Short Communication

A new subtype of hepatitis C virus genotype 1: complete genome and phylogenetic relationships of an Equatorial Guinea isolate

Maria Alma Bracho,[1]† Francy Yolima Carrillo-Cruz,[1]† Enrique Ortega,[2] Andrés Moya[1] and Fernando González-Candelas[1]

Correspondence
Fernando González-Candelas
fernando.gonzalez@uv.es

[1]Institut 'Cavanilles' de Biodiversitat i Biologia Evolutiva, Universitat de València, 46070 València, Spain

[2]Hospital General Universitari, Avinguda Tres Creus s/n, 46014 València, Spain

Received 7 November 2005
Accepted 7 February 2006

Hepatitis C virus (HCV) is the leading cause of chronic liver disease and is associated with hepatocellular carcinoma. However, there have been few studies on the distribution and genetic diversity of HCV isolates in non-developed countries. Here, the complete genome sequence of an HCV genotype 1 isolate from Equatorial Guinea is reported, the first complete HCV-1 genome of African origin. Phylogenetic analysis revealed that this sequence always grouped with sequences of genotype 1, but did not group clearly with any subtype described so far. An analysis of partial NS5B gene sequences with additional sequences of African origin also failed to find close similarities between the new sequence and any previously known isolate. Genetic divergence of the coding region of this new sequence with respect to the recognized subtypes of HCV-1 ranged from 20 to 22 %. It is proposed that this isolate is a representative of a new, distinct variant of HCV subtype 1.

Hepatitis caused by *Hepatitis C virus* (HCV) is a serious disease affecting more than 170 million people worldwide, approximately 3 % of the human population (WHO, 1999). Hepatitis C prevalence is far from being distributed evenly in human populations and varies widely depending on the risk population and the world region being studied (Memon & Memon, 2002). Its genome consists of a positive-sense RNA molecule of approximately 9·6 kb, which encodes a polyprotein of about 3000 aa. The nucleotide sequence of the HCV genome is highly variable and, based on genetic distances and phylogenetic analyses, all isolates are currently grouped into six genotypes [formerly denoted as clades by Robertson *et al.* (1998)], corresponding to the main branches in the phylogenetic tree. The same kind of analysis applied to isolates within each genotype has revealed several clusters of closely related sequences, denoted as subtypes (Simmonds *et al.*, 2005).

Simple estimates of genetic distances for the core, E1 and/or NS5B regions have been proposed to distinguish between types, subtypes and isolates (Simmonds *et al.*, 1993; Stuyver *et al.*, 1994). However, as more branches have been added to the HCV phylogenetic tree, type and/or subtype assignations often have become ambiguous. Moreover, intense sequence characterization of new partial HCV genomes has shown many instances in which genetic distance estimates were in between those corresponding to genotype- and subtype-delimiting ranges (Robertson *et al.*, 1998; Simmonds *et al.*, 2005). Hence, some confusion was generated when this distance-based criterion was used for classification and nomenclature of novel putative genotypes or subtypes (Jeannel *et al.*, 1998; Stuyver *et al.*, 1994; Tokita *et al.*, 1994a, b, 1995, 1996). In fact, recent proposals (Robertson *et al.*, 1998; Simmonds *et al.*, 2005) to establish a more comprehensible classification of HCV recommended that sequence-based classification and/or nomenclature should be based on extensive phylogenetic analysis rather than genetic distance, and preferably should be based on the complete sequence of the coding region for new genotypes. However, in order to assign a given sequence to a particular HCV subtype, Simmonds *et al.* (2005) recommended consistent phylogenetic grouping based on at least partial genomes, including both the core/E1 region and the NS5B region (nt 869–1296 and 8276–8615, respectively, in the H77 sequence reference; GenBank accession number AF009606), for which a considerable amount of sequence data are available. Nonetheless, due to the wide epidemiological use of subtype assignations, three or more examples of independently infected individuals with a new proposed subtype will be required for a definitive, subtype-specific designation (Simmonds *et al.*, 2005).

†These authors contributed equally to this work.

The GenBank/EMBL/DDBJ accession number for the complete HCV genome sequence described in this study is AJ851228.

Supplementary material is available in JGV Online.

The most widespread variants of genotype 1 isolated in Western countries can be classified clearly as subtype 1a or 1b. However, further molecular epidemiological research in some West and Central African countries has demonstrated that much greater diversity exists not only within this particular genotype, but also within genotypes 2 and 4 (Candotti *et al.*, 2003; Ndjomou *et al.*, 2003; Njouom *et al.*, 2003; Pasquier *et al.*, 2005; Simmonds *et al.*, 2005). Previous phylogenetic studies have shown that genotypes 1 and 4 form a statistically well-supported cluster (Salemi & Vandamme, 2002) in contrast to the entirely separate and independent branching at the base of the tree of the remaining genotypes. Moreover, phylogenetic analysis of partial sequences of highly variable HCV genomes isolated in West and Central Africa suggest that these two genotypes share a common origin in the region currently known as Cameroon and would have spread subsequently to other regions of Africa and the rest of the world (Ndjomou *et al.*, 2003). Almost certainly, more confident conclusions could be reached if complete genomes rather than partial genomes were used in phylogenetic (Salemi & Vandamme, 2002) and genome recombination studies (Colina *et al.*, 2004; Kalinina *et al.*, 2002), but complete genomes from African isolates are especially scarce: only two complete genomes belonging to genotype 4 (Bukh *et al.*, 1998) and genotype 5a (Chamberlain *et al.*, 1997) have been determined so far.

In the present study, we report the complete genome of an HCV isolate from a patient of African origin. Based on both genetic distance and phylogenetic analyses, this isolate could well be a representative of a new subtype within genotype 1. However, given the recent recommendations for a unified system of nomenclature (Simmonds *et al.*, 2005), no specific subtype designation can be applied to this case as yet.

A serum sample from a native Equatorial Guinean patient was obtained as part of an HCV molecular epidemiology study in the Comunitat Valenciana (Spain). The complete HCV genome (9481 nt; GenBank accession number AJ851228) from this patient's serum was obtained from six overlapping RT-PCR fragments, which subsequently were sequenced directly (see details in Supplementary Material, available in JGV Online). The initial amplification design consisted of only three overlapping fragments of ~3000 nt in the first-round PCR, followed by a hemi-nested second-round PCR, but this was achieved for only two of the three fragments. Although this design did not work successfully for one fragment, it could be a suitable approach for sequencing complete genomes, as the modification of conventional protocols required would not be extensive (Lu *et al.*, 2005; Rispeter *et al.*, 1997). Such minor modifications would probably involve a longer extension step during PCR and generously degenerate primers for a highly variable, first-read sequence, in order to ensure primer–cDNA hybridization.

Phylogenetic analysis of this new complete genome along with representatives of the six HCV genotypes described previously was restricted to the coding portion of the genome. Maximum-likelihood phylogenetic reconstructions (Fig. 1) revealed that the new sequence grouped clearly with HCV genotype 1 sequences, but grouped separately from any of the subtypes (1a, 1b and 1c) described for this genotype. According to this reconstruction, genotype 1 would consist of two sister clusters, one with subtypes 1a and 1c, with subtype 1b and the new sequence constituting a different cluster. Therefore, subtype 1b would be the closest to this

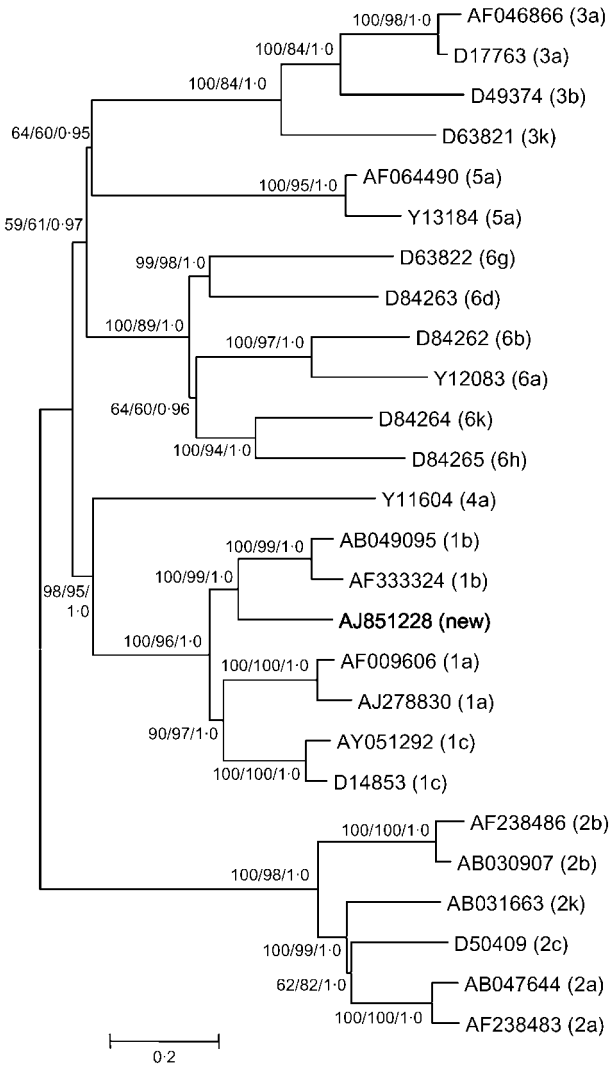

Fig. 1. Maximum-likelihood phylogenetic tree of complete genome sequences. The tree was obtained with PHYML using GTR+G+I for the new sequence and 25 complete genomes representative of all six HCV genotypes. Genotype and subtype labels, according to Simmonds *et al.* (2005), are indicated in parentheses next to the GenBank accession numbers. Support values of nodes were estimated by bootstrap (2000 replicates using neighbour joining with the maximum-likelihood distance), quartet puzzling (same evolution model, 10^4 puzzling steps) and Bayesian analysis (MrBayes 3.1.1, same model, 10^6 generations, 10% burn-in, 1000 sampled trees), respectively. Bar, 0·2 substitutions per nucleotide position.

new sequence (Fig. 1). Moreover, all of the nodes within genotype 1, including the one constituted by subtype 1b and the new sequence, were well-supported by three different measurements of node support (Fig. 1).

Genetic distances in the coding region showed that this sequence displayed 20–22 % nucleotide differences compared with representatives of the three confirmed subtypes of genotype 1. This range is similar to that separating these three subtypes (20–23 %) (data not shown). Moreover, when distances based on the maximum-likelihood nucleotide-substitution model were considered, the estimates of evolutionary distance between the new sequence and additional representatives of HCV subtypes 1a, 1b and 1c fell within approximately the same range as that separating these subtypes (Table 1). In agreement with the phylogenetic analysis, the shortest distances were found between the new sequence and the group of sequences belonging to subtype 1b.

In order to discard potential recombination events between different HCV subtype 1 genomes, phylogenetic reconstructions with the same reference sequences were also obtained separately for each gene (data not shown). With the exception of the NS4A gene, all of the topologies obtained were congruent with the one derived from the complete genome (Fig. 1), with subtype 1b sequences always appearing closest to the new sequence and subtypes 1a and 1c grouping together in a different cluster. In the discordant topology obtained with the NS4A gene, the new sequence occupied a basal position with respect to the other type 1 subtypes, although with low bootstrap support. This could be reflecting a low phylogenetic signal as a consequence of the short length of this gene (162 nt). Therefore, this new subtype 1 sequence did not correspond to any variant coming from a rare recombination event between genotypes (Colina *et al.*, 2004; Kalinina *et al.*, 2002) or genotype 1 subtypes.

A partial fragment (267 nt) of the NS5B gene was subjected to further phylogenetic analysis in which 34 additional sequences of genotype 1 were included, mostly from isolates of different African countries (Candotti *et al.*, 2003; Jeannel *et al.*, 1998; Ndjomou *et al.*, 2003) (Fig. 2). In general, sequences from the same country tended to group in well-supported clusters. This could reflect relatively recent epidemiological events, as indicated by the short genetic distances connecting these isolates. In the case of Cameroon sequences, many distant clusters were found to be dispersed along the tree, which is probably related to extensive sampling in an area with extremely high heterogeneity among HCV-1 isolates (Ndjomou *et al.*, 2003). The new sequence did not resemble closely any partial or complete sequence described previously, nor did it present any significant node support that could relate it to any group of sequences within the genotype 1 cluster. However, a poorly supported node connected the new sequence to a group of Cameroon sequences. It is worth noting that Equatorial Guinea is a small country located to the south of Cameroon. The new sequence was related more closely to subtype 1b than to subtypes 1a or 1c, coinciding with the results from the complete-genome analysis (Fig. 1). However, this relationship was less evident in the NS5B gene analysis, due to the almost-continuous branching of isolates along the phylogenetic tree.

RNA viruses are characterized by an extremely high genetic variability, even in samples from a single individual (Moya *et al.*, 2004; Simmonds, 2004). However, this variability apparently co-exists with clear discontinuities between genotypes, subtypes and even isolates. In HCV, these discontinuities allow virus populations to be classified into six genotypes and many different subtypes. To date, only three confirmed subtypes (1a–1c) and nine provisionally assigned subtypes (1d–1l) have been described for HCV genotype 1 (Simmonds *et al.*, 2005). As for other genotypes, confirmed subtypes of genotype 1 were established early on, both on the basis of genetic distances and on phylogenetic analysis of full-length sequences. Studies based on age-stratified seroprevalence data (Alter *et al.*, 1999) and phylogenetic analyses (Pybus *et al.*, 2001) indicate that HCV has recently spread rapidly in industrialized nations where subtypes 1a and 1b are considered prevalent. Clinical studies on treatment response or disease progression of such infections

Table 1. Mean genetic distances (lower-left matrix) between the new sequence and representative sequences from subtypes 1a, 1b and 1c and between these subtypes

The coding region of complete genomes was used for distance estimates. Standard errors are indicated in the upper-right matrix. The distance model was GTR+I+G with an assumed proportion of invariable sites of 0·43 and a shape parameter (α) of 1·16 for the Γ distribution of substitution rates at variable sites. The number of sequences included in each subtype group for the analysis is indicated in parentheses next to the subtype name.

	New sequence	Subtype 1a	Subtype 1b	Subtype 1c
New sequence		0·010	0·014	0·006
Subtype 1a (n=9)	0·656		0·014	0·016
Subtype 1b (n=10)	0·499	0·713		0·018
Subtype 1c (n=2)	0·688	0·588	0·742	

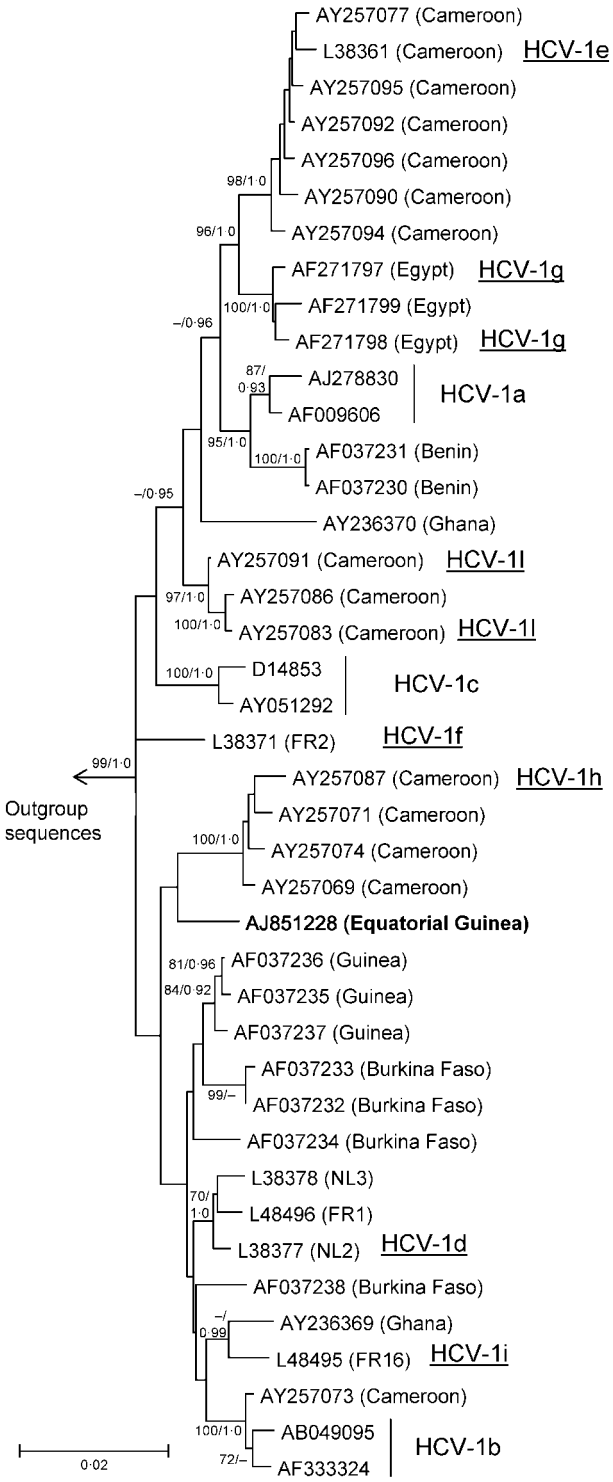

Fig. 2. Maximum-likelihood phylogenetic tree of partial NS5B gene sequences. The tree was obtained with PHYML using GTR+G+I for the new sequence, 25 genomes representative of all six HCV genotypes and 34 additional homologous sequences of genotype 1 cited by Jeannel *et al.* (1998), Candotti *et al.* (2003) and Ndjomou *et al.* (2003), sampled in African countries or of putative African origin. Only the genotype 1 clade is shown, with sequences from different subtypes marked (provisional subtypes are underlined). Nodes with bootstrap support >70 % (1000 replicates with PHYML) or Bayesian posterior probability >0·90 (MrBayes 3.1.1, same model, 10^6 generations, 10 % burn-in, 1000 sampled trees), respectively, are indicated. Bar, 0·02 substitutions per nucleotide position.

account for most of the genome information deposited in sequence databases, mainly as partial genome sequences. Like other highly prevalent subtypes, subtypes 1a and 1b have possibly been distributed widely through unsafe practices between intravenous drug users and through medical practices (i.e. blood transfusion, surgery, intravenous medication and haemodialysis). Moreover, transmission patterns and dated infection rates differ among countries and between viral subtypes (Nakano *et al.*, 2004).

In this study, we have revealed the existence of a new and distinct variant of HCV from Equatorial Guinea by sequencing and analysing its complete genome. This genome is as divergent from representative variants of confirmed genotype 1 subtypes as these are from each other. In consequence, we have provided evidence that this could be a truly new subtype. However, the consensus proposal for a unified system of naming HCV genotypes (Simmonds *et al.*, 2005) states that new subtype designation should be provided only in those cases where there is evidence of their spreading in identified transmission networks. According to this consensus, such evidence will qualify when the new subtype is found in 'several independently infected individuals'. In such cases, if distinctness is proved, then a provisional designation of a new subtype could be established through partial sequences (core/E1 and NS5B regions) or confirmed designation if at least one HCV isolate from the provisionally designated subtype is sequenced completely. However, as shown by the NS5B phylogenetic analysis of highly diverse isolates, many new sequences from this African area could potentially be designated provisional subtypes without being epidemiologically relevant. Therefore, we propose the generic name 'subtype 1' for the new variant reported here and, in order to avoid an increasing profusion of labels for provisional subtypes, we also suggest the same procedure for provisionally assigned subtypes (1d–1l), only considering them to be newly confirmed subtypes if they become epidemiologically or clinically relevant (Ndjomou *et al.*, 2003; Simmonds *et al.*, 2005).

Our current view of HCV variants distributed in separate subtypes within some of the major genotypes is probably the result of the recent epidemiological history of this virus. The most common subtypes come about through the extremely successful spread of a few variants that enter a 'fast track', spreading into a susceptible population whose new behavioural patterns help the virus to spread further (Simmonds, 2004). A clear example of this dynamic is the current spread of HCV subtype 1a in Western countries, where its prevalence is rising steadily (Pybus *et al.*, 2005; Thomson & Finch, 2005) thanks to having entered the

transmission route of intravenous drug users at the same time as a significant proportion of the general population has begun to engage in this risky behaviour.

If the genetic differences revealed between the new sequence reported here and the previously characterized subtypes are assumed to be of the same order in the whole genome as in the portion of the NS5B gene that we have analysed, including African samples from three different studies (Candotti *et al.*, 2003; Jeannel *et al.*, 1998; Ndjomou *et al.*, 2003), then it is likely that the current clear-cut separations between subtypes 1a, 1b and 1c will become diluted. This will come about because many intermediate forms will appear and a new range of subtypes will be defined following the new nomenclature rules (Simmonds *et al.*, 2005). The relevant question then will be: is such a division of any use? We believe that recognizing the intrinsically high variability of the genome sequence of RNA viruses should not necessarily be reflected in their endless subdivision into discrete categories, especially if these do not correspond to clear evolutionary splits or are not particularly relevant from a clinical or epidemiological point of view. In addition, conflicting reports on response, course, likelihood of becoming a chronic infection, etc. (Simmonds, 2004) for different subtypes/genotypes may be reconciled more easily if a more accurate reconstruction of the evolutionary history of HCV is obtained and such observations are freed from a biologically untenable view and are interpreted in an appropriate context.

Acknowledgements

This work was supported by Conselleria de Sanitat i Consum, Generalitat Valenciana, and projects BMC2001-3096 from MCYT, BFU2005-00503/BMC from MEC, and Grupos 03/204 from Agència Valenciana de Ciència i Tecnologia.

References

Alter, M. J., Kruszon-Moran, D., Nainan, O. V., McQuillan, G. M., Gao, F., Moyer, L. A., Kaslow, R. A. & Margolis, H. S. (1999). The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. *N Engl J Med* **341**, 556–562.

Bukh, J., Apgar, C. L., Engle, R., Govindarajan, S., Hegerich, P. A., Tellier, R., Wong, D. C., Elkins, R. & Kew, M. C. (1998). Experimental infection of chimpanzees with hepatitis C virus of genotype 5a: genetic analysis of the virus and generation of a standardized challenge pool. *J Infect Dis* **178**, 1193–1197.

Candotti, D., Temple, J., Sarkodie, F. & Allain, J.-P. (2003). Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. *J Virol* **77**, 7914–7923.

Chamberlain, R. W., Adams, N., Saeed, A. A., Simmonds, P. & Elliott, R. M. (1997). Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. *J Gen Virol* **78**, 1341–1347.

Colina, R., Casane, D., Vasquez, S., García-Aguirre, L., Chunga, A., Romero, H., Khan, B. & Cristina, J. (2004). Evidence of intratypic recombination in natural populations of hepatitis C virus. *J Gen Virol* **85**, 31–37.

Jeannel, D., Fretz, C., Traore, Y. & 9 other authors (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.

Kalinina, O., Norder, H., Mukomolov, S. & Magnius, L. O. (2002). A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *J Virol* **76**, 4034–4043.

Lu, L., Nakano, T., Smallwood, G. A., Heffron, T. G., Robertson, B. H. & Hagedorn, C. H. (2005). A refined long RT-PCR technique to amplify complete viral RNA genome sequences from clinical samples: application to a novel hepatitis C virus variant of genotype 6. *J Virol Methods* **126**, 139–148.

Memon, M. I. & Memon, M. A. (2002). Hepatitis C: an epidemiological review. *J Viral Hepat* **9**, 84–100.

Moya, A., Holmes, E. C. & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nat Rev Microbiol* **2**, 279–288.

Nakano, T., Lu, L., Liu, P. & Pybus, O. G. (2004). Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *J Infect Dis* **190**, 1098–1108.

Ndjomou, J., Pybus, O. G. & Matz, B. (2003). Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *J Gen Virol* **84**, 2333–2341.

Njouom, R., Pasquier, C., Ayouba, A. & 7 other authors (2003). Hepatitis C virus infection among pregnant women in Yaounde, Cameroon: prevalence, viremia, and genotypes. *J Med Virol* **69**, 384–390.

Pasquier, C., Njouom, R., Ayouba, A. & 7 other authors (2005). Distribution and heterogeneity of hepatitis C genotypes in hepatitis patients in Cameroon. *J Med Virol* **77**, 390–398.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. & Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323–2325.

Pybus, O. G., Cochrane, A., Holmes, E. C. & Simmonds, P. (2005). The hepatitis C virus epidemic among injecting drug users. *Infect Genet Evol* **5**, 131–139.

Rispeter, K., Lu, M., Lechner, S., Zibert, A. & Roggendorf, M. (1997). Cloning and characterization of a complete open reading frame of the hepatitis C virus genome in only two cDNA fragments. *J Gen Virol* **78**, 2751–2759.

Robertson, B., Myers, G., Howard, C. & 14 other authors (1998). Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. *Arch Virol* **143**, 2493–2503.

Salemi, M. & Vandamme, A.-M. (2002). Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. *J Mol Evol* **54**, 62–70.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus – 15 years on. *J Gen Virol* **85**, 3173–3188.

Simmonds, P., Holmes, E. C., Cha, T.-A. & 7 other authors (1993). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *J Gen Virol* **74**, 2391–2399.

Simmonds, P., Bukh, J., Combet, C. & 19 other authors (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology* **42**, 962–973.

Stuyver, L., van Arnhem, W., Wyseur, A., Hernandez, F., Delaporte, E. & Maertens, G. (1994). Classification of hepatitis C viruses based on phylogenetic analysis of the envelope 1 and nonstructural 5B regions and identification of five additional subtypes. *Proc Natl Acad Sci U S A* **91**, 10134–10138.

Thomson, B. J. & Finch, R. G. (2005). Hepatitis C virus infection. *Clin Microbiol Infect* **11**, 86–94.

Tokita, H., Okamoto, H., Tsuda, F. & 7 other authors (1994a). Hepatitis C virus variants from Vietnam are classifiable into the seventh, eighth, and ninth major genetic groups. *Proc Natl Acad Sci U S A* **91**, 11022–11026.

Tokita, H., Shrestha, S. M., Okamoto, H., Sakamoto, M., Horikita, M., Iizuka, H., Shrestha, S., Miyakawa, Y. & Mayumi, M. (1994b). Hepatitis C virus variants from Nepal with novel genotypes and their classification into the third major group. *J Gen Virol* **75**, 931–936.

Tokita, H., Okamoto, H., Luengrojanakul, P., Vareesangthip, K., Chainuvati, T., Iizuka, H., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1995). Hepatitis C virus variants from Thailand classifiable into five novel genotypes in the sixth (6b), seventh (7c, 7d) and ninth (9b, 9c) major genetic groups. *J Gen Virol* **76**, 2329–2335.

Tokita, H., Okamoto, H., Iizuka, H., Kishimoto, J., Tsuda, F., Lesmana, L. A., Miyakawa, Y. & Mayumi, M. (1996). Hepatitis C virus variants from Jakarta, Indonesia classifiable into novel genotypes in the second (2e and 2f), tenth (10a) and eleventh (11a) genetic groups. *J Gen Virol* **77**, 293–301.

WHO (1999). Global surveillance and control of hepatitis C. *J Viral Hepat* **6**, 35–47.