Identification of a single genotype of hepatitis G virus by comparison of one complete genome from a healthy carrier with eight from patients with hepatitis

Hai-Lin Wang, Yun-De Hou and Dong-Yan Jin

Institute of Virology, Chinese Academy of Preventive Medicine, 100 Yingxin Street, Beijing 100052, China

Different isolates of a putative hepatitis virus called hepatitis GB virus C or hepatitis G virus (HGV) have been cloned recently from patients with hepatitis. This virus has also been found commonly in healthy carriers. We have cloned and sequenced a complete HGV genome, designated HGVCN, from a healthy Chinese blood donor. HGVCN shares 85.8–90.0% nucleotide sequence identity and 95.4–97.5% amino acid identity with the eight available full-length HGV genomes. Furthermore, the majority (82.8%) of the nucleotide substitutions found in HGVCN were synonymous and a fairly uniform distribution of changes was found across the entire genome without identification of any hypervariable region. When compared with the African isolates GBVC and GBVC-EA, the HGVCN-encoded polyprotein contained a 31 amino acid N-terminal extension which was predicted to be a defective core-like sequence. The sequences of the HGV E1 and E2 proteins displayed unique motifs and were highly conserved. Phylogenetic analysis revealed that all nine complete HGV isolates were closely related and that HGVCN grouped with the other Chinese HGV isolate (HGVC964). Taken together, our findings suggest that there is one single genotype of HGV and that the HGV genome cloned from the healthy carrier is not significantly different from those derived from patient sera.

Introduction

Two closely related isolates of a novel virus, designated hepatitis GB virus C or hepatitis G virus (HGV), have recently been identified from patients with viral hepatitis (Leary et al., 1996; Linnen et al., 1996). HGV is a positive, single-stranded RNA virus with a genome about 9.3 kb in length. Its genomic organization resembles that of the Flaviviridae and it is distantly related to hepatitis C virus (HCV).

It is now well-established that HGV is a transfusion-transmitted blood-borne virus highly prevalent around the world and that persistent infection with HGV is common (Aikawa et al., 1996; Egawa et al., 1996; Masuko et al., 1996; Nakatsui et al., 1996; Stark et al., 1996; Tameda et al., 1996; Tsuda et al., 1996; Alter et al., 1997a,b; Brown et al., 1997; Dawson et al., 1997; Feucht et al., 1997; Kinoshita et al., 1997; Theodore & Lemon, 1997; Wang & Jin, 1997; Wu et al., 1997). The association of HGV with hepatitis is suggested by prevalence studies (Yoshida et al., 1995; Colombatto et al., 1996; Fiordalisi et al., 1996; Linnen et al., 1996; Tameda et al., 1996) and in one case by a prospective study of transfusion recipients (Linnen et al., 1996). However, more recent evidence suggests that HGV would probably account for only mild clinical disease at best (Alter, 1996; Egawa et al., 1996; Masuko et al., 1996; Tanaka et al., 1996; Wang et al., 1996; Alter et al., 1997a,b; Bralet et al., 1997; Haydon et al., 1997; Wang & Jin, 1997). Further studies are required to clarify these discrepancies and to reach a final conclusion on the pathogenesis of HGV. It has been speculated that there might be a fulminant HGV strain with crucial mutations in a particular region of the genome (Masuko et al., 1996; Heringlake et al., 1996). It remains to be determined whether this strain does exist in certain populations of hepatitis patients. On the other hand, as all available HGV genome sequences have been obtained from patients with hepatitis (Leary et al., 1996; Linnen et al., 1996; Erker et al., 1996), while HGV is frequently found in healthy carriers, it would be of great interest to obtain a complete HGV genome from a healthy individual and to compare its sequence with those published previously.
HGV shares with HCV not only sequence homology but also routes of transmission. Given the important implications of the profound genetic heterogeneity of HCV in diagnosis, pathogenesis, treatment and prevention (Bukh et al., 1995; Ohba et al., 1995; Simmonds, 1995), it is not surprising that the genomic variation of HGV is also an area of intense investigation (Fukushi et al., 1996; Kao et al., 1996; Brown et al., 1997; Pickering et al., 1997; Viazov et al., 1997; Wang & Jin, 1997; Wu et al., 1997). The quasispecies nature of HGV has been confirmed (Pickering et al., 1997; Viazov et al., 1997), but most reports document only modest sequence heterogeneity among different HGV isolates. The most remarkable findings in sequence analysis are the absence of a core-like region and hypervariable regions in the HGV genome (Erker et al., 1996; Linnen et al., 1996; Simons et al., 1996). However, studies of HGV variation have been based mainly on sequences of short genomic segments. Until now, only eight complete HGV genomes have been sequenced and extensively analysed (Leary et al., 1996; Linnen et al., 1996; Simons et al., 1996). Therefore, it will be necessary to obtain more full-length sequences from different geographical locations and to assess their genomic heterogeneity.

We have investigated the prevalence and genotype of HGV in Chinese professional blood donors and hepatitis patients (Wang & Jin, 1997). We have shown that HGV infection is frequently found in certain populations in China where other hepatitis viruses (A–E) are also hyperendemic. The Chinese HGV isolates share with each other and with the American prototype viruses striking nucleotide sequence identity (> 85%) in the NS3 region. Based on phylogenetic analysis of 446 nt NS3 sequences, we have proposed that there is only one predominant or major genotype of HGV in the world. In the present study, we strengthened and extended our findings by characterizing the complete genomic sequence of a Chinese HGV isolate cloned from a healthy blood donor.

Results and Discussion

Sequence comparison and variability

The complete genomic sequence of the Chinese HGV isolate HGVCN contained 9213 nt with a single long open reading frame coding for a 2873 amino acid polyprotein. Similarity searching via the BLAST and FASTA programs revealed that this is the ninth full-length HGV genome sequence deposited in the current GenBank database.

To assess the variability of HGV across the entire genome, we calculated the percentages of diverged nucleotide and amino acid residues between HGVCN and all different isolates (Table 1). It was noteworthy that the global and local sequence identity between them was exceedingly high and in the same range. Globally, HGVCN shared more than 85% identical nucleotides and more than 95% identical amino acids with all other HGV isolates. The distribution of the diverged residues was fairly even throughout the entire genome, although the E1 and E2 sequences were slightly more variable than those from the NS2–NS5 region. This was in sharp contrast to what is known for HCV (Table 1; compare the variability between HGVCN and HGV-PNF2161 with that between HCVN and HCV-1).

Our results obtained from comparison of nine full-length HGV genomes clearly indicate that HGVCN cloned from a healthy carrier was indistinguishable from other HGV isolates identified from patients with viral hepatitis. This implies that strain difference or mutations in the genome are unlikely to account for the contradictory documentation of the association of HGV with disease.

Given the much higher level of genotypic variations established for HCV (20–50% divergence for nucleotides and 15–45% for amino acids; Bukh et al., 1995; Ohba et al., 1995; Simmonds, 1995), we propose that all nine HGV isolates
compared in this study belong to the same genotype. We believe that this is the single HGV genotype currently identified around the world.

Our proposal is consistent with previously published data from many groups (Erker et al., 1996; Pickering et al., 1997; Viazov et al., 1997; Wang & Jin, 1997). However, it conflicts with two recent studies (Mukaide et al., 1997; Okamoto et al., 1997) which suggest three genotypes instead of one. One of these reports (Mukaide et al., 1997) is based only on analysis of 5’ non-coding sequence from 33 isolates. As previously discussed for HCV (Bukh et al., 1995; Ohba et al., 1995; Simmonds, 1995), final classification of these isolates into different genotypes should await additional sequence data. Analyses performed by the other group (Okamoto et al., 1997) are similar to those presented here. However, the divergence of nucleotide and peptide sequences either within or between the different groups described in that study falls within a narrow range (5–17% for nucleotide and 0–5% for peptide). Moreover, those groups are unlikely to display distinct biological features. In our opinion, the differences between them are not sufficiently significant to justify the division of genotypes.

**Frequency of silent substitutions**

Analysis of the genome sequence and variability of HGV strongly suggests that all HGV isolates available at this time are closely related and that the history of the molecular evolution of HGV and HCV is different. This was further corroborated by comparison of the frequencies of synonymous substitutions in the HGV and HCV genomes. In Table 2, we list side by side the frequencies of the synonymous substitutions found in the Chinese HGV and HCV isolates (HGVCN and HCVCN) when compared with the corresponding prototype viruses. Remarkably, most of the base substitutions (82.8% in HGV are silent and these silent substitutions occur more frequently in HGV than in HCV (59.3%). Similar results were obtained for HCVCN and HGVCN if different HGV and HCV strains were used as the reference (data not shown). These data imply that a selection for synonymous base substitutions was operative in the evolution of HGV. In other words, the intrinsic variability of the RNA genome of HGV was physically constrained by the proper functioning of the viral proteins. This is consistent with the neutral theory of molecular evolution. Additionally, silent changes in the HGV E1, E2 and NS5 regions are more frequent if compared with those in the corresponding regions of HCV, suggesting that

---

**Table 1. Sequence divergence between different isolates of HGV and HCV**

Percentages of diverged nucleotides and diverged amino acids are shown, the latter in parentheses. NCR, non-coding region.

<table>
<thead>
<tr>
<th></th>
<th>HGVCN/ HGVC964</th>
<th>HGVCN/ HGTw</th>
<th>HGVCN/ PNF2161</th>
<th>HGVCN/ GBVC-EA</th>
<th>HGVCN/ GBVC</th>
<th>HGVCN/ R10291</th>
<th>HGVCN/ GT110</th>
<th>HGVCN/ GT230</th>
<th>HCVN/ HCV-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’NCR C</td>
<td>18.6± 10.7</td>
<td>11.4± 11.7</td>
<td>16.9± 15.9</td>
<td>14.8± 5.4</td>
<td>15.9± 9.2</td>
<td>24.1± 22.4</td>
<td>27.7± 20.4</td>
<td>29.2± 27.0</td>
<td>21.6± 23.3</td>
</tr>
<tr>
<td>E1</td>
<td>8.8± 7.3</td>
<td>15.5± 6.8</td>
<td>14.5± 8.1</td>
<td>15.1± 7.8</td>
<td>14.2± 7.4</td>
<td>14.7± 11.7</td>
<td>15.1± 8.1</td>
<td>15.0± 7.9</td>
<td>21.1± 7.9</td>
</tr>
<tr>
<td>E2</td>
<td>10.1± 5.0</td>
<td>14.8± 7.3</td>
<td>16.3± 10.5</td>
<td>13.9± 5.5</td>
<td>14.5± 5.1</td>
<td>15.1± 7.5</td>
<td>16.3± 6.6</td>
<td>11.2± 4.6</td>
<td>21.6± 3.6</td>
</tr>
<tr>
<td>NS2</td>
<td>12.1± 8.3</td>
<td>15.1± 4.4</td>
<td>14.5± 4.4</td>
<td>16.8± 5.6</td>
<td>14.8± 3.6</td>
<td>14.4± 3.6</td>
<td>14.8± 2.8</td>
<td>10.1± 2.4</td>
<td>21.3± 2.6</td>
</tr>
<tr>
<td>NS3</td>
<td>9± 2.7</td>
<td>13.5± 2.3</td>
<td>13.1± 2.3</td>
<td>12.8± 2.1</td>
<td>14.1± 2.9</td>
<td>13.4± 2.3</td>
<td>13.7± 2.0</td>
<td>11.1± 2.1</td>
<td>21.6± 2.9</td>
</tr>
<tr>
<td>NS4</td>
<td>13.9± 2.9</td>
<td>14.5± 3.8</td>
<td>14.7± 4.2</td>
<td>13.3± 3.5</td>
<td>14.1± 3.2</td>
<td>13.9± 3.2</td>
<td>12.7± 3.2</td>
<td>9± 1.3</td>
<td>21.3± 3.0</td>
</tr>
<tr>
<td>NS5</td>
<td>9± 3.8</td>
<td>11.9± 2.5</td>
<td>12.4± 2.8</td>
<td>11.5± 2.6</td>
<td>11.9± 3.6</td>
<td>11.9± 3.1</td>
<td>11.2± 2.1</td>
<td>8.4± 1.2</td>
<td>21.3± 3.0</td>
</tr>
<tr>
<td>3’NCR*</td>
<td>6± 3</td>
<td>14.1± 4.6</td>
<td>13.9± 4.3</td>
<td>14.2± 4.3</td>
<td>13.9± 4.2</td>
<td>13.4± 3.5</td>
<td>10± 2.5</td>
<td>21± 15.2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11.2± 4.3</td>
<td>13.3± 3.8</td>
<td>14.1± 4.6</td>
<td>13.9± 4.3</td>
<td>14.2± 4.3</td>
<td>13.9± 4.2</td>
<td>13.4± 3.5</td>
<td>10± 2.5</td>
<td>21.6± 15.2</td>
</tr>
</tbody>
</table>

* The reported 3’NCR of HGVC964 is unrelated to any sequence in the database and its authenticity has not been established.

**Table 2. Frequencies of synonymous substitutions in HGVCN and HCVCN**

Numbers of synonymous substitutions divided by numbers of total substitutions (%) are shown.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>E1</th>
<th>E2</th>
<th>NS2</th>
<th>NS3</th>
<th>NS4</th>
<th>NS5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGVCN</td>
<td>71.6</td>
<td>65.2</td>
<td>84.5</td>
<td>90.7</td>
<td>86.3</td>
<td>88.9</td>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>HCVN</td>
<td>86.8</td>
<td>56.3</td>
<td>51.7</td>
<td>30.9</td>
<td>77.4</td>
<td>64.7</td>
<td>55.4</td>
<td>59.3</td>
</tr>
</tbody>
</table>
Fig. 1. Sequence alignment of the structural proteins (E1 and E2) from different HGV isolates. Dots indicate identity. Asterisks highlight the conserved cysteines. The putative N-glycosylation sites are underlined.
the positive selection for nonsynonymous changes that was operating on these HGV proteins was weak. This was further supported by the lack of hypervariable regions as shown later in this paper. Collectively, our results are compatible with the notion that HGV is a more adapted virus. This also explains the domination of a single HGV genotype.

**Sequence alignment of structural proteins**

While the HGV proteins were highly conserved in general, the structural proteins were slightly more divergent. The amino acid identity of the HGV E1 and E2 proteins ranged from 83.7% to 91.2% (Table 1). In addition, the nonsynonymous substitutions in the HGV E1 and E2 regions were less frequent than those in HCV (Table 2). However, it was still unclear whether some of the diverged nucleotides would cluster in particular subregions of the HGV structural proteins. Neither was it known whether the sequence of the HGV E1 and E2 proteins displayed unique features or motifs. To address these issues, we generated a multiple sequence alignment of the structural proteins from all different isolates of HGV (Fig. 1). The most striking variation was found in sequences following the initiating methionines. Isolates HGV-GT110 and HGV-R10291 contained a unique N-terminal extension of 93 and 68 amino acids, respectively. The HGVCN isolate shared with three other isolates (HGV-Iw, HGV-PNF2161 and HGVC964) a 31 residue extra domain at the N terminus. These extended sequences have been suggested to be the defective remains of an ancestral core-like protein (Linnen et al., 1996). In isolates GBVC, GBVC-EA and HGV-GT230, the core-like sequence was completely absent. These analyses confirmed the N-terminal heterogeneity of HGV, which could have arisen from single events of nucleotide insertion or deletion. The biological significance of this heterogeneity remains unknown. However, in view of the importance of the HCV core protein (Matsumoto et al., 1996), it would be of great interest to investigate whether the absence of a functional nucleocapsid in HGV is linked to changes in pathogenicity or attenuation of virulence.

It was clear from the sequence alignment (Fig. 1) that the diverged amino acids in HGV do not cluster in particular subregions and that a hypervariable region (HVR) similar to those found in HCV is lacking. The distribution of diverged amino acids was relatively uniform, but slightly more conserved or variable residues did exist. There were 27 cysteines and 4 N-linked glycosylation sites that were well-conserved in all aligned sequences. A cluster of 8 conserved cysteines was found at the N terminus of the HGV E1 protein whereas an equivalent cysteine-rich domain is absent in HCV.

**Phylogenetic analysis**

We performed parsimony analysis and generated a phylogenetic tree based on the progressive pairwise alignment of the full-length nucleotide sequences of all HGV isolates (Fig. 2). The HCV-1 sequence was included as an outgroup since previous studies have suggested the separation of HCV from HGV and GBVA as an earlier event (Leary et al., 1996; Erker et al., 1996). In this tree, HGVCN clustered with the other Chinese isolate, HGVC964, deposited in GenBank by Zhou et al. (1997). The bootstrap support for this cluster (99%) was statistically significant. The two Chinese isolates were closer to the Japanese isolate GT230 and they formed a separate grouping. This was compatible with a slightly higher identity between these isolates detected in sequence comparison (Table 1). On the other hand, the American isolates R10291 and PNF2161, the Japanese isolates HGV-Iw and GT110, and the east African isolate GBVC-EA were more closely related.

A distance matrix tree relating the nine HGV isolates was also constructed (data not shown). This tree was similar but not identical to the parsimony tree, suggesting that the major groupings were consistent but the relative genealogical relationship of the groups remained uncertain. Moreover, the evolutionary distances between different HGV isolates were very close to each other and this lends further support to the proposition of a single genotype of HGV.

In conclusion, our findings from sequence comparison and phylogenetic analysis of the nine full-length HGV genomes strongly suggest that these isolates are closely related and that there is one single genotype of HGV. Extensive comparisons of HGV and HCV sequences revealed that the evolutionary pathway of the two distantly related viruses was significantly different. We also provide evidence that the complete genome sequence of HGV cloned from a healthy carrier was indistinguishable from those isolated from patients with hepatitis. Further investigations are required to elucidate the molecular basis for the different pathogenic potential of HGV and HCV.

We thank Drs Chong-Bai Liu and Li Ruan for helpful discussions and for continuing support. We are grateful to Drs Mei-Yun Zhan, Yu Wang, and HCV.

**Fig. 2.** Parsimony tree of complete HGV genome sequences. Numbers on nodes are bootstrap confidence probabilities. An HCV-1 sequence was used as an outgroup to root the tree.
Ning-Shao Xia and co-workers for access to data prior to deposition in public databases. This work was initiated with financial support from the National Key Laboratory of Molecular Virology (grant MV96002 to H.-L.W. and D.-Y.J.). The joint effort to clone and sequence the complete HGVNCN genome was funded by the State Commission of Science and Technology (High-Tech R&D grant 863/102-07-02-07, to H.-L.W.), H.-L.W., Hong-Song Cheng, Ning-Shao Xia, Wen-Jie Tan, Gang Chen, Yu-Lan Liu, Jing Sun, Ding Zeng, Y.-D.H., D.-Y.J., Yu Wang and Mei-Yun Zhan are credited for this effort.

References


Sequence analysis of nine HGV genomes


Received 18 April 1997; Accepted 14 August 1997