COMPLETE GENOMES OF TWO HUMAN HEPATITIS A VIRUS ISOLATES FROM CHINA: ANALYSIS AND COMPARISON WITH OTHER ISOLATES

Y. HU*, N. HU, G. LIU

Department of Vaccine Research, Institute of Medical Biology, Chinese Academy of Medical Sciences, Peking Union of Medical College, 379 Jiaoling Road, Kunming, 650118 Yunnan, P.R. China

Received June 25, 2002; accepted September 23, 2002

Summary. – Complete sequences of the genomes of two wild type (wt) Human hepatitis A virus (HHAV) isolates, LU38 and LY6 from China were determined and compared with those of wt HHAV isolates AH1, AH2, AH3, FH1, FH2, FH3, GBM, HM175, LA and MBB. The genomes of both LU38 and LY6 consisting of 7477 nucleotides (nts) contained a 5'-non-translated region (5'-NTR, 733 nts), an open reading frame (ORF, 6681 nts), and a 3'-NTR (63 nts) followed by a poly(A)-tail. It encoded a polyprotein of 2227 amino acids (aa) Sequence comparison showed that LU38 shared the highest identities of 98.1% for nt (140 differences) and 99.2% for as (17 differences) with AH1, and the lowest identities of 91.4% for nt (741 differences) with HM175 and 98.1% for aa (43 differences) with GBM. LY6 shared the highest identities of 97.4% for nt (196 differences) and 98.7% for aa (28 differences) with H1 and the lowest identities of 91.2% for nt (642 differences) with HM175 and 97.7% for as (51 differences) with GBM. The subgenotyping revealed that the LU38 and LY6 isolates are of IA subgenotype. The phylogenetic analysis showed that LU38 is closest to AH1 and the LY6 to FH3, suggesting that the epidemiological link of hepatitis A (HA) had developed in China and Japan.

Key words: amino acid sequence; nucleotide sequence; Human hepatitis A virus, isolates; phylogenetic analysis; China

Introduction

HHAV infection is a worldwide health problem. HA is transmitted primarily by fecal or oral route and causes sporadic and epidemic acute hepatitis in humans (Lemon and Shapiro, 1994). HA is endemic in developing countries. In 1988, an explosive HA epidemic occurred in Shanghai Province, P.R. China, associated with consumption of raw clams, in which over 300,000 cases (4% of the total population) have been reported in two months (Kan *et al.*, 1989).

HHAV is a member of the species *Hepatitis A virus*, the genus *Hepatovirus*, the family *Picornaviridae*. HHAV is a non-enveloped icosahedral particle of 27–32 nm in diameter containing a single-stranded 7.5 kb positive-sense RNA genome with a long 5'-NTR and a short 3'-NTR with a poly(A) tail (Lemon and Robertson, 1993). Similarly to other picornaviruses, the HHAV genome contains a large ORF encoding a polyprotein of about 250 K, which is co- and post-translationally cleaved into smaller structural (2A, 2B, and 2C) and non-structural proteins (3A, 3B, 3C, and 3D) by a virus-encoded proteinase (Totsuka and Moritsugu, 1999).

To date, several groups have reported complete or partial nucleotide sequences of various wt isolates/strains of HHAV (Ticehurst *et al.*, 1983; Linemeyer *et al.*, 1985; Najarian *et al.*, 1985; Ovchinnikov *et al.*, 1985; Cohen *et al.*, 1987; Paul *et al.*, 1987; Graff *et al.*, 1994; Fujiwara *et al.*, 2001), which have been isolated from HA epidemics in diverse geographic regions, inclusive of the isolates from Shanghai, P.R. China,

*E-mail: huyunz@21cn.com; fax: +86871-8335334.

Abbreviations: aa = amino acid; HA = hepatitis A; HHAV = Human hepatitis A virus; IRES = internal ribosomal entry site; nt = nucleotide; 5'-NTR = 5'-non-translated region; 3'-NTR = 3'-non-translated region; ORF = open reading frame; PBS = phosphate-buffered saline pH 7.4; PCR = polymerase chain reaction; wt = wild type

where a fulminant HA has broken out in 1988 (Robertson *et al.*, 1992),

In order to elucidate genetic and epidemiological characteristics and molecular evolution of HHAV in China, we determined complete nucleotide sequences of genomes of the LU38 isolate from a fulminant HA epidemic and the LY6 isolate from a sporadic self-limited HA, and compared them as well as the deduced amino acid sequences with those of other 10 wt HHAV isolates from around the world.

Materials and Methods

Virus isolates. The wt HAV isolate LU38 was recovered from a stool specimen stored at -70°C originating from a patient with fulminant HA during an outbreak of HA epidemic in Nantong, Jiangsu Province (close to Shanghai where a similar outbreak had occurred in 1988). The other wt HHAV isolate YL6 originated from a patient with sporadic self-limited acute HA from LiaoYan, the Liao Ning Province in 2001.

cDNA synthesis and cloning. An antigen-capture polymerase chain reaction (PCR) (Jansen *et al.*, 1990) with some modifications was used to prepare cDNAs of LU38 and LY6 genomes. Sterile conical tubes (0.5 ml, Ependorf) were coated with human anti-HAV IgG (100 µl per tube) diluted 1:1,000 in 50 mmol/l sodium carbonate buffer pH 9.6. After 4 hrs at 37°C the unbound IgG was removed and 1% bovine serum albumin (150 µl per tube, Sigma) diluted in the same buffer as above was added. After 1 hr at 37°C the tubes were washed 3 times with phosphate-buffered saline pH 7.4 (PBS) containing 0.05% Tween 80 (300 µl per tube). A purified HHAV (100 µl per tube) was added and the tubes were incubated overnight at 4°C. Then they were washed 6 times with 40 mmol/l Tris pH 8.4 containing 40 mmol/l KCL and 7 mmol/l $MgCl_2$ (6 x 500 µl per tube) After addition of water (100 µl per tube) they were incubated at 95°C for 5 mins to disrupt the bound virus particles and to melt any secondary structures within the viral RNA. The first strand cDNA was synthesized using the Super-Script™ Preamplification System Kit (Gibco, Life Technologies) according to the manufacturer's instructions. Specific primers were used to produce subgenomic overlapping HHAV genome fragments covering the entire genome of an average length of about 1000 bp (Graff *et al.*,1994, Table 1). The clones of the HHAV fragments were obtained by PCR. The reaction mixture (50 µl) consisted of a PCR buffer (5 µl), 2.5 mmol/l dNTPs (8 µl), RT-PCR products (2 µl), positive-sense and negative-sense primers (300 nmoles), and 2.5 U of *Taq* DNA polymerase (Promega). The PCR parameters were as follows: 95°C for 5 mins and 30 cycles of 95°C for 30 secs (denaturation), 50°C for 30 secs (annealing), and 72°C for 60 or 90 secs (extension). After the PCR products (the HHAV fragments) were recovered and purified, they were inserted into the pGEM®-T vector (Promega). Competent *E. coli.* DH5a cells were transformed with the recombinant vectors. Three ampicillin-resistant clones were picked out for each fragment. The size of inserts in positive clones was estimated on the basis of the restriction sites at either end of the inserts. Plasmid preparations were made by use of the Wizard Plasmid Purification Kit (Promega).

Table 1. The primers used for PCR amplification of genomic RNAs of HHAV isolates LU38 and LY6

Clones	Primers	Sequences
A (0.8 kb)	5'RACE RTprimer	5'-G'CAGAATGAATC-3'
	5'RACE A1	5'-AGTCCGTTGATAGGACTGAG-3'
	5'RACE S1	5'-TGTTCTTCCTCAATATCTGCC-3'
	5'RACE A2	5'-TTCTAAGAAGACTCAGGGGG-3'
	5'RACE S2	5'-CTGGAAAATTCCTTGTTTGGCC-3'
B (0.5 kb)	B1	5'-GCTGAGGTACTCAGGGGC-3'
	B2	5'-AGGATAAACAGTCAAGGATGC-3'
C (1.1 kb)	C1	5'-ACATATGCAAGATTTGGCATTG-3'
	C2	5'-ATCCATAGCATGATAAAGAGG-3'
D (1.0 kb)	D1	5'-CCTGGATTTCTGACACTCC-3'
	D2	5'-CAGTGGATAACATGGCATTTG-3'
E (1.1 kb)	E1	5'-TCTGTCACAGAACAATCAGAG-3'
	E2	5'-AATCCCTGAACAAATGTCTCC-3'
F (1.2 kb)	F1	5'-TCCAGAATGATGGAGCTGAG-3'
	F2	5'-CTTCGACAAGCACTCCAAG-3'
G (1.2 kb)	G1	5'-AGTTCCTTAGTAATGACAGTTG-3'
	G2	5'-GCCATTGGATCAATCTCAGC-3'
H (1.1 kb)	H1	5'-AAGTGGAATTTTCTCAGTGTTC-3'
	H2	5'-GTCCAATCAAGTCAAGATTATC-3'
I (0.5 kb)	I1	5'-GATTCTCTGTTATGGAGATG-3'
	I2	5'-TTTTTTTTTTTTTTTTTTTTTTATTT-3'

*: phosphorylation.

DNA sequencing and analysis. The oligonucleotide primers specific for HAV and the primers corresponding to the T7/SP6 promoter region of pGEM®-T vector were used to sequence the identified HHAV fragments. A Taq Dye Deoxy Terminator Cycle Sequencing Kit and a 377 DNA Sequencer (Perkin Elmer) were used to determine nucleotide sequences. To eliminate possible errors in sequences caused by Taq polymerase in PCR, at least three clones of each HHAV fragment, derived from two individual PCR products, were sequenced. To determine correctly the sequence of the 5'-NTR of HHAV genome the respective cDNA fragment was obtained by use of a 5'-Full RACE Core Set (TaKaRa, Forhman *et al.*, 1988). Analysis and alignment of the obtained nucleotide sequences and deduced amino acid sequences were done using the OMEGA Sequence Analysis Program (Oxford Molecular).

Phylogenetic analysis. Multiple alignment of genome sequences of 12 HHAV isolates was done using the Clustal W Program (Thomson *et al.*, 1994). A phylogenetic tree was calculated from the genomes to determine the relationship of the isolates LU38 and LY6 to other 10 wt HHAV isolates by the neighbor joining method using the Vector NT1 Suite Software (Saitou and Nei, 1987).

Source of nucleotide sequences. The accession numbers of the sequences included in the analysis were as follows: AH1: AB020564; AH2: AB020564 AB020565; AH3: AB020564 AB020566; FH1: AB020567; FH2: AB020568; FH3: AB020569; HM175: M14707; LA: K02990; GBM: X75215; MBB: M20273; LU38: AF357222; LY6: AF485328.

Table 2. Difference numbers and identities of nucleotides and amino acids between genomes of HHAV isolates LU38 and LY6

Region	Nucleotides		Amino acids	
	Difference number	Identity (%)	Difference number	Identity (%)
Genome	171/7477	97.7		
5'-NTR	7/733	99.1		
ORF	168/6681	97.5	21/2227	99.1
VP4	0/69	100	0/23	100
VP2	16/666	97.6	0/222	100
VP3	24/738	96.7	3/246	98.8
VP1	25/822	96.9	1/274	99.6
2A	7/213	96.7	2/71	97.2
2B	26/753	96.5	6/251	98.0
2C	3/1005	99.7	1/335	99.7
3A	7/222	96.8	1/74	98.6
3B	3/69	95.7	1/23	95.7
3C	18/657	97.3	1/219	99.5
3D	39/1467	97.3	5/489	98.9
3'-NTR	0/63	100		

Difference number = different nt or aa/total nt or aa.

Results and Discussion

Comparison of nucleotide and amino acid sequences of HHAV isolates LU38 and LY6 with other wt HHAV isolates

Genomes of LU38 and LY6 consisted of 7477 nts excluding the poly(A) tail. The base composition of plus cDNAs of genomes of LU38 and LY6 were as follows: 29.01% and 29.02% of A, 16.25% and 16.14% of C, 21.97% and 22.08% of G, and 32.77% and 32.75% of T, respectively.

The genomes of LU38 and LY6 shared high identities of 97.7% and 99.1%, respectively, with 171 nt and 21 aa differences (Table 2). The 5'-NTR of HHAV contains a 5'-terminal hairpin, two pseudoknots, a pyrimidine-rich tract and an internal ribosomal entry site (IRES). The 5'-NTRs of LU38 and LY6 were 733 nt long with 7 nt differences. Their ORF consisted of 6681 nt , encoding 2227 aa. A hundred sixty-eight nt and 21 aa differences with identities of 97.5% (nt) and 99.1% (aa) were observed in the ORFs of LU38 and LY6 , respectively. A 65 nt and 4 aa differences in the structural protein region, and a 103 nt and 17 aa differences in the non-structural protein region were observed. There were no nt or aa differences in the VP4 region, 16 nt and no aa differences in the VP2 region, 24 nt and 3 aa differences in the VP3 region, and 25 nt and 1 aa differences in the VP1 region (Table 2). The identity in the non-structural protein region ranged from 95.7% to 99.7%, the highest one of 99.7% appeared in the 2C region, suggesting a higher conservation. 2B and 2C proteins were found to consist of 251 aa and 335 aa, respectively.

Both proteins playing certain roles in the replication of viral RNA are considered important for the adaptation of viruses to hosts (Graff *et al.*, 1994). 2C is considered to have helicase and NTPase activities and markers for guanidine resistance. 3A functions as a Vpg precursor, 3B is a genome-linked viral protein (Vpg), 3C is the sole protease for the HHAV protein processing, and 3D is an RNA-dependent RNA polymerase. Multiple mutations in 2B and 2C proteins are considered to contribute to the enhancement of virus replication in cooperation with 5'-NTR, 3A, 3B, 3C and 3D proteins.

In 2A, 2B, 2C, 3A, 3B, 3C and 3D regions, nt identities between LU38 and LY6 ranged from 96.5% to 99.7%, while aa identities ranged from 95.7% to 99.7%. The highest nt difference (39 nt) appearing in the 3D region resulted in 5 aa differences; 26 nt differences appearing in the 2B region resulted in 6 aa differences.

Subgenotyping of HHAV isolates LU38 and LY6

Alignment analysis of 168 nt in the VP1/2A region (nt 3024–3191) of LU38, LY6 and the wt HAV strain GBM of subgenotype IA revealed an identity of 95.8% for LU38 and 97.6% for LY6 with GBM (Table 2). Thus LU38 and LY6 were classified as HHAV isolates of subgenotype IA.

Comparison of nucleotide and amino acid sequences of genomes of HHAV isolates LU38 and LY6 with those of other wt HHAV isolates

Compared with the wt isolates AH1, AH2, AH3, FH1, FH2, FH3, GBM, LA, MBB and HM175 both LU38 and LY6 showed the lowest identities of 91.4% and 91.2% (741 nt and 661 nt differences) with HM175, respectively. LU38

Table 3. Identities of VP1/2A (nt 3024–3191) and genomes between LU38, LY6 and other HHAV isolates

Isolates	LU38 (%)		LY6 (%)	
	VP1/2A	Genome	VP1/2A	Genome
AH1	98.2	98.1	96.4	97.4
AH2	95.8	97.0	95.8	96.6
AH3	94.6	96.5	97.6	96.4
FH1	95.2	96.4	97	96.6
FH2	98.2	97.8	96.4	97.0
FH3	96.4	97.2	98.2	97.7
GBM	95.8	95.2	97.6	95.1
HM175	902.	91.4	90.2	91.2
LA	93.5	95.4	95.2	95.5
MBB	92.9	91.6	92.3	91.4

had the highest identity of 98.1% (140 nt differences) with AH1, but LY6 had the highest identity of 97.7% (172 nt differences) with FH3 (Table 3). In 5'-NTR, LU38 shared the lowest identity of 91.6% with GBM (630 nt differences) and the highest identity of 98.1% (140 nt differences) with AH1. LY6 had a similar identity of 91.4% (642 nt differences) with GBM, but the highest identity of 97.7% (172 nt differences) with FH3. The capsid protein regions were highly conserved among the 12 wt isolates. In the Vp2, Vp3 and Vp1 regions, LU38 and LY6 had high identities of 98.4% to 100% with AH1, AH2, AH3, FH1, FH2, FH3, GBM, MBB and HM175, and lower identities of 95.3% and 94.9% with LA (Table 3), respectively. In non-structural proteins, 2C and 3D regions were more variable, while 2B and 3C regions were more conserved (Table 3). The 3'-NTRs of LU38 and LY6 were the same as those of 6 isolates from Japan.

Phylogenetic relations of HHAV isolates LU38 and LY6 to other wt HHAV isolates

Phylogenetic analysis showed that the 12 wt HHAV isolates formed five subclusters. Subcluster I contained LU38, FH2 and AH1, subcluster II contained LY6 and FH3, subcluster III contained AH3, AH2 and FH1, subcluster IV contained LA and GBM, subcluster V contained MBB and HM175. LU38 was closest to the Japanese isolate AH1, and LY6 was closest to the Japanese isolate FH3. It suggested that phylogenetic relations of various wt HHAV isolates correlate with geographical regions of their origin (Fig. 1).

LU38 and LY6 are two wt HHAV isolates originating from different regions of China. LU38 was isolated in the Jiang Su Province in eastern China, while LY6 in the Liao Ning Province in northern China. The genomes of LU38 and LY6 without poly(A)-tails contained 7477 nt, and LU38 shared a high identity of 97.7% (nt) and 99.1% (aa) with LY6.

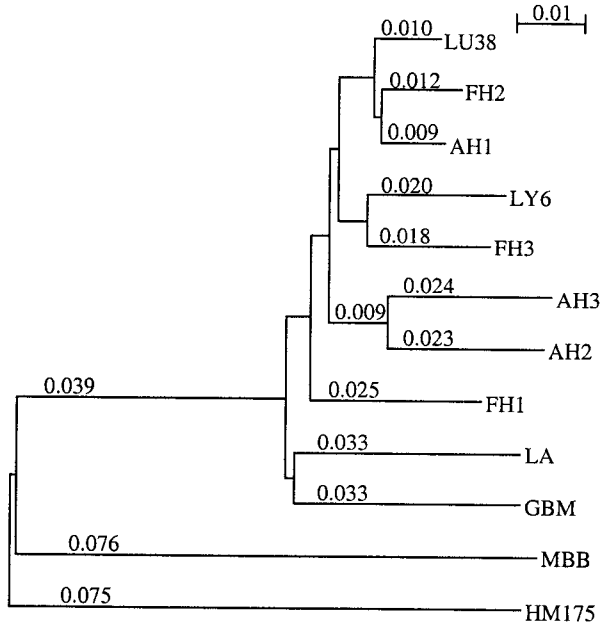

Fig. 1

Phylogenetic tree based on the genomes of 12 wt HHAV isolates

In 5'-NTR, Fuliwara *et al.* (2001) have reported that less nt substitutions were found in FH than AH, but our analysis showed no such a results in comparing LU38 and LY6 with AH, FH, HM175, LA, GBM and MBB.

Genomic sequences comparison of the HHAV isolates showed that the 3'-NTR region was very conserved, containing least differences. The greatest difference appeared in the 5'-NTR region, in which LU38 had the lowest identity of 91.4% with HM175 and the highest identity of 98.1% with AH1. LY6 had the lowest identity of 91.2% with HM175 and the highest identity of 97.4% with AH1. Moreover, in the IRES region of 5'-NTR, the identities were even higher than those in the entire 5'-NTR between LU38, LY6 and other isolates. It suggested the importance of the conserved IRES in different HHAV isolates for translation.

In the region of structural proteins the majority of aa differences between LU38, LY6 and other isolates appeared in capsid proteins. Within this region the highest difference was found in VP1. When the capsid regions of the 12 isolates were compared, about half (20 of 46) of the aa differences were clustered at the N-terminus of VP1, while the C-terminus of VP1 was conserved in accord with the relatedness of the antigenic sites. In the region of non-structural proteins the majority of aa identities between LU38, LY6 and other 10 isolates appeared in the 3C protein, implying that the 3C protein plays an important role in the processing of the polyprotein.

The subgenotyping revealed that LU38 and LY6 were of the subgenotype IA. The phylogenetic analysis and geographical incidence suggested that the epidemiological link of HA had developed in China and Japan.

References

Cohen IJ, Tichehurst RJ, Purcell HR, Burckler-White A, Baroudy BM (1987): Complete nucleotide sequence of wild-type hepatitis A virus: Comparison of different strains of hepatitis A virus and other picornaviruses. *J. Virol.* **61**, 50–59.

Forhman MA, Dush MK, Martin GR (1988): Rapid production of full-length cDNAs from rare trandcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002.

Fujiwara K, Yokosuka O, Fukai K, Imazeki F, Saisho H, Omata M (2001): Analysis of full-length hepatitis A virus genome in sera from patients with fulminant and self-limited acute type A hepatitis. *J. Hepatol.* **35**, 112–119.

Graff J, Normann A, Feinstone SM, Flehmig B (1994): Nucleotide sequence of wild-type hepatitis A virus GBM in comparison with two cell culture-adapted variants. *J. Virol.* **68**, 548–554.

Jansen RW, Newbold EJ, Lemon MS (1990): Molecular epidemiology of hepatitis A defined by an antigen-capture polymerase chain reaction method. *Proc. Natl. Acad. Sci. USA* **87**, 2867–2871.

Kan LY, Zhou TK, Fu TY (1989): An epidemiological study of a hepatitis A outbreak in Shanghai. *Chinese J. Infect. Dis.* **7**, 26–30 (in Chinese).

Lemon SM, Shapiro, CN (1994): The value of immunization against hepatitis A. *Infect. Dis.* **3**, 38–49.

Lemon SM, Robertson BH (1993): Current perspectives in the virology and molecular biology of hepatitis A virus. *Semin. Virol.* **4**, 285–295.

Linemeyer DL, Menke GJ, Martin-Gallardo A, Hughes VJ, Young A, Mitra WS (1985): Molecular cloning and partial sequencing of hepatitis A viral cDNA. *J. Virol.* **54**, 247–255.

Najarian R, Caput O, Gee W, Potter SJ, Renard A, Merryweather J, Van Nest G, Dina D (1985): Primary structure and gene organization of human hepatitis A virus. *Proc. Natl. Acad. Sci. USA* **82**, 2627–2631.

Ovchinnikov IA, Severdlov EV, Tsarev AS, Ariserian GS, Rokhlina OT, Chizhikov EV, Petrov AN, Prikhod'ko GG, Blinov MV, Vasilenko KS, Sandakhchier LS, Kusov II, Grabko IV, Fleer PG, Balaian SM, Drozdov GV (1985): Sequence of 3372 nucleotide units of RNA of the hepatitis A virus, coding the capsids VP4-VP1 and some nonstructural proteins. *Dokl. Akad. Nauk. SSSR* **285**, 1014–1018.

Paul VA, Tada H, von der Helm K, Wissel T, Kiehn R, Wimmer E, Deinhardt E (1987): The entire nucleotide sequence of the genome of human hepatitis A virus (isolate MBB). *Virus Res.* **8**, 153–171.

Robertson BH, Jasen RW, Khanna B, Totsuka A, Nainan OV, Siegl G, Widell A, Margolis HS, Isomura S, Ishizu HT, Moritsugu Y, Lemmon SM (1992): Genetic relatedness of hepatitis A virus strain recovered from different geographical regions. *J. Gen. Virol.* **73**, 1365–1377.

Saitou N, Nei M (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Thomson JD, Higgins DG, Gibson TJ (1994): Improving the sensitivity of progressive multiple sequence alignment through sequence weighing, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 2907–2920.

Totsuka A, Moritsugu D (1999): Hepatitis A virus proteins. *Intervirology* **42**, 63–68.

Ticehurst JR, Racanirllo VR, Baroundy BM, Saltimore D, Purcell RH, Feinstone SM (1983): Molecular cloning and characterization of hepatitis A virus cDNA. *Proc. Natl. Acad. Sci. USA* **80**, 5885–5889.