High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America

Nazle Mendonca Collaço Véras,1,2 Rebecca R. Gray,2 Luis Fernando de Macedo Brígido,3 Rosângela Rodrigues3 and Marco Salemi 2,4

Correspondence Marco Salemi salemi@pathology.ufl.edu

Received 23 November 2010
Accepted 25 March 2011

1Pós-Graduação em Biologia Molecular, Instituto de Biologia, Universidade de Brasília, Brasília 70919-900, DF, Brazil
2Department of Pathology, Immunology, and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL 32610, USA
3Retrovirus Laboratory, Virology Service, Adolfo Lutz Institute, Ave. Dr Arnaldo 355, São Paulo 01246-902, SP, Brazil
4Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA

Human immunodeficiency virus type 1 subtype C (HIV-1C) represents 30–65 % of HIV infections in southern Brazil, and isolated cases of HIV-1C infection have also been reported in Argentina, Uruguay, Paraguay and Venezuela. Phylogenetic studies have suggested that the Brazilian subtype C epidemic was initiated by the introduction of closely related strains. Nevertheless, because of sampling limitations, the point of entry and the timing of subtype C introduction into Brazil, as well as the origin of the founder lineage, remain controversial. The present study investigated the origin, spread and phylogeography of HIV-1C in South America. Phylogenetic analysis showed a well-supported monophyletic clade including all available strains from Brazil, Uruguay and Argentina. Only one lineage from Venezuela was unrelated to the epidemic involving the other three countries. Molecular clock and likelihood mapping analysis showed that HIV-1C introduction in Brazil dated back to the period 1960–1970, much earlier than previously thought, and was followed by a nearly simultaneous star-like outburst of viral lineages, indicating a subsequent rapid spread. Phylogeographic patterns suggested Paraná or Rio Grande do Sul as the possible entrance points of subtype C and an asymmetrical gene flow from Paraná to São Paulo, Santa Catarina and Rio Grande do Sul, as well as from Rio Grande do Sul to São Paulo fostered by the strong inter-connectivity between population centres in southern Brazil. The study illustrates how coupling phylogeography inference with geographical information system data is critical to understand the origin and dissemination of viral pathogens and potentially predict their future spread.

INTRODUCTION

More than two decades after the discovery of human immunodeficiency virus type 1 (HIV-1), researchers are still looking for effective methods to monitor and control the AIDS epidemic. Part of the challenge is the extensive HIV-1 inter- and intra-host genetic variability resulting in regional epidemics characterized by distinct evolutionary and population dynamic patterns (Holmes, 2008; Salemi et al., 2005a, 2008). Three major HIV-1 groups have been described: M (for main), O (for outlier), N (for non-M non-N, or new) and P (Plantier et al., 2009), with group M further classified into nine distinct subtypes (A–D, F–H, J and K), six subsubtypes (A1–A4, F1 and F2) and a number of circulating recombinant forms (Geretti, 2006; Meloni et al., 2004; Osmanov et al., 2002; Peeters, 2000; Takebe et al., 2004; Thomson et al., 2002; Vidal et al., 2006). HIV-1 subtype C (HIV-1C) is the most prevalent worldwide, contributing ~50 % of infections (Hemelaar et al., 2006).

One-third of the HIV-infected population (620 000 people) in South America lives in Brazil (UNAIDS/WHO, 2007), with an overall prevalence of 0.6 % among individuals aged 15–49 years (Szwarcwald & Souza Junior, 2006; UNAIDS/WHO, 2007). The Brazilian epidemic has been referred to as a mosaic of regional epidemics, reflecting the extension and socio-geographical diversity and regional heterogeneity of the country (Fonseca & Bastos, 2007). Nevertheless, the
south-eastern and southern regions together represent ~80% of the epidemic, showing the greatest incident rates since the 1980s (Brazil, 2008). In Brazil, HIV-1C was first identified in 1994 in Rio Grande do Sul and Sao Paulo (Osmanov et al., 1994) and since then its prevalence has been increasing in the southern region (Locateli et al., 2007).

Subtype C represents 30% of HIV infections in Paraná (Ferreira et al., 2008), 49–79% in Santa Catarina (Brigido et al., 2007; Locateli et al., 2007; Rodrigues et al., 2010) and 45% in Rio Grande do Sul (Brigido et al., 2007; Soares et al., 2005), and a few cases have also been described in Sao Paulo (Guimarães et al., 2002), Rio de Janeiro, Goias (Stefani et al., 2007) and Belem (Machado et al., 2009), suggesting a directional spread towards northern Brazil. More recently, isolated cases of HIV-1C infection were also reported in other South American countries (Carrion et al., 2004; Fontella et al., 2008; Rangel et al., 2009).

Phylogenetic studies have indicated that the Brazilian subtype C strains form a monophyletic cluster, suggesting that the epidemic was initiated by a limited introduction of very closely related strains (Bello et al., 2008; Fontella et al., 2008; Salemi et al., 2005b; Soares et al., 2005). Nevertheless, because of sampling limitations, the region of entry and the timing of subtype C introduction into Brazil, as well as the origin of the founder lineage, remain uncertain. Current estimates propose that the founder event occurred around the early 1980s, although this result was based on viral strains mostly sampled in Rio Grande do Sul state (Bello et al., 2008; Fontella et al., 2008; Salemi et al., 2005b; Soares et al., 2005). Several African countries have been suggested as the original source of the epidemic on the basis of phylogenetic and historical inference, including Burundi (Bello et al., 2008), Mozambique (de Macedo Brigido, 2009), Kenya and Ethiopia (Fontella et al., 2008; Soares et al., 2005). A recent analysis has suggested that, following its introduction into Brazil, subtype C was then disseminated from Brazil to other South American countries (Fontella et al., 2008), although only three non-Brazilian sequences were included.

The present study investigated the origin of HIV-1C in South America by calibrating an accurate molecular clock for sequences from Brazil, Argentina, Uruguay and Venezuela. Moreover, datasets containing sequences from different Brazilian regions in the north (Amazonas), south-east (Rio de Janeiro and Sao Paulo) and south ( Paraná, Santa Catarina and Rio Grande do Sul) were analysed and compared with geographical information system (GIS) data for the first time to characterize in depth the phylogeographic patterns of the Brazilian epidemic.

RESULTS

Substitution saturation analysis

HIV-1C datasets including available sequences from different Brazilian states (see Supplementary Fig. S1, available in JGV Online), as well as reference sequences from the HIV databases (http://www.hiv.lanl.gov/content/index), were generated for the HIV-1C p24 (gag), reverse transcriptase (RT) and gp41 (env) gene regions, respectively (Supplementary Tables S1 and S2, available in JGV Online, summarize all the sequences included in the datasets). No substitution saturation was detected using either the test described by Xia et al. (2003) (Supplementary Table S3, available in JGV Online) or transition and transversion versus divergence graphics (Supplementary Fig. S2, available in JGV Online) performed for every nucleotide sequence alignment used for p24, RT and gp41 phylogenetic and Bayesian analysis. Based on these results, all datasets showed the appropriate phylogenetic signal for consistent phylogenetic and molecular clock inferences.

Likelihood mapping analysis

The analysis including worldwide sequences displayed a significant star-like phylogenetic signal for p24 (31.4%), RT (27.3%) and gp41 (20.9%) (Fig. 1 and Supplementary Fig. S3, available in JGV Online). The elevated proportion of quartets in the centre of the likelihood maps analysing worldwide reference sequences (Fig. 1a–c) is consistent with the extensive spread HIV-1C, which is currently responsible for up to 50% of all HIV-1 infections (Hemelaar et al., 2006). Strikingly, however, the analysis showed...
including only South American sequences (Fig. 1d–f) showed an even higher star-like signal for p24 (49.2 %), RT (36.2 %) and gp41 (29.2 %). The nearly simultaneous outburst of multiple viral lineages characterizing HIV-1C phylogenies points to an exponential spread of the virus early after its initial introduction into the region.

High-resolution phylogenetics of HIV-1C in South America

p24, RT and gp41 maximum-likelihood (ML) and Bayesian genealogies were inferred to investigate the origin of HIV-1C in South America (Fig. 2 and Supplementary Fig. S4, available in JGV Online). In each gene, 98–100 % of the South American strains clustered within a monophyletic clade that always included sequences from Brazil, Argentina and Uruguay. With the exception of the gp41 ML tree, the clade statistical support result was highly significant for both ML (aLRT P values of 0.971 for p24, 0.908 for RT and 0.529 for gp41) and Bayesian (posterior probability \( P = 1 \) for each gene) analyses. In the p24 tree, the South American clade appeared to be closely related to sequences from Kenya and the Democratic Republic of Congo (Fig. 2a). In the RT tree, the South American clade did not cluster with any specific clade with statistical support but appeared to be interspersed among diverse African sequences. Only two Brazilian sequences from Sao Paulo (DT33SP01) and Paraná (DO2CT05) clustered outside the main clade and both were related to strains from Zambia (Fig. 2b), although the statistical support was weak (\( P < 0.5 \)). The only available sequence from Venezuela also clustered outside the South American clade and formed a supported monophyletic clade (\( P = 0.89 \)) with a strain from the Democratic Republic of Congo. Finally, the gp41 tree did not cluster with any specific clade with statistical support (Fig. 2c). Bayesian trees gave similar results (data not shown). Overall, the analyses confirmed the hypothesis that the subtype C epidemic in South America was initiated by a limited introduction of very closely related strains, probably originating from East Africa, although the exact country of origin could not be inferred unambiguously.

Time of the most recent common ancestor (TMRCA) of HIV-1C in South America

To investigate the timing of HIV-1C introduction into South America, a Bayesian molecular clock analysis was performed. Overall, HIV-1C median evolutionary rates within the South American datasets estimated according to different clock models gave similar results in different genes (Table 1). However, as the null hypothesis of a strict molecular clock was strongly rejected for all three genes, the relaxed molecular clock was employed to estimate \( T_{\text{MRCA}} \).

\( T_{\text{MRCA}} \) median and 95 % high posterior density (95 % HPD) intervals for the South American clade, as well as for HIV-1 group M and HIV-1C, were obtained for each gene (Table 2) using the constant coalescent prior. Other population priors (exponential and Bayesian skyline plots) usually result in poor mixing of the chains [effective sample size (ESS) < 200], even after 500 million steps. However, the overall median estimates of \( T_{\text{MRCA}} \) were similar to those obtained with the constant prior, although with wider confidence intervals (data not shown). Group M origin (~1930) (Korber et al., 2000; Salemi et al., 2001; Sharp & Hahn, 2008; Worobey et al., 2008) was also estimated as internal controls to verify the robustness of the Bayesian inference. Overall, the 95 % HPD intervals of subtype C and group M \( T_{\text{MRCA}} \) were consistent with previous results. Interestingly, analysis of the three gene datasets consistently traced the origin of the South American clade back to 1960–1970, at least one decade earlier than previously thought, with overlapping 95 % HPD intervals (Table 2) (Strimmer & von Haeseler, 1997).

HIV-1C phylogeography in Brazil

HIV-1C phylogeographic patterns in Brazil were investigated with a modified version of the Slatkin and Maddison method (Salemi et al., 2005a; Slatkin, 1989) and using the Bayesian phylogeography framework implemented in BEAST v.1.6.1. The analysis was based on the RT dataset because it was the only one including enough sequences from different Brazilian states (see Supplementary Fig. S1).

Four datasets were analysed, each including 56 randomly selected sequences from the states of Sao Paulo, Paraná, Santa Catarina and Rio Grande do Sul (14 sequences per state). As the relatively high star-like signal in the RT dataset (Fig. 1) can lead to uncertain phylogenetic relationships, especially near the root, a Bayesian approach was employed to take into account such an uncertainty, and all subsequent analyses were based on the posterior distribution of sampled trees (after the burn-in) of a Markov-chain Monte Carlo (MCMC). The most parsimonious reconstruction (MPR) of the ancestral characters in all trees showed Paraná and Rio Grande do Sul as the states of possible origin of the MRCA (four examples are given in Supplementary Fig. S5, available in JGV Online). Bayesian phylogeography confirmed the result by assigning nearly equal posterior probability to either Paraná or Rio Grande do Sul as the location of the root (data not shown). According to the metapopulation test, the null hypothesis of panmixia, i.e. absence of HIV-1C population subdivision among different states, could not be rejected for all datasets, indicating a relatively homogeneous epidemic across southern Brazil (Table 3). A similar analysis was carried out by also including viral strains from other South American countries. Such sequences appeared consistently on the terminal branches of each genealogy (data not shown), in perfect agreement with the hypothesis of the recent introduction of subtype C into Argentina and Uruguay from southern Brazil (Fontella et al., 2008). Given the limited number of non-Brazilian sequences
Fig. 2. ML phylogenetic analyses of HIV-1C. (a) ML p24 phylogeny (111 sequences); (b) ML RT phylogeny (367 sequences); (c) ML gp41 phylogeny (151 sequences). The South American monophyletic clade has been collapsed for display purposes (fully labelled trees are given in Supplementary Fig. S4). Tip branches were coloured according to the legend on the left of each tree. The country of origin of the African sequences most closely related to the South American ones is indicated by a two-letter code following the HIV databases convention (http://www.hiv.lanl.gov/content/index). The numbers along the monophyletic branches correspond to approximate likelihood-ratio test (aLRT) SH-like values. Trees were rooted using subtype A, B and J reference strains. Bars, branch lengths in nucleotide substitutions per site.
currently available, however, this result should not be overinterpreted.

The accessibility map (Fig. 3) shows the estimated time to travel from major urban centres (>500,000 inhabitants) to any point of interest in the map. City interconnectivity is very strong in south-eastern and southern Brazil, but not in central, north-eastern and northern geographical regions, which could explain why HIV-1C has remained confined so far to the southern region. The superimposition of HIV-1C mean gene flow observed in the Bayesian genealogies on the accessibility map (Fig. 3) showed a highly asymmetrical migration of subtype C among different Brazilian states (as indicated by the arrows in Fig. 3). Paraná appeared to be the hub of the epidemic from where the virus has spread radially to São Paulo, Santa Catarina and Rio Grande do Sul. A minor viral gene flow from Rio Grande do Sul to Santa Catarina and São Paulo was also evident. In general, the phylogeographic patterns indicated a higher net gene flow from north to south states (from Paraná to Santa Catarina and Rio Grande do Sul), representing 59% of all observed viral flow, than from south (Rio Grande do Sul, Santa Catarina and Paraná) to north states (São Paulo and Santa Catarina) (Fig. 4). Moreover, the accessibility map indicated that, whilst southern Brazilian states are, overall, highly interconnected, the accessibility tends to fade northward, with central and north-western Brazil characterized by largely disconnected areas, an unsurprising finding given the location of the Amazon forest.

Recently, de Oliveira et al. (2010) suggested that the South American HIV-1C epidemic originated from the East African subtype C via the UK. In order to evaluate this hypothesis, 53 RT sequences from the UK closely related to the South American clade (kindly provided by the authors of the paper) were included in the RT dataset, and ML phylogenetic as well as phylogeographic analyses were performed following the same procedure discussed above. According to the ML tree (Fig. 5), the majority of the UK sequences intermixed with the South American ones, thus corroborating the link between the two epidemics. In addition, sequences from the UK, Brazil and East Africa clustered basally to the South American monophyletic clade. However, the fact that subtype C lineages were first introduced in the UK and then in Brazil could not be confirmed. According to the phylogeographic analysis (Supplementary Fig. S6, available in JGV Online), no gene flow could be observed from the UK to Brazil, only from East Africa and Brazil to the UK. In conclusion, although a link between South American and the UK subtype C epidemics clearly exists, our analysis actually supports the scenario of a later introduction of HIV-1C into the UK from Brazil.

## DISCUSSION

The present study investigated the HIV-1C epidemic in South America. For the first time, three carefully selected HIV-1 datasets (p24, RT and gp41), covering six southern Brazilian states as well as currently available strains from neighbouring countries, were analysed. The application of high-resolution phylogenetics and phylogeography allowed an in-depth characterization of the epidemic, the re-evaluation of specific hypothesis and the formulation of new insights about the origin and dissemination patterns of the virus in the region. Overall, each genealogy confirmed the existence of a well-supported monophyletic clade including 98–100% of South American strains. The exact origin of HIV-1C, however, remains uncertain. The inconsistency among genealogies could be the result of a sampling bias in the currently available African strains and/or undersampling of the relevant ones such that the sequences most closely related to the founder strain may not have been sampled. Ultimately, the loss of phylogenetic information may make it impossible to ascertain the exact evolutionary relationships between South American and African strains.

### Table 1. Bayes factor (BF) between strict (null hypothesis) and relaxed (alternative hypothesis) molecular clock

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model†</th>
<th>Median evolutionary rate‡</th>
<th>Marginal likelihood</th>
<th>BF‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>p24</td>
<td>SC (H₀)</td>
<td>3.9×10⁻⁴ (1.8×10⁻⁷–1.3×10⁻³)</td>
<td>−1350.286</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC (H₁)</td>
<td>1.9×10⁻⁵ (7.0×10⁻⁸–3.6×10⁻⁷)</td>
<td>−1346.302</td>
<td>7.97</td>
</tr>
<tr>
<td></td>
<td>SC (H₀)</td>
<td>9.3×10⁻⁴ (2.3×10⁻⁴–1.6×10⁻³)</td>
<td>−4332.272</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC (H₁)</td>
<td>2.1×10⁻⁵ (7.0×10⁻⁸–3.6×10⁻⁷)</td>
<td>−4327.272</td>
<td>10.0</td>
</tr>
<tr>
<td>gp41</td>
<td>SC (H₀)</td>
<td>7.0×10⁻⁴ (1.7×10⁻⁴–1.2×10⁻³)</td>
<td>−7632.421</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC (H₁)</td>
<td>8.3×10⁻⁴ (1.2×10⁻⁴–1.8×10⁻³)</td>
<td>−7607.210</td>
<td>50.422</td>
</tr>
</tbody>
</table>

*The selected molecular clock model (H₀=null hypothesis, H₁=alternative hypothesis) is shown in bold.
**The 95% HPD intervals are shown in parentheses.
†BF>2 indicates positive evidence against the null hypothesis, 10>BF>6 indicates strong evidence against the null hypothesis and BF>10 indicates very strong evidence against the null model.
The molecular clock analysis dated the introduction of HIV-1C into South America in the 1960s–1970s, 15 years earlier than previous estimates (Bello et al., 2008; Salemi et al., 2005a; Santos et al., 2007). The discrepancy could be explained by the fact that our study was based on carefully selected datasets (according to the specific criteria outlined in Methods) to maximize the phylogenetic signal: noise ratio and the representation of sequences with accurately known sampling dates. Data from the Official Brazilian Demographic Data Center (http://www.ibge.gov.br/home/) show that the largest number of African immigrants to Brazil arrived during the 1970s, accounting for 32.4% of all African immigrants up to the year 2000. By comparison, only 9.4% of the immigrants arrived during 1980–1989. Taken together, the demographic data and molecular clock analysis strongly support the scenario of HIV-1C introduction into Brazil from Africa during the 1970s migration wave.

Shortly after HIV-1C introduction into either Paraná or Rio Grande do Sul (as inferred by phylogeographic patterns), likelihood mapping convincingly showed that the South American epidemic started to spread exponentially within the region. In contrast, the presence of a few Brazilian strains (<2% of all available sequences) clustering with African sequences outside the main South American monophyletic clade was too weak as evidence to support a continuous introduction of HIV-1C from Africa to Brazil. Such strains probably represent isolated cases of recently acquired infections rather than multiple, independently evolving regional epidemics, as suggested recently (Jones et al., 2009). Paraná appeared to be the epicentre of the epidemic, continuously exporting viruses to both northern and southern neighbouring states, but a substantial north-to-south net viral flow was also observed. In fact, the main gene flow within southern Brazilian states seems to follow the north-to-south axis, rejecting the hypothesis that the subtype C epidemic is migrating from southern to central Brazil, as suggested by Guimarães et al. (2002). The phylogeographic patterns were also in good agreement with the accessibility data showing a high degree of interconnections among southern Brazilian states, in contrast to central and north-western ones characterized by poorly connected areas. In fact, the accessibility map (Fig. 3) not only explains why the epidemic has almost exclusively been restricted so far to southern Brazil but also suggests that HIV-1C future spread is more likely to occur towards the north-eastern rather than the central or north-western states. Social and behavioural factors may also have contributed to the higher prevalence of subtype C in Santa Catarina and Rio Grande do Sul compared with São Paulo (Reiche et al., 2005). Several studies have shown that sexual transmission by vaginal intercourse is the main HIV-1 infection route in southern Brazil, especially for those infected by subtype C virus and BC recombinant forms (Dias et al., 2009; Soares et al., 2005). In comparison, HIV-1B infections in São Paulo, where this subtype is the most prevalent, are more associated with anal sex practices, and homosexuality is the major risk factor among HIV-seropositive male blood donors (de Almeida Neto et al., 2007). Such issues, however, need to be investigated further considering, for example, the observed South American subtype C gene flow to the UK and the subsequent spread of the epidemic within the epidemiological network of men who have sex with men (de Oliveira et al., 2010).

The well-supported clustering of Argentinean, Uruguayan and Brazilian strains within the South American clade resulted in excellent agreement with the accessibility data. Most of the territorial boundaries among these countries are represented by small rivers that are easy to cross, with no major geographical barrier that would impair the connection between Rio Grande do Sul (Brazil), Argentina and Uruguay (Brazil, 2010). The clustering of the sequence from Venezuela with a lineage from the Democratic Republic of Congo is too limited as a piece of evidence to assume that a separate subtype C epidemic is already occurring in that country. However, the fact that the Venezuelan strain did not cluster within the South

---

**Table 2.** \( T_{\text{MRCA}} \) median and 95% HPD interval estimates (relaxed molecular clock model) of HIV-1C according to different gene regions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>South American HIV-1C</th>
<th>HIV-1C</th>
<th>HIV-1 group M</th>
</tr>
</thead>
</table>

---

**Table 3.** HIV-1C metapopulation structure test

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Observed gene flow</th>
<th>95% CI</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>28–35</td>
<td>0.077</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>28–35</td>
<td>0.008</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>28–35</td>
<td>0.008</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>28–35</td>
<td>0.074</td>
</tr>
</tbody>
</table>

*Each dataset included 56 randomly selected sequences (14 from each Brazilian state: São Paulo, Paraná, Santa Catarina and Rio Grande do Sul).
†Number of migrations observed in the Bayesian tree.
§95% confidence interval of the distribution of migrations in 10 000 random trees.
\( P>0.01 \) indicates a non-significant population subdivision among different states.
American clade fits the hypothesis that the epidemic is unlikely to spread further north given the poor accessibility and natural geographical barrier of the Amazonas.

Understanding the origin and dissemination of different HIV-1 subtypes is essential to improve prevention and intervention strategies targeting specific geographical areas. HIV-1C accounts for 30–65% of the infections in southern Brazil. Although recent evidence suggests that its expansion may be declining (Bello et al., 2009), the virus keeps spreading within other South American countries (Carrion et al., 2004; Fontella et al., 2008; Rangel et al., 2009).

Phylogeography inference informed by GIS data, such as accessibility maps (Gray et al., 2009), allows a detailed reconstruction of the origin and dissemination patterns of HIV-1C in South America and may prove to be critical to analyse and predict the spread of other viral epidemics as well.

**METHODS**

**Datasets.** HIV-1C p24, RT and gp41 gene sequences from South America were downloaded from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/content/index). Viral sequences were selected according to the following inclusion criteria: (i) sequences had already been published in peer-reviewed journals (except for the new sequences from São Paulo described below); (ii) there was no uncertainty about the subtype assignment of each sequence and they were classified as non-recombinant; (iii) sequences were not epidemiologically linked by direct donor–recipient transmission; (iv) only one sequence per individual could be randomly selected; (v) the city/state of origin and sampling date were known and clearly established in the original publication; and (vi) RT sequences had to be from therapy-naïve subjects and/or with no mutation associated with drug resistance. Subtype C reference sequences worldwide were also chosen according to the same criteria to represent different countries, in particular Kenya, Mozambique, Ethiopia and Burundi, considered as the potential source of HIV-1C introduction into South America (de Macedo Brigido, 2009; Fontella et al., 2008; http://www.hiv.lanl.gov/content/index). Nine new RT sequences from São Paulo and Santa Catarina, sampled in 1998–2002,
**Fig. 5.** ML phylogenetic analyses of HIV-1C including sequences from the UK. Tip branches are coloured according to the key. The country of origin of the African sequences most closely related to the South American ones is indicated by a two-letter code following the HIV databases convention (http://www.hiv.lanl.gov/content/index). The numbers along the monophyletic branches correspond to aLRT SH-like values. Bars, nucleotide substitutions per site. The tree was rooted using subtype A, B and J reference strains.
and 15 new gp41 sequences from Sao Paulo, collected in 2009, were also included in the analysis (see next section).

Samples from Santa Catarina and Rio Grande do Sul were overrepresented in the RT dataset when compared with the number of available sequences from Rio de Janeiro, Sao Paulo and Parana. As the accuracy of coalescent analysis depends more on the number of informative sites than the number of sequences (Felsenstein, 2006), reduced datasets were generated by randomly removing some of these sequences. The final datasets are described in Supplementary Table S1. The full list of sequences with GenBank accession numbers, including worldwide reference sequences, is given in Supplementary Table S2. This study was approved by Institute Adolf Lutz Scientific and Ethical Committee (IAL-BM-09/08).

RNA extraction, amplification and sequencing. HIV-1 RNA was extracted from 19 plasma samples from Sao Paulo and five from Santa Catarina using a QIAamp viral RNA mini kit (Qiagen), according to the manufacturer’s protocol. cDNA was synthesized using SuperScript III enzyme (Life Technologies) in a final volume of 20 µl, at 50 °C, for 60 min. HIV-1 polymerase fragments spanning part of the RT (aa 1–235, relative to the HIV-1 HXB2 reference strain; GenBank no. K03455) and gp41 (Env aa 476–694) were obtained by nested PCR as described previously (Dachraoui et al., 2008; Rodrigues et al., 2005; Stuyver et al., 1997). Purified products were sequenced in an automated ABI 3100 Genetic Analyzer (Applied Biosystems). Sequences were assembled with Sequencher or Sequence Navigator software and submitted to quality control procedures (Brigido et al., 2007) to ensure a lack of sample mixtures and contamination.

Substitution saturation. Multiple sequence alignments were generated with the CLUSTAL algorithm (Thompson et al., 1994) and edited manually for optimization. The phylogenetic signal of the aligned nucleotide sequences was investigated using the method of Xia et al. (2003) implemented in the DAMBE program (Xia & Xie, 2001). This method is based on the concept of entropy in information theory. When the mean of observed entropy (H) is significantly smaller than the entropy when sequences have experienced full substitution saturation (HSS), then the sequences are under severe substitution saturation. However, sequences often fail to recover the true phylogeny long before the full substitution saturation is reached. Therefore, we can make use of the index of substitution rate, which is defined as \( I_s = H / H_{SS} \). If \( I_s \) is smaller than \( I_{ss} \) (the critical \( I_{ss} \) value at which the sequences will begin to fail to recover the true tree), we can conclude that the sequences have experienced severe substitution saturation and should not be used for phylogenetic reconstruction.

Substitution saturation was also evaluated by plotting the observed number of transitions and transversions against genetic distance for the \( n(n-1)/2 \) pairwise comparison in an alignment of \( n \) taxa using DAMBE (Xia & Xie, 2001). It is expected that transitions and transversions increase linearly with the genetic distance, with transitions being higher than transversions. Saturation is then reached when transversions outnumber transitions.

Likelihood mapping analysis. In this analysis, groups of four sequences (quartets), chosen randomly, were evaluated using ML. For each quartet, the three possible unrooted tree topologies were weighted. The likelihood weights were then plotted into a triangular surface. The fully resolved tree topologies were plotted in the three corners, which indicated the presence of a tree-like phylogenetic signal, and the unresolved quartets, indicating a star-like signal were shown in the central region of the triangle (Strimmer & von Haeseler, 1997). Likelihood mapping was performed using the TREE-PUZZLE program (Schmidt et al., 2002), by analysing 10000 random quartets for each dataset.

Phylogenetic analysis. ML phylogenetic trees were inferred with the PhyML program (Guindon & Gascuel, 2003; http://www.atgc-montpellier.fr/phyml/), using the general time reversible (GTR) + \( \Gamma + I \) nucleotide substitution model, which was selected with the hierarchical likelihood ratio test described by Swofford & Sullivan (2009). Subtype A (GenBank nos AF004885 and AF069670), B (GenBank nos U21135 and AY423387) and J (GenBank nos AF082394 and AF082395) sequences were used as outgroups. ML tree reliability was evaluated using the Shimodaira–Hasegawa (SH)-like aLRT (Anisimova & Gascuel, 2006), which compares the likelihoods of the best and the second-best alternative arrangements around the branch of interest. According to the type I error rate (test is significant! the branch is not corrected) analysis, the aLRT of an interior branch was almost exact for a cut-off value \( 0.9 \), and could be considered robust for 0.75.<P>0.9.

Bayesian genealogies were also inferred with the BEAST v.1.4.8 software package (Drummond & Rambaut, 2007; http://beast.bio.ed.ac.uk/Main_Page) using the same substitution model, a relaxed molecular clock (see below) and different coalescent priors (constant population size, exponential growth and Bayesian skyline plot). An MCMC chain was run for 100 million generations with sampling every 10000 generations. The results were visualized in Tracer v.1.4.1 (http://beast.bio.ed.ac.uk/Tracer). The ESS value for each parameter was >500, indicating sufficient mixing of the Markov chain. The maximum clade credibility (MCC) tree was then selected from the posterior tree distribution using TreeAnnotator v.4.1.8 available within the BEAST software package. Final trees were visualized and annotated with FigTree v.1.2.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Molecular clock analysis. To obtain a Bayesian estimate of HIV-1C origin in South America, the clade including sequences from Brazil, Argentina and Uruguay was constrained to be monophyletic. The evolutionary rate (nucleotide substitutions per site per year) and \( T_{MRCMA} \) (years) were inferred using sampled sequences at different time points by the MCMC approach implemented in BEAST. The analyses were performed with the same nucleotide substitution model and coalescent prior described in the previous section assuming a strict or a relaxed molecular clock. Separate analyses were performed using a tree and uniform prior for the root height with mean + SEM chosen according to previously published estimates for group M origin (Worobey et al., 2008). An MCMC was run for 100 million generations with sampling every 10 000 generations. The results were visualized in Tracer v.1.4.1. The ESS value for each parameter was >500, indicating sufficient mixing of the Markov chain.

Different clock models were compared by calculating the BF, which is the ratio of the marginal likelihoods (marginal with respect to the prior) of the two models being compared (Suchard et al., 2001). We calculated approximate marginal likelihoods for each coalescent model via importance sampling (1000 bootstraps) using the harmonic mean of the sampled likelihoods (with the posterior as the importance distribution). The difference (in log space) of marginal likelihood between any two models is \( \log BF \). Evidence against the null model (i.e. the one with lower marginal likelihood) was indicated by: \( 6 \geq 2 \log BF > 2 \) (positive), \( 10 > 2 \log BF > 6 \) (strong) and \( 2 \log BF > 10 \) (very strong). BF calculations were performed with Tracer v.1.4.1.

Phylogeographic analysis. For the phylogeography analysis, 14 sequences from each of the Brazilian states (Sao Paulo, Parana, Santa Catarina and Rio Grande do Sul) were selected randomly. The number of sequences representing each state was constrained to be the same to avoid potential biases in viral gene flow estimates due to unequal sampling sizes of different subpopulations (Salemi et al., 2005a; Slatkin, 1989). Four datasets each including a total of 56 randomly selected sequences were tested in order to confirm the reproducibility of results.
The hypothesis of metapopulation structure, i.e. the existence of different HIV-1C subpopulations in different Brazilian states, was tested with a modified version of the Slatkin and Maddison test (Salem et al., 2005a; Slatkin, 1989) using the MCC trees. A one-character matrix was obtained from the original dataset by assigning to each taxon in the tree a one-letter code, indicating its Brazilian geographical region of origin. The putative origin of each ancestral sequence in the tree was then inferred by finding the MPR of the ancestral character using either the ACCTRAN or DELTRAN option. The final tree length, i.e. the number of observed migrations in the genealogy, was computed and compared with the tree length distribution of 10,000 trees obtained by random joining-splitting. Observed genealogies significantly shorter than random trees ($P < 0.0001$) indicate the presence of subdivided populations with restricted gene flow. Calculations were carried out using MacClade v.4.06 (Maddison & Maddison, 2008). The viral gene flow (migrations) among different Brazilian states, as well as among Brazil, African states and the UK, was traced using the ‘State changes and stasis tool’ (MacClade software), which counts the number of changes in a tree for each pairwise character state. Viral gene-flow counts were traced for each of the four datasets and the mean was determined.

Ancestral reconstruction of the locations at the interior nodes was also tested in a Bayesian statistical framework implemented in BEAST v.1.6.1. A matrix of geographical locations was constructed based on the city of sampling for each sequence ($n=4$). A full model was used in which all six possible reversible exchange rates between locations were equally likely (flat prior). The relaxed molecular clock and a model of coalescent population growth (in which all six possible reversible exchange rates between locations were equally likely) were used for the analysis. The MCMC analysis was run until evidence of proper mixing (ESS>200) as described above. Results were analysed with Tracer v.1.4.1 and TreeAnnotator v.4.1.8.

ACKNOWLEDGEMENTS

We thank all the authors who confirmed the date and place of collection of GenBank sequences and Dr Robert J. Gifford for providing the sequences from the UK. N. M. C. V. is the recipient of a PhD scholarship by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). N. M. C. V. performed the phylogenetic and phylogeographic analysis; R. R. G. contributed to the phylogenetic analysis; L. F. M. B. contributed to the study design, supervised the initial part of the study, and revised and approved the final manuscript; R. R. obtained the new sequences analysed in the study; and M. S. contributed to study design, supervised all the analyses. N. M. C. V., R. R. G. and M. S. contributed to writing the manuscript.

REFERENCES


