Structural constraints on RNA virus evolution

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
Journal of Virology

Publisher Rights Statement:
Copyright © 1999, American Society for Microbiology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
The recently discovered hepatitis G virus (HGV) or GB virus C (GBV-C) is widely distributed in human populations, and homologues such as HGV/GBV-C_{CPZ} and GBV-A are found in a variety of different primate species. Both epidemiological and phylogenetic analyses support the hypothesis that GB viruses coevolved with their primate hosts, although their degree of sequence similarity appears incompatible with the high rate of sequence change of HGV/GBV-C over short observation periods. Comparison of complete coding sequences (8,500 bases) of different genotypes of HGV/GBV-C showed an excess of invariant synonymous sites (at 23% of all codons) compared with the frequency expected by chance (10%). To investigate the hypothesis that RNA secondary-structure formation through internal base pairing limited sequence variability at these sites, an algorithm was developed to detect covariant sites among HGV/GBV-C sequences of different genotypes. At least 35 covariant sites that were spatially associated with potential stem-loop structures were detected, whose positions correlated with positions in the genome that showed reductions in synonymous variability. Although the functional roles of the predicted secondary structures remain unclear, the restriction of sequence change imposed by secondary-structure formation provides a mechanism for differences in net rate of accumulation of nucleotide substitutions at different sites. However, the resulting disparity between short- and long-term rates of sequence change of HGV/GBV-C violates the assumptions of the “molecular clock.” This places a major restriction on the use of nucleotide or amino acid sequence comparisons to calculate times of divergence of other viruses evolving under the same structural constraints as GB viruses.

Although the flavivirus hepatitis G virus (HGV) or GB virus C (GBV-C) is newly discovered (17, 19), several observations suggest that it may have always infected humans and originated through coevolution with its primate hosts. First, it is widely distributed in human populations, with frequencies of active or past infection ranging from 5 to 15%, and its distribution extends even to highly isolated populations, such as indigenous tribes in Papua New Guinea and Central and South America. Second, although infection is frequently persistent and associated with high levels of circulating viremia, no evidence links HGV/GBV-C to any identifiable hepatic or nonhepatic disease, consistent with a process of mutual adaptation. Third, the geographical distribution of HGV/GBV-C variants reflects that of ancient human migrations (14). For example, sequences from the Far East are almost invariably genotype 3, and this genotype is otherwise found only in native inhabitants of North and South America. In contrast, Caucasian and other populations from India westward including Northern Africa are infected with genotype 2. Genotype 1 is confined to sub-Saharan Africa and shows the greatest overall sequence diversity (22, 25); particularly divergent variants have been recovered from Pygmy and other African populations (27a, 28). These genotype distributions can potentially be mapped to the emergence and migration of modern humans out of Africa 100,000 years ago (7, 28).

Finally, viruses closely related to HGV/GBV-C have been found in a variety of Old World and New World primate species, and their phylogenetic relationships mirror those of their hosts. HGV/GBV-C variants in wild-caught chimpanzees from Central and West Africa show around 27% nucleotide (15% amino acid) sequence divergence from HGV/GBV-C (1, 3). Distinct variants of HGV/GBV-C_{CPZ}, recovered from different subspecies of chimpanzees, differed at 19% of nucleotide sites (9.5% of amino acid sites) in NS5, a level of diversity greater than that found between the more divergent genotypes of HGV/GBV-C in humans (11% nucleotide and 3.3% amino acid divergence) and consistent with their likely greater population age. Even more divergent homologues of HGV/GBV-C, described as GBV-A, have been recovered from several species of New World primates, with 42% nucleotide (38% amino acid) sequence divergence in the NS5 region from homologous sequences of HGV/GBV-C from humans and chimpanzees (4, 16). Again mirroring host relationships, genetic variants of GBV-A differing from each other by around 25% are closely associated with different new world primate species (4, 5, 16). Observations of congruent sequence relationships between GB viruses and their primate host species are consistent with their coevolution.

However, this hypothesis is difficult to reconcile with the high short-term rate of sequence change of HGV/GBV-C, estimated at 3.9 \times 10^{-4} site/year over the whole genome (23), a rate comparable to that of other RNA viruses (e.g., 4 \times 10^{-4} in hepatitis C virus NS5 [26]). This rate appears incompatible with the lack of sequence diversity between HGV/GBV-C isolates in human populations, if the current distribution of genotypes derives from the migration of modern humans out of Africa. The rate of sequence change is also inconsistent with sequence relationships between human GB viruses and those found in other primate species. However, if GB viruses did not coevolve with their hosts, their current distribution can have originated only through multiple cross-racial and cross-species transmissions over the last few hundred years. This latter hypothesis does not accord with the geographical distribution of HGV/GBV-C genotypes in humans or with the congruent virus and host phylogenies. This explanation is also inconsistent with the recently described species barriers (GBV-A cannot be transmitted to chimpanzees, and HGV/GBV-C cannot be transmitted to New World primates [4a]).
A more radical explanation of the current data is that there are major differences between GB viruses and higher organisms in the constraints operating on sequence change. For vertebrates and other eukaryotes, there is a close concordance between fossil-based estimates of the times of species separation and the extent of divergence of nucleotide and amino acid sequences in a wide variety of nuclear genes (15). In contrast, restrictions on variability at certain sites in the HGV/GBV-C genome may prevent the accumulation of substitutions that limit the extent to which HGV/GBV-C can diversify over time. In the present study, we have compared sequences of different genotypes of HGV/GBV-C to identify such restrictions and to investigate whether these can be explained by mechanisms such as RNA secondary-structure formation.

MATERIALS AND METHODS

Nucleotide sequences. Currently available complete genomic sequences of HGV/GBV-C sequences of genotypes 1 to 3 (GenBank/EMBL accession numbers in parentheses) include the type 1 sequence, GBV-C (U63368); the type 2 sequence, HGV-J95 (U92161, U92162), R1291 (U35566), HGV-J1 (D97525), HGV-C (AF301329, GT20 (D90601), and CG0181 (AF003289); the type 2b sequence GBV-C(EA) (U67515); the type 3 sequences GT230 (D90601), CCG1828 (AF003293), HGV-VM71 (D87262), GSHS (D87262), GSHV (AF003293), HCV-GD (AF006500), HCVCN (U49695), B1IHC (AF003288), HGV-C (75356), and D87788 to D87715; and sequences of unclassified genotype, i.e., C112LC (AF003291) and G05BD (AF003292). Sequences were nucleotide from the start of the coding region after alignment. HGV/ GBV-C homologues in primate sequences were from chimpanzees (AP069691 to AF068913 [3]) and the New World primate species Sangvinsmys mystax, S. labialis, and Aotus trivirgatus (U34421, AF023424, and AF023425).

Coding sequences of serum albumin were obtained from the following mammalian species (GenBank accession numbers in parentheses): cow (Y17769), sheep (X170555), cat (X484421, germ (X405289), lemur (M73996), mouse (L59490), rat (M71083), hamster (X75029), and sheep (X70213).

Generation of simulated sequence data sets. The expected stochastic frequencies of invariant and variable codons and of covariant substitutions were determined by using control sequence data sets containing the same number of sequences generated in the following ways.

(i) Control set A. Nucleotide changes were introduced at random positions into 3 of the 47 representative HGV/GBV-C sequences (R120291) for the 17 representative sequences of types 1 to 3, and GT320 as a control for type 3 sequences) at a prespecified frequency (nucleotide divergence between sequences and does not alter the much greater constraint on the encoded amino acid sequence but also provide a mechanism that would inhibit the accumulation of substitutions at synonymous sites which are usually regarded as selectively neutral. Evidence for a restriction in variability at synonymous sites was provided by analysis of the distribution of variable and invariant nucleotide positions in HGV/GBV-C sequences (Fig. 2A). Among 17 HGV/ GBV-C sequences of genotypes 1, 2, and 3, a total of 622
codons were invariant at synonymous sites (23%) compared with a mean of 277 (10.0% ± 0.6%) between 10 independently generated simulated data sets of descendants of the HGV/GBV-C sequence, R10291 (control set A; Fig. 2B). These sequences had the same overall degree of divergence (J-C distance, 0.139; \(d_N\), 0.017; \(d_S\), 0.62; \(T_s/T_v\) ratio, 2.0) and had similar codon usage and base compositions biases at the codon positions 1, 2, and 3. In a second dataset of 17 genotype 3 sequences, the mean variability was lower (J-C distance, 0.10; \(d_N\), 0.016; \(d_S\), 0.43) and 30.3% codons were invariant at synonymous controls (Fig. 2C), 2.3 times the value in its matched control sequence data set. In contrast, comparisons of the mammalian coding sequences of alpha globin and albumin produced no evidence for a significant excess of invariant codon positions with respect to that expected by chance (ratios of 1.15 and 1.26, respectively, with respect to values of randomized control sequences).

Unequal codon usage does not explain the excess number of invariant synonymous sites in HGV/GBV-C sequences, because such biases are frequently more extreme in eukaryotic gene sequences such as alpha globin. For example, 36% of third-base sites among HGV/GBV-C sequences were G and 32% were C (total G+C content, 68%), compared with the more biased frequencies in alpha globin sequences of 32 and 51% (G+C content, 83%).

The distribution of invariant synonymous sites in the genome of HGV/GBV-C was not random (Fig. 3). Mean synonymous diversity over a sliding window of 50 codons ranged from 0.159 to 0.716, with extreme suppression of variability around nucleotide positions 1300, 4600, 6300, and 6700 (Fig. 3A). A much more restricted range of variability was observed in a matched simulated data set (control set A; range, 0.363 to 0.564 [Fig. 3B]). The pattern of variability across the genome was reproduced in a separate sequence data set of type 3 sequences, even though the underlying degree of sequence variability was lower (Fig. 3C). There was no evidence for the existence of evolutionarily conserved potential coding sequences (containing initiating methionine and stop codons) in either the +2 or +3 reading frames of the positive-sense genome or the +1 or +3 frames of the antisense genome, either in the parts of the genome showing low variability at synonymous sites or elsewhere (data not shown).

Secondary structure of RNA. The genomes of many RNA viruses, including HGV/GBV-C and HCV, contain regions that form internally base-paired stem-loop structures that play a role in RNA replication and translation through ribosomal binding to an internal ribosome entry site. While current descriptions of structured regions of viral genomes are generally confined to the untranslated regions at the extreme ends of viral genomes, there is also evidence for a series of stem-loop structures in the core gene of HCV (27), which may play a direct role in the translation of the HCV polyprotein (24). Secondary structures in the HCV core gene would also account for the marked suppression of synonymous substitutions in this region of the genome in the absence of convincing evidence for
an overlapping reading frame (10, 24). More extensive internal base pairing might therefore underlie the greater than expected frequency of invariant synonymous sites observed in the HGV/GBV-C genome.

Conventional methods for predicting RNA secondary structures are computationally intensive and are often based upon the calculation of minimum free energies (greatest internal base pairing) for each possible configuration of an RNA sequence. These secondary-structure predictions are sometimes independently supported by the occurrence of covariance, where a nucleotide substitution of an internally base-paired sequence is matched by a substitution in the paired sequence that preserves binding. However, to date, formal statistical methods to analyze the contribution of covariance detection in the identification of particular secondary-structure prediction have not been described, although recently a method was described in which thermodynamic structure prediction was combined with phylogenetic information to search for conserved structures in human immunodeficiency virus and HCV (9).

The length of the coding region of the HGV/GBV-C genome (8,500 bases) clearly poses problems for methods based only upon free-energy calculation because the number of possible internal base-paired combinations would be extremely large and similar in free energy. However, the existence of multiple divergent genomic sequences without substantial phylogenetic structure suggested to us that covariance might be used in place of free-energy calculations as the primary method to detect internal base pairings. An algorithm was used in which alignments of HGV/GBV-C sequences were systematically scanned for potential base pairings between variable sites.
and scored by the number of paired substitutions from the upstream and downstream consensus bases that maintained base pairing (covariance score). Candidate covariant sites were additionally scored on the length of sequence either side of the site that was capable of forming a stem-loop structure (Table 1). Analysis of a representative data set of 17 HGV/GBV-C sequences of all genotypes identified a numerical excess of sites (approximately 35 sites) with high combined covariance and stem-loop scores over the number detected in control sequences whose codon positions were randomly reordered to break any stem-loop structure (control set C), or sequences with randomly introduced variability at an equivalent level to that between epidemiologically unlinked HGV/GBV-C sequences (control set A). A separate analysis of a second data set containing 17 type 3 sequences also showed a similar excess of sites (30 sites) with high covariance and stem-loop values over control sequences (data not shown).

Five lines of evidence argue for the biological reality of these proposed structures. First, the spacing between covariant sites in HGV/GBV-C was nonrandom, since high covariance values were strongly associated with separations between upstream and downstream sites of less than 500 bases (Fig. 4A). The same association between spacing and covariance value was observed in a separate analysis of type 3 HGV/GBV-C sequences (data not shown) but was absent in control data sets A, B, or C (Fig. 4B; data not shown). Second, multiple covariant sites were frequently found in the same stem-loop structure, more often than expected by chance; an example of a stem-loop containing four covariant sites is shown in Fig. 5. Third, proposed structures in HGV/GBV-C sequences were found in homologous positions in the more distantly related HGV/GBV-C_TRO and GBV-A sequences by using a minimum-energy algorithm to predict RNA secondary structures between base positions 1 and 1,500 of their coding regions (RNADraw) (Fig. 6). Covariant sites between the available GBV-A sequences were found in this region. These structures were conserved despite the low degree of nucleotide sequence similarity between the GB viruses in this part of the E1 gene (the nucleotide sequence divergence ranged from 43 to 53% between positions 263 and 309 in the R10291 genome). Third-base positions in the predicted stem-loops for human HGV/GBV-C sequences and GBV-A were paired, and all covariant substitutions in the loop were synonymous.

Fourth, while potential stem-loop structures detected by the covariance algorithm were distributed throughout the HGV/GBV-C genome (Fig. 3A and C), there was a close correlation between these sites and regions where synonymous site variability was suppressed. For example, the region from positions 3000 to 3950 showed the greatest synonymous variability (peak value, >0.7) and was uninterrupted by detectable covariant sites. This region was bounded by the predicted stem-loop structures between positions 2964 and 3114 and positions 2958 and 3120 (5' end) and between positions 3894 and 3930 and positions 3897 and 3927 (3' end), and these coincided with deep declines in mean synonymous variability. There were no obvious associations between synonymous variability and sites of cleavage of the HGV/GBV-C polyprotein (Fig. 3A). Together, stem-loop structures predicted from covariant sites detected in both HGV/GBV-C sequence data sets accounted for at least 20% of the HGV/GBV-C genome. The 24 synonymous nucleotide substitutions observed in the longitudinal study of HGV/GBV-C sequence change over 8.4 years (23) were more frequently found in areas of the genome with higher mean synonymous variability (Fig. 3A).

Finally, the proposed extensive RNA secondary structure of the HGV/GBV-C genome was independently supported by the high free energies predicted by RNA folding. Although the formation of RNA secondary structures was favoured by the high G+C content of the genome (32% G and 27% C) and the excess of U over A residues (23 and 18%, respectively), the order of bases also contributed additional to the high free energy values observed for subgenomic regions. The mean free energy of eight overlapping 1,500-base fragments of the R10291 HGV/GBV-C sequence was 293 kJ/mol lower (range, 207 to 426 kJ/mol) than that of the same sequences in which nucleotides were randomly reordered (control set D). A similar difference in free energy was observed for the chimpanzee HGV/GBV-C_TRO sequence (mean reduction, 301 kJ/mol; range, 225 to 399 kJ/mol) or upon comparison of shorter subgenomic regions of R10291 (e.g., mean of 89 kJ/mol and range of 23 to 185 kJ/mol over a sequence length of 500 bases). In contrast, no significant difference in free energy was observed upon randomization of base order for 1,500 base regions of sequences where the identity of nucleotides was altered to change base pairing (G→C→U→A→G [control set B]). Similarly, no significant reduction was observed upon reordering the human gene sequences of alpha globin (−13 kJ/mol over a sequence length of 429 bases) or albumin (+13 kJ/mol over 1,845 bases).

Nucleotide order randomization also lowered the free energy for the reverse complement of the first 1,500 bases of the HGV/GBV-C genome (−119 kJ/mol), but to a lesser extent than that of the sense sequence (−294 kJ/mol), suggesting that the secondary structure of the positive strand was more relevant biologically. Secondary-structure predictions based on free energy provided independent evidence for the majority of the short-range stem-loop structures identified by covariance screening (Fig. 6) but did not support structures in which covariant sites were separated by more than 100 bases.
DISCUSSION

This study provides evidence for marked restrictions on the sequence variability of HGV/GBV-C that may limit the extent to which virus genome sequences can diverge over time. In contrast to mammalian genes, HGV/GBV-C sequences showed an increased frequency of invariant synonymous sites over that expected by chance, suggesting a functional role of certain nucleotide positions that was independent of the coding capacity for the HGV/GBV-C polyprotein and which did not result from the presence of overlapping gene sequences in other reading frames. A similar although less marked excess of invariant synonymous sites was recently reported in a separate analysis of swine vesicular disease virus and foot-and-mouth disease viruses sequences (8). The contribution of secondary-structure formation to the observed restricted variability of HGV/GBV-C was supported not only by the spatial association between covariant sites and stem-loop structures (Fig. 4; Table 1) but also by the contribution of base order to free-energy calculations of RNA folding. Similarly, randomization of the nucleotide order of swine vesicular disease virus 3B/C sequences led to a consistent although proportionately smaller reduction in free energy on folding (mean 7.6% reduction over 641 bases [24a], compared with 16% over an equivalent length of HGV/GBV-C).

In a folded genome, sequence change in regions that are internally base paired would require simultaneous nucleotide

FIG. 4. Frequency histogram of covariance scores between paired nucleotide sites according to their separation in the genome of 17 HGV/GBV-C sequences of genotypes 1, 2, and 3 (A) and control data set C (B). Solid bars represent frequencies of nucleotide pairs with covariance scores of 6 or greater. Frequencies were adjusted to account for the number of pairs compared.
substitutions on both sides of potential stem-loop structures to maintain base pairing and would lead to a much lower frequency of sequence change than in unpaired sites. The preferential accumulation of nucleotide changes at nonpaired sites may therefore account for the rapid sequence change of HGV/GBV-C over short periods, while further divergence, such as between different isolates of HGV/GBV-C or between different primate species, may occur only through covariant substitutions, which accumulate more slowly over longer periods (Fig. 1). For example, the substitution rate of HGV/GBV-C over 8.4 years, $4 \times 10^{-7}$ per site per year (23), predicts a frequency of covariant substitutions of less than $1.5 \times 10^{-7}$, which is comparable to the long-term rate of substitution observed between viruses infecting different primate species (such between HGV/GBV-C and GBV-A sequences [Fig. 1]). An even greater differential frequency ($1.1 \times 10^{-5}$ at covariant sites) would occur at nonsynonymous sites, and this greater restriction on nonsynonymous sequence change may contribute to the extreme $d_N/d_S$ ratio observed between HGV/GBV-C sequences (0.033) compared with that for coding sequences of other viruses and eukaryotes.

More precise matching of variable and nonvariable sites in the HGV/GBV-C genome with unpaired and paired nucleotides will require the completion of a model of its secondary structure. These investigations would include determining the relationship between secondary structure and the sites of nucleotide substitution in the HGV/GBV-C over short observation periods where high frequencies of sequence change at synonymous sites was observed (Fig. 4) (23). The base pairings predicted by covariance scanning will contribute to future attempts to determine the overall structure of HGV/GBV-C genome, since they restrict the remaining structure prediction by free-energy calculation to relatively short lengths of sequence. Such analyses would be assisted by the observation that all the sites identified by covariance were relatively closely spaced (<500 bases); this may be a general feature of HGV/GBV-C genomic RNA structure.

Structural constraints on sequence change independent of coding capacity may inhibit sequence change in other viruses with single-stranded genomes. For example, separate genotypes of simian immunodeficiency virus (SIV<sub>AGM</sub>) are associated with different subspecies of African green monkeys (11), while a similar association between genetic variants of SIV<sub>CPZ</sub> in different chimpanzee subspecies (Pan troglodytes troglodytes and P. troglodytes schweinfurthii) has recently been reported (6). If these associations also represent coevolution of viruses with their hosts, there is a similar discrepancy between their long- and short-term rates of sequence change. In the second example, the 37.7% sequence divergence which accumulated over a period of at least 0.5 million years between SIV variants infecting the troglodytes (SIV<sub>CPZ-US</sub> and SIV<sub>CPZ-GAB</sub>) and schweinfurthii (SIV<sub>CPZ-ANT</sub>) subspecies (21) represents a net rate of
sequence change of 3.4 × 10⁻⁷ per site per year. This rate is about 5,000 times lower than the rate (1.45794 SIMMONDS AND SMITH J. VIROL. 5794 approximately 5,000 times lower than the rate (1.45794 SIMMONDS AND SMITH J. VIROL.) inferred from comparison of different subtypes in the main (M) group of human immunodeficiency variants, which are considered to have originated over the last 50 years (based on a mean pairwise distance between single representative sequences of types A to J of 14.2% [29]). Despite the genetic and structural differences between flaviviruses and lentiviruses, the absolute values and differential between short- and long-term rates of sequence change of human immunodeficiency virus types 1 and 2 were remarkably similar to those predicted for HGV/GBV-C.

The functional role of the predicted RNA secondary structures in HGV/GBV-C remains unclear. Among several possibilities that may apply equally to other single-stranded RNA viruses, RNA folding may be required for packaging of the genome into the viral nucleocapsid or to protect the genome from RNA-degrading enzymes or it may be involved in the regulation of transcription or translation in analogous ways to the function of internal ribosome entry site structures in flaviviruses and picornaviruses. Recently, an internal stem-loop structure in the coding part of the human rhinovirus 14 genome sequence, comparable in size and free energy to those detected in HGV/GBV-C, was shown to be essential for human rhinovirus 14 negative-strand transcription (20). The long-range interactions between different genomic regions implied by these observations suggest an organized overall structure of the RNA genome, in which stem-loop structures may play an important structural or catalytic role in virus replication.

Irrespective of its functional significance, the evidence for markedly different substitution frequencies at different sites puts a fundamental limitation on the use of sequence divergence in the timing of virus origins. While the species association of HGV/GBV-C and related viruses in primates provides indirect evidence for their long-term rates of change, such inferences cannot be made for viruses with less close host associations. While it is possible that much of the genetic heterogeneity of RNA viruses originated very recently, our inability to measure long-term rates of sequence change or to understand the restrictions on sequence variability currently leaves open the possibility of very ancient origins of many of these RNA viruses. These findings suggest that the molecular clock, which has been of major value in the reconstruction of vertebrate and other eukaryotic phylogenies, cannot be simply applied to viruses with more unusual genomic structures.

REFERENCES