Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5′ non-coding region

Donald B. Smith,¹ Narcisa Cuceanu,¹ Fiona Davidson,² Lisa M. Jarvis,¹ John L. K. Mokili,¹ Saeed Hamid,³ Christopher A. Ludlam⁴ and Peter Simmonds¹

¹Department of Medical Microbiology, University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK
²Edinburgh and South East Scotland Blood Transfusion Service, Lauriston Place, Edinburgh EH3 9HB, UK
³Department of Medicine, The Aga Khan University Hospital, Stadium Road, Karachi 74800, Pakistan
⁴Department of Haematology, Royal Infirmary of Edinburgh, Lauriston Place, Edinburgh EH3 9YW, UK

We have investigated the ability of different sub-genomic fragments to reproduce the phylogenetic relationships observed between six complete genome sequences of GBV-C/hepatitis G virus (HGV). While similar relationships were observed following analysis of part of the 5′ non-coding region (5′ NCR), for the coding region they were not accurately reproduced for some large fragments or for the majority of fragments of 300 or 600 nucleotides. Analysis of 5′ NCR sequences from a large number of isolates, including newly obtained sequences from Pakistan, Zaire and Scotland, produced separate groupings of Asian, African and European/North American variants. These groupings are associated with specific polymorphisms in the 5′ NCR, many of which were covariant and consistent with a proposed secondary structure for this region. The relatively low level of amino acid sequence variation observed between these geographically and phylogenetically defined groups of variants suggests that they are unlikely to display significant biological differences.

Introduction

Investigation of the genetic heterogeneity of GBV-C/hepatitis G virus (HGV) (Simons et al., 1995; Linnen et al., 1996) is at an early stage. This newly discovered human virus has a genome structure related to that of hepatitis C virus (HCV) with a 5′ non-coding region (5′ NCR) capable of acting as an internal ribosome entry site (IRES) (Simons et al., 1996) followed by a long open reading frame capable of encoding presumed structural (E1 and E2) and nonstructural (NS2, NS3, NS4a, NS4b, NS5a and NS5b) proteins (Simons et al., 1995; Linnen et al., 1996). GBV-C/HGV RNA can be detected in 2–4% of blood donors throughout the world (Jarvis et al., 1996; Linnen et al., 1996; Moaven et al., 1996; Schlüeter et al., 1996; Stark et al., 1996), but an association with disease in infected individuals has yet to be established.

Analysis of the 5′ NCR has led to the suggestion that variants of GBV-C/HGV can be divided into three or more ‘genotypes’ (Fukushi et al., 1996; Muerhoff et al., 1996). Some of these groupings are associated with the geographical origin of the variant, but otherwise nothing is known about the virological or clinical implications of genomic differences between them. Attempts to demonstrate phylogenetic groupings of GBV-C/HGV isolates based on sequence analysis of subgenomic coding regions have generally been unsuccessful. For example, comparison of sequences within the NS3 helicase region for isolates from around the world failed to differentiate between Asian, African and European/North American isolates (Tsuda et al., 1996; Pickering et al., 1997), and as much diversity was observed within isolates from Taiwan as between isolates from different continents (Kao et al., 1996). Similarly, analysis of a fragment of NS5b failed to distinguish between African and Asian isolates, and as much variation was observed between variants from Russia or Germany as between isolates from different continents (Viazov et al., 1997). In neither case was there a difference in the distribution of sequence distances within groups compared to between groups. In contrast, for HCV three non-overlapping distributions corresponding to virus type, subtype and isolate have been described for coding regions (Simmonds et al., 1993a, 1994), but these groupings are less clear from analysis.

Author for correspondence: Donald Smith.
Fax +44 131 650 6531; e-mail Donald.B.Smith@ed.ac.uk

The GenBank accession numbers of the sequences reported in this paper are AF003152–AF003186.

0001-4671 © 1997 SGM
of the 5’NCR (Simmonds et al., 1993b; Ohba et al., 1995), probably because of its extreme conservation and the existence of covariant substitutions (Smith et al., 1995).

In this study we have used the complete genome sequences of six different GBV-C/HGV isolates to investigate the ability of different subgenomic regions to reproducing the phylogenetic relationships displayed by comparison of the complete sequences. While these relationships are not reproduced by analysis of short fragments of coding regions, they can be reproduced by comparison of 5’NCR sequences. We describe the features of variability within the 5’NCR that may be useful in the development of methods to discriminate between different variant groupings.

Methods

Samples. Sera were obtained from 14 patients infected with GBV-C/HGV with chronic hepatitis from Pakistan, 12 of whom were co-infected with HCV or hepatitis B virus, and from nine women from Zaïre who were co-infected with human immunodeficiency virus. Samples of plasma were also available from 12 haemophiliacs previously treated in Edinburgh with locally manufactured clotting factor concentrates not inactivated for enveloped viruses (Jarvis et al., 1996). Two haemophiliacs had received commercially produced concentrates, one exclusively (Ed 3), while the other (Ed 81) had also received locally produced, but inactivated concentrates. Samples were stored at −70 °C before extraction of RNA using proteinase-K–Sarkosyl and phenol–chloroform extraction as described previously (Jarvis et al., 1994).

RT–PCR and sequencing. Purified RNA was reverse-transcribed and amplified by nested RT–PCR using primers derived from the 5’NCR of GBV-C/HGV: S4571 – sense outer (positions −445 to −428, see below for method of numbering), S4572 – sense inner (positions −419 to −399), S4573 – antisense inner (positions −70 to −97) and S4574 – antisense outer (positions −22 to −42) (Jarvis et al., 1996). RNA from the haemophiliac samples was also amplified with primers T1721 – sense outer (5’ GGGCACAACGACGCCCACGTACGGTC 3’), positions 181 to 203 (Jarvis et al., 1996). Reverse transcription was performed using avian myeloblastosis virus reverse transcriptase (Promega) at 42 °C for 30 min, while conditions for PCR were hot start at 80 °C for 2.5 min followed by 30 cycles of 94 °C for 18 s, 55 °C for 21 s and 72 °C for 90 s. Secondary PCR was carried out using 1 µl of the primary PCR product, and amplified products were visualized after electrophoresis through a 2% agarose gel stained with ethidium bromide.

Nucleotide sequences. PCR products were re-amplified from primary products using primers S4572 and biotinylated S4573 and sequenced directly from magnetically separated single strands after immobilization on streptavidin-coated beads (Dynabeads, Dynal). Secondary PCR products produced using primers T1722 and T1723 were purified and cloned using the LiGATor kit (R&D Systems), and sequenced using T7 DNA polymerase (Sequenase, USB) and both sense and antisense plasmid primers. The consensus sequence of one to five clones was used for phylogenetic analysis. Other nucleotide sequences obtained from GenBank are identified by their accession number, or by isolate name for the six complete genome sequences (GBV-C, U36380; PNF2161, U44402; R10291, U45966; GBV-C(EA), U63715; HGV-C964, U75356; and HGV-Iw, D87255).

Numbering of nucleotide positions. The true 5’ terminus of the GBV-C/HGV genome is currently unknown. Although the sequence of PNF2161 extends the furthest upstream, the existence of additional 5’ sequences is suggested by the lack of a well-defined 5’-terminal hairpin structure similar to those defined for GBV-A (Simons et al., 1996) and for different genotypes of HCV (Smith et al., 1995). Because of this uncertainty we have adopted a system similar to that commonly used for describing nucleotide positions within the HCV genome. Positions are given relative to the AUG codon near the start of the long open reading frame which follows the multiple stem–loop structures and the polypyrimidine tract of the internal ribosome entry site (IRES) (Simons et al., 1996). This AUG (positions 524–526 of Simons et al., 1996) is likely to represent the initiation codon for translation of the GBV-C/HGV polyprotein because of its position relative to the IRES, the initiation of in vitro translation products at this site (Simons et al., 1996), and the lack of a conserved upstream open reading frame or of an alternative in-frame AUG codon (Muerhoff et al., 1996). Positions in the coding region are numbered from the AUG of the prototype sequence (GBV-C) while positions in the 5’NCR are given negative numbers relative to this AUG. The 5’NCR contains 11 sites of insertion relative to the GBV-C sequence in our data set, and these positions are un-numbered.

Sequence analysis. Sequences were aligned using Esee (version 1.09, Eric Cabot) or Simmonic Software (version 1.0, Peter Simmonds). Phylogenetic trees were produced using the MEGA package (Kumar et al., 1993). Free energies for stem–loop structures were obtained using the program FOLDRNA in the GCG package (Genetics Computer Group, Wisconsin, USA, accessed via the SEQNET facility of Daresbury).

Results

Phylogenetic analysis of complete coding sequences

Three major ‘types’ of GBV-C/HGV have been described based on analysis of the 5’NCR, two of which were further divided into ‘subtypes’ (Muerhoff et al., 1996). This analysis included three variants for which complete genome sequences are available, namely GBV-C (‘type 1’), and PNF2161 and R10291 (‘subtype 2a’). Three additional complete genome sequences are now available, and we investigated their phylogenetic relationships through analysis of their coding regions (Fig. 1a). HGV-Iw, isolated from a Japanese hepatitis patient (Shao et al., 1996), is more closely related to PNF2161 than to R10291, and grouped with these sequences in 100% of bootstrap re-sampling replications, and so is ‘subtype 2a’. GBV-C(EA), isolated from an East African child, is more divergent but grouped with the ‘subtype 2a’ variants in 100% of bootstrap replications, and shares a higher degree of sequence identity with these sequences than with GBV-C (Erker et al., 1996). Comparison of the Chinese variant HGV964 with the other five complete genome sequences revealed two areas where numerous nucleotide and amino acid substitutions occur (positions 5917–5948, and 8502 onwards). The strong conservation of five out of the six sequences available for these regions suggests that sequencing errors could be responsible for the extreme variation of HGV964 in these regions. A perfect alignment of the amino acid sequence
Fig. 1. Phylogenetic analysis of six complete GBV-C/HGV genomes. (a) Phylogenetic tree of the coding region (positions 1–8502) of six complete GBV-C/HGV genomes. The percentage of bootstrap re-sampling replicates (n = 5000) in which branches 'A' and 'B' were observed is indicated. (b) Bootstrap support for phylogenetic groupings based on analysis of subgenomic fragments from the coding region of complete genome sequences. Phylogenetic analysis was performed on the complete coding region and on progressively smaller subgenomic fragments from 4200 to 300 nt. The percentage of bootstrap replicates (n = 500) in which PNF2161, R10291 and HGV-Iw shared a common branch (A) or in which these sequences and GBV-C(EA) shared a common branch (B) is indicated by different patterns. Bootstrap frequencies of less than 70% are regarded as not providing evidence for the phylogenetic grouping. Nucleotide positions are numbered relative to the AUG initiation codon of GBV-C.
encoded by positions 5917–5948 can be obtained by shifting the HGVC964 sequence by one nucleotide. Comparisons between the complete genome sequences have therefore been made by including this frameshift and by removing the second discrepant region for which no obvious realignment could be found. These alterations had little effect on the topography of the trees generated and HGVC964 did not group with any of the other sequences. We present evidence below from analysis of 5′NCR sequences that GBV-C(EA) belongs to ‘subtype 2b’ and that HGVC964 represents a new group of variants distinct from the Japanese variants which form ‘type 3’.

We next investigated the extent to which analysis of different subgenomic fragments supported the grouping of isolates PNF2161, R10291 and GBV-Iw (A), or the grouping of these isolates and GBV-C(EA) (B) in bootstrap re-sampling replications (Fig. 1b). Two 1800 nucleotide (nt) fragments in the 3′-half of the coding region failed to support the grouping of PNF2161, R10291 and GBV-Iw sequences, while one or both groupings were not supported by bootstrap re-sampling in 3/6 (50%) 1200 nt fragments, in 8/14 (57%) 600 nt fragments or in 23/28 (82%) 300 nt fragments. The five 300 nt regions that did provide bootstrap support for both groupings were positions 301–600 (E1), 601–900 (E2), 1801–2100 (NS2), 4201–4500 (NS3) and 4801–5100 (NS4a).

**Phylogenetic analysis of 5′NCR sequences**

Because of the difficulties associated with phylogenetic analysis of coding regions, and since analysis of the 5′NCR has previously been shown to discriminate between GBV-C/HGV isolates from different parts of the world (Muerhoff et al., 1996), we tested the ability of fragments of the 5′NCR to reproduce the phylogenetic relationships observed amongst complete coding sequences (Fig. 2). Comparison of the largest region for which sequences were available for all six sequences (positions −456 to −1) failed to provide support for the

---

**Fig. 2.** Level of bootstrap support for phylogenetic groupings based on analysis of fragments of the 5′NCR. The level of bootstrap support (500 replications) for the grouping of PNF2161, R10291 and HGV-Iw (A) or of these three sequences and that of GBV-C(EA) (B) is indicated by different patterns. The 5′ and 3′ boundaries of the regions analysed are indicated at the bottom of the figure, numbered relative to the initiator AUG of GBV-C.
grouping of PNF2161 and HGV-Iw with R10291 (A), or for the grouping of these sequences with GBV-C(EA) (B). These groupings were also not observed for fragments between −325 and −1, but were supported by analysis of 5′NCR fragments between positions −402 to −246. This represents less than a third of the 5′NCR, and is much smaller than that of coding regions giving similar degrees of bootstrap support (from 300 to 4200 nt depending on the region).

Geographical distribution of GBV-C/HGV groupings

In order to test the ability of phylogenetic analysis of a restricted portions of the 5′NCR to discriminate between GBV-C/HGV isolates from different geographical regions, we analysed sequences available from GenBank and new sequences obtained from pregnant women from Zaïre (n = 9), patients with chronic hepatitis from Pakistan (n = 14) and haemophiliacs treated in Edinburgh (n = 12). Separate clusters of African, Asian and European/North American sequences were observed consistently when 5′-terminal regions of the 5′NCR were analysed. For the region −366 to −235, virus sequences from Zaïre were similar to those present in published sequences from West Africa, while those from Pakistan and from most of the Edinburgh haemophiliacs grouped with sequences previously described from infected individuals in Europe and North America (Fig. 3). The only exceptions were two haemophiliacs infected with variants related to the African (Ed 3) or Asian groupings (Ed 81), both of whom had received commercial factor concentrates, and in contrast to the other haemophiliacs who had only received factor concentrates manufactured from blood donations collected locally. Two other exceptions from the general pattern were GBV-C(EA) from a child in East Africa with acute non-A–E hepatitis, and HGV-Iw from a Japanese hepatitis patient. Complete genome sequences are available for both of these isolates, and phylogenetic analysis of the full-length coding sequences (Fig. 1a) is consistent with the groupings obtained by analysis of 5′NCR sequences. Phylogenetic analysis of fragments of a similar size from the 3′ terminus of the 5′NCR did not consistently produce the same groupings of isolates as those based on geographical origin (data not shown). These observations suggest that analysis of the region −366 to −235 accurately reproduces the phylogenetic relationship of GBV-C/HGV isolates, consistent with the analysis of similar regions from the complete genome sequences (Fig. 2).

5 NCR polymorphisms correlated with different phylogenetic groupings

We next searched for polymorphisms in the 5′NCR that were associated with these phylogenetic groupings since they might simplify the identification of GBV-C/HGV groups from sequence information, or allow the development of methods of identification that avoid the need for sequence analysis. Polymorphisms that are strongly correlated with phylogenetic groupings are clustered in three separate regions of the 5′NCR (Fig. 4), and the majority of these were consistent with a proposed secondary structure for GBV-C/HGV 5′NCR (Simons et al., 1996) since substitutions either maintained proposed base pairings, or occurred in unpaired regions. Group-specific polymorphisms are also present outside these regions (positions −65, −103, −104, −108, −149, −306, −307), but the remaining polymorphic positions vary without regard to group. For example, substitutions within the unpaired polypyrimidine stretch between positions −21 to −10 generally are confined to alteration between U or C residues and no group-specific pattern can be discerned. Another such region is that between positions −147 to −120, where despite the considerable sequence variability, all sequences are capable of forming stem–loop structures with free energies of between −42 kJ and −89 kJ (Fig. 5). Some substitutions in this region are group-specific, but at the majority of polymorphic sites, substitutions occur in all of the phylogenetic groups. Similarly, most substitutions between positions −185 and −177 do not correlate with the phylogenetic groupings based on analysis of complete genome sequences, but the ability to form a short stem–loop structure (region IVa’ of Simons et al., 1996) is conserved.

Discussion

Variation of GBV-C/HGV coding regions

Our phylogenetic analysis of six complete genome sequences of GBV-C/HGV has revealed several unexpected findings. While certain phylogenetic groupings are apparent from the analysis of the complete coding region, these groupings are only poorly supported when smaller genome segments are compared. These observations provide an explanation for the previous inability to observe consistent phylogenetic relationships between GBV-C/HGV isolates from different parts of the world based on analysis of 118 nt within NS3 (positions 3767–3884) (Kao et al., 1996; Pickering et al., 1997) or from the analysis of 354 nt within NS5b (positions 6742–7095) (Viazov et al., 1997). Analysis of similar genome regions from complete sequences revealed that neither a 300 nt fragment of NS3 nor a 1200 nt fragment of NS5 were able to reproduce the phylogenetic relationships of the complete coding sequence. This finding suggests that the use of short coding sequences such as 118 nt or less of NS3 in investigations of the epidemiology of GBV-C/HGV (Berg et al., 1996; Heringlake et al., 1996; Masuko et al., 1996; Schmidt et al., 1996; Schreier et al., 1996; Tsuda et al., 1996) may not be reliable for the reconstruction of phylogenetic relationships. The finding that analysis of relatively long coding regions is required to distinguish between different groups of GBV-C/HGV variants differs markedly from the situation for HCV, for which phylogenetic groupings observed by analysis of complete virus genomes are closely mirrored by those observed for subgenomic fragments throughout the genome,
Fig. 3. For legend see facing page.

is equivalent to a distance of 0.02
Variation of GBV-C/HGV

<table>
<thead>
<tr>
<th>Accession</th>
<th>Region</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J02237</td>
<td>5'NCR 1</td>
<td>Group-specific NCR polymorphisms. Sequences for the regions 390 to 320 and 297 to 235 are grouped on the basis of phylogenetic analysis of positions 366 to 235. Sequence identity with PNF2161 is indicated by dots, missing information by spaces and ambiguous or mixed nucleotides by question marks.</td>
</tr>
</tbody>
</table>

Fig. 4. Group-specific 5’NCR polymorphisms. Sequences for the regions 390 to 320 and 297 to 235 are grouped on the basis of phylogenetic analysis of positions 366 to 235. Sequence identity with PNF2161 is indicated by dots, missing information by spaces and ambiguous or mixed nucleotides by question marks.

Fig. 3. Consensus phylogenetic tree of geographically disparate isolates. The tree was generated by comparison of positions 366 to 235 of the 5’NCR. The percentage of bootstrap replicates in which major groupings were observed amongst 500 replicates is indicated. Sequences are identified by accession numbers or, for the sequences reported in this study, by the prefixes Pak for Pakistan, Ed for Edinburgh and Zai for Zaire. The geographical origin of each sequence is indicated at the right.
and as small as 222 nt (Simmonds et al., 1994, 1996; Ohba et al., 1995).

**Variation of the GBV-C/HGV 5′ NCR**

The extent of variation is unequal across the 5′NCR, with some areas being relatively well-conserved and others being more variable, as previously noted (Erker et al., 1996; Muerhoff et al., 1996). However, in contrast to the lack of discrimination of short fragments of coding regions, analysis of 5′-terminal fragments of the 5′NCR can be used to reproduce the phylogenetic groupings displayed by complete genome segments (Fig. 2). This conclusion is also supported by analysis of a large set of sequences (n = 89), since isolates generally grouped according to their geographical origin (Fig. 3). On the basis of this analysis the following groups can be distinguished: Group 1, a diverse collection of African sequences, Group 3 consisting of relatively similar Asian sequences, Group 2 consisting of sequences from North America, Europe and the Indian subcontinent, as well as HGV-Iw from Japan and GBV-C(EA) from East Africa. This last isolate together with some other European and North American isolates may represent a subgroup (Group 2b) since they share an unusual pattern of substitution in the 5′NCR (Fig. 4). Similarly, there is evidence for subgroupings amongst the diverse sequences comprising Group 1. A single Chinese sequence (GBV-C 964) may represent a distinct group (referred to here as group 4) since analysis of the complete coding sequence confirms that it is distinct from Groups 1 and 2. Although comparison with the complete coding sequences of Group 3 isolates is not yet possible, this sequence contains several unusual substitutions in the 5′NCR (Fig. 4).

The finding that 5′NCR polymorphisms are associated with phylogenetic groupings obtained by analysis of complete coding regions and that group according to geographical origin makes it possible to develop rapid methods for identifying GBV-C/HGV isolates, similar to those developed for the genotyping of HCV (Stuyver et al., 1993; McOmish et al., 1994). Group-specific polymorphisms are present throughout the 5′NCR, but they are most frequent between positions −490 and −235 (n = 39) where they are concentrated into three clusters (Fig. 4). Sequence information for the region between positions −490 and −459 may be more difficult to obtain because very little upstream sequence information is currently available from which to design universal primers for RT–PCR amplification. Although there are nine group-specific polymorphisms between positions −234 and −1, this region also contains three highly polymorphic regions that vary without regard to geographical grouping, and so phylogenetic analysis of this region fails to reproduce the relationships of complete coding sequences (Fig. 2) or to produce geographical groupings (data not shown). Similarly, a previous study failed to observe an association between phylogenetic groupings and geographical origin for the region −267 to +17.
(Pickering et al., 1997). Consequently, the optimal region for identifying the group of GBV-C/HGV isolates at present appears to be the region between positions –490 and –235, encompassing three blocks of group-specific polymorphisms.

**Significance of phylogenetic groupings**

The existence of distinct groups of GBV-C/HGV variants is implied by phylogenetic analysis of complete genome sequences, and by the presence of polymorphisms in the 5’NCR that correlate with geographical origin. However, the biological significance of these groupings is presently unclear, especially since representatives of different GBV-C/HGV Groups 1, 2 and 4 differ over their complete coding regions at only 3.7–4% of amino acid positions. This level of amino acid sequence variation is much less than that observed for individual genes between different serotypes of vesicular stomatitis virus (30–50%), poliovirus (20–30%) or dengue virus (23–38%), or between different types of paramyxoviruses (10–17%) or subtypes of HIV (12–36%). Relative to HCV, variation is at the lower end of the range observed between complete polyprotein sequences of epidemiologically unrelated isolates of the same HCV subtype (3.5–8.1%), and much less than that observed between different HCV subtypes (13–19%) or between different HCV types (23–29%). While different types of HCV show substantial variation in antigenicity and differ in clinical features such as the frequency with which infection can be eliminated by interferon treatment, similar differences have not been associated with the more minor differences between HCV subtypes such as subtypes 1a and 1b, or the still more subtle differences between isolates of the same subtype. Comparison with HCV and other viruses would therefore suggest that the limited diversity between GBV-C/HGV variants is unlikely to be associated with major biological or clinical differences. For this reason we have used the descriptive terms ‘group’ and ‘subgroup’ rather than the terms ‘type’ or ‘subtype’ (Berg et al., 1996; Fukushi et al., 1996; Muerhoff et al., 1996; Schreier et al., 1996) that suggest that variation between phylogenetic groups is significant.

While there is relative conservation of the GBV-C/HGV polyprotein between different groups of variants, their sequences still vary by 11–13% overall because of substitution at synonymous sites. The ratio of nonsynonymous to synonymous evolutionary distances between different variant groups ranges from 0.029 to 0.054 (mean 0.04), compared with between 0.1 and 0.14 for different representatives of HCV subtype 1a. The reason for this difference between GBV-C/HGV and HCV is not known, but presumably reflects differences in selective pressures related to their interaction with the host immune system during infection. Evolutionary distances at synonymous sites between variants of GBV-C/HGV (0.56–0.7) are greater than those observed between different isolates of HCV subtype 1a (0.14–0.17) or subtype 1b (0.24–0.36) but less than those observed between different subtypes (1.02–1.65). For HCV, evolutionary distances at synonymous sites have been used to date the time of divergence of different subtypes at more than 300 years ago (Smith et al., 1997). A similar estimate is not possible for GBV-C/HGV since the rate of evolution is currently unknown, but the substantial synonymous distances between sequences from different variant groups as well as their strong correlation with geographical origin, suggest that divergence of these groups is not a recent phenomenon.

We are grateful to Sue Graham for technical assistance, to staff at the Royal Infirmary of Edinburgh Haemophilia Centre for collecting samples, to Sergei Viazov for providing data before publication, and to Scott Muerhoff for clarifying the origin of certain variants. This work was supported by a grant from the Wellcome Trust to D.B.S. and P.S., and by a grant to N.C. from the Darwin Trust. P.S. is a Darwin Research Fellow.

**References**


McOmish, F., Yap, P. L., Dow, B. C., Pollett, E. A. C., Seed, C., Keller, A. J., Cobain, T. J., Krusius, T., Kolho, E., Naukkarinen, R., Lin, C., Lai,


Received 24 January 1997; Accepted 25 March 1997