THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0890

Reconstructing the origins of human hepatitis viruses

P. Simmonds

Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh EH9 1QH, UK (*peter.simmonds@ed.ac.uk*)

Infections with hepatitis B and C viruses (HBV, HCV) are widespread in human populations throughout the world, and are major causes of chronic liver disease and liver cancer. HBV, HCV and the related hepatitis G virus or GB virus C (referred to here as HGV/GBV-C) are capable of establishing persistent, frequently lifelong infections characterized by high levels of continuous replication. All three viruses show substantial genetic heterogeneity, which has allowed each to be classified into a number of distinct genotypes that have different geographical distributions and associations with different risk groups for infection. Information on their past transmission and epidemiology might be obtained by estimation of the time of divergence of the different genotypes of HCV, HBV and HGV/GBV-C using knowledge of their rates of sequence change. While information on the latter is limited to short observation periods and is therefore subject to considerable error and uncertainty, the relatively recent times of origin for genotype of each virus predicted by this method (HCV, 500–2000 years; HBV, 3000 years; HGV/GBV-C, 200 years) are quite incompatible with their epidemiological distributions in human populations. They also cannot easily be reconciled with the recent evidence for species-associated variants of HBV and HGV/GBV-C in a range of non-human primates. The apparent conservatism of viruses over long periods implied by their epidemiological distributions instead suggests that nucleotide sequence change may be subject to constraints peculiar to viruses with single-stranded genomes, or with overlapping reading frames that defy attempts to reconstruct evolution according to the principles of the 'molecular clock'. Large population sizes and intense selection pressures that optimize fitness may be additional factors that set virus evolution apart from that of their hosts.

Keywords: hepatitis viruses; hepatitis B virus; hepatitis C virus; evolution; origins; primates

1. INTRODUCTION

As illustrated by many of the papers in this issue, a wide range of different approaches have been taken to discover more about the origins of emerging pathogens, none more so than trying to understand the appearance of the large number of new virus and bacterial infections in humans. In this review, I shall attempt to review the methods and models available for reconstruction of virus histories, and how these might be applied to recover the history of human hepatitis viruses. It is perhaps a reflection of the difficulty of this exercise that much of this is highly speculative, although the trawl through the existing evidence provides a number of leads for future bioinformatic and epidemiological investigations of viruses infecting humans and other primates.

As their name suggests, the subjects of this review are united in their ability to cause hepatitis in humans. In other respects, they are distinct, both virologically and epidemiologically. Hepatitis A and E viruses (HAV, HEV) cause acute, resolving hepatitis and are transmitted by faecal–oral routes; hepatitis B and C viruses (HBV, HCV) and D virus are transmitted principally through blood contact, and show a number of adaptations to allow them to establish persistent infection in their hosts. The virological diversity of human hepatitis viruses is reflected by their classification into a wide range of different virus families, from picornaviruses (HAV), caliciviruses (HEV) and flaviviruses (HCV) for those with single-stranded RNA genomes to the retrovirus-related Pararetroviridae into which HBV and related viruses in other animals and birds have been assigned. This review cannot therefore propose a unified scheme for the origins of human hepatitis viruses, but provides a set of markedly contrasting vignettes of how different approaches at reconstruction may be applied.

2. EVOLUTIONARY HISTORIES

Many different sources of data can be used in the reconstruction of evolutionary origins of organisms. These include historical descriptions, archaeological material, the fossil record, the existence of homology (shared descent) among other contemporary organisms, and finally the nature of the virus–host interaction.

All human hepatitis viruses were discovered within the last 35 years, so any historical description older than this is necessarily confined to the clinical features of infection. Unfortunately, the symptoms and clinical signs associated with viral hepatitis are indistinguishable from hepatitis that arises from other infectious and non-infectious causes. It is unlikely that any amount of detective work in written archives will ever recover unequivocal evidence of the past occurrence of viral hepatitis in humans. Turning to archaeological data, clinical specimens suitable for recovery of viruses by isolation or by polymerase chain reaction older than 30 years are rare and restricted, so it is difficult to obtain any direct evidence for the type of viruses that may have existed in the past. Attempts to detect human hepatitis viruses from older archaeological material have been unsuccessful so far. Particularly problematic for most human hepatitis viruses is the instability of RNA in organic material.

Reconstruction of virus histories by analysis of their current distributions and genetic relatedness is also fraught with problems. Virus histories are inextricably linked to those of their hosts, which themselves may be uncertain in detail or impossible to meaningfully reconstruct for viruses that can frequently cross species barriers. Viruses recombine with one another, with other viruses and with the genomes of the cells they infect and they may undergo major genome rearrangements and changes in replication strategy.

Rates of sequence change in viruses, particularly those with RNA genomes, are invariably much greater than those of their hosts, and this presents a number of problems in evolutionary reconstruction. To persist within an infected individual or a host population, most RNA viruses must replicate continuously. The time from infection to production of progeny viruses may typically take one to three days, over which period several copyings of the virus genome occur. For a virus such as HCV, there may therefore be 100 000 genome replications over a period corresponding to a human generation; over this time, only 33 (female) or 200 (male) cellular divisions separate a human egg from its gametes. Compounding this difference in replication frequency, RNA viruses generally encode their own nucleic-acid-replicating enzymes, which typically lack proofreading activity and therefore produce far greater numbers of mutations per replication cycle than that of their hosts. Combined, these two factors probably underlie the greater than million times rate of sequence change compared with other organisms. Even over relatively short evolutionary periods, fixation of the large number of mutations could make evolutionary relationships between viruses difficult or impossible to recover. Viruses may be subjected to intense selection pressures to evade the host's immune response and antiviral treatment and to adapt to new hosts on crossing species barriers. Rapid, adaptive changes are favoured in viruses by the frequently large population sizes in an infected organism and the high mutant frequency generated during replication (see § 6).

Despite the rapid sequence change of RNA viruses, HBV and retroviruses, a more modest, potentially achievable goal in virus reconstruction is to estimate the times of divergence of contemporary genetic variants of a virus, such as the genotypes of HBV and HCV. This is achieved by application of the principle of the 'molecular clock', which is based on the repeatable observation that the degree of both nucleotide and encoded amino-acid sequence divergence (calculated using relatively simple corrections for multiple substitutions) between homologous genes in different species remains proportional to their time of divergence. For example, sequence change in the α-chain of haemoglobin is relatively constant over extremely long periods of vertebrate evolution, and between sequences that have become extremely divergent. Nucleotide sequences from various mammals and marsupials which differ from each other by up to 40% show an inferred rate of sequence change ($0.5-1.7 \times 10^{-9}$ per site per year); this is little different from that observed over the approximately ten-times shorter period of ape speciation ($0.8-1.5 \times 10^{-9}$ per site per year). The simplest, although not the only conclusion that can be drawn from these observations, is that whatever functional constraints there may be on the encoded protein (in this case, in enzymatic activity), these are insignificant compared with the flexibility with which amino-acid substitutions can be introduced into the protein sequence, without evident change in fitness of the organism. However, irrespective of its mechanism, the importance of the molecular clock in evolutionary reconstruction lies in its ability to predict from the rate of sequence change of a gene (even over a short observation period) the time of divergence of other, more distantly related species. Given the lack of other evidence to construct virus histories, the molecular clock has been enthusiastically adopted as the method to calculate times of divergence of genetic variants of a wide variety of different viruses (Suzuki & Gojobori 1997; Zanotto *et al.* 1996; Bollyky & Holmes 1999; Zhang *et al.* 1999; Korber *et al.* 2000).

While these studies are of considerable value, the strict application of the principle of the molecular clock to the reconstruction of virus histories can produce many difficulties and incompatibilities with observational data. As I shall discuss in the context of the evolution of HBV, and hepatitis G virus or GB virus C (referred to here as HGV/GBV-C), some viruses do not appear to evolve in a manner comparable to that of animals and plants. Conventional methods for estimating evolutionary distances appear to fail to detect the resulting rate heterogeneity at different nucleotide sites, and substantially underestimate times of divergence. As will be discussed, viruses, for all their mutability and extreme population dynamics may be far more conservative, and older than has so far been recognized.

The final information that may contribute to understanding virus origins derives from the nature of the interaction of the virus with its host. Virus–host relationships range from commensal to parasitic. In many cases, the effect of the virus on its host is to become less harmful over time through adaptive changes in both parties. For example, herpesviruses, for which there is genetic evidence for a process of coevolution over the period of mammalian speciation (McGeoch *et al.* 2000), are exquisitely adapted to persist in their hosts without causing significant disease, but to remain infectious to allow transmission down the generations. In contrast, herpesvirus B, the homologue of herpes simplex virus in macaques, can cause fatal infection in humans. The pathogenicity of recently emerging viruses, such as human immunodeficiency virus (HIV)-1 and -2, Ebola and hantavirus, probably reflects similar failures of mutual adaptation. In such cases, viruses recently acquired through zoonotic transmission events are more pathogenic than viruses always present in humans.

In the following sections, available information on the epidemiology, pathogenicity, rate of sequence change, genetic variability, genotype distributions, and the existence of homologous viruses in non-human primates will be combined to provide a composite analysis of the various competing theories for the origins of HGV/GBV-C, HBV and HCV. As a forewarning, much of the following will strike many readers as absurdly speculative. The review, however, does achieve one of its original purposes in highlighting the considerable amount of information missing from this type of endeavour.

3. HGV/GBV-C

(a) *Background*

The name HGV/GBV-C remains as an ugly acronym for the virus independently but simultaneously discovered in 1995 (Linnen *et al*. 1996; Leary *et al*. 1996). The description of the virus as HGV virus is doubly unfortunate as there is no evidence that it causes hepatitis in its natural host (humans) either acutely or after long-term carriage. It is therefore something of an interloper in this review, but included because it demonstrates the approaches that can be taken to reconstruct virus origins most clearly.

HGV/GBV-C is distantly related to HCV and other flaviviruses in the *Hepacivirus* genus. Infection is found widely in human populations, with frequencies of active or past infection ranging from 5% to up to 50%. This distribution extends even to highly isolated populations, such as indigenous tribes people in Papua New Guinea, sub-Saharan Africa and Central and South America (Smith *et al*. 2000; Tanaka *et al*. 1998*a*,*b*; Mison *et al*. 2000). Infection is frequently persistent and associated with high levels of circulating viraemia, although no evidence links HGV/GBV-C to any identifiable hepatic or non-hepatic disease.

Variants of HGV/GBV-C show quite limited sequence variability, with nucleotide sequences differing from each other by a maximum of 13% (figure 1). HGV/GBV-C has been tentatively classified into four or five genotypes based on these sequence relationships (Smith *et al*. 2000; Mison *et al*. 2000; Sathar *et al*. 1999; Muerhoff *et al*. 1996; Mukaide *et al*. 1997), although the variants lack the clear phylogenetic groupings that underpin the genotype classification of other viruses such as HCV (see §5), perhaps reflected in the extreme conservation of the encoded amino-acid sequence throughout the genome. Expressed numerically, sequence divergence at non-synonymous sites (dN; i.e. sites where nucleotide substitutions alter the encoded amino acid) is at least 50 times less than the variability found at synonymous (i.e. silent) sites (dS) (Simmonds & Smith 1999). Most coding sequences show biases towards synonymous variability (i.e. they show a ratio of dN/dS significantly less than 1), but there are few known coding sequences with ratios approaching 0.02 (or less) that are found in HGV/GBV-C. The underlying reasons for this restricted variability are currently unclear.

Homologues of HGV/GBV-C are widely distributed in primates (Charrel *et al*. 1999; Adams *et al*. 1998; Leary *et al*. 1997; Bukh & Apgar 1997; Birkenmeyer *et al*. 1998) (figure 2). HGV/GBV-C variants more divergent than those found in humans have been found in different subspecies of wild-caught chimpanzees from Central and western Africa (Adams *et al*. 1998). Even more divergent homologues of HGV/GBV-C, described as GBV-A, have been recovered from several species of New World primates (Bukh & Apgar 1997; Erker *et al*. 1998; Leary *et al*. 1997). Again mirroring host relationships, genetic variants of GBV-A differing from each other by *ca*. 25% are closely associated with different New World primate species.

(b) *HGV/GBV-C: theories of origins*

The historical and archaeological record for HGV/GBV-C infection in humans is blank, as would be expected for a virus where infection is asymptomatic. The principal information we have to reconstruct its evolutionary history therefore has to be based on sequence comparisons of existing viruses.

Sequence comparison of HGV/GBV-C in samples collected 8 years apart from an individual acutely infected with HGV/GBV-C indicated a rate of 3.9×10^{-4} nucleotide substitutions per site per year over the whole genome (Nakao *et al*. 1997). This rate is comparable to that of other RNA viruses (e.g. 4×10^{-4} for HCV (Smith *et al*. 1997*b*); 1.4×10^{-4} for HIV-1 (Zhu *et al*. 1998)). If this rate of sequence change is maintained over longer periods, then the genotypes found in humans would have originated from a common ancestor approximately 170–180 years ago. This estimate can be refined by extrapolation of the rate of sequence change at synonymous sites, as this is not influenced by the functional constraints that appear to operate on the amino-acid sequence of HGV/GBV-C (see §3a). Using a synonymous substitution rate of 1.1×10^{-2} per site per year, and Jukes–Cantor-corrected synonymous distances between genotypes of 0.59–0.64, a slightly earlier time of divergence of 225–240 years ago is estimated. Further refinement of this estimate by evolutionary distances that allow for different frequencies of transitions and transversion at synonymous sites have little further effect on this estimate. Extrapolating further, nucleotide sequence differences between HGV/GBV-C variants infecting humans, chimpanzees and New World primates imply times of divergence 600–1000 years ago.

These relatively recent times of origin of human genotypes of HGV/GBV-C and of the species-associated variants in primates make little sense epidemiologically. The estimate for the time of origin of human genotypes implies epidemic, global spread over the last 200–300 years, while the variants found in the wild in chimpanzees and in New World primates in South America would therefore require a transmission chain of infection operative over the last millennium. It is difficult to imagine by what route HGV/GBV-C could have spread so far over this time-scale. Even more puzzling is the distribution in primates. The implied cross-species transmissions are incompatible with observations that GB viruses obtained from New World primates are non-infectious in chimpanzees, nor can human HGV/GBV-C infect New World primates (J. Bukh, personal communication). A transmission chain linking these divergent primate species in recent evolutionary history is therefore unlikely indeed.

Putting aside the predictions from the molecular clock, the broad distribution in human populations, and its

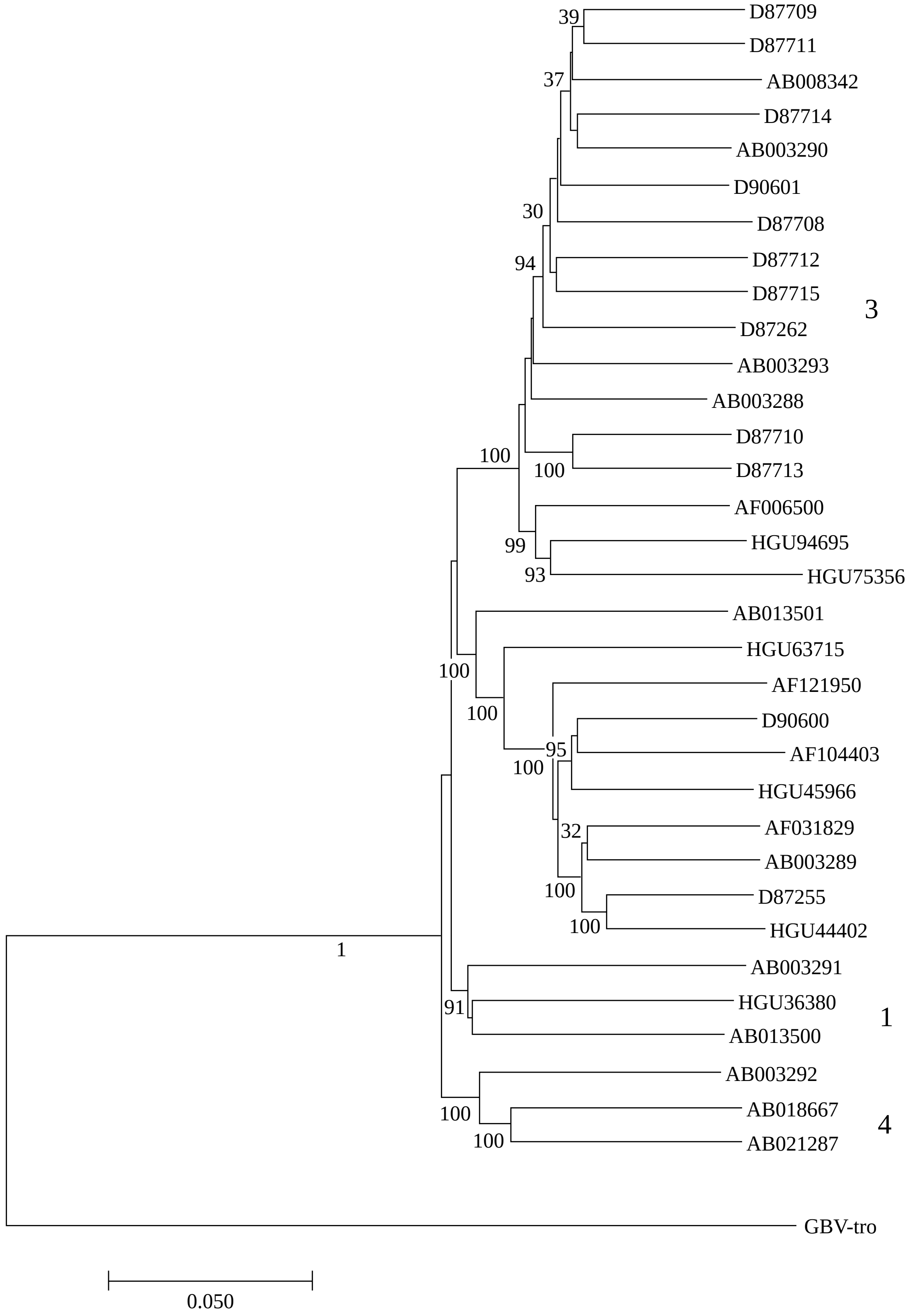

Figure 1. Phylogenetic analysis of complete genomic sequences of HGV/GBV-C (Smith *et al.* 2000), and their provisional classification into four genotypes (the complete sequence of the proposed genotype 5 is currently unavailable). The tree was generated using Jukes–Cantor distances calculated using the MEGA package (Kumar *et al.* 1993), with the more distantly related chimpanzee sequence HGV/GBV-Ccpz as an outgroup. Numbers on branches indicate number of bootstrap re-samplings from 100 supporting observed phylogeny, restricted to values of 70% or greater; *p* distances indicated on the scale bar.

apparent non-pathogenicity, are much more consistent with the long-term presence and close host association of HGV/GBV-C with humans. Further evidence for this hypothesis is provided by the geographical distribution of HGV/GBV-C genotypes among indigenous populations in different parts of the world. In all cases, these are congruent with the distributions expected if HGV/GBV-C was already present in modern human populations as they migrated out of Africa 100 000–150 000 years ago (Tucker *et al.* 1999; Konomi *et al.* 1999; Liu *et al.* 2000; Tanaka *et al.* 1998*a*,*b*; Mison *et al.* 2000; Katayama *et al.* 1997; Gonzalez Perez *et al.* 1997). For example, sequences

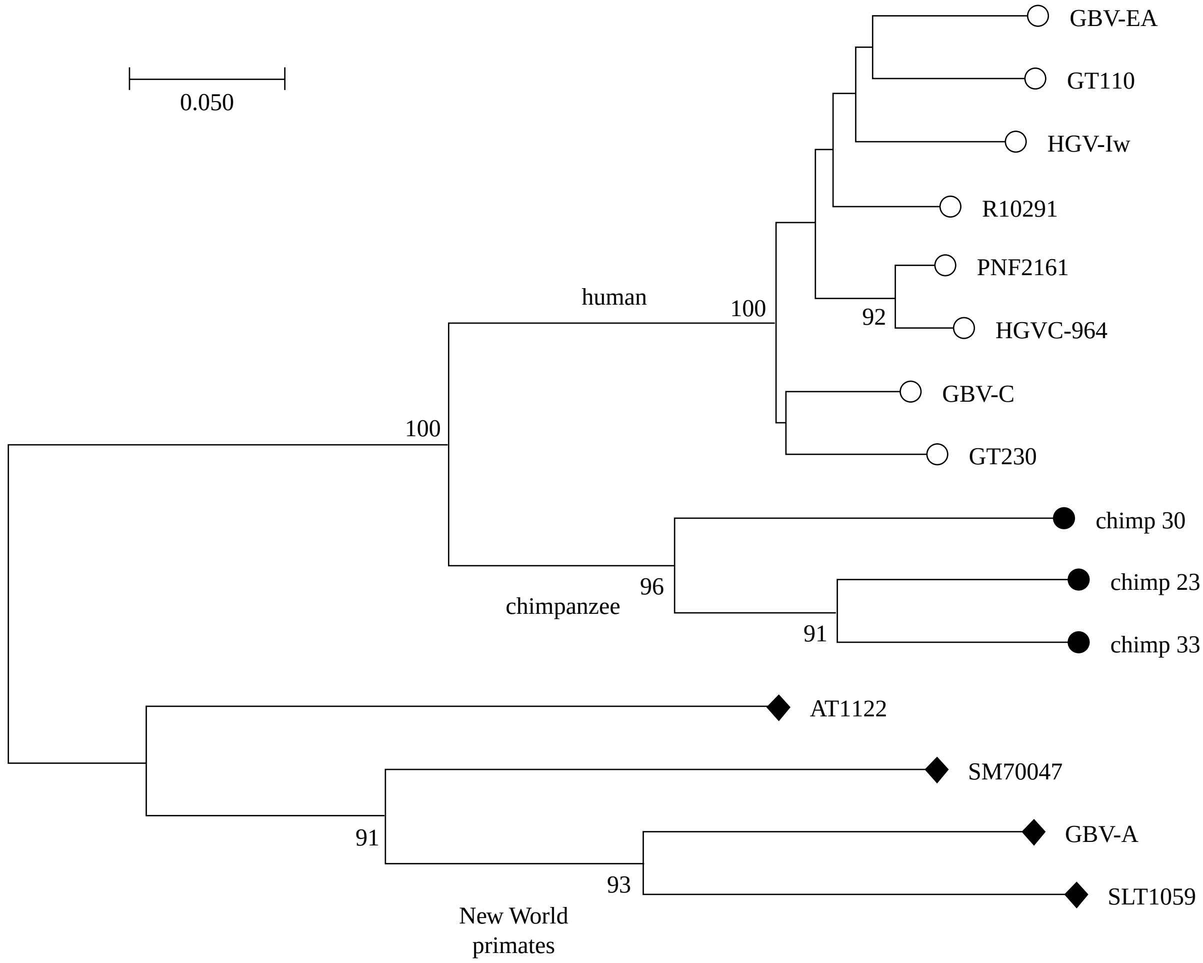

Figure 2. Unrooted phylogenetic tree of sequences from the NS5 region of HGV/GBV-C from humans (open circles), HGV/GBV-Ccpz from chimpanzees (solid circles) and GBV-A sequences recovered from different New World primate species (solid diamonds) (Adams *et al.* 1998). The branching order (but not scale) of GB virus sequences from different primates was congruent with the genetic relatedness of their host species. Note the greater sequence divergence between HGV/GBV-Ccpz variants recovered from *troglodytes* and *verus* subspecies of chimpanzees than found between human HGV/GBV-C genotypes.

from the populations in the Far East are almost invariably genotype 3, and this genotype is otherwise only found in native inhabitants of North and South America (Konomi *et al.* 1999; Tanaka *et al.* 1998*b*; Gonzalez Perez *et al.* 1997). In contrast, Caucasian and other populations from India westwards including northern Africa are infected with genotype 2. Genotype 1 is confined to sub-Saharan Africa, and shows the greatest overall sequence diversity (Liu *et al.* 2000; Smith *et al.* 1997*a*; Muerhoff *et al.* 1997); particularly divergent variants have been recovered from Pygmy and other African populations (Sathar *et al.* 1999; Tanaka *et al.* 1998*a*).

The long association of HGV/GBV-C in humans is mirrored in other primates, where the distribution of homologues of HGV/GBV-C is congruent with host relationships (Charrel *et al.* 1999; Adams *et al.* 1998; Leary *et al.* 1997; Bukh & Apgar 1997; Birkenmeyer *et al.* 1998) (figure 2). These repeated examples of phylogenetic congruency between GB viruses and their primate host species are clearly consistent with the hypothesis of virus and host coevolution. However, this hypothesis is quite incompatible with the recent dates of divergence predicted by the molecular clock. For example, sequence divergence of 13–14% (and 0.6–0.65 at synonymous sites) between HGV/GBV-C variants that diverged over 100 000 years ago implies a sustained rate of measurable sequence change several orders of magnitude lower than observed over 8 years (figure 3). Similarly, the rate of sequence change over the interval in which Old and New World primate species evolved is even more discrepant from this short-term rate (approximately 10 000-fold lower).

Structural analysis of the genome of HGV/GBV-C has suggested a possible resolution between these short- and long-term rates of sequence change (Cuceanu *et al.* 2001; Simmonds & Smith 1999). We have suggested that the extensive RNA secondary structure formed on folding of the single-stranded genome may constrain sequence changes in base-paired regions and therefore lead to a marked underestimation of the frequency of multiple substitutions that occurred on comparison of more divergent sequences. This could reproduce the apparent 'slowing' of sequence change implied from the sequence relationships between HGV/GBV-C genotypes, and between GB viruses infecting different primates (figure 3).

Our evidence that RNA secondary structure influences sequence evolution of HGV/GBV-C is based upon several

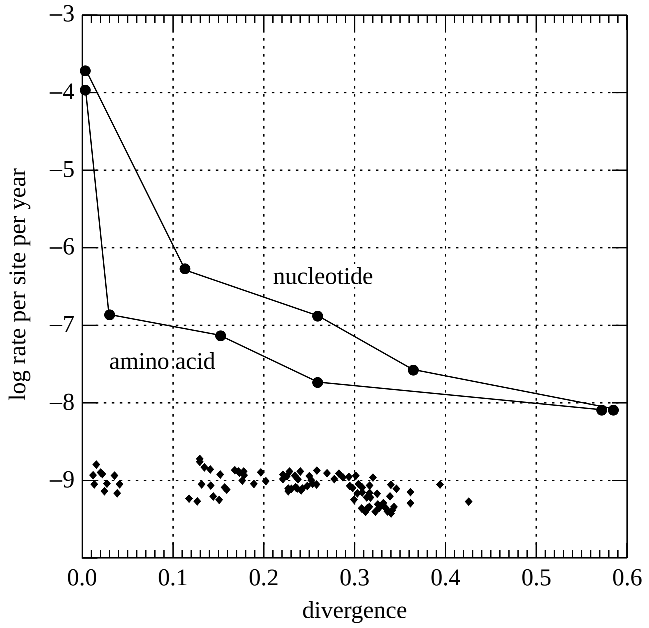

Figure 3. Comparison of inferred rate of sequence change of HGV/GBV-C with degree of amino-acid sequence variation divergence in the NS5 gene. Data points from left to right represent the following divergence times: (i) 8.4 years: time-course in HGV/GBV-C-infected individual (Nakao *et al.* 1997); (ii) 100 000 years: divergence of modern humans; (iii) 1.6 Myr: divergence of *troglodytes* and *verus* subspecies of chimpanzees (*Pan troglodytes*; Morin *et al.* 1994); (iv) seven Myr: divergence of humans and chimpanzees (Jones *et al.* 1992; Morin *et al.* 1994); (v) 35 Myr: divergence of Old (human and chimpanzees) and New World primates (*Sanguinis mystax* and *labiatus*, *Aotus trivirgatus*; Jones *et al.* 1992). Sequences compared were from the NS5 region of the genome (amino-acid positions 2498–2561 in sequence PNF2161 [U44402]), with divergences and rates based on Jukes–Cantor distances. The rate of sequence change measured over 8 years (4×10^{-4} substitutions per site per year) was over 10 000 times more rapid than the rate of divergence of the period of New and Old World primate separation. For comparison, sequence divergence between pairs of α-globin genes of mammals, placentals and birds (solid diamonds) are plotted against the rate of amino-acid sequence change using times of speciation derived from palaeontological records (Kumar & Hedges 1998). In contrast to HGV/GBV-C, rates of sequence change were relatively constant irrespective of their degree of divergence. Over a wide range of separation times (1–310 Myr), rates of sequence change remained within the range $0.5–1.7 \times 10^{-9}$ nucleotide substitutions per site per year.

unusual features in the distribution of sequence variability in coding sequences (Simmonds & Smith 1999). Quite apart from the extreme conservation of the encoded amino-acid sequence of the HGV/GBV-C polyprotein (remarked upon in § 3a), we also obtained evidence that a large proportion of synonymous sites in the coding part of the HGV/GBV-C genome is also unexpectedly invariant; comparison of complete coding sequences (8600 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%). This carries the necessary implication that there are fitness constraints on sequence change of HGV/GBV-C over and above the coding function of the genome. As described in detail elsewhere (Cuceanu *et al.* 2001; Simmonds & Smith 1999), it now appears likely that the RNA genome of HGV/GBV-C forms a complex and extensive secondary structure through internal base pairing. The high free energy on folding, the existence of multiple covariant sites, and the conservation of specific stem-loops between quite divergent GB virus sequences (such as HGV/GBV-C, chimpanzee HGV/GBV-C (HGV/GBV-Ccpz) and GBV-A) all point to (an) evolutionarily conserved function(s) for the predicted secondary structure. What this actually is, and the extent to which it may be found in other viruses with single-stranded genomes, remains to be determined.

Nonetheless, the restrictions imposed by secondary structure provide an important clue towards understanding how the coevolution hypothesis for HCV/GBV-C can be reconciled with the observation of its rapid rate of sequence change over short observation periods. For example, there may be a class of synonymous sites which are in non-base-paired parts of the genome, and where sequence change may be relatively unconstrained. These substitutions may therefore be fixed at the frequency predicted from the measured short-term rate of sequence change (Nakao *et al.* 1997). A different class of nucleotide sites that participate in internal base-pairing could be under greater constraint if the resulting secondary structure influenced the fitness of the virus. Substitutions may therefore only occur if simultaneous compensatory changes occur to retain base-pairing in the stem-loop. Indeed, in our analysis of HGV/GBV-C sequences, covariant sites in the predicted stem-loop were found throughout the coding part of the genome, and rivalled the 5′-untranslated regions of HCV and HGV/GBV-C in frequency and complexity. As the third base positions (normally synonymous) are usually opposite each other in the predicted stem-loops of HGV/GBV-C, the frequency at which covariant substitutions may occur is the substitution frequency squared (i.e. approximately $10^{-7}–10^{-8}$ substitutions per site per year). This is indeed quite similar to the rate at which sequence divergence accumulates over the longer periods of human dispersal and primate speciation (figure 3).

It seems therefore as if after the rapid accumulation and saturation of substitutions at unpaired sites, further diversification can only occur at the much slower rate required by paired changes that retain secondary structure. Indeed, the extreme conservation of the amino-acid sequence may reflect the even greater difficulty of simultaneous sequence change at opposite, non-synonymous base-paired sites; both amino-acid changes would have to be neutral or beneficial to HGV/GBV-C for the covariant change to be fixed in the virus population. The extreme dN/dS ratio mentioned above (§ 3a) may therefore result more from the peculiar constraints imposed by the requirement to maintain RNA secondary structure, rather than functional or structural conservatism of the encoded proteins.

Further analysis of free energy on folding, frequencies of covariant sites and other manifestations of internal base-pairing should also be determined for other viruses with single-stranded genomes, to investigate whether the constraints imposed by secondary structure formation represent a more general principle of virus evolution.

4. HBV

(a) *Background*

HBV chronically infects *ca.* 5% of the human population. HBV is transmitted by sexual contact and by parenteral exposure, although it is thought that mother-to-child perinatal transmission, and the establishment of a lifelong highly infectious carrier state is responsible for the observed high rates of endemicity in high-prevalence regions such as southern and eastern Asia, sub-Saharan Africa and among indigenous peoples in Central and South America.

HBV is classified in the Hepadnaviridae, and contains a partly double-stranded DNA genome of approximately 3200 bases. HBV replicates via an RNA intermediate anti-genome sequence, encoding a potentially error-prone polymerase enzyme with both reverse transcriptase and DNA polymerase activities. An unusual feature of the HBV genome is the presence of multiple overlapping reading frames for the genes encoding the core, polymerase and surface antigen genes; 67% of the genome is multiply coding, and therefore lacks what would be conventionally regarded as synonymous and non-synonymous sites. It is additionally probable that regions of the genome that are non-coding may be involved in a variety of secondary structure interactions necessary for circularization and transcription.

HBV is currently classified into seven genotypes differing from each other by nucleotide sequence distances of *ca.* 10–13% (figure 4). Genotypes A and D have global distributions, genotypes B and C are found predominantly in eastern and southeastern Asia, genotype E is predominant in western Africa, and the most divergent genotype F is found exclusively among indigenous peoples in Central and South America (Arauz-Ruiz *et al.* 1997; Norder *et al.* 1994). Genotype G is identified in too few samples at present for its distribution to be determined.

(b) *Theories of origins of HBV*

HBV contains a polymerase enzyme without proof-reading activity, and error frequencies during RNA or DNA copying are likely to be of the order measured for the related retroviruses, and for other RNA viruses. Measurement of the rate of sequence change of HBV is complicated by the existence of overlapping reading and the lack of synonymous sites in most of the coding sequence. Compounding this difficulty is the evidence that many amino-acid changes, particularly in the pre-core region, have a positive selective value, and may occur as an immune-evasion strategy. In a recent study (Hannoun *et al.* 2000), individuals with hepatitis and who had cleared HBe antigen (HBeAg) (interpreted as evidence of a vigorous immune response to HBV) showed a mean 12-fold greater nucleotide substitution rate than individuals who were apparently immunotolerized (mild hepatitis, HBeAg-positive). Taking the latter group as representing the evolution of HBV in the absence of immune pressure, HBV shows a substitution rate of 2.1 (range of 0–13) $\times 10^{-5}$ substitutions per site per year over a mean observation period of 22 years (range of 20–35 years). As it is generally HBeAg-positive carriers who transmit HBV infection between human generations, this rate is likely to be the most appropriate for extrapolating substitution rates over longer periods. On this basis, the human genotypes of HBV would have originated from a common ancestor *ca.* 2300–3100 years ago. How this predicted time of divergence fits with the various theories for the origin of HBV is explored in the remainder of this section.

A bold account for the geographical distribution of HBV genotypes proposed that HBV originated from the Americas, and spread into the Old World over the last 400 years after contact from Europeans during colonization (Bollyky *et al.* 1997). The existence of the various human HBV genotypes was interpreted as the diversification of HBV after their geographical dispersal over the 400-year period, although the date calculated from the molecular clock implies that a much longer period would be required (over 2000 years). Although it is possible that the human genotypes pre-existed before spread, there is no evidence for infection other than with genotype F in contemporary indigenous populations in South America. Secondly, it would be more difficult to account for the differences in geographical distributions of HBV genotypes in the Old World if they had not diverged *in situ*.

However, the main difficulty with accepting the 'out of South America' hypothesis is the recent discovery that HBV is widely distributed in Old World primate species, such as chimpanzees, orang-utans and gibbons. Although primate infection was often originally dismissed as the result of accidental transmission from humans to captive animals (Lanford *et al.* 1998; Norder *et al.* 1996; Vaudin *et al.* 1988; Zuckerman *et al.* 1978), it has now become more firmly established that West African chimpanzees are infected with HBV in the wild (Hu *et al.* 2000; Takahashi *et al.* 2000; MacDonald *et al.* 2000) (figure 4). This chimpanzee-specific genotype of HBV showed *ca.* 11% divergence from the human genotypes A–E and G. HBV infection has also recently been reported in wild-caught gibbons and orang-utans in southeastern Asia, both of which harbour specific genotypes of HBV equidistant from each other, from chimpanzee and human genotypes (Grethe *et al.* 2000; Warren *et al.* 1999). Sticking to the predictions of the molecular clock, primate-associated genotypes of HBV would also have originated within the last 2000–3000 years. As humans are the only species known to have travelled between the various continents over this period, it would be necessary to propose that humans were the vectors for the spread of HBV into other species. As discussed above (§ 3) for HGV/GBV-C, this is plainly absurd.

Since the evidence for primate infections in the wild became widely known, the difficulty has been to provide alternative hypotheses for the evolutionary history of HBV. The main difficulty with any account arises from the unexpected observation of equivalent sequence relationships between human genotypes A–E and G to each other and to the primate-species-associated genotypes of HBV. It is also difficult to rationally fit into any scheme the outlier human HBV genotype F and the even more divergent HBV variant obtained from a captive woolly monkey, a New World primate (Lanford *et al.* 1998).

At least one group has proposed that the evolutionary history of HBV corresponds to the spread of anatomically modern humans as they migrated from Africa *ca.*

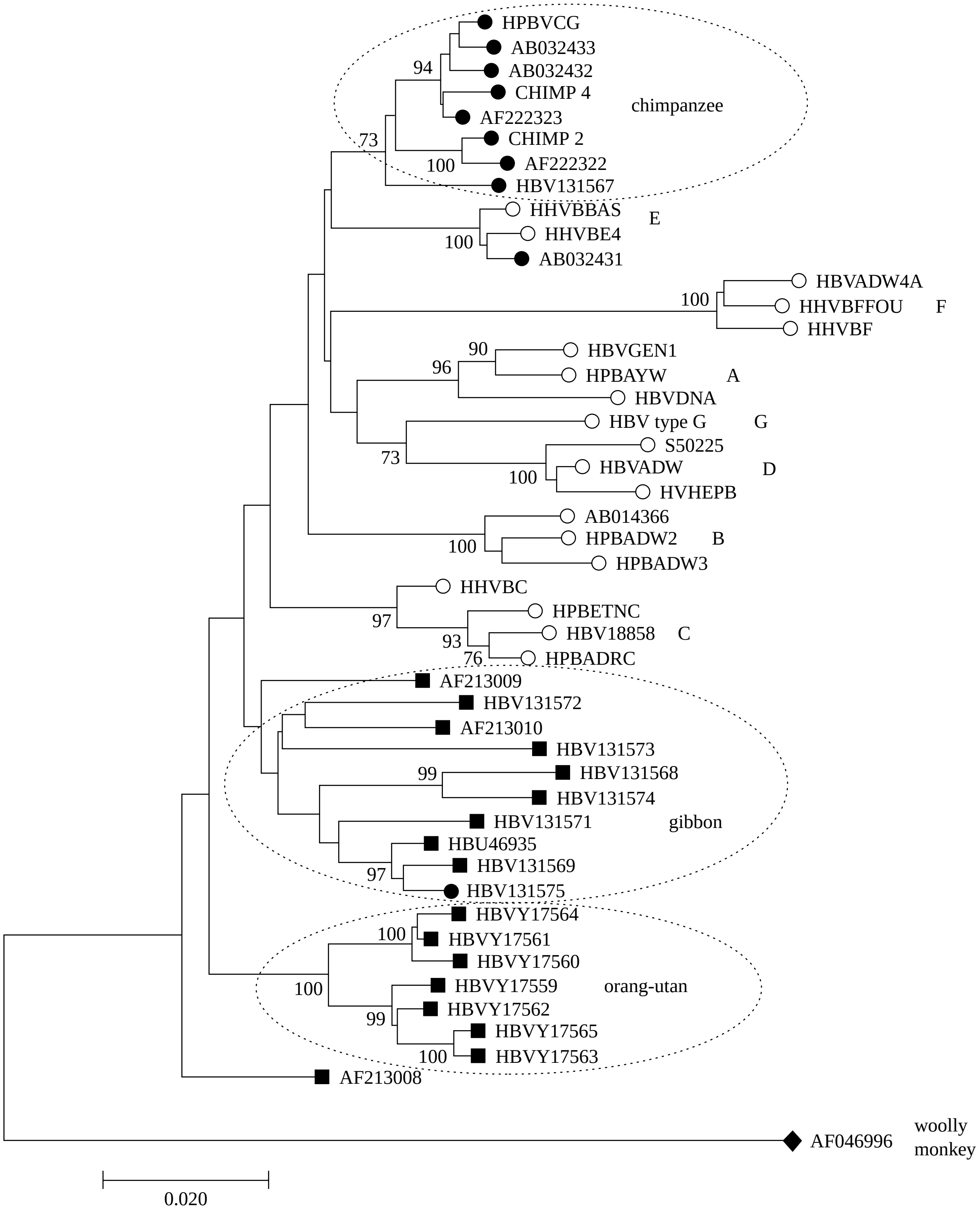

Figure 4. Phylogenetic analysis of HBsAg gene sequences of representative sequences of human genotypes A–G (open circles) and HBV sequences recovered from chimpanzees and gorillas (solid circles), gibbons and orang-utans (solid squares) and the New World woolly monkey (solid diamond). Note the intermixing and approximately equal sequence divergence between human genotypes A–E and G with sequences recovered from different primate species. See legend to figure 2 for details of tree construction. Exceptions to the human and other primate species associations are as follows: one of the sequences in the chimpanzee clade originated from a gorilla (AJ131567); a chimpanzee sequence (HBV131567) groups with gibbons; another chimpanzee sequence (AB032431) groups in human genotype E. Finally, sequence AF213008 from a gibbon groups separately from all other sequences. It is not impossible that laboratory error or contamination, or inadvertent transmission of HBV between species in captivity, can explain these discrepancies, although it is also possible that the claimed species or genotype associations reflect limitations on current epidemiological data.

100 000 years ago (Norder *et al.* 1994; Magnius & Norder 1995). However, unlike HGV/GBV-C, the phylogeny of HBV genotypes in no way corresponds to genetic relationships between human population groups. For example, the presence of genotype F in native American populations is inconsistent with the presence of genotypes B and C in Mongoloid northeast Asians, who are genetically their nearest relatives. Indeed, there is little relationship

between HBV genotype distributions with any of the other human population groups (southeast Asians, Caucasians and African populations). As indicated above, the other incongruity is the existence of species-specific genotypes of HBV in chimpanzees, gibbons and orang-utans, and the way in which they are intermixed with human genotypes. This is quite unexpected, as the primate viruses should be much more divergent from human variants and from each given the much longer period of co-speciation of primate species (10–15 million years (Myr)).

The third hypothesis for HBV origins that we recently discussed in a preliminary way (MacDonald *et al.* 2000) argues that variants found in chimpanzees, gibbons, orang-utans and in the New World primate woolly monkey species are those which co-speciated over 10–35 Myr. In this case, the outlying position of the woolly monkey HBV sequence and the equal divergence of HBV variants from Old World primate species is (approximately) consistent with host phylogeny and fossil-based estimates for their relative times of divergence. If co-speciation occurred, then the long-term rate of sequence change of HBV would range from 3 to 5×10^{-9} nucleotide changes per site per year, a challenge to reconcile with its short-term rate of sequence change. Interestingly, a range of much more genetically divergent hepadnaviruses infects rodents in North and South America, such as the woodchuck (*Marmota monax*), ground squirrel (*Spermophilus beecheyi*) and arctic ground squirrel (*S. parryii*). These viruses may be a manifestation of an equivalent process of coevolution over even longer periods. Remarkably, the genetic distance between primate and rodent HBV variants after their divergence *ca.* 110 Myr ago indicates a rate of sequence change of 6.8×10^{-9} changes per site per year, bizarrely similar to the rate operating over primate co-evolution. HBV homologues in rodents are unable to infect primates (and vice versa), an observation that implies a considerable evolutionary gulf between the various mammalian HBV variants. This is quite incompatible with the prediction from the molecular clock for a time of divergence between primate and rodent HBV variants of only 10 000 years.

If HBV coevolved in primates then the existence of numerous equally distinct human genotypes of HBV would require a different explanation. It is possible that human HBV infection arose many times through contact with different primates infected with species-specific genotypes (equivalent to those found in chimpanzees, gibbons and orang-utans). In some ways, this scenario corresponds to that believed to underlie the origins of HIV infection in humans. Infection with HIV-1 is likely to have originated through at least three separate cross-species transmissions from chimpanzees (Gao *et al.* 1999), while human infection with HIV-2 in western Africa arose independently several times through contact with sooty mangabeys (Feng *et al.* 1992). A primate origin for human HBV infection is indeed supported by the observation that the areas of high HBV prevalence in humans are those in which contact and cross-species transmission from primates is most likely (South America, sub-Saharan Africa and southeastern Asia). Indeed, certain HBV genotypes are specific to these three areas (F, E and B/C, respectively). The problem with the theory for a primate origin is that, to date, no HBV genotypes are shared between primates and humans; if HBV genotypes A–G originated in primates, then the actual species involved in transmission to humans remain unidentified.

At this stage, it would be unwise to accept any of the above hypotheses for the origin of HBV as proven or even probable. However, the lack of sequence diversity in viruses with such widely different epidemiological distributions provides another example of a virus evolving in a markedly different way from that of higher organisms.

5. HCV

(a) *Background*

Infection with HCV has become established as a major cause of chronic liver disease in Western countries. Its spread in these populations is poorly understood, although it is known to be transmitted by blood contact, and has particularly targeted risk groups such as injecting drug users (IDUs), and in the past, recipients of blood transfusions and blood products. HCV infection is frequently persistent, and sets in train an inexorable course of slowly progressive liver disease. HCV contains a positive-sense RNA genome and is classified with HGV/GBV-C as a flavivirus; its host range is confined to humans and close primate relatives.

(b) *Emergence of HCV?*

Currently it is estimated that *ca.* 0.5% of the UK population is infected with HCV. Much of the evidence for the recent spread of HCV derives indirectly from descriptions of current genotype frequencies of HCV in different risk groups and populations. HCV can be classified into a number of distinct genotypes, whose distribution varies both geographically and between risk groups (reviewed in Simmonds 1998). Currently known variants of HCV collected from different parts of the world can be divided into six main ‘genotypes’, many of which contain more closely related variants (figure 5). Each of the six main genotypes of HCV is approximately equally divergent from each other, differing at 31–34% of nucleotide positions on pairwise comparison of complete genomic sequences, and leading to *ca.* 30% amino-acid sequence divergence between the encoded polyproteins.

In Europe, genotypes 1b and 2 are widely distributed, particularly in older age groups, while those infected through drug use are more likely to be infected with genotypes 3a and 1a. The observation of genotypes associated with drug use in Europe that are distinct from those found in individuals infected through other routes suggests infection of IDUs originated through a geographically large transmission network largely distinct from other HCV-infected individuals.

(c) *Theories of the origins of HCV*

Relatively accurate rates for the rate of sequence change of HCV have been determined for sequences such as those in NS5, which are not under immune-mediated selection pressures. Comparison of HCV variants infecting individuals inadvertently infected with HCV-contaminated anti-D immunoglobulin 17–20 years previously indicated rates of sequence change of 4.1 and 7.1×10^{-4} per site per year in the NS5 and envelope (E1) regions of the genome (Smith *et al.* 1997*b*). This figure is

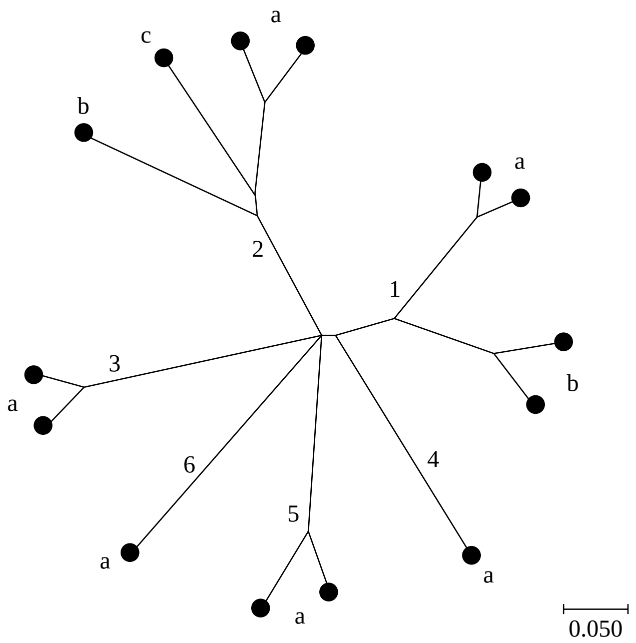

Figure 5. Sequence relationships of complete genomic sequences of the common variants of HCV found in Western countries (Chamberlain *et al.* 1997) shown as an unrooted tree, and their classification into six genotypes (first tier) and subtypes a, b, etc. (second tier). The nomenclature of the HCV genotypes follows the consensus proposal for classification of HCV (Simmonds *et al.* 1994). See legend to figure 2 for details of tree construction. Genotype 1a is widely distributed in Northern Europe and the USA associated with IDUs. Genotype 1b is the commonest genotype worldwide in older age groups, with risk factors generally ill-defined. Genotype 2 is found predominantly in older HCV-infected individuals from Mediterranean countries and the Far East. Genotype 3 is principally associated with IDUs, particularly those from Europe. Genotype 4 is widely distributed in the Middle East and is associated with past medical treatments (e.g. Bilharzia injections). Genotype 5 is found commonly only in South Africa. Genotype 6 is found in IDUs in Hong Kong and Vietnam, and more recently in Australia.

similar to that of HGV/GBV-C (§ 3) and other RNA viruses.

Assuming this rate of sequence change is maintained over longer periods, the diversity of variants within each of the genotypes associated with the risk groups for HCV infection was made; these included types 1a, 1b and 3a in Western countries. For type 1b, 40 NS5 sequences from epidemiologically unrelated individuals in Europe, USA, Asia and Japan showed a distribution of pairwise distances approximately four times greater than those between anti-D recipients, indicating a time of divergence *ca.* 60–70 years ago (Smith *et al.* 1997*b*). The absence of any country or region-specific phylogenetic groupings further implies that the initial spread of type 1b occurred relatively rapidly, and became disseminated throughout many of the world's populations over a short period. Recent epidemic spread of HCV is consistent with the star-like shape of the phylogenetic tree analysed using mid-depth analysis (Holmes *et al.* 1995).

The diversity of sequences among type 3a variants was more restricted than type 1b, suggesting a more recent dissemination (40 years based upon distances in NS5). The genetic evidence of relatively recent spread of genotypes such as 3a into IDUs is consistent with the epidemiological evidence for the widespread increase in needle-sharing drug abuse since the 1960s, while the greater diversity of types 1b and subtypes of type 2 implies earlier, different modes of transmission, consistent with a range of other risk factors identifiable in older HCV-infected individuals.

Although these and other ongoing molecular epidemiology studies of HCV sequence diversity appear successful in documenting the relatively recent spread of HCV, the problems and controversy associated with the analysis of HGV/GBV-C and HBV sequences suggest the need for caution with reconstruction of its earlier history. The problem with HCV is that, apart from studies of genotype distributions, there is little other information currently available that would help towards identifying the source of the current epidemic HCV infection in the West. Furthermore, unlike HBV and HGV/GBV-C, there is currently no evidence for the existence of homologues of HCV in non-human primates, apart from GBV-B discussed below.

Following the format of previous sections and assuming the validity of the molecular clock, we could extrapolate the time of divergence of the main genotypes of HCV from the rate measured over 20 years. This ranges from 500 to 2000 years using different methods to allow for different substitution rates at synonymous and non-synonymous sites, and of transitions and transversion (Smith *et al.* 1997*b*). Whether or not this estimate is correct is currently impossible to determine.

Other information on the origin of HCV derives indirectly from investigations of genotype distributions in non-Western countries. These are poorly documented, particularly in sub-Saharan Africa. What information is available, however, indicates a quite different pattern of sequence variability of the virus. For example, in West African countries, such as Gambia, Ghana, Burkina-Faso, Benin and Guinea (Ruggieri *et al.* 1996; Wansbrough Jones *et al.* 1998; Mellor *et al.* 1995; Jeannel *et al.* 1998), small-scale surveys have indicated a predominance of infection with genotype 2. In contrast to Western countries, these type 2 infections are characterized by considerable sequence diversity, with different individuals each being infected with different subtypes, each in turn distinct from the 2a, 2b and 2c subtypes found in the West. Similarly, genotype 4 infection in Central Africa (the Congo, Gabon, Central African Republic) is also characterized by extreme subtype diversity (Xu *et al.* 1994; Menendez *et al.* 1999; Stuyver *et al.* 1993; Fretz *et al.* 1995; Bukh *et al.* 1993), quite different from the epidemic pattern of type 4a infection in Egypt and elsewhere in the Middle East. Combining a large number of geographical surveys, HCV variants from five different regions in Africa and southeastern Asia contain areas of great subtype diversity (figure 6). The accumulated sequence information from these various surveys produces a phylogenetic tree quite different from those constructed from samples collected in Western countries (figures 5 and 6). Most markedly different is the diversity of HCV sequences at the level of subtypes. It rather looks as if the subtype tier of HCV variability that was first apparent on

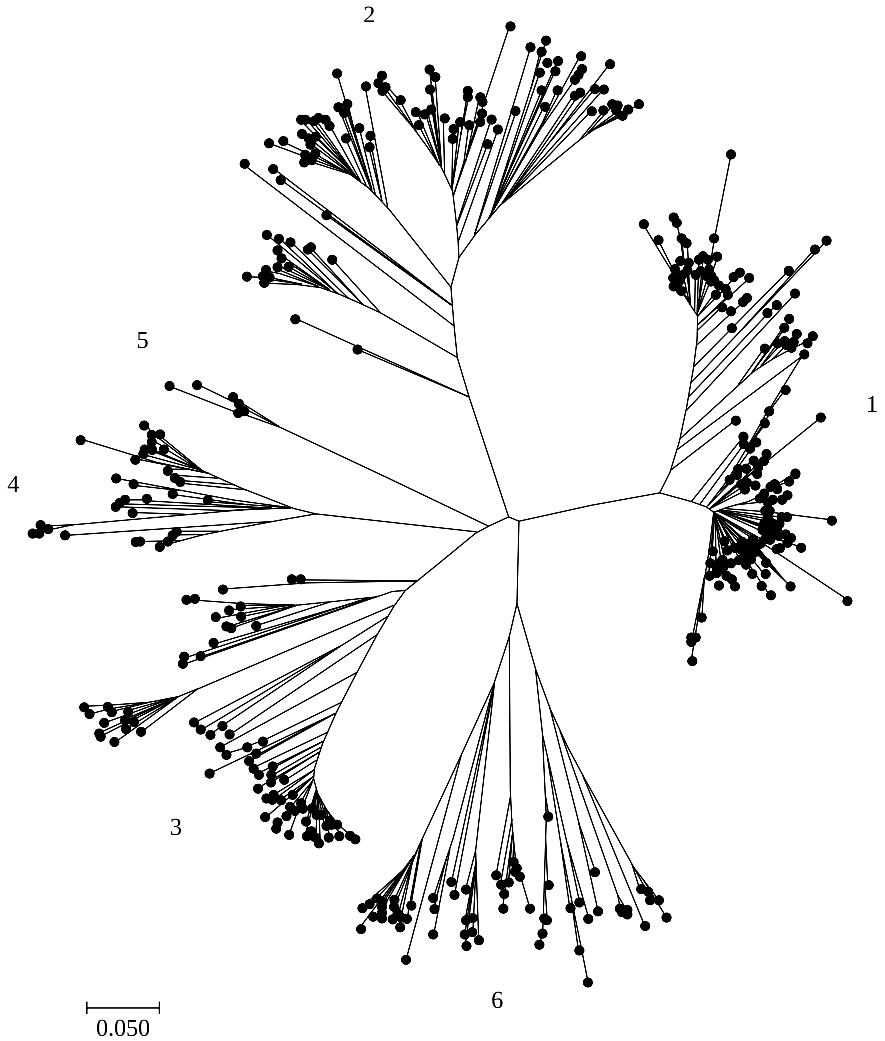

Figure 6. Unrooted phylogenetic analysis of nucleotide sequences from part of the HCV NS5B region amplified from HCV-infected individuals, including those from sub-Saharan Africa and southeastern Asia. Note the much greater diversity and number of subtypes in genotypes 1, 2, 3, 4 and 6 from those found in Western countries. See legend to figure 2 for details of tree construction.

comparison of HCV sequences from Western countries simply reflects the residual founder effect of certain HCV variants that sparked its recent epidemic spread.

For another virus, HIV-1, the great sequence diversity in Central Africa provides genetic evidence for the original source of the subsequent worldwide epidemic. By analogy, it could be imagined that the equivalent areas for HCV represent sites of origin for each of the genotypes that subsequently spread into Western countries. As with HIV-1, the sheer range of HCV sequences in sub-Saharan Africa and southeastern Asia argues for the long-term presence of HCV in human populations from these areas. Extending the analogy, the demographic and epidemiological factors that led to the global spread of HIV-1 in the last 30–40 years could be the same ones underlying the epidemic spread of HCV over the same period, although it is clear that spread of certain genotypes of HCV, such as types 1b and subtypes of genotype 2 occurred earlier. This hypothesis of HCV origins differs in being multifocal; southern Asia appears to harbour the greatest diversity of genotypes 3 and 6, while types 1, 2, 4 and probably 5 would be of African origin. To substantiate this speculative hypothesis, we clearly need much more information on the epidemiology of HCV infection in these areas. In particular, it would be important to identify the transmission routes between individuals that would maintain HCV infection in a human population for the long periods implied by its genetic variability. There are few, if any, other examples of viruses infecting populations where the principal route of transmission is parenteral.

Intriguingly, the areas of greatest genetic diversity of HCV are also those of greatest HBV prevalence. Perhaps if we understood more about the origins of HBV we might be able to explain this coincidental distribution of HCV diversity and the existence of shared factors that may maintain infections with both viruses in these communities. Pursuing the analogy with the origins of HIV-1 (and possibly HBV), there would also be a place for more intensive investigation of infection with HCV or related viruses in Old World primates.

6. CONCLUDING REMARKS

In many respects (and certainly for many readers), this review has achieved very little. It documents the lack of direct information on the origins of hepatitis viruses, and the speculative nature of much of the indirect evidence. What I hope it has achieved is a reassessment of the validity of the molecular clock, which has been seized upon by many researchers as possibly the only way to reconstruct virus histories. As demonstrated for HGV/GBV-C and HBV, extrapolation of rates of sequence change to longer periods produces absurdly recent predicted times of their origins. Quite apart from the difficulty in explaining the presence of these viruses in far-flung, often highly isolated human populations, it cannot account for their distributions in wild-caught primates between which cross-species transmission is unlikely historically or impossible because of biological species barriers.

Despite its shortcomings, much of the indirect evidence for virus origins that derives from studies of genotype distributions points to very long-term virus–host interrelationships. In the case of HGV/GBV-C, it appears that recognizably similar viruses remain over periods of primate speciation, a quite different outcome from the conventional view of RNA viruses, where properties of extremely rapid sequence change, ability to adapt, and ephemeral nature have always been emphasized (Holland *et al*. 1982). My own investigations of this discrepancy have concentrated on the constraints on sequence change, such as RNA secondary structure formation in HGV/GBV-C, which may lead to significant underestimation of evolutionary distances between more distantly related variants. Other authors have documented the unusual complications imposed by the use of multiple reading frames by HBV, and the lack of conventional synonymous sites in most of the genome (Mizokami *et al*. 1997).

Aside from these specific issues, a more general principle governing the evolution of viruses that is distinct from that of large, multicellular organisms is population size. As discussed much more expertly elsewhere, the problem with becoming multicellular and big is the inevitable restriction this places on the size of populations between which meaningful selection can occur. Absurdly large genome sizes, for the most part packed with repetitive, mobile elements and other junk DNA, introns and often nonsensical redundancy in gene function, biochemical pathways and the immune system together indicate that the process of selection remains weak in the overall context of genome change. It seems additionally that the lack of effective selection may extend to coding sequences; the reason that the molecular clock appears to operate in animal and plants is probably that there is very little effective selection pressure to prevent the fixation of mutations that have relatively minor effects on organism fitness. As a result, gene sequences can diverge more or less randomly during speciation, and reproduce the linear relationship between time and degree of sequence divergence that is the central tenet of the theory.

Contrast the situation of these higher eukaryotes with bacteria and viruses. These are characterized by small genome sizes, an almost universal lack of introns or gene reduplications, optimization of codon usage in bacteria, and in the extreme cases of some viruses like HBV, such extreme economy in coding sequences that most of the genome contains multiple reading frames. This apparent optimization of replication fitness may be facilitated by their large population sizes within a given environment.

Although population size and high mutation rates have generally been seen as factors enhancing the adaptive ability of viruses to cope with new pressures (such as antiviral treatment), in some circumstances the same factors may produce the opposite effect in stable environments. Because of large population size, and the ability of fitter mutants to rapidly replace entire virus populations in the infected individual, viruses capable of establishing persistent infections (such as HCV, HBV and HGV/GBV-C) may become highly optimized for the environment in which they replicate. Mutants with sequence changes that had even a marginal harmful effect on virus fitness, such as a conservative amino-acid substitution (or in the case of HGV/GBV-C, a synonymous substitution that disrupted secondary structure), could be effectively driven out by the large number of competing members of the population pool. Similar substitutions, for example in the haemoglobin gene of an elephant living in a small breeding group would, in contrast, have no significant impact on its reproductive fitness, and would be as likely to become fixed in the elephant population as neutral or even beneficial mutations.

Selection pressures on virus populations during persistent virus infections may be able to drive out much of the variability associated with changes in viral phenotype. The extraordinary conservatism and evolutionary stasis suggested from the reconstruction of the evolutionary histories of hepatitis viruses may paradoxically be the result of those factors that promote virus evolution and diversification in other, less stable environments. Effectively, some viruses may have found their fitness peak for a particular host, and neither transmission bottlenecks nor population drift are necessarily able to drive them from that peak. The high mutation rate of hepatitis viruses may provide them with the means for rapid re-establishment of the original, fitness-optimized population. Large population sizes, the intense selection pressures that operate within them, and high mutation rates that promote convergence to fitness peaks, may be the factors that set virus evolution apart from that of their hosts.

REFERENCES

Adams, N. J., Prescott, L. E., Jarvis, L. M., Lewis, J. C. M., McClure, M. O., Smith, D. B. & Simmonds, P. 1998 Detection of a novel flavivirus related to hepatitis G virus/GB virus C in chimpanzees. *J. Gen. Virol.* **79**, 1871–1877.

Arauz-Ruiz, P., Norder, H., Visona, K. A. & Magnius, L. O. 1997 Genotype F prevails in HBV infected patients of Hispanic origin in Central America and may carry the precore stop mutant. *J. Med. Virol.* **51**, 305–312.

Birkenmeyer, L. G., Desai, S. M., Muerhoff, A. S., Leary, T. P., Simons, J. N., Montes, C. C. & Mushahwar, I. K. 1998 Isolation of a GB virus-related genome from a chimpanzee. *J. Med. Virol.* **56**, 44–51.

Bollyky, P. L. & Holmes, E. C. 1999 Reconstructing the complex evolutionary history of hepatitis B virus. *J. Mol. Evol.* **49**, 130–141.

Bollyky, P. L., Rambaut, A., Grassly, N., Carman, W. F. & Holmes, E. C. 1997 Hepatitis B virus has a New World evolutionary origin. *Hepatology* **26**, 765.

Bukh, J. & Apgar, C. L. 1997 Five new or recently discovered (GBV-A) virus species are indigenous to New World monkeys and may constitute a separate genus of the Flaviviridae. *Virology* **229**, 429–436.

Bukh, J., Purcell, R. H. & Miller, R. H. 1993 At least 12 genotypes of hepatitis C virus predicted by sequence analysis of the putative El gene of isolates collected worldwide. *Proc. Natl Acad. Sci. USA* **90**, 8234–8238.

Chamberlain, R. W., Adams, N. J., Taylor, L. A., Simmonds, P. & Elliott, R. M. 1997 The complete coding sequence of hepatitis C virus genotype 5a, the predominant genotype in South Africa. *Biochem. Biophys. Res. Commun.* **236**, 44–49.

Charrel, R. N., de Micco, P. & de Lamballerie, X. 1999 Phylogenetic analysis of GB viruses A and C: evidence for cospeciation between virus isolates and their primate hosts. *J. Gen. Virol.* **80**, 2329–2335.

Cuceanu, N., Tuplin, A. & Simmonds, P. 2001 Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′ end of the hepatitis G virus/GB virus-C genome. *J. Gen. Virol.* **82**, 713–722.

Erker, J. C., Desai, S. M., Leary, T. P., Chalmers, M. L., Montes, C. C. & Mushahwar, I. K. 1998 Genomic analysis of two GB virus A variants isolated from captive monkeys. *J. Gen. Virol.* **79**, 41–45.

Feng, G., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Greene, B. M., Sharp, P. M., Shaw, G. M. & Hahn, B. H. 1992 Human infection by genetically diverse SIVsm-related HIV-2 in west Africa. *Nature* **358**, 495–499.

Fretz, C., Jeannel, D., Stuyver, L., Herve, V., Lunel, F., Boudifa, A., Mathiot, C., de The, G. & Fournel, J. J. 1995 HCV infection in a rural population of the Central African Republic (CAR): evidence for three additional subtypes of genotype 4. *J. Med. Virol.* **47**, 435–437.

Gao, F. (and 11 others) 1999 Origins of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441.

Gonzalez Perez, M. A., Norder, H., Bergstrom, A., Lopez, E., Visona, K. A. & Magnius, L. O. 1997 High prevalence of GB virus C strains genetically related to strains with Asian origin in Nicaraguan hemophiliacs. *J. Med. Virol.* **52**, 149–155.

Grethe, S., Heckel, J. O., Rietschel, W. & Hufert, F. T. 2000 Molecular epidemiology of hepatitis B virus variants in nonhuman primates. *J. Virol.* **74**, 5377–5381.

Hannoun, C., Horal, P. & Lindh, M. 2000 Long-term mutation rates in the hepatitis B virus genome. *J. Gen. Virol.* **81**, 75–83.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. & Van de Pol, S. 1982 Rapid evolution of RNA genomes. *Science* **215**, 1577–1585.

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P. & Harvey, P. H. 1995 Revealing the history of infectious disease epidemics using phylogenetic trees. *Phil. Trans. R. Soc. Lond.* B **349**, 33–40.

Hu, X., Margolis, H. S., Purcell, R. H., Ebert, J. & Robertson, B. H. 2000 Identification of hepatitis B virus indigenous to chimpanzees. *Proc. Natl Acad. Sci. USA* **97**, 1661–1664.

Jeannel, D. (and 11 others) 1998 Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in west Africa. *J. Med. Virol.* **55**, 92–97.

Jones, S., Martin, R. & Pilbeam, D. 1992 *Human evolution*. Cambridge University Press.

Katayama, Y., Apichartpiyakul, C., Handajani, R., Ishido, S. & Hotta, H. 1997 GB virus C hepatitis G virus (GBV-C/HGV) infection in Chiang Mai, Thailand, and identification of variants on the basis of 5′-untranslated region sequences. *Arch. Virol.* **142**, 2433–2445.

Konomi, N., Miyoshi, C., La Fuente Zerain, C., Li, T. C., Arakawa, Y. & Abe, K. 1999 Epidemiology of hepatitis B, C, E, and G virus infections and molecular analysis of hepatitis G virus isolates in Bolivia. *J. Clin. Microbiol.* **37**, 3291–3295.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. & Bhattacharya, T. 2000 Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796.

Kumar, S. & Hedges, S. B. 1998 A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920.

Kumar, S., Tamura, K. & Nei, M. 1994 *MEGA: Molecular Evolutionary Genetics Analysis* software for computers. *Comput. Appl. Biosci.*, **10**, 189–191.

Lanford, R. E., Chavez, D., Brasky, K. M., Burns III, R. B. & Rico-Hesse, R. 1998 Isolation of a hepadnavirus from the woolly monkey, a New World primate. *Proc. Natl Acad. Sci. USA* **95**, 5757–5761.

Leary, T. P., Muerhoff, A. S., Simons, J. N., Pilot Matias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. G., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. 1996 Sequence and genomic organization of GBV-C: a novel member of the Flaviviridae associated with human non-A–E hepatitis. *J. Med. Virol.* **48**, 60–67.

Leary, T. P., Desai, S. M., Erker, J. C. & Mushahwar, I. K. 1997 The sequence and genomic organization of a GB virus A variant isolated from captive tamarins. *J. Gen. Virol.* **78**, 2307–2313.

Linnen, J. (and 29 others) 1996 Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* **271**, 505–508.

Liu, H. F., Muyembe-Tamfum, J. J., Dahan, K., Desmyter, J. & Goubau, P. 2000 High prevalence of GB virus C/hepatitis G virus in Kinshasa, Democratic Republic of Congo: a phylogenetic analysis. *J. Med. Virol.* **60**, 159–165.

MacDonald, D. M., Holmes, E. C., Lewis, J. C. & Simmonds, P. 2000 Detection of hepatitis B virus infection in wild-born chimpanzees (*Pan troglodytes verus*): phylogenetic relationships with human and other primate genotypes. *J. Virol.* **74**, 4253–4257.

Magnius, L. O. & Norder, H. 1995 Subtypes, genotypes and molecular epidemiology of the hepatitis B virus as reflected by sequence variability of the *S*-gene. *Intervirology* **38**, 24–34.

McGeoch, D. J., Dolan, A. & Ralph, A. C. 2000 Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74**, 10 401–10 406.

Mellor, J., Holmes, E. C., Jarvis, L. M., Yap, P. L., Simmonds, P. & International Collaborators 1995 Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification. *J. Gen. Virol.* **76**, 2493–2507.

Menendez, C. (and 10 others) 1999 Molecular evidence of mother-to-infant transmission of hepatitis G virus among women without known risk factors for parenteral infections. *J. Clin. Microbiol.* **37**, 2333–2336.

Mison, L., Hyland, C., Poidinger, M., Borthwick, I., Faoagali, J., Aeno, U. & Gowans, E. 2000 Hepatitis G virus genotypes in Australia, Papua New Guinea and the Solomon Islands: a possible new Pacific type identified. *J. Gastroenterol. Hepatol.* **15**, 952–956.

Mizokami, M., Orito, E., Ohba, K., Ikeo, K., Lau, J. Y. & Gojobori, T. 1997 Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.* **44**(Suppl. 1), S83–S90.

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J. & Woodruff, D. S. 1994 Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science* **265**, 1193–1201.

Muerhoff, A. S., Simons, J. N., Leary, T. P., Erker, J. C., Chalmers, M. L., Pilot-Matias, T. J., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. 1996 Sequence heterogeneity within the 5′-terminal region of the hepatitis GB virus C genome and evidence for genotypes. *J. Hepatol.* **25**, 379–384.

Muerhoff, A. S., Smith, D. B., Leary, T. P., Erker, J. C., Desai, S. M. & Mushahwar, I. K. 1997 Identification of GB virus C variants by phylogenetic analysis of 5′-untranslated and coding region sequences. *J. Virol.* **71**, 6501–6508.

Mukaide, M. (and 15 others) 1997 Three different GB virus C/hepatitis G virus genotypes—phylogenetic analysis and a genotyping assay based on restriction fragment length polymorphism. *FEBS Lett.* **407**, 51–58.

Nakao, H., Okamoto, H., Fukuda, M., Tsuda, F., Mitsui, T., Masuko, K., Lizuka, H., Miyakawa, Y. & Mayumi, M. 1997 Mutation rate of GB virus C hepatitis G virus over the entire genome and in subgenomic regions. *Virology* **233**, 43–50.

Norder, H., Courouce, A. M. & Magnius, L. O. 1994 Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology* **198**, 489–503.

Norder, H., Ebert, J. W., Fields, H. A., Mushahwar, I. K. & Magnius, L. O. 1996 Complete sequencing of a gibbon hepatitis B virus genome reveals a unique genotype distantly related to the chimpanzee hepatitis B virus. *Virology* **218**, 214–223.

Ruggieri, A., Argentini, C., Kouruma, F., Chionne, P., D'Ugo, E., Spada, E., Dettori, S., Sabbatani, S. & Rapicetta, M. 1996 Heterogeneity of hepatitis C virus genotype 2 variants in west central Africa (Guinea Conakry). *J. Gen. Virol.* **77**, 2073–2076.

Sathar, M. A., Soni, P. N., Pegoraro, R., Simmonds, P., Smith, D. B., Dhillon, A. P. & Dusheiko, G. M. 1999 A new variant of GB virus C/hepatitis G virus (GBV-C/HGV) from South Africa. *Virus Res.* **64**, 151–160.

Simmonds, P. 1998 Variability of the hepatitis C virus genome. In *Hepatitis C virus*, 2nd edn (ed. H. W. Reesink), pp. 38–63. Basel, Switzerland: Karger.

Simmonds, P. & Smith, D. B. 1999 Structural constraints on RNA virus evolution. *J. Virol.* **73**, 5787–5794.

Simmonds, P. (and 45 others) 1994 A proposed system for the nomenclature of hepatitis C viral genotypes. *Hepatology* **19**, 1321–1324.

Smith, D. B., Cuceanu, N., Davidson, F., Jarvis, L. M., Mokili, J. L. K., Hamid, S., Ludlam, C. A. & Simmonds, P. 1997*a* Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5′ non-coding region. *J. Gen. Virol.* **78**, 1533–1542.

Smith, D. B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P. L. & Simmonds, P. 1997*b* The origin of hepatitis C virus genotypes. *J. Gen. Virol.* **78**, 321–328.

Smith, D. B., Basaras, M., Frost, S., Haydon, D., Cuceanu, N., Prescott, L., Kamenka, C., Millband, D., Sathar, M. A. & Simmonds, P. 2000 Phylogenetic analysis of GBV-C/hepatitis G virus. *J. Gen. Virol.* **81**, 769–780.

Stuyver, L., Rossau, R., Wyseur, A., Duhamel, M., Van der Borght, B., Van Heuverswyn, H. & Maertens, G. 1993 Typing of hepatitis C virus isolates and characterization of new subtypes using a line probe assay. *J. Gen. Virol.* **74**, 1093–1102.

Suzuki, Y. & Gojobori, T. 1997 The origin and evolution of Ebola and Marburg viruses. *Mol. Biol. Evol.* **14**, 800–806.

Takahashi, K., Brotman, B., Usuda, S., Mishiro, S. & Prince, A. M. 2000 Full-genome sequence analyses of hepatitis B virus (HBV) strains recovered from chimpanzees infected in the wild: implications for an origin of HBV. *Virology* **267**, 58–64.

Tanaka, Y. (and 13 others) 1998*a* African origin of GB virus C hepatitis G virus. *FEBS Lett.* **423**, 143–148.

Tanaka, Y. (and 12 others) 1998*b* GB virus C/hepatitis G virus infection among Colombian native Indians. *Am. J. Trop. Med. Hyg.* **59**, 462–467.

Tucker, T. J., Smuts, H., Eickhaus, P., Robson, S. C. & Kirsch, R. E. 1999 Molecular characterization of the 5′ non-coding region of South African GBV-C/HGV isolates: major deletion and evidence for a fourth genotype. *J. Med. Virol.* **59**, 52–59.

Vaudin, M., Wolstenholme, A. J., Tsiquaye, K. N., Zuckerman, A. J. & Harrison, T. J. 1988 The complete nucleotide sequence of the genome of a hepatitis B virus isolated from a naturally infected chimpanzee. *J. Gen. Virol.* **69**, 1383–1389.

Wansbrough Jones, M. H., Frimpong, E., Cant, B., Harris, K., Evans, M. R. W. & Teo, C. G. 1998 Prevalence and genotype of hepatitis C virus infection in pregnant women and blood donors in Ghana. *Trans. R. Soc. Trop. Med. Hyg.* **92**, 496–499.

Warren, K. S., Heeney, J. L., Swan, R. A., Heriyanto & Verschoor, E. J. 1999 A new group of hepadnaviruses naturally infecting orangutans (*Pongo pygmaeus*). *J. Virol.* **73**, 7860–7865.

Xu, L. Z., Larzul, D., Delaporte, E., Brechot, C. & Kremsdorf, D. 1994 Hepatitis C virus genotype 4 is highly prevalent in central Africa (Gabon). *J. Gen. Virol.* **75**, 2393–2398.

Zanotto, P. M. D., Gould, E. A., Gao, G. F., Harvey, P. H. & Holmes, E. C. 1996 Population dynamics of flaviviruses revealed by molecular phylogenies. *Proc. Natl Acad. Sci. USA* **93**, 548–553.

Zhang, G., Haydon, D. T., Knowles, N. J. & McCauley, J. W. 1999 Molecular evolution of swine vesicular disease virus. *J. Gen. Virol.* **80**, 639–651.

Zhu, T. F., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M. & Ho, D. D. 1998 An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597.

Zuckerman, A. J., Thornton, A., Howard, C. R., Tsiquaye, K. N., Jones, D. M. & Brambell, M. R. 1978 Hepatitis B outbreak among chimpanzees at the London Zoo. *The Lancet* **2**, 652–654.