The sequence and genomic organization of a GB virus A variant isolated from captive tamarins

Thomas P. Leary, Suresh M. Desai, James C. Erker and Isa K. Mushahwar

Virus Discovery Group, Experimental Biology Research, Abbott Laboratories, Dept. 90D, Bldg. L3, 1401 Sheridan Road, North Chicago, IL 60064, USA

Recently, gene fragments of several novel variants of GB virus A were isolated from the serum of distinct monkey species that had not been experimentally inoculated with an infectious agent. These variants appeared to be species-specific in that sequences isolated within a species were virtually identical, though sequences were strikingly different when compared between each species. In the present study, the nucleotide sequence of one of these variants, GBV-A_{lab}, was extended to near-genome length. Similar to the other GB viruses, GBV-A_{lab} appears to encode a single large polyprotein of 2967 amino acids that is post-translationally cleaved by cellular and viral proteases into the individual viral proteins. The structural proteins are found at the N-terminal end of the polyprotein, while the nonstructural proteins are found at the C terminus. Amino acid sequence comparisons of the large polyprotein demonstrate that GBV-A_{lab} is 74% identical to GBV-A and 48% identical to GBV-C, sharing only marginal identity with GBV-B and HCV-1 at 27%. Examination of the GBV-A_{lab} polyprotein reveals that structural motifs are conserved for a protease, a helicase and a replicase. Phylogenetic analysis of the polyprotein confirms previous results that GBV-A_{lab} is a member of the Flaviviridae, distinct from GBV-B and HCV, though more closely related to GBV-A and GBV-C.

Introduction

The GB agent was originally characterized by Deinhardt et al. (1967) in tamarins (Saguinus sp.) inoculated with the serum of a patient experiencing acute hepatitis in the third day of jaundice. Acute hepatitis in these animals was demonstrated by elevations in serum alanine aminotransferase levels, as well as liver histopathology. Acute phase serum from the inoculated animals could be further passaged in tamarins, allowing the viral origin of the agent to be established (Deinhardt et al., 1967; Schlauder et al., 1995a). Studies designed to isolate the infectious agent responsible for causing hepatitis in the tamarin model revealed the presence of two distinct viruses (Simons et al., 1995b), termed GB virus A and B (GBV-A and GBV-B, respectively).

Both GBV-A and GBV-B have positive-strand, linear RNA genomes approximately 10 kb in length (Simons et al., 1995b). Each virus potentially encodes a large polyprotein that appears to be post-translationally cleaved into the individual viral proteins (Muerhoff et al., 1995). The structural genes are located at the N terminus of the polyprotein, while the nonstructural genes are found at the C terminus. Conserved within these viral proteins are structural motifs for a helicase and an RNA-dependent RNA polymerase, each associated with the supergroup II helicases and replicases of positive-strand RNA viruses, respectively. Phylogenetic analyses demonstrated that each virus is a member of the Flaviviridae, most closely related to the hepatitis C virus (HCV) group (Muerhoff et al., 1995). Polyprotein sequence comparisons of GBV-A and GBV-B show 27% identity to one another and approximately 28% identity to HCV-1, suggesting that GBV-A and GBV-B are representatives of two new genera within the Flaviviridae. PCR studies using degenerate oligonucleotide primers designed in conserved helicase sequences in GBV-A, GBV-B and HCV were used to isolate an additional Flaviviridae-like virus, GBV-C, from human serum (Leary et al., 1996b; Simons et al., 1995a).

Utilizing oligonucleotide primers designed to the 5' nontranslated region (NTR) of GBV-A, we have recently reported the isolation of species-specific GBV-A variants from the serum of tamarins (Saguinus labiatus), mystax (Saguinus
and owl monkeys (Aotus trivirgatus), termed GBV-
A<sub>lab</sub>, GBV-A<sub>xic</sub> and GBV-A<sub>tr</sub>, respectively (Leary et al., 1996a). The variants are virtually identical when isolated within a given species (> 95%). However, when sequences between species are compared, dramatic differences are observed, with only 52–79% identity to the GBV-A prototype virus within the 5<sup>′</sup> NTR. These observations have since been confirmed and extended, and five species-specific variants of GBV-A have now been described (Bukh & Apgar, 1997). In the present study, we describe the genomic extension and organization of GBV-A<sub>lab</sub> report a detailed analysis of the viral genome, the viral proteins and compare GBV-A<sub>lab</sub> with other members of the Flaviviridae.

**Methods**

Serum from a tamarin (T1059) previously identified as harbouring a variant of GBV-A (Leary et al., 1996a) was used as the nucleic acid source material. Total nucleic acids were extracted from 50 µl of serum using the DNA/RNA Isolation Kit (US Biochemical), and then subjected to reverse transcription (50 µl volume) in the presence of random hexamers with a Perkin-Elmer RNA PCR Kit as directed by the manufacturer. Sequential genome extension reactions (25 µl volume) were performed by anchored ‘touchdown’ PCR as previously described (Roux, 1994). The final genomic extension utilized a specifically primed cdNA as template. Following amplification, gene products were cloned into pT7Blue T-vector (Novagen), and then each strand was sequenced with an Applied Biosystems model 373 DNA sequencer.

Nucleotide sequences were compiled utilizing the Fragment Assembly Program of the Wisconsin Sequence Analysis Package (Version 8, September 1994, Genetics Computer Group, Madison, Wisconsin, USA). Sequence alignments were performed with the GAP or PILEUP programs of the same package with the default settings in place for gap creation and gap extension. Phylogenetic distances between pairs of the amino acid sequences were determined by the PROTDIST program (Felsenstein, 1993) of the PHYLIP package (version 3.5c). These computed distances were utilized for the construction of phylogenetic trees using the program FITCH with TREEVIEW (Page, 1996) producing the final output.

GenBank accession numbers of viral sequences used in the alignments are as follows: GBV-A, U22303; GBV-B, U22304; GBV-C, U36380; HCV-1 (genotype 1a), M62321; HCV/JK1 (genotype 1b), X61596; HCV/J6 (genotype 2a), D00944; HCV/J8 (genotype 2b), D10988; HCV3A (genotype 3a), D28917.

**Results and Discussion**

**Genomic analyses of GBV-A<sub>lab</sub>**

We have previously reported that oligonucleotide primers specific for the 5′ NTR of GBV-A could amplify a virus gene product from tamarins not experimentally inoculated with an infectious agent (Leary et al., 1996a; Schlauder et al., 1995b). Though these sequences (termed GBV-A<sub>lab</sub>) are virtually identical to one another when isolated from tamarins (greater than 96%), they are only 76% identical to sequences within the GBV-A 5′ NTR. To demonstrate that these sequences were part of a larger viral entity, the gene fragment was sequentially extended by ‘anchored’ PCR. The sense primers in these reactions were gene specific, designed to match known regions of the GBV-A<sub>lab</sub> genome, while the antisense primers were GBV-A specific, designed within regions of high homology between GBV-A and GBV-C (Table 1). We reasoned that regions of homology between GBV-A and GBV-C, two viruses that are 52% identical at the nucleotide level, would also be conserved in GBV-A<sub>lab</sub>. Six sequential PCR reactions were performed allowing the genome to be extended to 9550 nucleotides in length, 103 nucleotides short of GBV-A. Based on identity plots with GBV-A, 59 nucleotides may exist on the 5′ end of GBV-A<sub>lab</sub> which have yet to be deciphered, while an additional 84 nucleotides may exist at the 3′ end of the GBV-A<sub>lab</sub> genome. Thus, the GBV-A<sub>lab</sub> genome appears slightly larger than GBV-A by approximately 40 nucleotides, for the most part within the coding region of the virus. Based on the existing sequences, GBV-A<sub>lab</sub> is 67-9% identical to GBV-A and 51-9% identical to GBV-C at the nucleotide level.

As had been shown with the other GB viruses, a large open reading frame (ORF) appears to be encoded by GBV-A<sub>lab</sub> (2967 amino acids). Within the ORF, GBV-A<sub>lab</sub> is most similar to GBV-A, being 74% identical. Like GBV-A, GBV-A<sub>lab</sub> is 48% identical at the amino acid level to GBV-C (Leary et al., 1996b), and shares only limited identity (27%) to GBV-B and HCV-1. When the putative viral polyprotein cleavage products of GBV-A<sub>lab</sub> are compared to those of GBV-A or GBV-C, extensive regions of identity are found within the individual proteins (Table 2). Particularly high identities, when compared to the overall average, are found within the NS3 and NS5 regions; these regions encode a protease/helicase and replicase, respectively, in HCV. Detailed analyses of the putative NS3 helicase and the NS5B replicase regions are shown in Fig. 1. For the most part, those amino acid residues held in common between the GBV-A, GBV-B, GBV-C and HCV-1 helicases (Fig. 1A) are also conserved within the region of analysis in GBV-A<sub>lab</sub>. Of interest is that all residues typically defining the supergroup II helicases of positive-strand RNA viruses (Koonin & Dolja, 1993) are absolutely conserved between these viruses, including GBV-A<sub>lab</sub>. Correspondingly, all common residues within the NS5B replicases of GBV-A, GBV-B, GBV-C and HCV-1 are also conserved within this region of GBV-A<sub>lab</sub> (Fig. 1B). Similarly, those residues that define the supergroup II replicases of positive-strand RNA viruses (Koonin & Dolja, 1993) are also conserved between these five viruses. Thus, both within the helicase and replicase regions, each of these viruses is very highly conserved, indicative of the functional importance of these domains to virus replication.

Unlike GBV-B and HCV (Muerhoff et al., 1995; Choo et al., 1991), GBV-A<sub>lab</sub> lacks a highly basic, core-like protein at the N terminus of the polyprotein. Of interest, both GBV-A and GBV-C also lack a core-like protein in this region (Simons et al., 1996). Examination of the GBV-A<sub>lab</sub> genome in each of the six translational reading frames failed to uncover the presence of a
potential, appropriately sized basic protein. Thus, GBV-\textsubscript{A\textsubscript{lab}} does not appear to encode a core-like protein. Approximately 508 nucleotides exist upstream of the GBV-\textsubscript{A\textsubscript{lab}} translation start codon, in the presumed 5' NTR. This region is very similar in size to both GBV-A and GBV-C, each of which possess an internal ribosome entry site (IRES) within the 5' NTR (Simons \textit{et al}, 1996) that allows for cap-independent translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to conclusively demonstrate the presence of an IRES translation of the large polyprotein. Though empirical data are needed to concl...
replication is unknown. Within the GBV-A$_{lab}$ E2 protein, three potential N-linked glycosylation sites are present, though one of these motifs has a proline residue in the second position and thus may not be glycosylated (Gavel & von Heijne, 1990). Of note, each of these three motifs is conserved within GBV-A$_{lab}$ (Erker et al., 1996). Like GBV-A and GBV-C, the envelope genes of GBV-A$_{lab}$ are unique from GBV-B and HCV in the relatively low number of potential N-linked glycosylation sites that exist. This would result in the relative molecular mass of these proteins being substantially smaller than those of HCV and GBV-B. It is possible that significant O-linked glycosylation of the GBV-A$_{lab}$ envelope proteins occurs, though this does not appear to be the case in the GBV-C E2 that has been expressed in recombinant systems (Pilot-Matias et al., 1996a). It remains to be determined whether the lack of glycosylation is a disadvantage to the virulence or infectivity of these viruses. Finally, each of the 19 cysteine residues in GBV-A$_{lab}$ E2 is conserved within the GBV-A E2, and 18 of these are conserved within GBV-C E2, suggesting that these proteins fold in a very similar manner despite the high degree of divergence observed at the amino acid level.

In the HCV system, cleavage of the nonstructural gene products occurs by way of two distinct viral proteases. The first of these is a zinc-dependent thiol protease encoded within an NS2/NS3 gene product, and is responsible for the autocatalytic cleavage of NS2 from NS3 (Grakoui et al., 1993b). Amino acid residues 952 (His) and 993 (Cys) within the HCV NS2 protein constitute the active site of this enzyme. Though minimal conservation is found between GBV-A$_{lab}$, GBV-A and GBV-C within the NS2 region (Table 2), appropriately spaced His and Cys residues are found at positions 886 and 927 of the GBV-A$_{lab}$ polyprotein, similar to GBV-A (Muerhoff et al., 1995), GBV-C (Leary et al., 1996b) and HCV-1 (Grakoui et al., 1993a). This would suggest a similar mode of processing by GBV-A$_{lab}$ at the NS2/NS3 junction, the scissile bond which is predicted to occur between amino acids 962 and 963 of the polyprotein.

The protease responsible for cleavage of the remaining HCV polyprotein is a serine protease encoded within the first one-third of the NS3 protein, utilizing the NS4A gene product as a co-factor (Failla et al., 1994; Lin et al., 1994). Comparable to other serine proteases, the functional site of this enzyme is a triad (Grakoui et al., 1993a) comprising amino acid residues 1083 (His), 1107 (Asp) and 1165 (Ser) in HCV. As shown in Table 2, identity in the NS3 region between GBV-A$_{lab}$, GBV-A and GBV-C is substantially greater than the average identity between the entire polyproteins. Further, as is the case with GBV-A and GBV-C (Muerhoff et al., 1995; Leary et al., 1996b),
appropriately spaced His, Asp and Ser residues are conserved at positions 1018, 1042 and 1098 of the GBV-A<sub>lab</sub> polyprotein, respectively. Thus, it would appear that GBV-A<sub>lab</sub> processes the downstream nonstructural gene products by way of a virally encoded serine protease much like HCV.

In the HCV system, the NS3 serine protease cleaves the polyprotein at specific sites in which the scissile bond is flanked by Cys/Thr at the P1 position and Ser/Ala at the P1′ position (Grakoui et al., 1993a). Additionally, an acidic residue is usually found in the P6 position, though this residue has not been shown to be an absolute requirement (Bartenschlager et al., 1995). Previous studies of GBV-A suggested that Thr or Ala occurs in the P1 position, while Ser is found in the P1′ position (Muerhoff et al., 1995). Amino acid sequence alignments of the predicted large polyproteins from the GB viruses and several distinct types of HCV were used to calculate the phylogenetic distances between the polyproteins. When comparing the various HCV types, differences range between 0–174 and 0–396 amino acid replacements per position, whereas GBV-A<sub>lab</sub> and GBV-A differ by 0–338 amino acid replacements per position (data not shown). Based on this analysis, it is clear that GBV-A<sub>lab</sub> and GBV-A are types of the same virus, comparable to what was observed when this same analysis was performed on the 5′ NTR of these viruses which suggested that they were at least subtypes of one another (Leary et al., 1996a). Buhk & Apgar (1997) have confirmed this observation, though they have described two distinct variants of GBV-A isolated from S. labiatus. Comparisons of GBV-A<sub>lab</sub>, GBV-A and GBV-C to GBV-B or the HCV isolates unambiguously differentiate these viruses from one another (> 2-2 substitutions per position). As would be expected from the identity values (Table 2), the differences between GBV-A<sub>lab</sub>, GBV-A and GBV-C are much smaller (0-338–0-989).

The calculated distances were used to generate the unrooted phylogenetic tree in Fig. 3. The HCV isolates group to a single major branch of the tree, as does GBV-B, GBV-A<sub>lab</sub>, GBV-A and GBV-C group on a third major branch of the tree. This branch is further subdivided, distinguishing GBV-A<sub>lab</sub> and GBV-A from GBV-C. The distances between GBV-A<sub>lab</sub> and GBV-A are roughly equivalent to those observed between the various HCV isolates. Again, these data support the hypotheses that GBV-A<sub>lab</sub> and GBV-A are types of the same virus, distinct from the other GB viruses or HCV, though clearly within the Flaviviridae.

In this report, we describe the sequence and genomic organization of a novel variant of GBV-A, occurring in captive tamarins that have not been knowingly inoculated with an infectious agent. Like the other GB viruses, GBV-A<sub>lab</sub> has an RNA genome of approximately 10 kb in length which possesses a large open reading frame in excess of 2900 amino acids that is presumably translated as a single polyprotein. The polyprotein appears to be processed into the individual viral proteins similar to the other Flaviviridae in that both cellular and viral proteases are involved. Within the individual viral proteins, conserved protease, helicase and replicate motifs are...
Fig. 3. Phylogenetic analysis of Flaviviridae polyproteins. Evolutionary distances were used to generate the unrooted phylogenetic tree as described in Methods. Scale is provided in amino acid replacements per position.

present. Phylogenetic analyses demonstrate that GBV-A$_{lab}$ is a member of the Flaviviridae, most closely related to GBV-A, though branching on the phylogenetic tree with GBV-C.

It is intriguing to note that apparently healthy animals harbour a blood-borne virus in the absence of disease. GBV-A has been transmitted in tamarins independent of GBV-B, yet these animals show no conspicuous illness or elevation in liver enzyme levels (Schlauder et al., 1995a). These animals remain persistently infected with the virus and do not develop antibody responses to recombinant antigens expressed in E. coli-based systems (Pilot-Matias et al., 1996b). Further, these viruses are very highly conserved at the nucleic acid and amino acid levels, indicating an absence of immune selective pressures. This is in stark contrast to HCV, another persistent virus which seems to escape immune clearance. Perhaps viral persistence needs to be evaluated at two levels, one in which the virus escapes immune clearance, while in the other, the virus escapes immune detection. Supporting the later argument is the human immune response to trans, or that prevents it from arising, may allow a partial understanding of viral persistence.

The authors wish to thank Michelle Chalmers and Julie Yamaguchi for excellent technical assistance, and Dr Scott Muerhoff for critical review of the manuscript.

References


Received 17 March 1997; Accepted 6 June 1997