contributed to enhance viral replication with 5' UTR and P3 proteins and to express the cytopathic phenotype<sup>[29]</sup>. In 2A, 2B and 2C proteins, 2C protein was the most variable between twelve isolates, DL3 shared the lowest identity of 94.9 % (17 amino acid differences) with MBB and the highest identity of 98.8 % with AH3 and FH3 (only 4 amino acid differences). In contrast, 2B was more conservable. The 99.2 % identities of DL3 with AH1 and AH2 and the 98.8 % identities of DL3 with AH3 and GBM revealed high conservation of 2B protein among wild-type viruses. It has been reported that a little more amino acid substitutions seemed to be found in FH than in AH in 2B region<sup>[18]</sup>. This research also showed that more amino acid substitutions were found in FH than that in AH when compared with DL3, suggesting the possible association between the severity of hepatitis A and amino acid substitutions in 2B region.

The HAV P3 region encoded 3A, 3B, 3C and 3D proteins with 74, 23, 219 and 489 amino acids, respectively. 3A functioned as pre-Vpg, 3B was a genome-linked viral protein (Vpg), 3C was the sole protease for HAV protein processing. The recombinant HAV 3C prepared in *E. coli* catalyzed the putative cleavage sites and released mature or intermediate proteins, VP0, VP3, VP1-2A, VP1, 2A, 2B, 2BC, 2C, 3ABC, 3AB, 3C and 3D<sup>[30-34]</sup>. 3D was an RNA-dependent RNA polymerase. By comparison, 3B protein showed highest amino acid sequence homology, with 100 % identities among eleven isolates except 95.7 % identity between DL3 and GBM. Compared with 3D proteins, 3C protein showed higher amino acid sequence identities in all twelve isolates. In this protein, DL3 shared the 96.3 % identity with SLF88, 98.6 % with AH3 and GBM, and the highest identity 99.1 % with the other eight isolates. The comparison indicated that high conservation of 3C protein in wild-type HAV played an important role in the replication of virus. 3D among isolates of genotype I showed no significant variation. But the 3D polymerase of SLF88 showed low identity with that of DL3, with only 93.3 % identity (33 amino acid differences). GBM showed a little higher variability among the eleven reference isolates in P3 region, especially for 3A and 3D. The identities for these two proteins between DL3 and GBM were 93.2 % (5 amino acid differences) and 97.8 % (11 amino acid differences), respectively. Even for 3B, GBM had 1 amino acid difference with DL3.

### 3' Untranslated region

The 3' UTR of DL3 was 63-nucleotide long, the same as those of 9 isolates, 1 nucleotide less than that of LA, and 14 nucleotides more than that of SLF88. All twelve isolates had two stop codons, separated by 6 nucleotides. The 3' UTR

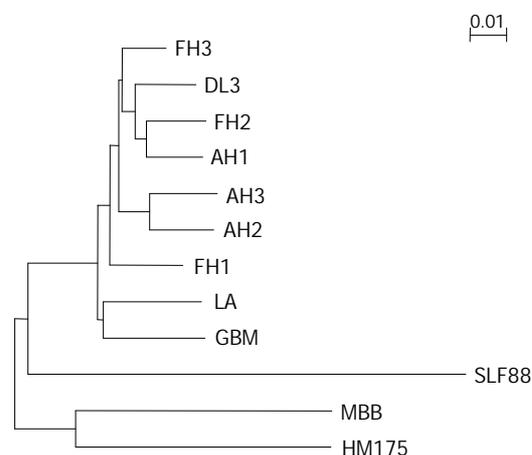
exhibited the great identity in nucleotide sequence between DL3 and isolates from Japan, GBM and HM175, with only 1 or 2 nucleotide difference. But for DL3 and the other 3 isolates, identities were lower, especially 71.4% between DL3 and SLF88.

### Subgenotype classification

Alignment analysis of 168 nucleotides in VP1/2A (nucleotide 3 024- 3 191) of DL3 with other eleven wild-type isolates revealed that DL3 had >92.5 % identities with six Japanese isolates, LA and GBM (IA subgenotype), <92.5 % identity with HM175 (IB subgenotype), and only 83.9 % identity with SLF88 (the newly determined isolate of genotype VII). According to these comparison, DL3 was classified as subgenotype IA. Something strange was that comparison of the same region between DL3 and MBB showed 94 % identity, suggesting DL3 to be subgenotype IB. To solve this paradox, the VP1 amino terminus (nucleotide 2 208- 2 468) of DL3 was compared with isolate MBB. The result showed DL3 belonged to subgenotype IA.

### Phylogenetic relationships

Phylogenetic relations of various HAV wild-type isolates were analyzed on the genomes of twelve isolates (Figure 2). The result showed that twelve wild-type isolates formed four main subclusters, subcluster I contained DL3, AH1, FH2 and FH3, subcluster II contained AH2 and AH3, subcluster III contained LA and GBM, and subcluster IV contained MBB and HM175. Subcluster I, II, III and FH1 composed of subgenotype IA cluster. SLF88 located near this cluster. By this phylogenetic analysis DL3 isolate showed closer phylogenetic affiliation to AH1 and FH2. Besides, the comparison of full-length genome of nine isolates showed DL3 had higher identity with AH1 and FH2 than those with GBM, LA, HM175, MBB and SLF88. The analysis suggested that phylogenetic relations of various HAV wild-type isolates correlated with geographical region.



**Figure 2** Phylogenetic relations of DL3 full-length genome to those of other HAV.

## DISCUSSION

Although hepatitis A remains a disease of major public health importance in China, the study on the genome of wild-type HAV isolate in China has not been reported. Here, we analyzed the full-length genome of a wild-type HAV isolate DL3 in China. The genome sequence of DL3 contained 7 476 nucleotides, dividing into three regions, 732-nucleotide 5' UTR, 6 681-nucleotide ORF coding for a polyprotein of 2 227 amino acid, 63-nucleotide 3' UTR.

Sequence comparisons of genomes showed DL3 shared the highest nucleotide identity (97.5 %) with AH1 and amino

acid identity (99.2 %) with AH2 and FH3, and lowest nucleotide identity (85.7 %) and amino acid identity (96.8 %) with SLF88. But the nucleotide sequence and amino acid sequence differences were not evenly distributed along the genome in all isolates. The 5' UTR region showed high conservatism, and even SLF88 is 88.5 % identical to DL3, an identity higher than that of full-length genome. This suggests the importance of this region for HAV. The 3' UTR region showed high identities between DL3 and other isolates of genotype I. When compared with SLF88, DL3 had high divergence of 28.6 % with SLF88, which lead to 18 nucleotide changes, including 14 deletions and 4 mutations. Although the exact function of 3' UTR is unknown, several studies have revealed that sequence elements within the 3' UTR are expected to play an important role in initiation and regulation of synthesis of HAV RNA<sup>[35, 36]</sup>. The high heterogeneity between DL3 and SLF88 in the 3' UTR may suggest the difference of the replicating characteristics between DL3 and SLF88.

In structural region, the highest difference (13 amino acid differences, between DL3 and LA) existed in VP1. But this high divergence was due to 1 amino acid insertion into LA. For DL3 and other eleven isolates, high homology was observed, and the high variability of VP1 reported early was not observed. In nonstructural region the majority of amino acid identities between DL3 and other eleven isolates appeared in 3B and 3C proteins, implying that both proteins play important roles in the replication of virus and the processing of polyprotein. Differences between DL3 and other isolates of subgenotype IA in 2C region were a little more than those in 2B region. But differences between DL3 and other isolates of subgenotype IB or SLF88 in 2C region were much more than those in 2B region. SLF88 had 33 amino acid mutations with DL3 in 3D, suggesting the polymerase of SLF88 may have different efficiency compared with DL3.

Based on the higher nucleotide sequence identities of DL3 with AH1 and FH2 in the whole 7 476-nucleotide genome (97.5 % and 97.3, respectively), 732-nucleotide 5' UTR region (98.0 % and 98.2 %, respectively), 63-nucleotide 3' UTR region (98.4 % for both), 168-nucleotide region of subgenotype classification (98.2 %), 6681-nucleotide ORF (97.5 % and 97.2 %, respectively), DL3 is found to be in phylogenesis closest to AH1 and FH2. So we proposed that DL3, AH1 and FH2 probably represent different isolates of the same strain. Moreover, it is noteworthy that Japan where AH1 and FH2 were isolated geographically close to China, where DL3 was isolated. The phylogenetic affiliation and geographical closure suggested the epidemiological link of hepatitis A happened in Dalian and Japan.

By comparison of VP1/2A junction, genotypes were defined by at least 15 %, and subgenotypes differed by at least 7.5 %. According to this, we defined DL3 as subgenotype IA by comparisons with isolates HM175, AH1, AH2, AH3, FH1, FH2 and FH3. Interestingly, comparisons of this region with isolate MBB showed DL3 to be subgenotype IB. To solve this paradox, the VP1 amino terminus of DL3 was compared with isolate MBB. The comparison showed DL3 belonged to subgenotype IA. This suggests that it may not be so accurate to define a genotype only by comparison of VP1/2A junction. And the more accurate result may be achieved by comparison of more nucleic acid fragments.

## REFERENCES

- 1 **Lemon SM.** Type A hepatitis: new developments in an old disease. *N Engl J Med* 1985; **313**: 1059-1067
- 2 **Cuthbert JA.** Hepatitis A: old and new. *Clin Microbiol Rev* 2001; **14**: 38-58
- 3 **Minor PD.** Picornaviridae. In Francki RIB, Fauquet CM, Kundson

- DL, Brown F, eds. Archives of Virology, Supplement 2. New York: Springer-Verlag 1991: 320-326
- 4 **Marvil P**, Knowles NJ, Mockett APA, Britton P, Brown TDK, Cavanagh D. Avian encephalomyelitis virus is a picornavirus and is most closely related to hepatitis A virus. *J Gen Virol* 1999; **80**: 653-662
  - 5 **Weitz M**, Siegl G. Hepatitis A virus: structure and molecular virology. In: Zuckerman AJ, Thomas HC. Viral hepatitis. 2<sup>nd</sup> edition. London: Churchill Livingstone 2001: 15-27
  - 6 **Totsuka A**, Moritsugu Y. Hepatitis A virus proteins. *Intervirology* 1999; **42**: 63-68
  - 7 **Robertson BH**, Jansen RW, Khanna B, Totsuka A, Nainan OV, Siegl G, Widell A, Margolis HS, Isomura S, Ito K, Ishizu T, Moritsugu Y, Lemon SM. Genetic relatedness of hepatitis A virus strains recovered from different geographical regions. *J Gen Virol* 1992; **73**: 1365-1377
  - 8 **Ticehurst JR**, Racaniello VR, Baroudy BM, Baltimore D, Purcell RH, Feinstone SM. Molecular cloning and characterization of hepatitis A virus cDNA. *Proc Natl Acad Sci USA* 1983; **80**: 5885-5889
  - 9 **Linemeyer DL**, Menke JG, Martin-Gallardo A, Hughes JV, Young A, Mitra SW. Molecular cloning and partial sequencing of hepatitis A viral cDNA. *J Virol* 1985; **54**: 247-255
  - 10 **Najarian R**, Caput D, Gee W, Potter SJ, Renard A, Merryweather J, Van Nest G, Dina D. Primary structure and gene organization of human hepatitis A virus. *Proc Natl Acad Sci USA* 1985; **82**: 2627-2631
  - 11 **Venuti A**, Di Russo C, Del Grosso N, Patti A-M, Ruggeri F, Stasio RP, Martiniello GM, Pagnotti P, Degener AM, Midulla M, Pana A, Perez-Bercoff R. Isolation and molecular cloning of a fast-growing isolate of human hepatitis A virus from its double-stranded replicative form. *J Virol* 1985; **56**: 579-588
  - 12 **Robertson BH**, Brown VK, Bradley DW. Nucleic acid sequence of the VP1 region of attenuated MS-1 hepatitis A virus. *Virus Res* 1987; **8**: 309-316
  - 13 **Cohen JJ**, Ticehurst JR, Purcell RH, Buckler-White A, Baroudy BM. Complete nucleotide sequence of wild-type hepatitis A virus: Comparison of different isolates of hepatitis A virus and other picornaviruses. *J Virol* 1987; **61**: 50-59
  - 14 **Paul AV**, Tada H, von der Helm K, Wissel T, Kiehn R, Wimmer E, Deinhardt F. The entire nucleotide sequence of the genome of human hepatitis A virus ( isolate MBB ). *Virus Res* 1987; **8**: 153-171
  - 15 **Khanna B**, Spelbring JE, Innis BL, Robertson BH. Characterization of a genetic variant of human hepatitis A virus. *J Med Virol* 1992; **36**: 118-124
  - 16 **Graff J**, Normann A, Feinstone SM, Flehmig B. Nucleotide sequence of wild-type hepatitis A virus GBM in comparison with two cell culture-adapted variants. *J Virol* 1994; **68**: 548-554
  - 17 **Beneduce F**, Pisani G, Divizia M, Pana A, Morace G. Complete nucleotide sequence of a cytopathic hepatitis A virus isolate isolated in Italy. *Virus Res* 1995; **36**: 299-309
  - 18 **Fujiwara K**, Yokosuka O, Fukai K, Imazeki F, Saisho H, Omata M. Analysis of full-length hepatitis A virus genome in sera from patients with fulminant and self-limited acute type A hepatitis. *J Hepatol* 2001; **35**: 112-119
  - 19 **Ching KZ**, Nakano T, Chapman LE, Demby A, Robertson BH. Genetic characterization of wild-type genotype VII hepatitis A virus. *J Gen Virol* 2002; **83**: 53-60
  - 20 **Bishop NE**, Hugo DL, Borovec SV, Anderson DA. Rapid and efficient purification of hepatitis A virus from cell culture. *J Virol Med* 1994; **47**: 203-216
  - 21 **Jansen RW**, Siegl G, Lemon SM. Molecular epidemiology of human hepatitis A virus defined by an antigen-capture polymerase chain reaction method. *Proc Natl Acad Sci USA* 1990; **87**: 2867-2871
  - 22 **Frohman MA**, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* 1988; **85**: 8998-9002
  - 23 **Thompson JD**, Higgins DG, Gibson TJ. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; **22**: 4673-4680
  - 24 **Saitou N**, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406-425
  - 25 **Brown EA**, Day SP, Jansen RW, Lemon SM. The 5' nontranslated region of hepatitis A virus RNA: secondary structure and elements required for translation *in vitro*. *J Virol* 1991; **65**: 5828-5838
  - 26 **Brown EA**, Zajac AJ, Lemon SM. *In vitro* characterization of an internal ribosome entry site (IRES) present within the 5' nontranslated region of hepatitis A virus RNA: comparison with the IRES of encephalomyocarditis virus. *J Virol* 1994; **68**: 1066-1074
  - 27 **Cohen L**, Benichou D, Martin A. Analysis of deletion mutants indicated that the 2A polypeptide of hepatitis A virus participates in virion morphogenesis. *J Virol* 2002; **76**: 7495-7505
  - 28 **Emerson SU**, Huang YK, Purcell RH. 2B and 2C mutations are essential but mutations throughout the genome of HAV contribute to adaptation to cell culture. *Virology* 1993; **194**: 475-480
  - 29 **Zhang H**, Chao SF, Ping LH, Grace K, Clarke B, Lemon SM. An infectious cDNA clone of a cytopathic hepatitis A virus: genomic regions associated with rapid replication and cytopathic effect. *Virology* 1995; **212**: 686-697
  - 30 **Schultheiss T**, Kusov YY, Gauss-Muller V. Proteinase 3C of hepatitis A virus ( HAV ) cleaves the HAV polyprotein P2-P3 at all sites including VP1/2A and 2A/2B. *Virology* 1994; **198**: 275-281
  - 31 **Schultheiss T**, Sommergruber W, Kusov Y, Gauss-Muller V. Cleavage specificity of purified recombinant hepatitis A virus 3C proteinase on natural substrates. *J Virol* 1995; **69**: 1727-1733
  - 32 **Probst C**, Jecht M, Gauss-Muller V. Processing of proteinase precursors and their effect on hepatitis A virus particle formation. *J Virol* 1998; **72**: 8013-8020
  - 33 **Harmon SA**, Updike W, Jia XY, Summers DF, Ehrenfeld E. Polyprotein processing in cis and in trans by hepatitis A virus 3C protease cloned and expressed in Escherichia coli. *J Virol* 1992; **66**: 5242-5247
  - 34 **Malcolm BA**, Chin SM, Jewell DA, Stratton-Thomas JR, Thudium KB, Ralston R, Rosenberg S. Expression and characterization of recombinant hepatitis A virus 3C proteinase. *Biochemistry* 1992; **31**: 3358-3363
  - 35 **Nuesch JP**, Weitz M, Siegl G. Proteins specifically binding to the 3' untranslated region of hepatitis A virus RNA in persistently infected cells. *Arch Virol* 1993; **128**: 65-79
  - 36 **Kusov Y**, Weitz M, Dollenmeier G, Gauss-Muller V, Siegl G. RNA-protein interactions at the 3' end of the hepatitis A virus RNA. *J Virol* 1996; **70**: 1890-1897