Journal of General Virology (2009), **90**, 2086–2096 **DOI** 10.1099/vir.0.011569-0

Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa

Peter V. Markov,[1] Jacques Pepin,[2] Eric Frost,[2] Sylvie Deslandes,[2] Annie-Claude Labbé[3] and Oliver G. Pybus[1]

Correspondence
Oliver G. Pybus
oliver.pybus@zoo.ox.ac.uk

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[2]Department of Microbiology and Infectious Diseases, University of Sherbrooke, Sherbrooke, Canada

[3]University of Montreal, Montreal, Canada

Understanding the origin and nature of hepatitis C virus (HCV) genetic diversity is critical for improving treatment and vaccine design, and such diversity is the sole source of information about the virus' epidemic history prior to its identification 20 years ago. In this paper, we study the molecular epidemiology of HCV genotype 2 in its region of endemic origin, west and central Africa. Our analysis includes 56 new and highly diverse HCV isolates sampled from infected individuals in Guinea-Bissau. By combining phylogenetic, geographical and epidemiological information, we find a previously unappreciated geographical structure in the diversity of HCV genotype 2, pointing to a history of eastwards spatial spread from the west African coast to Cameroon that took place over several centuries. Molecular clock analysis dates the common ancestor of HCV in Guinea-Bissau to 1470 (1414–1582). The phylogenetic position of isolates from Madagascar and Martinique suggests a role for the historical slave trade in the global dissemination of HCV and of the epidemic subtypes 2a and 2c. Coalescent-based estimates of epidemic growth indicate a rapid 20th-century spread of HCV genotype 2 in Cameroon that is absent in Guinea-Bissau. We discuss this contrast in the context of possible parenteral HCV exposure during public-health campaigns undertaken during the colonial era.

Received 9 March 2009
Accepted 25 May 2009

INTRODUCTION

Hepatitis C virus (HCV) is an important human pathogen, with 170 million chronic infections globally and 2–4 million incident cases each year (World Health Organization, 1999; Perz *et al.*, 2004). Due to its late manifestations – liver cirrhosis and hepatocellular carcinoma (Seeff, 2000) – HCV is responsible for substantial morbidity and mortality, claiming the lives of approximately 9000 people each year in the USA alone (CDC, 1998).

HCV evolves very rapidly, resulting in formidable genetic diversity that is classified into six genotypes (1–6), each further divided into subtypes, e.g. 1a, 1b, 1c etc. (Simmonds *et al.*, 1993, 2005). Some subtypes are found only in particular regions, whilst others are distributed globally. Geographically restricted subtypes that exhibit high local genetic variation, long-term local persistence and low transmission rates are termed 'endemic' and are typically found in the tropics (Pybus *et al.*, 2007). For example, endemic HCV genotype 3 is distributed throughout south Asia (Mellor *et al.*, 1995), whilst genotype 6 is found in south-east Asia (Pybus *et al.*, 2009), genotypes 1 and 2 in west Africa (Ruggieri *et al.*, 1996; Jeannel *et al.*, 1998; Wansbrough-Jones *et al.*, 1998; Candotti *et al.*, 2003) and genotype 4 in central Africa and the Middle East (Mellor *et al.*, 1995; Ndjomou *et al.*, 2003). In contrast, 'epidemic' HCV subtypes (i.e. 1a, 1b, 2a, 2b, 2c and 3a) show lower diversity at any one location, a reflection of their quick spread during the 20th century via effective transmission routes such as blood transfusions, injecting drug use and invasive medical procedures (Smith *et al.*, 1997; Pybus *et al.*, 2001, 2005). An intermediate epidemiological pattern occurs where particular local strains have been amplified after the 1900s, usually by large-scale health interventions generating high local prevalence without substantial dissemination elsewhere, resulting in so-called 'local epidemic' strains, most notably subtype 4a in Egypt (Ray *et al.*, 2000).

Here, we focus on HCV in west Africa, where highly divergent genotype 2 strains are found (Jeannel *et al.*, 1998;

The GenBank/EMBL/DDBJ accession numbers for the sequences newly generated in this study are GQ153856–GQ153911. Details are available with the online version of this paper.

Details of all other GenBank accession numbers used and a supplementary figure showing molecular clock phylogenies are also available with the online version of this paper.

Candotti *et al.*, 2003). Genotype 2 has been reported from all countries along the African Atlantic coast, from Senegal to Cameroon, and evolutionary analysis suggests that the strain is several centuries old (Simmonds, 2001). It is almost the only HCV genotype seen in Guinea-Bissau (Plamondon *et al.*, 2007), is predominant in Ghana (Candotti *et al.*, 2003) and is found, in various proportions together with genotype 1, in Guinea, Burkina Faso and Benin (Jeannel *et al.*, 1998). In Cameroon, genotype 2 is a minor genotype in comparison with genotypes 1 and 4 (Nkengasong *et al.*, 1995; Ndjomou *et al.*, 2002; Simmonds, 2004; Pasquier *et al.*, 2005). Genotype 2 genetic diversity declines towards Cameroon, such that Cameroonian genotype 2 strains form a single clade within the more diverse group of west African sequences, suggesting likely migration of one particular genotype 2 lineage from west to central Africa (Ndjomou *et al.*, 2003; Pasquier *et al.*, 2005). In contrast, genotypes 1 and 4 probably originated in central Africa and spread westwards and north-eastwards, respectively (Wansbrough-Jones *et al.*, 1998; Ndjomou *et al.*, 2003; Pasquier *et al.*, 2005).

Investigation of the spatial distribution of HCV diversity enables us to understand the past geographical spread of the infection, shedding light on the tangle of demographic, social and biological factors that have given rise to current patterns of diversity, and elucidating previously unrecognized routes of ongoing transmission. Evolutionary, phylogenetic and coalescent-based analyses of pathogen genomes are now established tools in molecular epidemiology, and especially relevant for recent epidemics with limited historical surveillance data, such as human immunodeficiency virus and HCV (e.g. Pybus *et al.*, 2001; Worobey *et al.*, 2008). In the context of HCV in Africa, such methods have been used to reconstruct the epidemic history of subtype 4a in Egypt (Pybus *et al.*, 2003) and the local spread of genotypes 1, 2 and 4 in Cameroon (Njouom *et al.*, 2007; Pouillot *et al.*, 2008). In both instances, coalescent analyses provided evidence of a historical increase in HCV spread, correlated temporally with a proposed source of parenteral transmission – mass-treatment campaigns against schistosomiasis in Egypt and against yaws and syphilis in Cameroon (Pépin & Labbé, 2008). Crucially, for HCV in Egypt it was possible to validate the genetic analysis against compelling epidemiological evidence (Frank *et al.*, 2000). In addition to its epidemiological importance, HCV genetic diversity is of clinical relevance, as HCV genotypes vary in their susceptibility to treatment with ribavirin plus interferon (Yoshioka *et al.*, 1992; Chemello *et al.*, 1994; Zein, 2000) and in the effectiveness of the immune response that they elicit (Kimura *et al.*, 2000). Furthermore, the enormous genetic diversity of HCV is a major obstacle to designing universally and sustainably efficacious vaccines.

Here, we report a comprehensive phylogenetic and evolutionary study of HCV genotype 2 in Africa, which includes the first isolation and genetic characterization of HCV infections from Guinea-Bissau. We employed phylogeographical methods that reconstruct the synchronous geographical dispersal and genetic diversification of the virus. We find that our new isolates from Guinea-Bissau are unexpectedly highly diverse and, in addition, we detect a previously unrecognized spatial structuring of HCV diversity, indicating an eastwards expansion of the virus over several centuries in Africa. We also find evidence that the slave trade played a role in the past global dissemination of HCV genotype 2. Our results show that the past epidemic behaviour of the virus in Guinea-Bissau differs substantially from that reported previously for Cameroon (Njouom *et al.*, 2007) and we discuss our results in the context of possible risk factors, such as mass-treatment campaigns.

METHODS

Sample collection and sequencing. The study received ethical approval from the Guinea-Bissau Ministry of Health and the institutional review board of Centre Hospitalier Universitaire de Sherbrooke, Canada. Participants aged $\geqslant$50 years who gave consent were recruited from January to March 2005 in Bissau City, Guinea-Bissau. Capillary blood was deposited onto filter papers. Samples were screened for HCV antibodies by using Detect-HCV v. 3 (Adaltis). Non-reactive samples were considered HCV-seronegative. Reactive samples were further tested with Ortho HCV 3.0 ELISA Test (Ortho-Clinical) and Monolisa Anti-HCV Plus v. 2 (Bio-Rad). Samples reactive with all three ELISAs were considered HCV-positive, while discordant results samples were further tested by INNO-LIA HCV Score (Innogenetics). See Plamondon *et al.* (2007) for further details of the study population.

RT-PCR amplification of the NS5B region from HCV-seropositive samples was attempted. Reverse transcription and an external amplification cycle were performed with 5 µl RNA extract and primers EF101F (5′-TTCTCGTATGATACCCGCTGYTTTGA) and HCVNS5Rnb (5′-TACCTGGTCATAGCCTCCGTGAAGGCTC). An aliquot (1 µl) of the RT-PCR product was amplified by nested PCR under the same amplification conditions using primers HCVNS5F2p (5′-TATGATACCCGCTGCTTTGACTC,G/I,AC), HCVNS5R2c (5′-CTGGTCATAGCCTCCGTGAAGGCTCTCAGG) and HCVNS5R2d (5′-CTGGTCATAGCCTCCGTGAAGGCTCGTAGG). Amplicons were sequenced by using the HCVNS5F2p primer. Chromatograms were inspected visually and edited manually by using 4Peaks (http://mekentosj.com). Each sequence was analysed by HCV-BLAST (Kuiken *et al.*, 2005) to determine the closest matching HCV reference sequences. Our new isolates were added to GenBank under accession numbers GQ153856–GQ153911 (see Supplementary Table S2, available in JGV Online).

Phylogenetic and phylogeographical analysis. In total, we obtained 56 new sequences from infected individuals in Bissau, spanning 290 nt of the NS5B gene [HCV H77 location, 8284–8604 (Kuiken *et al.*, 2005)]. All available HCV genotype 2 reference sequences spanning the same region were obtained from the HCV Sequence Database (Kuiken *et al.*, 2005). In total, 190 genotype 2 sequences, including the Guinea-Bissau isolates, reference sequences from parts of continental Africa, Madagascar, the Caribbean island of Martinique and three representative isolates from each of the global epidemic subtypes 2a, 2b and 2c, were aligned manually by using Se-Al (http://tree.bio.ed.ac.uk). To investigate viral diversity and ancestral relationships, we estimated a maximum-likelihood (ML) phylogeny in GARLI v. 0.95 (Zwickl, 2006). The Hasegawa–Kishino–Yano (HKY) nucleotide-substitution model (Hasegawa *et al.*, 1985)

was used, with a gamma-distributed model of among-site rate variation, and the tree was mid-point-rooted. To reconstruct the timescale of genotype 2 movement, we estimated a strict molecular clock phylogeny under the above-mentioned substitution model, using Bayesian Markov Chain Monte Carlo (MCMC) inference as implemented in BEAST v. 1.4.6 (Drummond & Rambaut, 2007). We used the previously estimated NS5B gene substitution rate of 0.0005 substitutions per site year^{-1} (Pybus *et al.*, 2001; Mizokami *et al.*, 2006) for evolutionary rate in the MCMC analyses. Each MCMC run contained 100 million states, sampled once every 10 000 states. MCMC convergence and effective sample sizes were checked by using Tracer v. 1.4 (Rambaut & Drummond, 2007). For both phylogenies, likely ancestral lineage states were reconstructed by using a parsimony approach (Slatkin & Maddison, 1989) and the trees were annotated correspondingly by using FigTree (http://tree.bio.ed.ac.uk).

To examine qualitatively the relationship between geographical origin and phylogenetic position of these genotype 2 sequences, we arranged the branches in the ML phylogeny of all continental African sequences in order of increasing number of nodes between the tips and the root of the tree within the limits of the inferred topology. We then plotted the sequential position of each taxon in the tree against the distance of its sampling location from the westernmost sampling location on the continent. We also collected available estimates of the regional prevalence of genotype 2 from published literature.

Coalescent analyses. We used coalescent-based population genetic analyses to estimate the epidemic history of HCV in Guinea-Bissau. A subset of the initial alignment was created, containing all Guinea-Bissau isolates from the paraphyletic west African clade. The effective number of infections through time was estimated by using the Bayesian skyline plot approach (Drummond *et al.*, 2005) as implemented in BEAST v. 1.4.6 (Drummond & Rambaut, 2007). Population and evolutionary parameters were jointly estimated by using Bayesian MCMC inference; each run contained 10 million states, sampled once every 10 000 states. MCMC convergence and effective sample sizes were checked by using Tracer v. 1.4. The Bayesian MCMC approach that we used explicitly incorporates phylogenetic uncertainty when estimating divergence times, hence the confidence limits for dates that we report represent the statistical uncertainty arising from the length of sequences used.

Robustness analyses. We performed further analyses to investigate the statistical robustness of our evolutionary analysis results. First, we performed a Bayesian MCMC test of phylogeographical structure by using the program BaTS (Parker *et al.*, 2008). This program calculates parsimony score (PS) and association index (AI) tests of geographical structure (Slatkin & Maddison, 1989; Wang *et al.*, 2001) whilst explicitly including phylogenetic uncertainty. A significant result from BaTS is therefore robust to the short sequence length used. Second, to obtain significant support for our main phylogenetic clusters, we created a concatenated alignment of HCV genotype 2 isolates that had been sequenced in multiple genome regions (n=22). A Bayesian MCMC strict molecular clock phylogeny was estimated from this phylogeny, using the same approach as described above for the NS5B alignment. Third, to test the robustness of our coalescent estimates of epidemic history, we repeated the above coalescent analysis using Cameroonian genotype 2 sequences that were cropped to the same 252 nt subgenomic region as our Guinea-Bissau isolates. The results obtained from these cropped sequences could then be compared directly with those obtained by using exactly the same method on uncropped, longer sequences published previously (Njouom *et al.*, 2007).

RESULTS

Fig. 1 shows the estimated ML phylogeny of all available African genotype 2 sequences. The tree exhibits very distinct spatial structure, such that sequences sampled from any given area of the African Atlantic coast cluster together. The most ancestral group, which we term the 'Guinea–Gambia' cluster, is highly diverse, paraphyletic and consists predominantly of sequences sampled in Guinea-Bissau, but also a few from neighbouring Senegal, Guinea and the Gambia (labelled green in Figs 1–3). The second cluster contains sequences from the three adjacent countries of Burkina Faso, Benin and Ghana (the 'Benin–Ghana' cluster; labelled blue in Figs 1–3), whilst the third cluster contains sequences from Cameroon and the Central African Republic (the 'central African' cluster; labelled red in Figs 1–3). The central African cluster is nested within the Benin–Ghana cluster, which in turn is nested within the basal Guinea–Gambia clade.

The central African clade contains only sequences from that region, and few sequences from central Africa are found anywhere else on the tree. The Benin–Ghana cluster is similar in genetic diversity to the central African one, but contains strains from many other locations, which appear mainly as singleton lineages that are scattered among the local strains of the Benin–Ghana cluster. The Guinea–Gambia cluster exhibits the largest genetic diversity and contains a number of isolates from other locations, most of which are gathered in three major 'emigrant' clades (Fig. 1).

Sequences from Madagascar appear at three distinct positions in the tree. The majority of Malagasy sequences belong to one of the aforementioned emigrant clades in the Guinea–Gambia cluster. Although it contains only seven isolates, this group is highly diverse. Martinique sequences appear as singleton lineages at various locations on the phylogeny, but always in close association with sequences from the Benin–Ghana region. The global epidemic subtype 2b clustered within the Guinea–Gambia clade, whilst epidemic subtypes 2a and 2c both originated from the Benin–Ghana region (Fig. 1).

In Fig. 2, we present the relationship between the cladistic position of each sequence and the distance of its sampling place from the westernmost sampling location. For clarity, strains from outside continental Africa are not shown in this figure. The plot clearly demonstrates a distinct spatial trend, with sequences sampled from more westerly locations tending to be closer to the tree root than those sampled further east. The prevalence of genotype 2 relative to other HCV genotypes also declines from west to east.

With a few minor exceptions, the topology of the molecular clock tree (Fig. 3) agreed closely with that of the ML tree (Fig. 1). The most recent common ancestor (MRCA) of all genotype 2 sequences included in our analysis was dated to the year 1091 [95 % credible intervals (CIs), 709–1228]. The date of the MRCA of the primary continental lineage of genotype 2 was estimated to be 1470

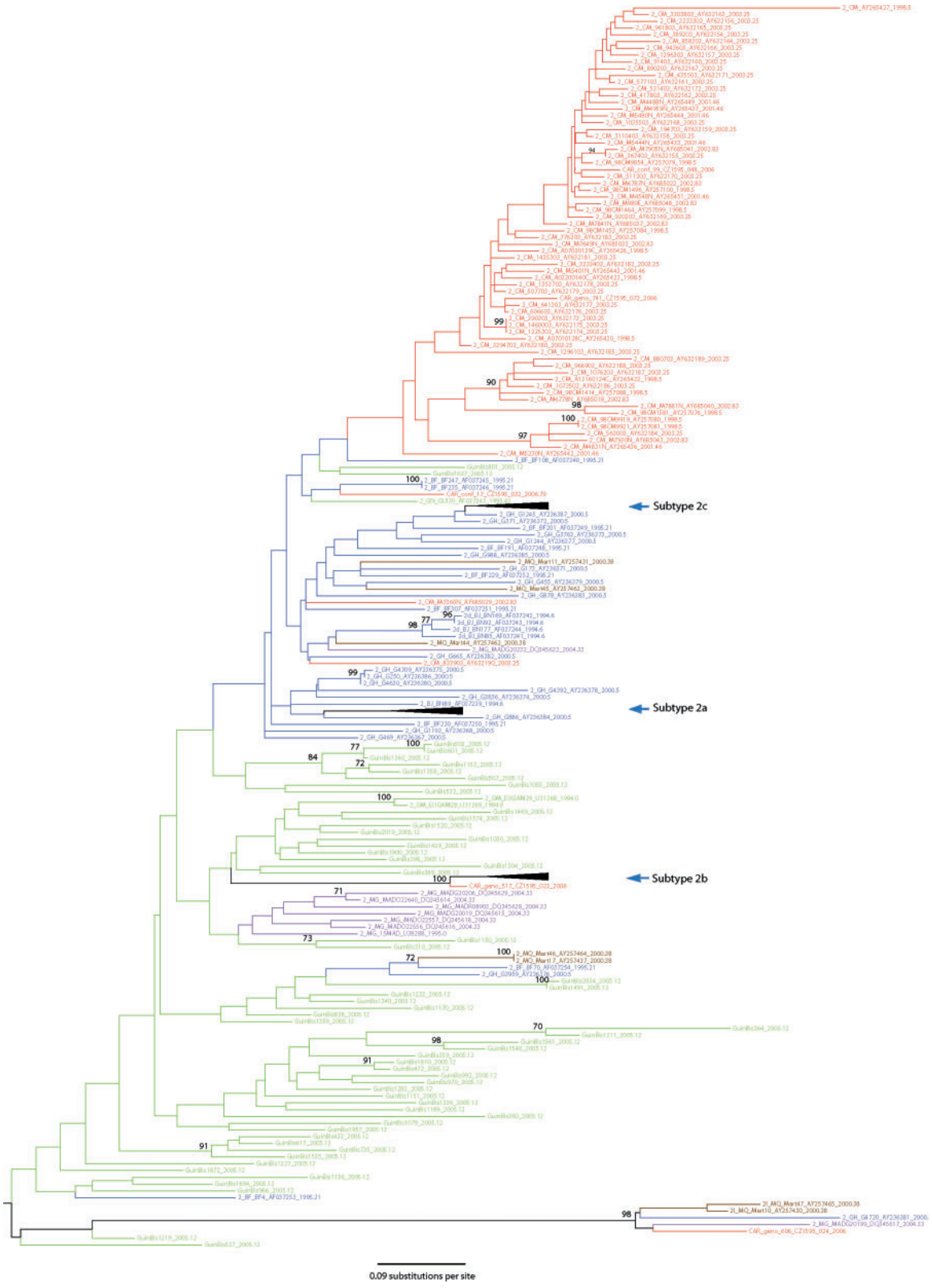

Fig. 1. ML phylogeny of HCV genotype 2 in Africa, estimated without a molecular clock; the tree is mid-point-rooted. The positions of global subtypes 2a, 2b and 2c are indicated. Sequences are coloured according to their sampling location (red, Cameroon and Central African Republic; blue, Ghana, Benin and Burkina Faso; green, Guinea-Bissau, Guinea, the Gambia and Senegal; purple, Madagascar; brown, Martinique). Ancestral branch locations were reconstructed by using a parsimony approach, with ambiguous branches coloured black. See Methods for further details. Bootstrap values >70 are noted above well-supported nodes. Bar, 0.09 substitutions per site.

(95 % CIs, 1326–1541) – this is also the age of the Guinea–Gambia cluster. The MRCAs of the Benin–Ghana and central African clusters have similar dates, estimated to be 1627 (1556–1680) and 1637 (1611–1743), respectively. The estimated date of the Malagasy MRCA was between approximately 1640 and 1760. The molecular clock analysis placed the three epidemic subtypes within the same regional clusters as the ML analysis, with very high statistical support for these groupings, although the central African and Benin–Ghana clusters were placed as sister lineages (rather than the latter being paraphyletic with respect to the former). There was strong support for seven Malagasy sequences forming a cluster and for that cluster deriving from the Guinea–Gambia clade. Every Martinique sequence in the tree was found to group with strains sampled from the Benin–Ghana region, sometimes with a high degree of statistical support; this pairing was found in several places in the tree, not just within the Benin–Ghana cluster. The molecular clock analysis can be used to provide an upper estimate of the dates of the Martinique/Benin–Ghana divergence events, which were 1731 (1669–1794), 1733 (1671–1779), 1738 (1701–1825), 1799 (1709–1804) and 1856 (1767–1886), respectively.

Both phylogenetic analyses included a highly genetically divergent clade with a diverse geographical profile (Figs 1 and 3). Although sequences from this group clustered together very strongly, the large distance from these sequences to the rest of the tree results in very weak phylogenetic signal, hence the position of this divergent lineage within the phylogeny is highly uncertain and we can make no inferences about it at present.

Fig. 4 displays the estimated Bayesian skyline plots for genotype 2 in Guinea-Bissau and Cameroon. The epidemic history of HCV in Guinea-Bissau is one of logistic growth (Fig. 4a). The effective number of infections rose slowly between the early 17th and early 18th centuries, before moving into an exponential phase between 1850 and 1900. The growth rate subsequently slowed during the 20th century. Due to the comparatively short sequences analysed, the CIs of the skyline plot are wide from the end of the 19th century onwards, such that a model of constant exponential growth (with no 20th-century slow-down) cannot be definitively ruled out. Our demographic analysis of the data from Cameroon showed that the effective number of infections was low from the late 17th century until the 1920s. The skyline plot then rises remarkably rapidly between the 1930s and 1940s, before ceasing growth after 1950.

Robustness analyses

Table 1 presents the results of our Bayesian MCMC tests of phylogeographical structure. For all locations, the AI and PS statistics both significantly reject the null hypothesis of no association between sampling location and phylogeny. Therefore, the geographical clustering that we observe is robust to the sequence length used. In addition, a small set of genotype 2 isolates sequenced in multiple genome regions provided strong statistical support (see Supplementary Fig. S1, available in JGV Online) for the existence and positions of the major clusters identified in Figs 1 and 3. Lastly, to test whether our Bayesian skyline plots results are robust to the sequence length used, we shortened NS5B sequences from Cameroon to match the length of our Guinea-Bissau sequences and repeated the skyline plot analysis. Our ability to accurately infer an explosive growth event in Cameroon during the first half of 20th century [shown previously by Njouom *et al.* (2007) using non-shortened sequences] demonstrates that the lack of epidemic growth in Guinea-Bissau during the last century is not an artefact of low statistical power.

DISCUSSION

Our phylogenetic analysis revealed previously unrecognized geographical structure in the genetic diversity of HCV genotype 2 in Africa, indicating considerable isolation among locations through time with respect to factors instrumental in historical HCV transmission. The grouping of endemic HCV strains by location has recently been demonstrated for genotype 6 in east Asia (Pybus *et al.*, 2009), but the results presented here go further, in demonstrating a geographical cline that exhibits genetic isolation by distance. Lineage movements between all pairs of adjacent locations are observed, but not between the non-adjacent central Africa and Guinea–Gambia locations, consistent with a model of isolation by distance.

Several lines of evidence support a west African origin of genotype 2, with subsequent expansion eastwards (Ndjomou *et al.*, 2003). First, the Guinea–Gambia cluster is the most phylogenetically basal and genetically diverse, whilst the central African cluster is the least so (Figs 1 and 2). Second, the molecular clock results show that HCV is a more recent presence in Benin–Ghana and central Africa than in Guinea–Gambia (Fig. 3). Our timing of the MRCA of the central African cluster closely matches previous estimates (Pouillot *et al.*, 2008). Third, several sequences within the Guinea–Gambia and Benin–Ghana clusters have been sampled from other locations, whereas none of the central African cluster sequences were sampled outside central Africa. Fourth, genotype 2 prevalence relative to that of other genotypes declines in an eastwards direction. We can be less certain about the nesting of the central African cluster within the Benin–Ghana clade, which is observed in the ML tree, but not with strong bootstrap support (Fig. 1). The dates of the MRCAs of the Benin–Ghana and central African clusters are not statistically significantly different from each other, hence we cannot reject with certainty the hypothesis of simultaneous introduction of HCV from Guinea–Gambia to both regions.

Sequences sampled in Martinique always grouped with those from the Benin–Ghana area. These associations

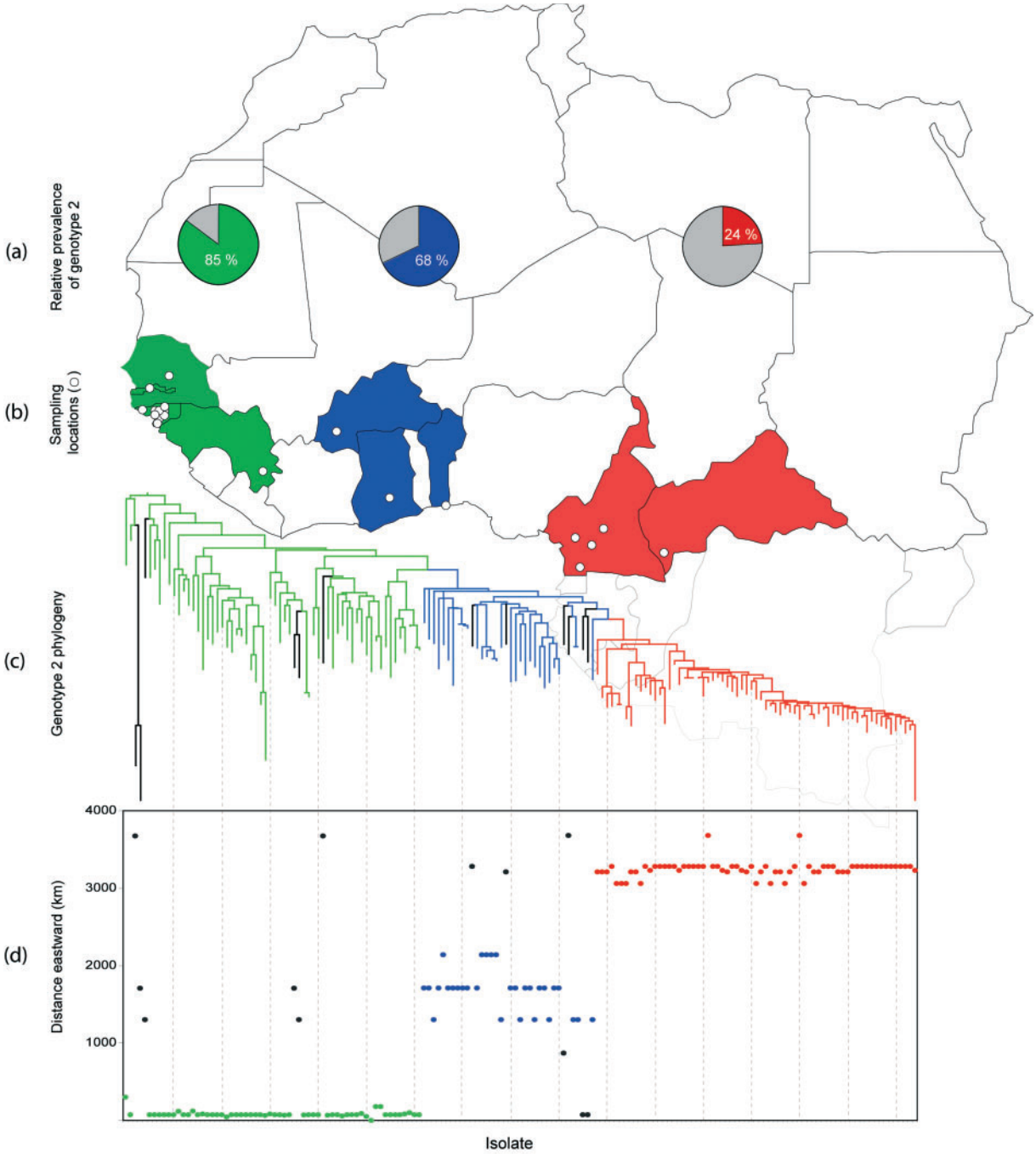

Fig. 2. Phylogeography of HCV genotype 2 in west and central Africa. See Fig. 1 for geographical colouring-scheme details. (a) Prevalence of genotype 2 relative to other HCV genotypes (grey). (b) Countries (coloured) and locations (○) where isolates were sampled. (c) ML phylogeny of continental African sequences. Black branches represent emigrant sequences sampled from locations outside their respective cluster. (d) Number of nodes in the phylogeny from the tree root to each sequence, plotted against the geographical distance in km from the westernmost sampling point.

feature on the ML tree (Fig. 1) and some also have high Bayesian posterior probability support (Fig. 3). This grouping is observed across the entire west African clade and not just in the Benin–Ghana cluster, suggesting that infections moved to the New World via Benin–Ghana, even when they originated from Guinea–Gambia. These observations are in accord with historical records, which indicate that the majority of the Martinique population was drawn from the Bight of Benin, the Bight of Biafra and the Gold Coast (present-day Ghana, Togo, Benin and Nigeria; Eltis

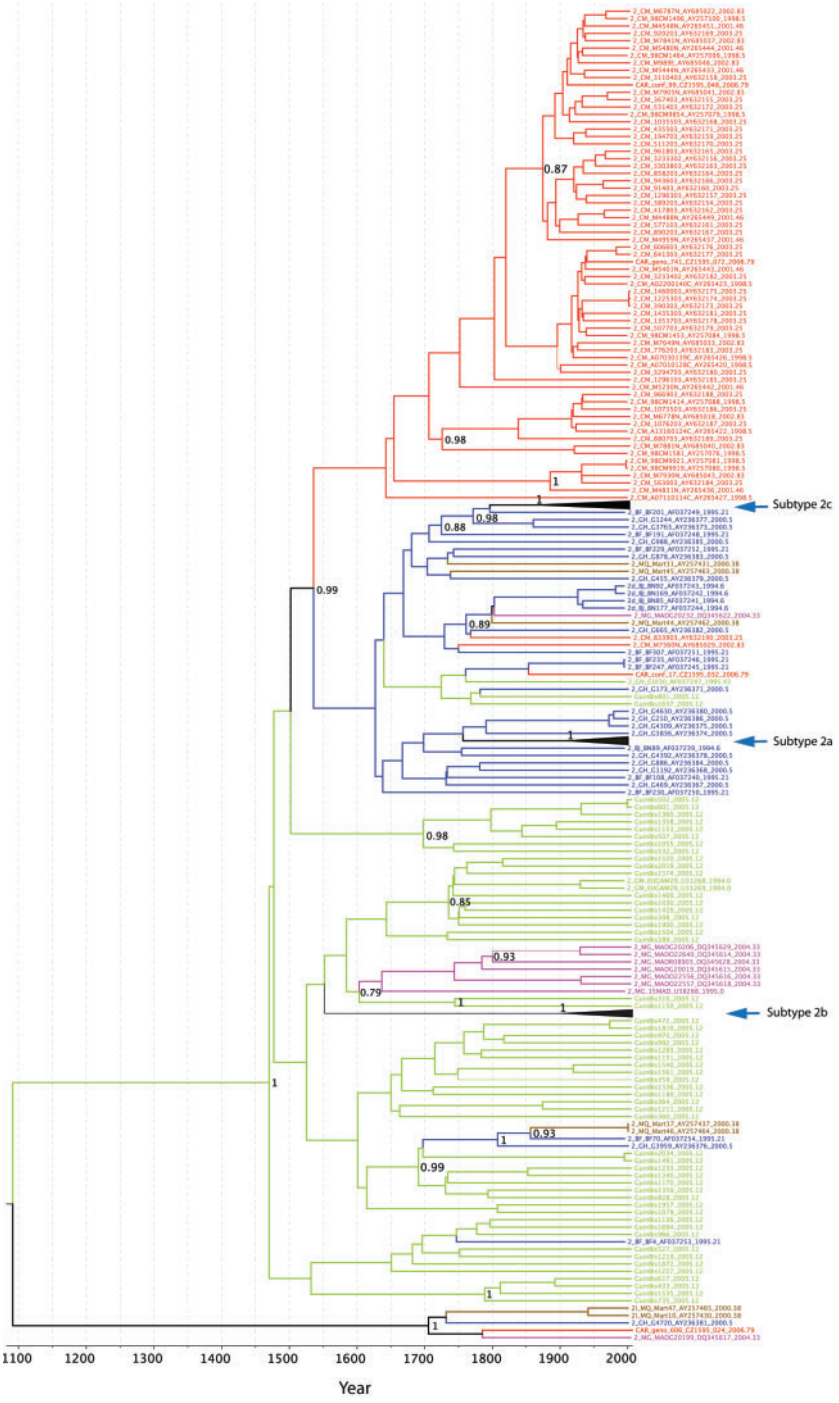

Fig. 3. Molecular clock phylogeny of HCV genotype 2 in Africa. See Fig. 1 for geographical colouring-scheme details. The positions of global subtypes 2a, 2b and 2c are indicated. Ancestral branch locations were reconstructed by using a parsimony approach, with ambiguous branches coloured black. Posterior probability values >0.8 are noted above well-supported nodes. See Methods for further details.

et al., 1999) during the transatlantic slave trade. The molecular clock results (Fig. 3) are consistent with the hypothesis of slave trade-associated viral migration, as they demonstrate that HCV genotype 2 was already widespread in west and central Africa at the peak of slave movement during the 17th and 18th centuries. In addition, the molecular clock places an upper limit on the dates of each Martinique/Benin–Ghana divergence event, none of which appears to occur earlier than the 17th century. It is interesting to note that the epidemic subtypes 2a and 2c both appear to have originated from the Benin–Ghana area. It is therefore likely that the slave trade has played a historical role in the global dissemination of HCV genotype 2. A similar role has previously been proposed for the transcontinental transmission of yellow fever virus prior to mass global travel (Bryant *et al.*, 2007).

The genetic similarity of genotype 2 sequences from Madagascar to isolates from the remote west African coast is puzzling (Figs 1 and 3). Seven Malagasy isolates are

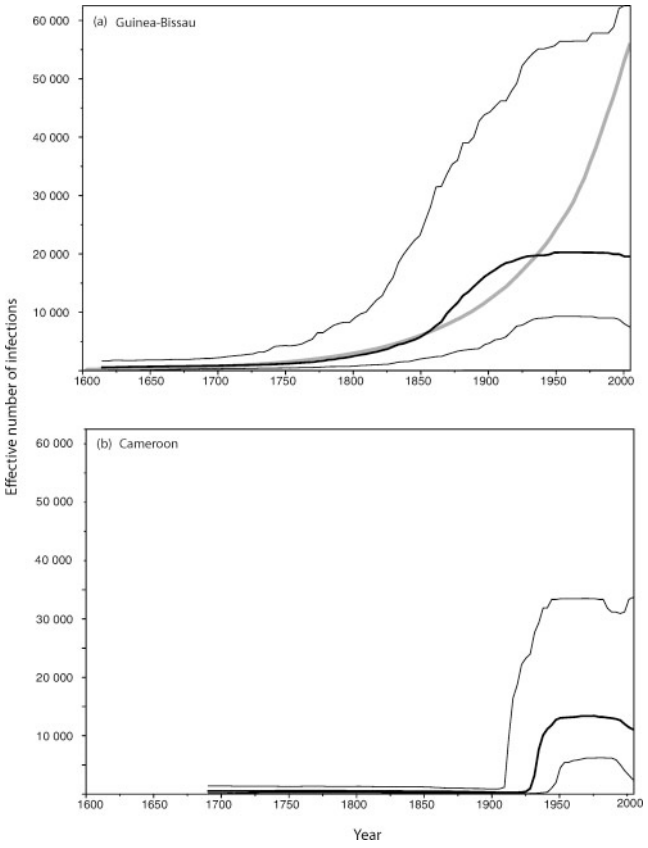

Fig. 4. Estimated effective number of infections through time in (a) Guinea-Bissau and (b) Cameroon, estimated by using the Bayesian skyline plot approach. The thick black line represents the mean estimate and the thin black lines represent the 95 % CIs of the estimate. The grey line represents the effective number of infections estimated under a simple parametric model of exponential growth.

found in a single clade within the Guinea–Gambia cluster, suggesting that a substantial proportion of the genotype 2 infections on the island are descended from a single introduction event. This clade is very diverse, indicating a historical introduction some time during the 17th or 18th centuries (Fig. 3). The presence of two other singleton Malagasy lineages suggests repeated introduction of genotype 2 to Madagascar – these two lineages cluster closely with isolates from Martinique and the Benin–Ghana area. If non-African genotype 2 sequences are included in phylogenetic analysis, then some isolates from the southern French cities of Marseilles and Toulouse group with the above-mentioned strains and with the larger cluster of seven Malagasy isolates (data not shown). These observations support an epidemiological link between the Caribbean, Benin–Ghana and southern France that provides an alternative and more recent potential route of viral migration to the slave trade. Further studies are clearly needed to investigate the role of France in the spread of HCV among Francophone regions. Military conscripts from west African colonies, called Tirailleurs Sénégalais, may provide one avenue of future research: tens of thousands of both Tirailleurs and Malagasy soldiers fought in the French colonial army in the trenches of World War I (Clayton, 1988; Echenberg, 1991), when African troops were also stationed for training in large camps throughout southern France (Deroo & Champeaux, 2006).

Njouom *et al.* (2007) used the Bayesian skyline plot method to reconstruct the spread of three different HCV genotypes in Cameroon, each of which underwent rapid epidemic spread between 1910 and 1970, whilst Pouillot *et al.* (2008) similarly inferred that genotype 2 spread rapidly in Cameroon from 1920 to 1950. Our findings are consistent with these previous results (Fig. 4), although our wide CIs do not allow us to discriminate the exact timings of past growth events. The past epidemic history of HCV in Guinea-Bissau is very different, with a substantial proportion of epidemic growth predating the 20th century. Although we could not definitively reject continued exponential growth during the 20th century, HCV epidemic growth in Guinea-Bissau appears considerably slower than that in Cameroon. A similar pattern of gradual pre-20th-century growth was found in a study of west African countries that did not include Guinea-Bissau (Pouillot *et al.*, 2008). The pre-20th-century spread of HCV in Guinea-Bissau is entirely consistent with gradual growth towards an endemic equilibrium following the infection's introduction, as has been argued previously for endemic genotype 4 (Pybus *et al.*, 2001). The transmission routes that maintained the stable endemic transmission of HCV prior to the 20th century are, as yet, unknown [discussed further by Pybus *et al.* (2007)].

We suggest that the differential epidemic histories of HCV genotype 2 in the two countries probably result from historical differences in the large-scale administration of intravenous antimicrobial drugs, decades before the risk of transmission of blood-borne viruses was understood. After World War I, medical care in Cameroun Français was provided mostly by military doctors, and public-health interventions aimed to cover the whole population (Gouvernement Français, 1921–1938, 1947–1957). From the 1920s on, mobile teams comprehensively screened areas endemic for African trypanosomiasis, and those found to be infected were treated locally with intravenous and intramuscular drugs (Pépin & Labbé, 2008). A quickly decreasing incidence led to the adoption of this model for the control of other infections. Between the late 1920s and 1960, cases of yaws and syphilis were actively sought and treated parenterally with arsenical drugs (Pépin & Labbé, 2008). In the southern forested areas, where yaws was more common, >10 % of the population was often treated each year, and high HCV prevalences (>40 %) among the elderly have been found in these regions (Nerrienet *et al.*, 2005; Njouom *et al.*, 2007). Up to 20 000 leprosy patients were also treated with intravenous or intramuscular drugs from the mid-1930s (Pépin & Labbé, 2008).

Table 1. Bayesian MCMC tests of phylogeographical structure

Performed using BaTS (Parker *et al.*, 2008), available from http://evolve.zoo.ox.ac.uk.

Statistic/geographical cluster	Observed value (95 % CI)	Expected value* (95 % CI)	*P*-value
Association index (AI)			
Guinea–Gambia	0.52 (0.1–0.91)	8.99 (7.5–10.47)	<0.01
Benin–Ghana	1.92 (1.27–2.57)	6.23 (5.42–6.93)	<0.01
Central African	1.05 (0.54–1.54)	10.01 (8.75–11.44)	<0.01
Madagascar	0.42 (0.05–0.63)	1.86 (1.44–2.30)	<0.01
Martinique	0.79 (0.35–1.17)	1.48 (1.11–1.90)	<0.01
Parsimony score (PS)			
Guinea–Gambia	8.37 (7.0–9.0)	50.64 (47.05–54.15)	<0.01
Benin–Ghana	15.95 (14.0–18.0)	31.90 (30.07–33.46)	<0.01
Central African	6.25 (5.0–7.0)	57.80 (53.91–61.92)	<0.01
Madagascar	3.05 (3.0–3.0)	8.86 (8.23–9.00)	<0.01
Martinique	4.84 (4.0–5.0)	6.87 (6.34–7.00)	<0.02

*Expected AI or PS value under the null hypothesis (the hypothesis of no association between phylogeny and location, i.e. no clustering of isolates by sampling location).

In contrast, the health system before the mid-1940s in Portuguese Guinea (now Guinea-Bissau) was more directed towards protecting the health of the European colonists and their Guinean employees. For instance, 54 712 cases of sleeping sickness were detected and treated in Cameroon in 1928 (Jamot, 1932), compared with <30 annual cases in Portuguese Guinea at that time (Ferreira, 1961a). The same magnitude of disparity characterized the effort dedicated to the treatment of yaws, syphilis, leprosy and trypanosomiasis in the two countries (Gouvernement Français, 1921–1938; Ferreira, 1961b; Pépin & Labbé, 2008). The first organized public-health interventions started in 1946 with a programme for sleeping-sickness control that diagnosed 404 cases (Ferreira, 1961b), rising to a peak of 2169 cases in 1952 and decreasing thereafter (Ferreira, 1961c). Treatment of treponemal infections started on a small scale in 1953, but arsenicals were little used (Pinto, 1955). Thus, the 25 year delay in organizing public-health interventions in Portuguese Guinea, combined with a lower incidence of yaws and trypanosomiasis in this drier land, resulted in a much lower proportion of the population receiving intravenous injections than in Cameroun Français, and a reduced opportunity for iatrogenic HCV transmission.

The genetic sequences that we used to investigate the epidemic history of HCV genotype 2 were comparatively short. This meant that we were able to include many previously reported database reference strains of similar length, thereby achieving a more comprehensive and continental-scale analysis. The disadvantage of using such data is that they contain less phylogenetic information, which is reflected in the low bootstrap scores for many internal nodes and in the highly uncertain placing of one divergent lineage. We were also unable to assign new subtype designations reliably to our diverse Guinea-Bissau isolates. Despite these shortcomings, we could obtain useful and robust inferences because we employed a Bayesian analysis framework that takes into account various sources of uncertainty – the 95 % CIs that we report include the phylogenetic uncertainty in our data. We also undertook three additional analyses to ensure the accuracy of our results. In each case, the additional analyses provided further statistical support for our main observations, and demonstrated that our conclusions are robust to the length of the sequences employed.

Given the very high diversity of HCV observed in a small country like Guinea-Bissau and the lack of sequence data from many larger territories in the region, we conclude that there is enormous HCV diversity yet to be discovered. For example, a small survey of HCV in Nigeria (Agwale *et al.*, 2004) reported three 'non-subtypable' genotype 2 infections (out of 12 HCV-positive individuals), which agrees with our eastwards trend in genotype prevalence (Fig. 2) and suggests that endemic HCV genotype 2 is present in the country. The lack of HCV sequence data from Nigeria is particularly unfortunate, as it is the most populous country in Africa and occupies a geographically central position within the range of endemic HCV genotype 2.

ACKNOWLEDGEMENTS

We are indebted to Mireille Plamondon and Alfredo Claudino Alves for their important role in obtaining the Guinea-Bissau samples and to Abby Harrison for help with checking chromatograms. Thanks to Joe Parker and Aris Katzourakis for helpful feedback and discussion. P. V. M is funded by the BBSRC and Merton College.

REFERENCES

Agwale, S. M., Tanimoto, L., Womack, C., Odama, L., Leung, K., Duey, D., Negedu-Momoh, R., Audu, I., Mohammed, S. B. & other authors (2004). Prevalence of HCV coinfection in HIV-infected individuals in

Nigeria and characterization of HCV genotypes. *J Clin Virol* **31** (Supp. 1), S3–S6.

Bryant, J. E., Holmes, E. C. & Barrett, A. D. (2007). Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. *PLoS Pathog* **3**, e75.

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

CDC (1998). Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. *MMWR Morb Mortal Wkly Rep* **47**(**RR-19**), 1–39.

Chemello, L., Alberti, A., Rose, K. & Simmonds, P. (1994). Hepatitis C serotype and response to interferon therapy. *N Engl J Med* **330**, 143.

Clayton, A. (1988). *France, Soldiers and Africa*. London: Brassey's Defence Publishers.

Deroo, E. & Champeaux, A. (2006). *La Force Noire. Gloire et Infortunes d'une Légende Coloniale*. Paris: Tallandier (in French).

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214.

Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**, 1185–1192.

Echenberg, M. (1991). *Colonial Conscripts: The Tirailleurs Sénégalais in French West Africa, 1857–1960*. Portsmouth, NH: Heinemann Educational Books.

Eltis, D., Behrendt, S. D., Richardson, D. & Klein, H. S. (editors) **(1999).** *The Trans-Atlantic Slave Trade – A Database on CD-ROM*. New York: Cambridge University Press.

Ferreira, F. S. (1961a). História da doença do sono na Guiné portuguesa: IV – período de 1927 a 1932. *Boletim Cultural da Guiné Portuguesa* **16**, 139–157 (in Portuguese).

Ferreira, F. S. (1961b). História da doença do sono na Guiné Portuguesa: V – período de 1933 a 1946. *Boletim Cultural da Guiné Portuguesa* **16**, 313–347 (in Portuguese).

Ferreira, F. S. (1961c). História da doença do sono na Guiné Portuguesa: VII – período de 1947 a 1956. *Boletim Cultural da Guiné Portuguesa* **16**, 569–606 (in Portuguese).

Frank, C., Mohamed, M. K., Strickland, G. T., Lavanchy, D., Arthur, R. R., Magder, L. S., El Khoby, T., Abdel-Wahab, Y., Aly Ohn, E. S. & other authors (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *Lancet* **355**, 887–891.

Gouvernement Français (1921–1938). *Rapport Annuel adressé par le Gouvernement Français au Conseil de la Société des Nations conformément à l'article 22 du pacte sur l'Administration sous mandat du Territoire du Cameroun pour l'année: 1921–1938*. Geneva, Switzerland: Office des Nations Unies à Genève (in French).

Gouvernement Français (1947–1957). *Rapport Annuel du Gouvernement Français aux Nations-Unies sur l'Administration du Cameroun placé sous la tutelle de la France*. Geneva, Switzerland: Office des Nations Unies à Genève (in French).

Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160–174.

Jamot, E. (1932). La lutte contre la maladie du sommeil au Cameroun. *Ann Inst Pasteur (Paris)* **48**, 481–539 (in French).

Jeannel, D., Fretz, C., Traore, Y., Kohdjo, N., Bigot, A., Pê Gamy, E., Jourdan, G., Kourouma, K., Maertens, G. & 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.

Kimura, Y., Hayashida, K., Ishibashi, H., Niho, Y. & Yanagi, Y. (2000). Antibody-free virion titer greatly differs between hepatitis C virus genotypes. *J Med Virol* **61**, 37–43.

Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. (2005). The Los Alamos HCV sequence database. *Bioinformatics* **21**, 379–384.

Mellor, J., Holmes, E. C., Jarvis, L. M., Yap, P. L. & Simmonds, P. (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.

Mizokami, M., Tanaka, Y. & Miyakawa, Y. (2006). Spread times of hepatitis C virus estimated by the molecular clock differ among Japan, the United States and Egypt in reflection of their distinct socio-economic backgrounds. *Intervirology* **49**, 28–36.

Ndjomou, J., Kupfer, B., Kochan, B., Zekeng, L., Kaptue, L. & Matz, B. (2002). Hepatitis C virus infection and genotypes among human immunodeficiency virus high-risk groups in Cameroon. *J Med Virol* **66**, 179–186.

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.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C. & other authors (2005). Hepatitis C virus infection in Cameroon: a cohort-effect. *J Med Virol* **76**, 208–214.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. & Pouillot, R. (2007). The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infect Genet Evol* **7**, 361–367.

Nkengasong, J. N., Nyambi, P., Claeys, H., De Beenhouwer, H., Collart, J.-P., Ayuk, J. & Ndumbe, P. (1995). Predominantly hepatitis C virus genotypes 1 and 2 are found in Cameroon. *J Infect Dis* **171**, 1380–1381.

Parker, J., Rambaut, A. & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**, 239–246.

Pasquier, C., Njouom, R., Ayouba, A., Dubois, M., Sartre, M. T., Vessière, A., Timba, I., Thonnon, J., Izopet, J. & Nerrienet, E. (2005). Distribution and heterogeneity of hepatitis C genotypes in hepatitis patients in Cameroon. *J Med Virol* **77**, 390–398.

Pépin, J. & Labbé, A. C. (2008). Noble goals, unforeseen consequences: control of tropical diseases in colonial central Africa and the iatrogenic transmission of blood-borne viruses. *Trop Med Int Health* **13**, 744–753.

Perz, J. F., Farrington, L. A., Pecoraro, C., Hutin, Y. J. F., Armstrong, G. L. & Bell, B. P. (2004). Estimated global prevalence of hepatitis C virus infection. Presented at the *42nd Annual Meeting of the Infectious Diseases Society of America, 30 September–3 October 2004, Boston, MA, USA*. Arlington, VA: Infectious Diseases Society of America.

Pinto, A. R. (1955). Relatório sobre a actividade da missão permanente de estudo e combate da doença do sono e outros endemas na Guiné Portuguesa: referente ao ano de 1955. *An Inst Med Trop (Lisb)* **13**, 275–332 (in Portuguese).

Plamondon, M., Labbé, A. C., Frost, E., Deslandes, S., Alves, A. C., Bastien, N. & Pépin, J. (2007). Hepatitis C virus infection in Guinea-Bissau: a sexually transmitted genotype 2 with parenteral amplification? *PLoS One* **2**, e372.

Pouillot, R., Lachenal, G., Pybus, O. G., Rousset, D. & Njouom, R. (2008). Variable epidemic histories of hepatitis C virus genotype 2 infection in West Africa and Cameroon. *Infect Genet Evol* **8**, 676–681.

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., Drummond, A. J., Nakano, T., Robertson, B. H. & Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol Biol Evol* **20**, 381–387.

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.

Pybus, O. G., Markov, P. V., Wu, A. & Tatem, A. (2007). Investigating the endemic transmission of the hepatitis C virus. *Int J Parasitol* **37**, 839–849.

Pybus, O. G., Barnes, E., Taggart, R., Lemey, P., Markov, P. V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I. & other authors (2009). Genetic history of the hepatitis C virus in East Asia. *J Virol* **83**, 1071–1082.

Rambaut, A. & Drummond, A. J. (2007). Tracer v. 1.4. http://beast.bio.ed.ac.uk/Tracer

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J. & Thomas, D. L. (2000). Genetic epidemiology of hepatitis C virus throughout Egypt. *J Infect Dis* **182**, 698–707.

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.

Seeff, L. B. (2000). Hepatitis C. In *Natural History of Hepatitis C*. Edited by T. J. Liang & J. H. Hoofnagle. San Diego, CA: Academic Press.

Simmonds, P. (2001). The origin and evolution of hepatitis viruses in humans. *J Gen Virol* **82**, 693–712.

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., Chan, S. W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J. & Urdea, M. S. (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., Deleage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C. & other authors (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology* **42**, 962–973.

Slatkin, M. & Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613.

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

Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E. & Simmonds, P. (2001). Identification of shared populations of human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *J Virol* **75**, 11686–11699.

Wansbrough-Jones, M. H., Frimpong, E., Cant, B., Harris, K., Evans, M. & Teo, C. (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.

WHO (1999). Hepatitis C – global prevalence (update). *Wkly Epidemiol Rec* **49**, 10.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J. J., Kabongo, J. M., Kalengayi, R. M. & other authors (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664.

Yoshioka, K., Kakumu, S., Wakita, T., Ishikawa, T., Itoh, Y., Takayanagi, M., Higashi, Y., Shibata, M. & Morishima, T. (1992). Detection of hepatitis C virus by polymerase chain reaction and response to interferon-alpha therapy: relationship to genotypes of hepatitis C virus. *Hepatology* **16**, 293–299.

Zein, N. N. (2000). Clinical significance of hepatitis C virus genotypes. *Clin Microbiol Rev* **13**, 223–235.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD dissertation, The University of Texas at Austin, TX, USA.