JOURNAL OF VIROLOGY, Sept. 1997, p. 6501–6508 Vol. 71, No. 9
0022-538X/97/$04.00+0

Identification of GB Virus C Variants by Phylogenetic Analysis of 5′-Untranslated and Coding Region Sequences

A. SCOTT MUERHOFF,[1]* DONALD B. SMITH,[2] THOMAS P. LEARY,[1] JAMES C. ERKER,[1] SURESH M. DESAI,[1] AND ISA K. MUSHAHWAR[1]

Virus Discovery Group, Abbott Laboratories, North Chicago, Illinois 60064,[1] and Department of Medical Microbiology, University of Edinburgh Medical School, Edinburgh EH8 9AG, Scotland, United Kingdom[2]

Received 20 February 1997/Accepted 19 May 1997

Phylogenetic analysis of 44 GB virus C (GBV-C) 5′-untranslated region (5′-UTR) sequences from 37 individuals suggested the presence of GBV-C genotypes (A. S. Muerhoff, J. N. Simons, T. P. Leary, J. C. Erker, M. L. Chalmers, T. J. Pilot-Matias, G. J. Dawson, S. M. Desai, and I. K. Mushahwar, J. Hepatol. 25:379–384, 1996) that correlated with geographic origin: type 1, 2a and 2b, and 3 isolates are found predominantly in West Africa, the United States and Europe, and Japan, respectively. We have extended our analysis to include 5′-UTR sequences from 129 globally distributed GBV-C isolates and sequences from the second envelope protein (E2) gene and nonstructural (NS) regions 3 and 5b from a subset of these isolates. Bootstrap analysis of a 157-nucleotide segment of the 5′-UTR from 129 sequences provided weak support for the existence of the four major groups of GBV-C isolates previously described, although phylogenetic analysis of a 374-nucleotide segment of the 5′-UTR from 83 isolates provided stronger support. Thus, the groups of GBV-C variants previously identified upon analysis of the entire 5′-UTR can be distinguished by analysis of the shorter, 374-nucleotide region from the 5′-UTR. In contrast, independent analysis of the E2, NS3, or NS5b region sequences does not identify groups of GBV-C variants that correlate with geographic origin. However, bootstrap analysis of these coding sequences, when linked to form colinear sequences, demonstrates that longer coding regions can produce GBV-C groupings that are similar to that determined from 5′-UTR sequence analysis. The inability to distinguish between GBV-C variants by using small segments of coding sequence suggests that the GBV-C genome is constrained. As a result of these constraints, there is a high degree of nucleotide and amino acid sequence conservation between isolates from widely separated geographic areas. Hence, substitutions at many nucleotide positions are not tolerated, so that substitutions at the positions which can change are saturated, thereby obscuring the evolutionary relationships.

GB virus C (GBV-C) is a newly identified positive-strand RNA virus belonging to the *Flaviviridae* family (18, 34). The virus infects humans and is parenterally and vertically transmitted (6, 9, 20). On the basis of phylogenetic analysis of helicase and replicase genes, as well as the entire polyprotein sequence, GBV-C is most closely related to GB virus A followed by hepatitis C virus (HCV) (18, 38). Its genomic organization is similar to HCV; however, unlike HCV, GBV-C does not appear to encode a nucleocapsid or core protein at the 5′ end of the genome. Instead, the virus encodes a putative envelope glycoprotein, similar to the first envelope protein (E1) of GBV-A, GBV-B, and HCV, as the initial polypeptide of the viral polyprotein. The location of the GBV-C initiator methionine has been determined to be within a few amino acids of the putative E1 signal sequence (33). It has also been shown that GBV-C possesses an internal ribosome entry site and that sequences beginning at the 5′ terminus of the virus and extending into the amino-terminal coding region of the E1 gene are essential for internal ribosome entry site function (33). In addition, studies of the nucleotide sequences of the GBV-C 5′-terminal region (approximately 600 nucleotides [nt]) from 37 individuals did not identify additional sequences, indicating that artifacts due to reverse transcription-PCR amplification were not responsible for the deletion of sequences in this region. Each of the isolates possessed a single conserved AUG codon within three codons of the presumed E1 signal sequence (23). These observations suggest that the 5′ terminus of the GBV-C genome lacks a gene encoding a core polypeptide.

Phylogenetic analysis of 44 GBV-C 5′-untranslated region (5′-UTR) sequences demonstrated the presence of five groups of isolates which correlated with their geographic origin (23). Groups 1a and 1b were found predominantly in Western Africa, while groups 2a and 2b were found in the United States and Europe. Group 3 isolates were found mainly in Japan. Similar observations were made recently by Fukushi et al. (10), who compared 5′-UTR sequences from six Japanese isolates with those from GBV-C and hepatitis G virus (21). However, there is some uncertainty about the significance of these groupings, since similar analysis of a portion of the nonstructural region 3 (NS3) gene of GBV-C isolates failed to demonstrate the same groupings (23). This may have resulted from the small number of sequences analyzed, the limited number of positions in the alignment (only 135 nt), or an unusual property of the region analyzed. Alternatively, the groupings obtained from comparison of 5′-UTR sequences may not be reflected in phylogenetic relationships in the coding region. To address these issues, we obtained the nucleotide sequences from the E2, NS3, and NS5b regions of various GBV-C isolates for which 5′-UTR sequences were already available. We report here a more complete investigation into the phylogenetic relationships between sequences of different isolates from the 5′-UTR and three different coding regions.

* Corresponding author. Mailing address: Department 90D, Bldg. L3, Abbott Laboratories, 1401 Sheridan Rd., North Chicago, IL 60064-4000. Phone: (847) 938-1077. Fax: (847) 938-6042. E-mail: muerhoffs@randb.abbott.com.

MATERIALS AND METHODS

Extraction of human sera and cDNA synthesis. Sera from individuals previously shown to be infected with GBV-C were used as the source of viral RNA. Briefly, RNA was extracted from 25 μl of human serum with commercially available nucleic acid extraction kits as directed by the manufacturer (USB total nucleic acid extraction kit and Qiagen HCV RNA isolation kit). Total nucleic acid or RNA was reverse-transcribed with random hexamers in a final volume of 25 μl with the Perkin-Elmer RT-PCR kit. cDNAs were either used immediately in PCRs or stored at −20°C. One-fifth of the cDNA was amplified in subsequent PCRs with primers (final concentration, typically 0.5 to 1.0 μM) from the 5′-UTR and the E2, NS3, and NS5b regions of the GBV-C genome as described below. All PCRs were performed with the GeneAmp RNA PCR kit (Perkin-Elmer).

PCR amplification of the 5′-UTR. For the majority of the samples, amplification of the GBV-C 5′-UTR was performed by a previously described method (23). Briefly, two oligonucleotide primers, S1 (5′-CACTGGGTGCAAGCCCCAGAA-3′) and GBVCE1wb2 (5′-CAGGGCGCAACAGTTTGTGAG-3′) were used to amplify approximately 600 nt from the 5′ end of the GBV-C genome. Other GBV-C 5′-UTR-specific primers located internal to the S1 and GBVCE1wb2 primers (22) were used to amplify sequences when the S1-GBVCE1wb2 primer combination was unsuccessful. Thus, in some cases, only a portion of the 5′-UTR sequence was obtained. Thermocycling conditions were as previously described (22).

PCR amplification of E2 sequences. PCR amplification of a portion of the GBV-C E2 gene was performed in a volume of 12.5 μl with either of two sets of primers: GBV-C-E2-5′ (5′-AGCAGCGTATTGTCATGGTCTTC-3′) and GBV-C-E2-3′ (5′-GTACAGCTGGCAGAGCCAACTGG-3′), or GBV-C-E2s1 (5′-CCTCCTCCTGCTGGAGCAGCG-3′) and GBV-C-E2a1 (5′-GATCCCAGTGCGGCTATGGTACAG-3′). Second-round PCR (25 μl) was performed with nested primers GBV-C-E2s2 (5′-CGTATTGTCATGGTCTTCCTCCTG-3′) and GBV-C-E2a2 (5′-CAAACCAAGACACAGAACCCCAC-3′). The region of the E2 gene amplified (357 bp) corresponds to the putative amino terminus of the protein (7). Both first- and second-round PCR were performed by the touchdown PCR protocol described previously (28).

PCR amplification of NS3 sequences. PCR amplification of GBV-C NS3 region sequences (188 bp) was performed with 5 μl of cDNA (out of 25 μl prepared) in a volume of 25 μl as previously described (19).

PCR amplification of NS5b sequences. PCR amplification of GBV-C NS5b region sequences (402 bp) was performed in a volume of 25 μl with 2.0 mM $MgCl_2$ and 0.5 μM each oligonucleotide primer: gbvc-ns5-5 (5′-AGGAGGCAATAAGGACTGTTAGGC-3′) and gbvc-ns5-3 (5′-CTGTCGAAGCAAGTGGCATCCAC-3′). Amplification was performed by the touchdown PCR method as described previously (28).

Sequencing and sequence analysis. PCR products were separated by electrophoresis through a 2% agarose gel and then excised and purified with the QIAEX gel extraction kit (Qiagen). Purified PCR products were sequenced directly on an ABI 373 DNA sequencer with ABI sequencing ready reaction kit (Perkin-Elmer) and GBV-C gene-specific primers. When insufficient PCR product was obtained for direct sequencing, the amplicons were cloned, two or three clones from each were sequenced, and consensus sequences were then generated; when variation between clones was significant, they were analyzed independently. Nucleotide and amino acid sequences were aligned with PILEUP (Wisconsin Sequence Analysis Package; Genetics Computer Group, Madison, Wis.).

Evolutionary distances between sequences were determined with the DNADIST program (Kimura two-parameter method) of the PHYLIP package, version 3.5c (8). The analysis was performed for all sites within the 5′-UTR. The computed distances were used for the construction of phylogenetic trees by using the neighbor-joining method of the program NEIGHBOR. This method of analysis was used, as opposed to the distance matrix program FITCH or maximum-likelihood or parsimony method, because of computational limitations. However, while the neighbor-joining method does not necessarily produce the minimum-evolution tree, it does produce a tree with the correct topology (17, 29).

The robustness of the trees was assessed by bootstrap resampling (1,000 data sets) of the multiple-sequence alignments with the programs SEQBOOT, DNADIST, and NEIGHBOR. The consensus tree was calculated with CONSENSE. Bootstrap values of less than 70% are regarded as not providing evidence for the phylogenetic grouping. The trees were produced with the PHYLIP program RETREE with the midpoint rooting option. The final graphical output was created with the program TREEVIEW (25).

GBV-C genome analysis. Windows analysis of synonymous and nonsynonymous distances between complete coding sequences were computed with the program Windows (13). The groups of sequences compared were as follows: (i) GBV-C (GBV-C, U36380; PNF2161, U44402; R10291, U45966; GBV-C(EA), U63715; HGVC964, U75356; HGV-Iw, D87255) (ii) isolates of HCV subtype 1b (BK, M58335; J, D90208; J483, D01217; JTA, D01171; JKIG, X61596; C2, D10934; GENOM, L02836; UNKCDS, M96362; GENANTI, M84754; L2, U01214; N, S62220; 1S, D50483; 2S, D50485; 3S, D50484; PP, D30613; HD-1, U45476), and (iii) subtypes of type 1 (1a: PT, M62231; 1b: BK, M58335; 1c: G9, D14853). Alignment of the coding sequences of different variants of GBV-C revealed two regions (6412 to 6444, and 8500 onwards) where the isolate HGVC964 differs markedly on both the nucleotide and amino acid levels. The first of these regions can be aligned by assuming that a frameshift has occurred in the sequence of HGVC964, but no obvious realignment is possible for the second region, and so this extreme 3′ region of the coding sequence was excluded from the Windows analysis.

Nucleotide sequence accession numbers. The sequences reported in this paper have been deposited in the GenBank data base (accession no. A007950 to A008113).

RESULTS

Phylogenetic analysis of 5′-UTR sequences. We previously demonstrated that GBV-C isolates could be separated into five groups or genotypes by comparison of nucleotide sequences of the 5′-UTR (23). To investigate the validity of these groupings, we expanded our data set to 129 GBV-C 5′-UTR sequences of different length from 120 individuals (Tables 1 and 3; Fig. 1) and performed phylogenetic analysis. Analysis of the largest region of overlap available for all 129 sequences, i.e., 157 nucleotides (positions 137 to 292; numbering based upon the GBV-C prototype), suggested the presence four distinct groups (1, 2a, 2b, and 3) of sequences (Fig. 2). Bootstrap resampling of the data revealed robust segregation of a group including the sequences previously classified as genotype 1 (23) and gave weaker support for group 2 (73%) and 3 (73%, if isolate 140 is considered to be a group 3 isolate). There was, however, no support for the grouping of sequences including those classified earlier as genotype 2a (41% of trees), 2b (59% of trees), or genotype 3 (44%). Stronger support for the presence of the three major groups and two subgroups was obtained by phylogenetic analysis of 83 GBV-C 5′-UTR sequences of 374 nt in length (nt 79 to 447 of GBV-C [Fig. 3]), where groups 1 and 2 were observed in 100% and 99% of bootstrap replicates and groups 2a, 2b, and 3 were supported by 69 to 75% of replicates. There was one anomaly in that isolate 140 (China) segregated with group 2 isolates upon analysis of the longer 5′-UTR segment instead of with group 3 sequences when the shorter segment was analyzed (compare Fig. 2 and 3). Regardless of the grouping of isolate 140, phylogenetic analysis of either the 157- or 374-nt segment of the 5′-UTR results in bootstrap support of the presence of three major groups of GBV-C isolates. Analysis of the 374-nt region supported the separation of group 2 into subgroups 2a and 2b.

5′-UTR sequences contributing to the observed groupings. Bootstrap resampling of our previous alignment, which included the entire GBV-C 5′-UTR region from nt 34 to 630 (44 sequences) (23), provided very strong support for the presence of four groups of sequences: 1, 2a, 2b, and 3, with 100, 98, 93, and 100% bootstrap values for each of these groups, respectively (see Fig. 5). To determine which regions of the 5′-UTR were contributing to the groupings observed by analysis of the entire 5′-UTR and to determine if there was any effect of sampling artifact on the groupings obtained, we performed phylogenetic analysis on other regions of the alignment of 44 sequences described above, including the regions that produced the trees shown in Fig. 2 and 3. As shown in Fig. 5, bootstrap analysis of either the 137 to 292 or 79 to 477 regions supported the existence of three major groups of sequences, although bootstrap support for the existence of subgroups 2a and 2b was obtained only by analysis of the 79 to 477 region. Analysis of 259 positions from this alignment of 44 sequences (nt 34 to 292) resulted in the segregation of groups 1, 2a, 2b, and 3 with bootstrap values of 100, 80, 82, and 99%, respectively (see Fig. 5). Analysis of the 339 positions downstream of position 292 of this same alignment provided bootstrap support for groups 1, 2, and 3, but not for the 2a and 2b subgroups. Regardless of the length or relative position of the 5′-UTR

TABLE 1. GBV-C isolates, geographic origin, and grouping

Isolate[a]	Origin[b]	Group[c]	Region(s)[d]	Isolate	Origin	Group	Region(s)	Isolate	Origin	Group	Region(s)
1	USA	2a	1	49	France	2a	1	97	USA	2a	1
2	USA	2a	1, 2, 3	50	France	2a	1	98	USA	2a	1
3	Greece	2a	1, 4	51	France	2a	1	99	USA	2a	1
4	Italy	2a	1	52	France	1b	1	100	USA	2a	1
5	USA	2a	1	53	France	2a	1	101	USA	2b	1
6	USA	2a	1, 2, 3, 4	54	France	2a	1	102	USA	3	1
7	USA	2a	1, 3	55	France	2b	1	103	USA	3	1
8	USA	2a	1, 2	56	France	2a	1	104	USA	1	1
9	USA	2a	1, 2, 3, 4	57	France	2a	1	105	USA	1	1
10	USA	2a	1	58	France	3	1	106	USA	2a	1
11	USA	2a	1	59	France	2a	1	107	USA	1	1
12	Greece	2b	1	60	France	2a	1	108	USA	3	1
13	Greece	2b	1	61	France	2a	1	109	USA	3	1
14	USA	2b	1, 2, 3, 4	62	France	2a	1	110	USA	2a	1
15	USA	2b	1, 2, 3	63	France	2a	1	111	USA	1	1
16	E. Africa	2b	1, 2, 3, 4	64	France	2b	1	112	USA	2a	1
17	Greece	2b	1, 2, 3, 4	65	France	2b	1	113	USA	3	1
18	USA	2b	1, 2, 3, 4	66	France	2a	1	114	USA	3	1
19	USA	2b	1, 2, 3, 4	67	Germany	2a	1, 2	115	USA	2a	1
20	USA	2b	1	68	Germany	2a	1, 2	116	USA	2a	1
21	Japan	3	1, 2, 3, 4	69	Germany	2b	1, 2	117	USA	3	1
22	Japan	3	1, 2, 3, 4	70	Ghana	1	1, 3	118	USA	3	1
23	Ghana	1	1, 2, 3, 4	71	Greece	2a	1	119	Vietnam	3	1
24	Ghana	1	1	72	Greece	2a	1	120	Vietnam	2a	1, 2
25	Ghana	1	1, 2, 4	73	Japan	2b	1	121	Vietnam	3	1, 2
26	Ghana	1	1	74	Japan	3	1	122	Ghana	?	2, 3, 4
27	Ghana	1	1	75	Japan	2a	1	123	USA	ND[e]	4
28	Ghana	1	1, 2	76	Spain	2a	1	124	USA	ND	3
29	Ghana	1	1	77	Spain	2a	1	125	Canada	ND	3
30	Ghana	1	1, 2, 3, 4	78	UK	2a	1	127	USA	ND	3
31	Ghana	1	1, 2, 3, 4	79	UK	2a	1	128	USA	ND	3
32	Ghana	1	1, 4	80	USA	2a	1	129	USA	ND	3
33	Ghana	1	1, 2, 4	81	USA	2a	1	130	Italy	ND	3
34	Ghana	1	1, 2, 4	82	USA	2a	1	131	Italy	ND	3
35	Ghana	1	1	83	USA	2a	1, 4	132	USA	ND	3
36	Ghana	1	1, 4	84	USA	2a	1	133	USA	ND	3
37	Ghana	1	1, 4	85	USA	2a	1, 4	134	Japan	3	1
38	Ghana	1	1	86	USA	2a	1, 4	135	Japan	3	1
39	Ghana	1	1, 3	87	USA	1	1, 2, 4	136	Japan	3	1
40	Ghana	1	1, 4	88	USA	2a	1	137	Japan	3	1
41	Ghana	1	1	89	USA	2a	1	138	Japan	3	1
42	Ghana	1	1, 4	90	USA	2a	1	139	Japan	3	1
43	USA	2a	1, 2, 3, 4	91	USA	2b	1	140	China	3?	1, 2, 3, 4
44	USA	2a	1, 2, 3, 4	92	USA	2a	1	141	Japan	2a	1, 2, 3, 4
45	Argentina	2a	1	93	USA	3	1				
46	Argentina	2b	1	94	USA	2a	1				
47	Argentina	2a	1	95	USA	2b	1				
48	France	2a	1	96	USA	2b	1				

[a] Isolates for which the complete genome sequence is available are as follows: 16, GBV-C(EA), U63715; 30, GBV-C prototype, U36380; 43, HGV PNF2161, U44402; 44, HGV R10291, U45966; 140, Chinese isolate HGVC964, U75356; and 141, Japanese isolate HGV-Iw, D87255 (31). The 5′-UTR sequences from isolates 1 to 42 are from reference 23. NS3 helicase sequences from isolates 7, 16, 18, 30, and 70 were reported previously (34). Isolates 134 to 139 are from reference 10. Isolates 48 to 51 were obtained from a single individual over a 4-year period (patient 8 [1]). Isolates 76 and 77 were obtained from a single individual, 3 years apart (patient 4 [12]). Isolates 5, 10, 11, and 20 were erroneously reported earlier as being from Europe (23) but were actually from the United States.

[b] USA, United States; UK, United Kingdom.

[c] Group designations are based on analysis of 5′-UTR sequences as shown in Fig. 3 and as previously defined (23).

[d] Region refers to the GBV-C genomic segment(s) sequenced, where 1 is the 5′-UTR, 2 is E2, 3 is NS3, and 4 is NS5b.

[e] ND, not determined.

segment we analyzed, there was no support for different groupings of sequences other than the three major groups observed; that is, sequences that grouped together by analysis of the entire 5′-UTR also grouped together by analysis of any of the shorter 5′-UTR segments examined. The only exception was isolate 140 from China (compare Fig. 2 and 3). We have found that analysis of the entire 5′-UTR sequence provides the strongest bootstrap support for identification of three groups and two subgroups of GBV-C sequences (Fig. 4).

5′-UTR-defined groupings and geographic origin. The distribution of GBV-C isolates by group assignment (defined by the analysis presented in Fig. 3) and geographic origin is shown in Table 2. While there is no absolute correlation between country or geographic region of origin, certain trends in the worldwide distribution of GBV-C groups are apparent. Group 2a was the prevailing variant in the United States and Europe, accounting for 57 and 71% of all isolates, respectively. This group was also found in Argentina, Japan, and Vietnam. The

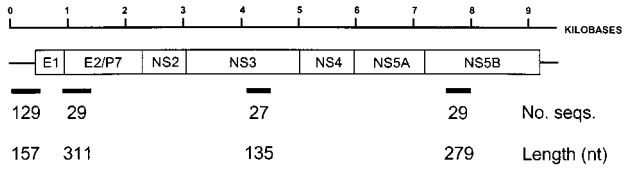

FIG. 1. GBV-C genome map. Coding-region sequences are indicated by the open rectangle. The 5′- and 3′-UTR are depicted by the lines at either end of the putative viral polyprotein. The relative size and position of the putative structural and nonstructural (NS) genes are indicated. The positions along the genome, length, and number of sequences obtained from the noncoding and coding regions for phylogenetic analysis are shown below the map.

majority of group 1 isolates were from Ghana, with five group 1 isolates from the United States but only one from Europe. Group 3 isolates were predominant in Japan but were also found in the United States and more rarely in Europe. Group 2b isolates occur with similar frequency to group 3 isolates in the United States although group 2b isolates were found much more often than group 3 isolates among the European isolates examined. The geographic distribution of GBV-C groups determined by phylogenetic analysis strongly correlates with the geographic distribution of sequence motifs (i.e., polymorphisms) observed within the 5′-UTR, as we described previously (23). Thus, GBV-C groups are defined by phylogenetic analysis of the 5′-UTR and by the orderly geographic occurrence of sequence substitutions within the 5′-UTR. It is clear that GBV-C is globally distributed, with certain groups predominating in particular geographic regions.

FIG. 2. Consensus phylogenetic tree of GBV-C isolates based on 157 positions within the 5′-UTR (nt 137 to 292). The tree was generated from an alignment of 129 sequences including the 44 sequences previously reported (23). The sequences are identified by numbers corresponding to those in Table 1. The internal node numbers represent the bootstrap values (expressed as a percentage of all trees) obtained from 1,000 replicates. The use of this portion of the 5′-UTR shows that GBV-C sequences separate into four major groups, 1, 2a, 2b, and 3, as indicated in boldface type next to the corresponding major branch. The tree was rooted by using the midpoint of the longest branch, and branch lengths are proportional to the evolutionary distance between sequences. A distance scale in nucleotide substitutions per position is shown.

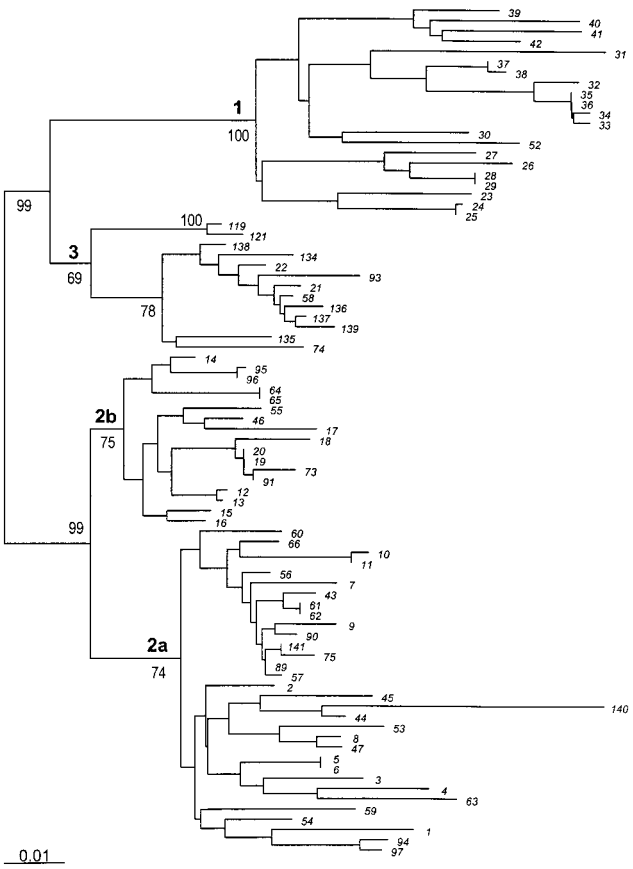

FIG. 3. Consensus phylogenetic tree of 83 GBV-C isolates by using 374 nucleotides within the 5′-UTR (nucleotides 79 to 447).

Phylogenetic analysis of coding-region sequences. Analysis of short coding regions from the E2, NS3, or NS5b gene (Fig. 1; Table 3) from a subset of isolates failed to produce the same phylogenetic groupings as observed in the 5′-UTR (data not shown). For all three coding regions, a single distribution of evolutionary distances was observed between different isolates. Sequences from different 5′-UTR groupings were intermingled, and any grouping of sequences observed was not supported by bootstrap resampling. For example, analysis of 29 E2 sequences revealed that most 5′-UTR-defined group 2a sequences grouped together with a high bootstrap value (94%) but other 2a sequences grouped with group 2b, while group 1 and 3 sequences were mixed (data not shown).

One possible explanation for this inconsistent grouping of sequences is that the coding regions compared were too short. To address this possibility, we created longer stretches of coding sequence by combining the E2, NS3, and NS5b nucleotide sequences from 16 isolates. Phylogenetic analysis of these artificially joined sequences produced the consensus tree shown in Fig. 4. Isolates classified as group 2a by 5′-UTR analysis grouped together with 99% bootstrap support. Group 2b iso-

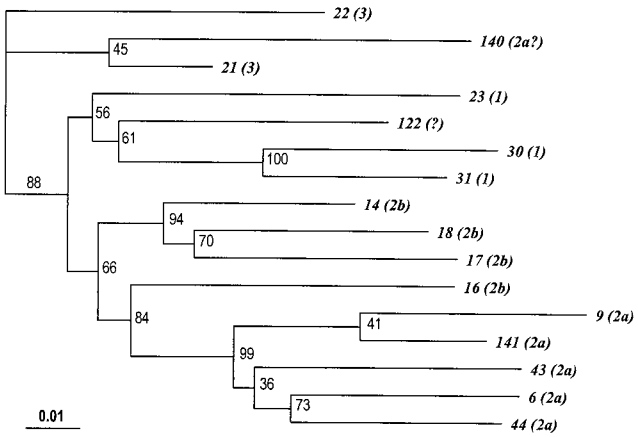

FIG. 4. Consensus phylogenetic tree of 16 GBV-C isolates obtained with concatenated E2, NS3, and NS5b sequences. Sequences from these separate regions of the viral genome were linked together to form colinear sequences for each of the 16 isolates. Phylogenetic analysis was then performed as described in Materials and Methods. Bootstrap values obtained from 1,000 resamplings of the data are shown. The 5′-UTR-determined groupings, based on the consensus tree shown in Fig. 2, are shown in parentheses next to each isolate number. The 5′-UTR sequence is not available for isolate 122; therefore, the group designation is not shown.

lates also grouped together in 94% of the trees, although isolate 16 [GBV-C(EA) (7)] appears to separate from both 2a and 2b isolates in 84% of the trees. Isolate 140 (China) segregated with the group 3 isolates in 88% of the trees, while group 1 isolates shared a common branch in only 56% of the trees. Phylogenetic analysis of the E2, NS3, and NS5b sequences of these 16 isolates as separate regions did not support the existence of GBV-C groups (1,000 bootstrap resamplings [data not shown]). Thus, the grouping of the 16 isolates does not result from the reduced number of sequences in the analysis. With the exception of the low bootstrap support for group 1 isolates and the altered grouping of isolate 140, the groups defined by analysis of the joined E2, NS3, and NS5b sequences were similar to those obtained upon analysis of the long 5-UTR region (Fig. 3). Therefore, relatively long regions of coding sequence are required to distinguish the groupings defined by 5′-UTR sequence analysis.

TABLE 2. Geographic distribution and 5′-UTR groupings of GBV-C isolates from 120 individuals[a]

Origin	No. of isolates of type:				
	1	2a	2b	3	Total
United States	5	31	9	9	54
Europe	1	20	6	1	28
Japan		2	1	9	12
China				1?	1
Vietnam		1		2	3
West Africa	18				18
East Africa			1?		1
Argentina		2	1		3

[a] The table shows only isolates that are unique; i.e., multiple sequences from the same individual of the same group are counted only once (Table 1). The total number of unique GBV-C isolates from each region is shown in the far right column. The group assignment of the Chinese and East African isolates are uncertain.

TABLE 3. GBV-C genomic regions analyzed: number of sequences, length of sequence, and relative position along the genome

Region	n[a]	Length[b]	Location[c]
5′-UTR			
Short	129	157	137–292
Long	83	374	79–447
E2	29	311	1119–1429
NS3	27	135	4273–4407
NS5b	29	279	7711–7989
E2+NS3+NS5b[d]	16	730	

[a] n is the number of sequences examined.
[b] Length refers to the number of positions in the alignment and may be longer than the region amplified due to the insertion of gaps in some sequences to optimize the alignment.
[c] Location is based on the numbering of the prototype GBV-C isolate (GenBank accession no. U36380).
[d] The lengths of the E2 and NS3 regions are as shown for their separate analysis. The NS5b sequences used in the combined coding region analysis were 8 nt longer than that used for independent analysis of NS5b sequences.

DISCUSSION

The investigation of sequence diversity among different isolates of GBV-C is important because variants may differ in their patterns of serologic reactivity, pathogenicity, virulence, responses to therapy, etc. Such differences have been reported for genotypes of HCV, where patients infected with genotype 1 respond less well to interferon treatment than do patients infected with genotype 2 (reviewed in reference 5). However, at present the extent to which different isolates of GBV-C can be grouped into distinct genotypes is uncertain. Phylogenetic analysis of 44 GBV-C 5′-UTR sequences from 37 individuals has provided evidence for the presence of distinct GBV-C groups that correlate with geographic origin (23), and similar findings have been reported by others (10, 11). However, similar groupings are not observed after analysis of NS3 (2, 3, 14, 16, 26, 30, 36) or NS5a (27, 37) sequences. This inconsistency could arise because the regions considered were too short, or too variant, or because distinct genotypes of GBV-C do not exist.

To clarify the extent of variation among different isolates of GBV-C, we have compared the 5′-UTR, E2, NS3, and NS5b sequences (Fig. 1) in a set of isolates from a variety of geographic locations. The phylogenetic relationships observed for

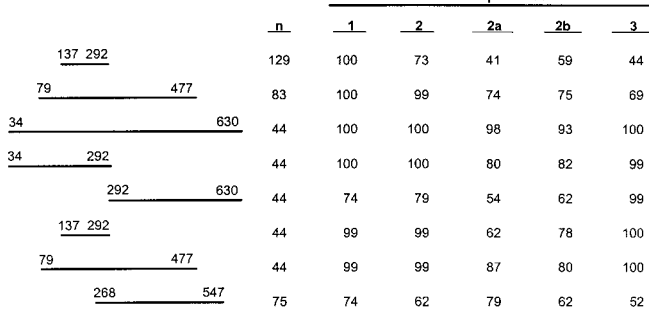

FIG. 5. Bootstrap values for the GBV-C groups and subgroups obtained upon phylogenetic analysis of various regions of the GBV-C 5′-UTR. The position and relative length of the 5′-UTR segment analyzed is depicted on the left. The numbering is based on the GBV-C prototype, where the 5′-most nucleotide is position 1. Bootstrap values are the results of 1,000 resamplings of the data and are expressed as percentages. n is the number of sequences in each data set. The values obtained for the data sets consisting of 129 or 83 isolates are from Fig. 2 and 3. The data sets consisting of 44 isolates are derived from the previously reported alignment (23).

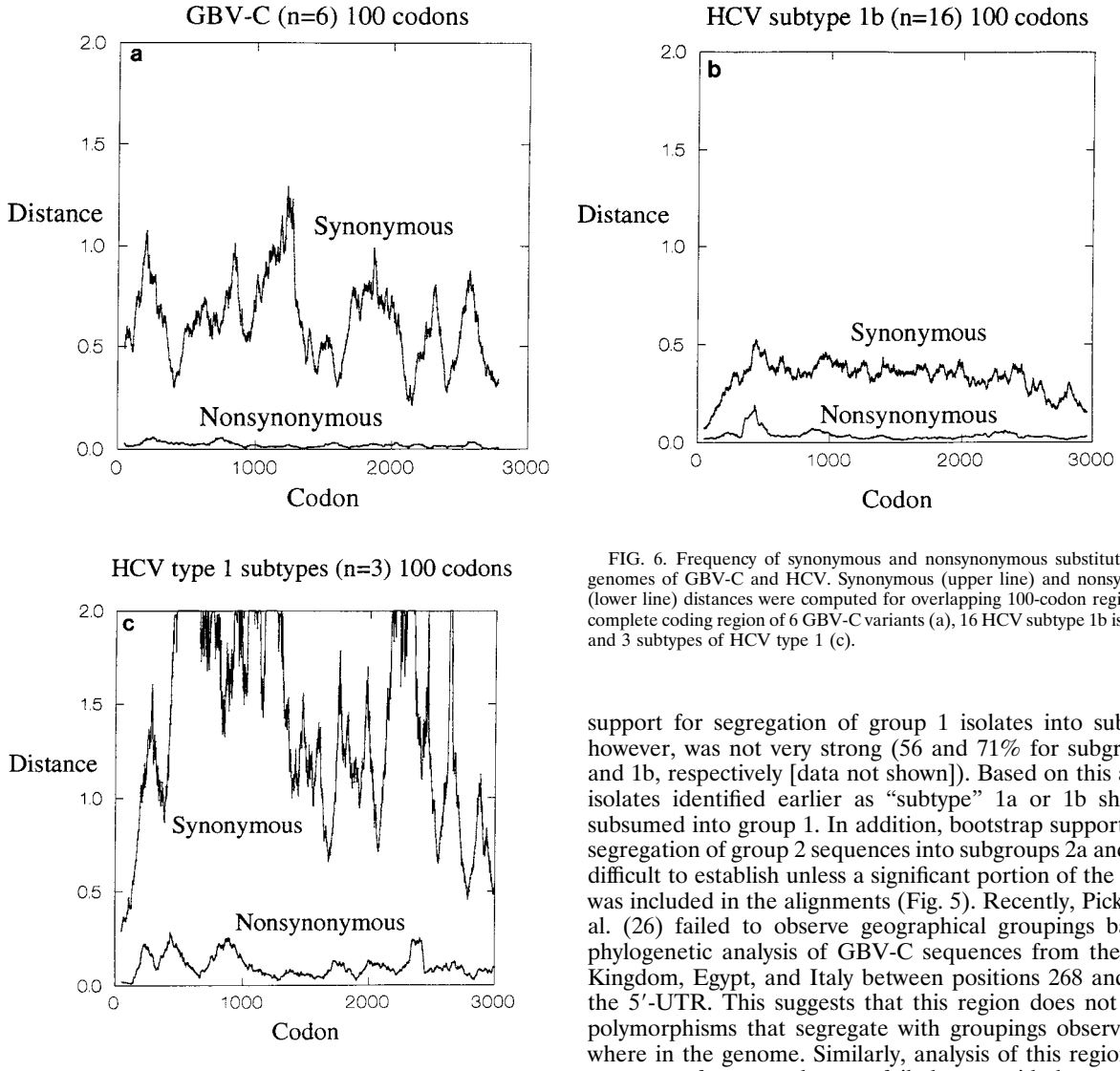

FIG. 6. Frequency of synonymous and nonsynonymous substitution across genomes of GBV-C and HCV. Synonymous (upper line) and nonsynonymous (lower line) distances were computed for overlapping 100-codon regions of the complete coding region of 6 GBV-C variants (a), 16 HCV subtype 1b isolates (b), and 3 subtypes of HCV type 1 (c).

5′-UTR sequences (Fig. 2 and 3) were not observed when analysis was confined to a single coding region. For example, the groupings observed by analysis of 135 nt from the NS3 region of 26 isolates were different from those obtained by 5′-UTR analysis and did not correlate with geographic origin. The same was true for a larger data set (130 isolates) over a shorter segment within NS3 (length of sequence overlap, 80 nt [data not shown]). However, when E2, NS3, and NS5b sequences from 16 isolates were concatenated, phylogenetic analysis produced similar groupings of sequences to those observed in the analysis of 5′-UTR sequences of these isolates (Fig. 4). This suggests that phylogenetic analysis of short segments of coding sequence is not sufficient to identify GBV-C genotypes. Thus, conclusions regarding the presence of GBV-C strains or genotypes based on analysis of short segments within the NS3 region alone, or other coding regions, may not accurately reflect the true phylogenetic relationships between the isolates.

Our analysis demonstrated that three groups of GBV-C variants were consistently identified by analysis of any of the 5′-UTR segments we examined. Previously, we had separated group 1 isolates into subgroups 1a and 1b (23). The bootstrap support for segregation of group 1 isolates into subgroups, however, was not very strong (56 and 71% for subgroups 1a and 1b, respectively [data not shown]). Based on this analysis, isolates identified earlier as "subtype" 1a or 1b should be subsumed into group 1. In addition, bootstrap support for the segregation of group 2 sequences into subgroups 2a and 2b was difficult to establish unless a significant portion of the 5′-UTR was included in the alignments (Fig. 5). Recently, Pickering et al. (26) failed to observe geographical groupings based on phylogenetic analysis of GBV-C sequences from the United Kingdom, Egypt, and Italy between positions 268 and 547 of the 5′-UTR. This suggests that this region does not contain polymorphisms that segregate with groupings observed elsewhere in the genome. Similarly, analysis of this region for 75 sequences from our data set failed to provide bootstrap support for groups 2 and 3 (Fig. 5).

Analysis of the short region of the 5′-UTR (nt 137 to 292) from the Chinese isolate (isolate 140) suggests that it may segregate with group 3 isolates (Fig. 2), while analysis of a longer region results in its segregation with group 2a isolates (Fig. 3). The tree obtained from examination of E2, NS3, and NS5b sequences (Fig. 4) demonstrates that this isolate segregates with group 3 sequences (86% of trees). This discrepancy could be resolved if the analysis was extended to longer coding regions or to a 5′-UTR sequence that included the extreme 5′ region, which contains several group-specific substitutions (23). Unfortunately, the genome sequence of the Chinese isolate obtained from GenBank is 78 nt shorter than the GBV-C prototype sequence; hence, these 5′-terminal sequences are not available for analysis. Similarly, isolate 16, also known as GBV-C(EA) (7), segregated with subgroup 2b isolates upon analysis of the long 5′-UTR segment (Fig. 3), in accordance with our previous analysis of the entire 5′-UTR of this isolate (23). However, this grouping was not supported by analysis of the adjoined coding region sequences (Fig. 4). The ambiguous results observed for isolates 16 and 140 could be attributable to template shuffling during RT-PCR or to recombination of viral genomes, assuming that the individuals were infected with

more than one GBV-C genotype. However, there is no evidence available to suggest that either individual was at risk for multiple infections (due to frequent transfusions, for example), which would increase the possibility that such genome interactions could occur. It has been shown that recombination between HCV genotypes is extremely rare (15, 35), and since GBV-C and HCV share similar geographic restriction of genotypes, it is possible that recombination between GBV-C types is equally rare. Alternatively, the discordant groupings of the Chinese and East African isolates may indicate that accurate phylogenetic analysis of GBV-C sequences may require even longer coding-region sequences than that used here or may require carefully defined portions of the 5′-UTR. Confirmation of the groupings obtained by our analysis of coding- and noncoding-region sequences will have to await the availability of additional full-length GBV-C genomic sequences from widely separated geographic areas.

Our comparison of phylogenetic analysis of GBV-C with different coding and noncoding regions indicates a very different pattern of variation from that documented in HCV, where consistent phylogenetic relationships are observed for many different subgenomic regions including E1 and NS5b (4, 24, 32). Part of the reason for this difference may lie in the relative conservation of GBV-C isolates in nucleotide and amino acid sequence in comparison to that between HCV types (7). We have investigated this issue more closely by analyzing the distances between six GBV-C complete genome sequences at synonymous and nonsynonymous sites for different regions of the genome (Fig. 6a). Throughout the GBV-C genome, nonsynonymous distances (d_N) were much lower than synonymous distances (d_S), with a mean d_N-to-d_S ratio of 0.033. Nonsynonymous distances were relatively constant throughout the genome, with only slight increase in variability within the E2 and NS2 genes and no evident hypervariable region, as noted previously (7). Comparison of this pattern of variability with that observed for 16 different isolates of HCV subtype 1b (Fig. 6b) revealed that nonsynonymous distances were similar to or lower than those between different isolates of GBV-C despite the divergence at synonymous sites being equal or greater for the GBV-C sequences, and consequently the d_N-to-d_S ratio for HCV subtype 1b was higher (0.112). Similarly, synonymous distance between HCV subtype 1 isolates partially overlapped with those between GBV-C sequences, but nonsynonymous distances were lower for the GBV-C sequences (Fig. 6c). These more detailed comparisons strengthen the suggestion that there is strong selection against nonsynonymous substitution throughout the GBV-C genome. This is reflected in the high degree of polyprotein sequence conservation even among isolates from widely separated geographic regions.

The observation that nonsynonymous substitutions are strongly suppressed in the GBV-C genome implies that analysis of amino acid sequences (as opposed to nucleotide sequences) will require the analysis of longer sequences in reconstructing the phylogenetic relationships between different variants. However, this does not fully explain why analysis of large coding regions is required to distinguish between different groups of GBV-C variants defined by sequence relationships in the 5′-UTR. Different subtypes of HCV are closer to saturation at synonymous sites than are variants of GBV-C (Fig. 6a and c), and yet they can be distinguished by analysis of synonymous substitutions within a 222-nt fragment of NS5b or a 300-nt fragment of E1 (data not shown). Hence, the inability to reliably distinguish between GBV-C variants by using fragments of this size suggests that the GBV-C genome is subject to constraint at both synonymous and nonsynonymous sites. Synonymous substitutions might be constrained because of unusually strong codon preferences or because of noncoding functions of the RNA genome. As a result of these constraints, it may be that substitutions at many nucleotide positions cannot be tolerated, so that substitutions at those positions which can change have become saturated, thereby obscuring phylogenetic relationships.

ACKNOWLEDGMENTS

We acknowledge the competent technical assistance of Michelle Chalmers, Jennifer Lund, and Julie Yamaguchi. We also thank Peter Simmonds for valuable discussions during the preparation of the manuscript.

REFERENCES

1. **Allain, J.-P., S. H. Dailey, Y. Laurian, D. S. Vallari, A. Rafowicz, S. Desai, and S. G. Devare.** 1991. Evidence for persistent hepatitis C virus (HCV) infection in hemophiliacs. J. Clin. Invest. **88:**1672–1679.
2. **Berg, T., U. Dirla, U. Maumann, H.-G. Heuft, S. Kuther, H. Lobeck, E. Schreier, and U. Hopf.** 1996. Responsiveness to interferon alpha treatment in patients with chronic hepatitic C coinfected with hepatitis G virus. J. Hepatol. **25:**763–768.
3. **Brown, K. E., S. Wong, M. Buu, T. V. Binh, T. V. Be, and N. S. Young.** 1997. High prevalence of GB virus C/hepatitis G virus in healthy persons in Ho Chi Minh City, Vietnam. J. Infect. Dis. **175:**450–453.
4. **Bukh, J., R. H. Miller, and R. H. Purcell.** 1995. Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes. Semin. Liver Dis. **5:**41–63.
5. **Dawson, G. J., A. S. Muerhoff, L. Birkenmeyer, T. Leary, and S. Desai.** 1996. Molecular biology of hepatitis viruses. Curr. Hepatol. **16:**83–141.
6. **Dawson, G. J., G. G. Schlauder, T. J. Pilot-Matias, D. Thiele, T. P. Leary, P. Murphy, J. E. Rosenblatt, J. N. Simons, F. E. A. Martinson, R. A. Gutierrez, J. R. Lentino, C. Pachucki, A. S. Muerhoff, A. Widell, G. Tegtmeier, S. Desai, and I. K. Mushahwar.** 1996. Prevalence studies of GB virus-C using reverse-transcriptase-polymerase chain reaction. J. Med. Virol. **50:**97–103.
7. **Erker, J. C., J. N. Simons, A. S. Muerhoff, T. P. Leary, M. L. Chalmers, S. M. Desai, and I. K. Mushahwar.** 1996. Molecular cloning and characterization of a GB virus C isolates from a patient with non-A-E hepatitis. J. Gen. Virol. **77:**2713–2720.
8. **Felsenstein, J.** 1993. PHYLIP inference package, version 3.5c. Department of Genetics, University of Washington, Seattle.
9. **Feucht, H.-H., B. Zollner, S. Polywka, and R. Laufs.** 1996. Vertical transmission of hepatitis G. Lancet **347:**615–616.
10. **Fukushi, S., C. Kurihara, N. Ishiyama, H. Okamura, F. B. Hoshino, A. Oya, and K. Katayama.** 1996. Nucleotide sequence of the 5′ noncoding region of hepatitis G virus isolated from Japanese patients: comparison with reported isolates. Biochem. Biophys. Res. Commun. **226:**314–318.
11. **Gonzalez-Perez, M. A., H. Norder, A. Bergstrom, E. Lopez, K. A. Visona, and L. O. Magnius.** 1997. High prevalence of GB virus C strains genetically related to strains with Asian origins in Nicaraguan hemophiliacs. J. Med. Virol. **52:**149–155.
12. **Guerrero, E., A. Guerrero, L. Gil, R. Montes, J. Mateos, M. Cunningham, D. Vallari, J. Casey, S. Watanabe, B. Zeck, S. Desai, and S. Devare.** 1994. Serological response to hepatitis C virus (HCV) in serial bleeds from hemodialysis patients, p. 485–488. *In* K. Nishioka, H. Suzuki, S. Mishiro, and T. Oda (ed.), Viral hepatitis and liver disease. Springer-Verlag, New York, N.Y.
13. **Ina, Y.** 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol. Evol. **40:**190–226.
14. **Kao, J.-H., P.-J. Chen, S.-C. Hsiang, W. Chen, and D.-S. Chen.** 1996. Phylogenetic analysis of GB virus C: comparison of isolates from Africa, North America, and Taiwan. J. Infect. Dis. **174:**410–413.
15. **Kato, N., Y. Ootsutama, T. Tanaka, M. Nakagawa, T. Nakazawa, K. Muraiso, S. Ohkoshi, M. Hijikata, and K. Shimotohno.** 1992. Marked sequence diversity in the putative envelope proteins of hepatitis C viruses. Virus Res. **22:**107–123.
16. **Kinoshita, T., K. Miyake, H. Nakao, T. Tanake, F. Tsuda, H. Okamoto, Y. Miyakawa, and M. Mayumi.** 1997. Molecular investigation of GB virus C infection in hemophiliacs in Japan. J. Infect. Dis. **175:**454–457.
17. **Kuhner, M. K., and J. Felsenstein.** 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11:**459–468.
18. **Leary, T. P., A. S. Muerhoff, J. N. Simons, T. J. Pilot-Matias, J. C. Erker, M. C. Chalmers, G. G. Schlauder, G. J. Dawson, S. M. Desai, and I. K. Mushahwar.** 1996. The sequence and genomic orginization of GBV-C: a novel member of the *Flaviviridae* associated with human non A–E hepatitis. J. Med. Virol. **48:**60–67.
19. **Leary, T. P., A. S. Muerhoff, J. N. Simons, T. J. Pilot-Matias, J. C. Erker, M. L. Chalmers, G. G. Schlauder, G. J. Dawson, S. M. Desai, and I. K. Mushahwar.** 1996. Consensus oligonucleotide primers for the detection of GB virus C in human cryptogenic hepatitis. J. Virol. Methods **56:**119–121.

20. **Lin, H.-H., J.-H. Kao, P.-J. Chen, and D.-S. Chen.** 1996. Mechanism of vertical transmission of hepatitis G. Lancet **347:**1116.
21. **Linnen, J., J. Wages, Z.-Y. Zhang-Keck, K. E. Fry, K. Z. Krawczynski, H. Alter, E. Koonin, M. Gallagher, M. Alter, S. Hadziyannis, P. Karayiannis, K. Fung, Y. Nakatsuji, J. W.-K. Shih, L. Young, M. P. Jr., C. Hoover, J. Fernandez, S. Chen, J.-C. Zou, T. Morris, K. C. Hyams, S. Ismay, J. D. Lifson, G. Hess, S. K. H. Foung, H. Thomas, D. Bradley, H. Margolis, and J. P. Kim.** 1996. Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. Science **271:**505–508.
22. **Muerhoff, A. S., J. N. Simons, J. C. Erker, S. M. Desai, and I. K. Mushahwar.** 1996. Conserved nucleotide sequences within the GB Virus C 5′ untranslated region: design of PCR primers for detection of viral RNA. J. Virol. Methods **62:**55–62.
23. **Muerhoff, A. S., J. N. Simons, T. P. Leary, J. C. Erker, M. L. Chalmers, T. J. Pilot-Matias, G. J. Dawson, S. M. Desai, and I. K. Mushahwar.** 1996. Sequence heterogeneity within the 5′-terminal region of the hepatitis GB virus C genome and evidence for genotypes. J. Hepatol. **25:**379–384.
24. **Ohba, K.-I., M. Mizokami, T. Ohno, K. Suzuki, E. Orito, Y. Ina, J. Y. N. Lau, and T. Gojobori.** 1995. Classification of hepatitis C virus into major types and subtypes based on molecular evolutionary analysis. Virus Res. **36:**201–214.
25. **Page, R. D. M.** 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:**357.
26. **Pickering, J. M., H. C. Thomas, and P. Karayiannis.** 1997. Genetic diversity between hepatitis G virus isolates: analysis of nucleotide variation in the NS-3 and putative 'core' peptide genes. J. Gen. Virol. **78:**53–60.
27. **Pujol, F. H., Y. E. Khudyakov, M. E. Cong, L. Blitz-Dorfman, C. L. Loueiro, F. Capriles, S. Beker, F. Liprandi, and H. A. Fields.** 1996. Unpublished data.
28. **Roux, K. H.** 1994. Using mismatched primer-template pairs in touchdown PCR. Bio/Techniques **16:**812–814.
29. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for the reconstruction of phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
30. **Schreier, E., M. Hohne, U. Kunkel, T. Berg, and U. Hopf.** 1996. Hepatitis GBV-C sequences in patients infected with HCV contaminated anti-D immunoglobulin and among i.v. drug users in Germany. J. Hepatol. **25:**385–389.
31. **Shao, L., H. Shinzawa, K. Ishikawa, X. Zhang, M. Ishibashi, H. Misawa, N. Yamada, H. Togashi, and T. Takahashi.** 1996. Sequence of hepatitis G virus genome isolated from a Japanese patient with non-A-E hepatitis: amplification and cloning by long reverse transcription PCR. Biochem. Biophys. Res. Commun. **228:**785–791.
32. **Simmonds, P., D. B. Smith, F. McOmish, P. L. Yap, J. Kolberg, M. S. Urdea, and E. C. Holmes.** 1994. Identification of genotypes of hepatitis C virus by sequence comparisons in the core, E1 and NS-5 regions. J. Gen. Virol. **75:**1053–1061.
33. **Simons, J. N., S. M. Desai, D. E. Schultz, S. M. Lemon, and I. K. Mushahwar.** 1996. Translation initiation in GB viruses A and C: evidence for internal ribosome entry and implications on genome organization. J. Virol. **70:**6126–6135.
34. **Simons, J. N., T. P. Leary, G. J. Dawson, T. J. Pilot-Matias, A. S. Muerhoff, G. G. Schlauder, S. M. Desai, and I. K. Mushahwar.** 1995. Isolation of novel virus-like sequences associated with human hepatitis. Nat. Med. **1:**564–569.
35. **Smith, D. B., J. Mellor, L. M. Jarvis, F. Davidson, J. Kolberg, M. Urdea, P.-L. Yap, and P. Simmonds.** 1995. Variation in the hepatitis C virus 5′ non-coding region: implications for secondary structure, virus detection and typing. J. Gen. Virol. **76:**1749–1761.
36. **Tsuda, F., S. Hadiwandowo, N. Sawada, M. Fukuda, T. Tanaka, H. Okamoto, Y. Miyakawa, and M. Mayumi.** 1996. Infection with GB virus C (GBV-C) in patients with chronic liver disease or on maintainance hemodialysis in Indonesia. J. Med. Virol. **49:**248–252.
37. **Viazov, S.** 1996. Unpublished data.
38. **Zuckerman, A. J.** 1996. Alphabet of hepatitis viruses. Lancet **347:**558–559.